VIROLOGICA SINICA 2014, 29 (1): 40-47
DOI 10.1007/s12250-014-3389-y

RESEARCH ARTICLE

Differential selection in HIV-1 gp120 between subtype B and East Asian variant B'

Stefan Dang[1], Yan Wang[2], Bettina Budeus[1], Jens Verheyen[3], Rongge Yang[2], Daniel Hoffmann[1]✉

1. Research Group Bioinformatics, Center of Medical Biotechnology and Faculty of Biology, University of Duisburg-Essen, Essen 45117, Germany;
2. AIDS and HIV Research Group, State Key Laboratory of Virology, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan 430071, China;
3. Institute of Virology, University of Duisburg-Essen, Essen 45117, Germany

HIV-1 evolves strongly and undergoes geographic differentiation as it spreads in diverse host populations around the world. For instance, distinct genomic backgrounds can be observed between the pandemic subtype B, prevalent in Europe and North-America, and its offspring clade B' in East Asia. Here we ask whether this differentiation affects the selection pressure experienced by the virus. To answer this question we evaluate selection pressure on the HIV-1 envelope protein gp120 at the level of individual codons using a simple and fast estimation method based on the ratio k_a/k_s of amino acid changes to synonymous changes. To validate the approach we compare results to those from a state-of-the-art mixed-effect method. The agreement is acceptable, but the analysis also demonstrates some limitations of the simpler approach. Further, we find similar distributions of codons under stabilizing and directional selection pressure in gp120 for subtypes B and B' with more directional selection pressure in variable loops and more stabilizing selection in the constant regions. Focusing on codons with increased k_a/k_s values in B', we show that these codons are scattered over the whole of gp120, with remarkable clusters of higher density in regions flanking the variable loops. We identify a significant statistical association of glycosylation sites and codons with increased k_a/k_s values.

KEYWORDS human immunodeficiency virus 1; selection pressure; genomic background

A defining feature of HIV-1 is its ability to adapt to the immune systems of its hosts by evolving new variant viruses. Some of the evolved variations became practically fixed in certain geographical regions and led to the emergence of characteristic regional variants, as e.g. in the case of the East Asian subtype B' (Graf M, et al., 1998; Li Z, et al., 2012a), likely an offspring of the pandemic subtype B. Variants such as B' constitute novel genomic backgrounds of the retrovirus on which natural selection and neutral evolution will then operate anew. The effect of the different backgrounds on the further course of retroviral evolution is unknown. It can be envisioned that for different genomic backgrounds the evolution of resistance to anti-retroviral drugs is different, or that for the same HLA type immune escape paths depend on genomic background. A case in point is the observation of novel mutation patterns associated with drug resistance in East Asian clade B' (Deng X, et al., 2008).

In the present study we test the hypothesis that selection pressure depends on the retroviral genomic background. Similar studies have demonstrated that this is the case for the *env* gene and the reverse transcriptase in various subtypes and host populations (Choisy M, et al., 2004; Travers S A, et al., 2005; Pond S L, et al., 2006). Here, we specifically analyze selection pressure on the viral envelope protein gp120 of pandemic subtype B and the related East Asian subtype B'. We have chosen gp120 as this protein is exposed to a particularly strong selection pressure by the host immune system, and this might lead also to strong differences in selection pressure be-

Received: 6 October 2013, Accepted: 20 December 2013,
Published ahead of print: 16 January 2014
✉ Corresponding author.
Phone: +49-201-1834391, Fax: +49-201-1833437,
Email: daniel.hoffmann@uni-due.de

tween B and B'. Further, we analyzed B and B' as these clades can, on one hand, be distinguished clearly (Wang Y, et al., 2013b), while on the other hand they are quite closely related, so that differences could be traced back to specific genomic patterns.

One way to quantify selection pressure is to evaluate for the codons of a protein coding gene the ratio of non-synonymous nucleotide mutations (leading to a different amino acid) to synonymous mutations of codons (not leading to a different amino acid), often termed d_n/d_s or k_a/k_s (Nei M, et al., 1986; Li W H, 1993). We distinguish *directional* selection, pushing a population away from an established state, and *stabilizing* selection, tying a population to an established state. When no selection pressure is present, we observe *neutral* evolution (Kimura M, 1968). For directional selection, the k_a/k_s ratio should take a higher value (amino acid mutations favored), for stabilizing selection a lower value (synonymous mutations favored). Often the methods for estimating k_a/k_s were applied on a per gene basis, implying averaging over the codons of the studied gene. However, it is possible that within a gene there are positions experiencing directional selection and other positions under stabilizing selection, so that averaging over all codons of a gene may lead to canceling of contributions from different codons, and hence to an underestimation of selection pressure. Therefore, other methods have been developed that allow estimation of selection pressure on a per codon basis. Often these methods employ complex probabilistic models to explain the observed mutations in a gene on the background of a phylogenetic tree (Nielsen R, et al., 1998; Huelsenbeck J P, et al., 2006; Murrell B, et al., 2012). While these methods promise accurate results, they usually require relatively costly calculations. A fast and simple approximation was developed by Chen L (2004). This method counts synonymous and non-synonymous mutations with respect to a reference sequence, and corrects k_a/k_s ratios by a null model derived from the data. The use of a single reference corresponds to the assumption of a star-like phylogeny, as discussed by Chen L (2006).

In the present work we first compare results for selection pressure in the HIV-1 gp120 of the East Asian subtype B' from the MEME method (Murrell B, et al., 2012) with results from the much simpler approach by Chen L (2004). MEME uses a complex mixed-effect model that can account for variation of selection over time and between branches of a phylogenetic tree. If results agree, this would justify using the simpler method, especially for larger data sets where computational demands for the complex model may exceed the resources.

The original method by Chen L (2004) not only estimated directional and stabilizing selection pressure in a codon-wise fashion, but also included a significance criterion for directional selection. Here we slightly extend the method to also allow for significance assessment for stabilizing selection. The extended method is then applied to gp120 of subtype B and of the East Asian variant B', and results discussed in terms of the structure of gp120. Finally, we analyze codons that have a increased values of k_a/k_s in B' compared to B. These patterns show an interesting distribution in the structure of gp120, and we find a statistical association of glycosylation sites with these codons.

MATERIALS AND METHODS

Sequence data

All gp120 sequences were retrieved from the HIV database (http://hiv.lanl.gov) at Los Alamos National Laboratory (LANL). We included only one sequence per patient and excluded sequences that were flagged as "problematic". A total of 2212 sequences of subtype B (including B') were retrieved; we call this sequence set B_{total}. Following Wang Y (2013b), sequences with an L/W-pattern in V3 were assigned to clade B'. With this criterion, the B_{total} set of 2212 sequence was split into 120 B' sequences and 2092 B sequences (excluding B'). The latter set of 2092 sequences is used as representative for subtype B if not explicitly stated otherwise, especially in our comparisons with B'.

For the computation of k_a/k_s values, we used two reference sequences, namely HXB2 (GenBank K03455) (Ratner L, et al., 1985), a commonly used reference sequence for subtype B, and RL42 (GenBank U71182.1) (Graf M, et al., 1998), a widely used reference sequence for the East Asian variant B'. Throughout the paper, all codon numbers refer to gp120 in HXB2.

For the comparison between the two methods for the identification of codons under selection, the smaller B' set of sequences was first translated into amino acids with transeq (Rice P, et al., 2000), the sequences aligned with MAFFT (Katoh K, et al., 2002), and reversely transcribed with revtrans (Wernersson R, et al., 2003). In this way we obtained a codon-wise alignment for 118 B' sequences (two of the 120 original sequences could not be processed).

Prior to the analysis of differential selection pressure between B and B', the B_{total} set of sequences was aligned with MAFFT (Katoh K, et al., 2002) using default parameters.

Computation of selection pressure

The reference method MEME was accessed via the server offered by the MEME authors (http://www.datamonkey.org, Delport W (2010)). The codon-wise alignment of B' sequences described above was submitted to the MEME method using default codon substitution bias model and

significance level (last access October 5, 2013).

We used the kaksCodon method described in Chen L (2004) and implemented in R-package CorMut (Li Z, et al., 2012b), with one extension. Chen L (2004) use a log-odds ratio (*LOD*) as indicator of whether a value $k_a/k_s > 1$ is significant evidence for directional selection:

$$LOD_{dir} = \log_{10} P(i \geqslant N_Y \mid N, q, (\frac{k_a}{k_s})_{corr}) = 1 \quad (1)$$

$$= \log_{10} \sum_{i=N_Y}^{N} \binom{N}{i} q^i (1-q)^{N-i} \quad (2)$$

with P the probability of having at the respective codon the observed number N_Y or more mutations to amino acid residue Y, N the total number of observed mutations at this site, q the *a priori* probability of a mutation to Y for a given null model including average transition and transversion probabilities estimated from the data, and $(k_a/k_s)_{corr}$ the k_a/k_s ratio, corrected by division by the the value of k_a/k_s expected under the null model. The null model is detailed in Chen L (2004).

Computing the LOD_{dir} according to Eq (1) corresponds to application of a right-tailed binomial test, e.g. with LOD_{dir}=3 corresponding to a p-value of 10^{-3}. This makes only sense as a significance criterion for directional selection, where the number of observed mutations is greater than expected under the null model.

We extended this significance indicator by an analogous quantity LOD_{stabil} for stabilizing selection:

$$LOD_{stabil} = \log_{10} P(i \leqslant N_Y \mid N, q, (\frac{k_a}{k_s})_{corr}) = 1 \quad (3)$$

$$= \log_{10} \sum_{i=0}^{N_Y} \binom{N}{i} q^i (1-q)^{N-i} \quad (4)$$

corresponding to the application of a left-tailed binomial test. In practice, we obtained a combined *LOD* for both stabilizing and directional selection as $LOD = -\log_{10} p$ with the p-value resulting from a two-tailed binomial test for given N, N_Y, q.

In the following we drop the index *corr* of k_a/k_s, i.e. k_a/k_s is meant as symbol for a ratio of the naïve ratio of k_a and k_s divided by the corresponding ratio under the null model described in Chen L (2004). This is also the output of function kaksCodon in R-package CorMut (Li Z, et al., 2012b), implementing the method proposed by Chen L (2004). We used version 1.2.0 of CorMut, downloaded from http://www. bioconductor.org.

In this paper, codons with $k_a/k_s > 1$ (or $\log(k_a/k_s) > 0$) and $LOD > 3$ are considered to be under significant *directional* selection, codons with $k_a/k_s < 1$ (or $\log(k_a/k_s) < 0$) and $LOD > 3$ are considered to be under significant *stabilizing* selection.

All statistical analyses were conducted with the R software (R Development Core Team, 2006), version 3.

RESULTS

Comparison of selection pressure estimates for gp120

We first compared the codon-wise selection pressures estimated by the MEME method (Murrell B, et al., 2012) with its more complex and powerful model, with results obtained by using the faster and simpler method of Chen L (2004) (in the following called "kaksCodon"), as implemented in the kaksCodon function of R-package CorMut (Li Z, et al., 2012b).

In the B' alignment covering 629 codons, we identified with kaksCodon 38 codons under directional selection with log-odds ratio $LOD > 3$, compared to 113 codons under significant directional selection $p \leqslant 0.05$ found with MEME. 25 codons were identified by both methods. The difference between 113 codons from MEME and 38 from kaksCodon is likely due to the more complex statistical model of MEME, which is more sensitive than the simple counting method in kaksCodon. The latter lacks power when applied to the comparatively small number of B' sequences. But as long as we find a reasonable number of positions, this high false negative rate may be tolerated if we aim at finding out whether there is differential selection at all. It is unclear whether the 13 codons predicted by kaksCodon but not by MEME are false positives. Irrespective of this question, the predictions for directional selection by both methods are highly significantly associated (a Fisher exact test yields $p = 1.6\times10^{-11}$). Moreover, the power of the simple counting method should increase with the number of sequences, while the computational effort remains small. Thus we consider kaksCodon an acceptable method for the fast screening for positions under selection.

Selection pressure and structure of gp120

Apart from checking agreement of the kaksCodon method with a reference method such as MEME, it is also important to test the results of kaksCodon against independent information. For instance it is plausible that parts of gp120 with the greatest exposure to the immune system, e.g. to antibodies, will experience strong directional selection pressure, while parts of gp120 responsible for the interior architecture of the protein will be subject to stabilizing selection pressure. We expect to find this pattern in the data, irrespective of whether we evaluate sequences of subtype B or of East Asian variant B'. Indeed, Figure 1 and 2 show agreement between structural elements and type of selection pressure. The N-terminal signal peptide and the variable loops V1 to V5 are predominantly under directional selection, and it is also there where the maximum values of k_a/k_s are

reached. Conversely, the constant regions C1 to C5 are predominantly under stabilizing selection, and it is there where the positions under strongest stabilizing selection lie. However, the picture is not black-and-white: e.g. some positions in V2 or V3 are under strong stabilizing selection, while in C3 there are relatively many positions under directional selection.

This overvall picture is the same for the subtype B data (Figure 1) and for the B' data (Figure 2), although the smaller size of the B' data leads to less statistical power and thus predicts significant selection for fewer codons.

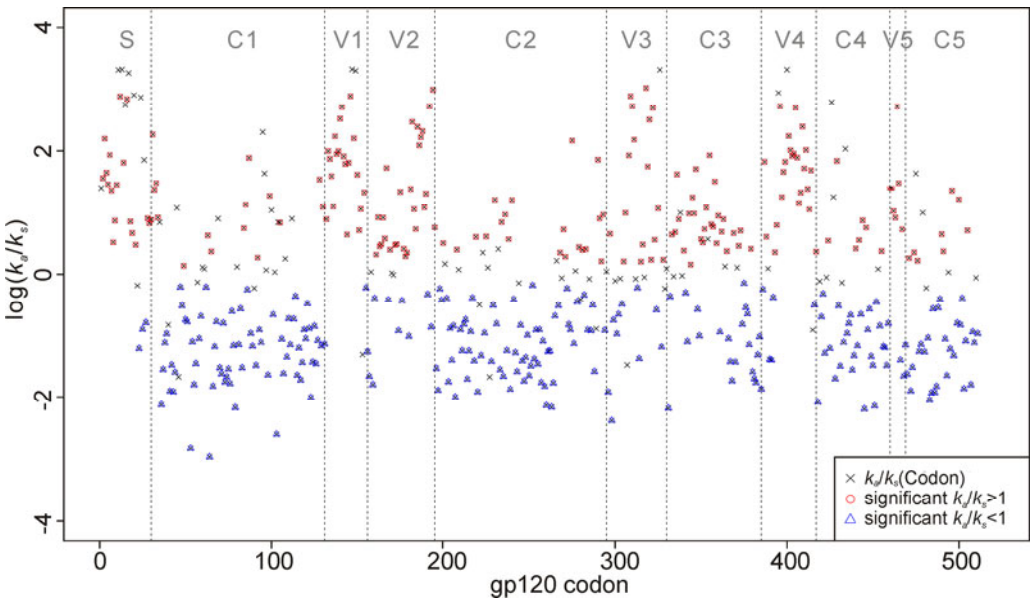

Figure 1. kaksCodon analysis along gp120 of subtype B. The plot gives the value of log(k_a/k_s) for all codons (black crosses). Codons under significant directional selection are additionally marked by a red circle, codons under significant stabilizing selection by a blue triangle. Vertical dashed lines indicate boundaries of gp120 structural elements.

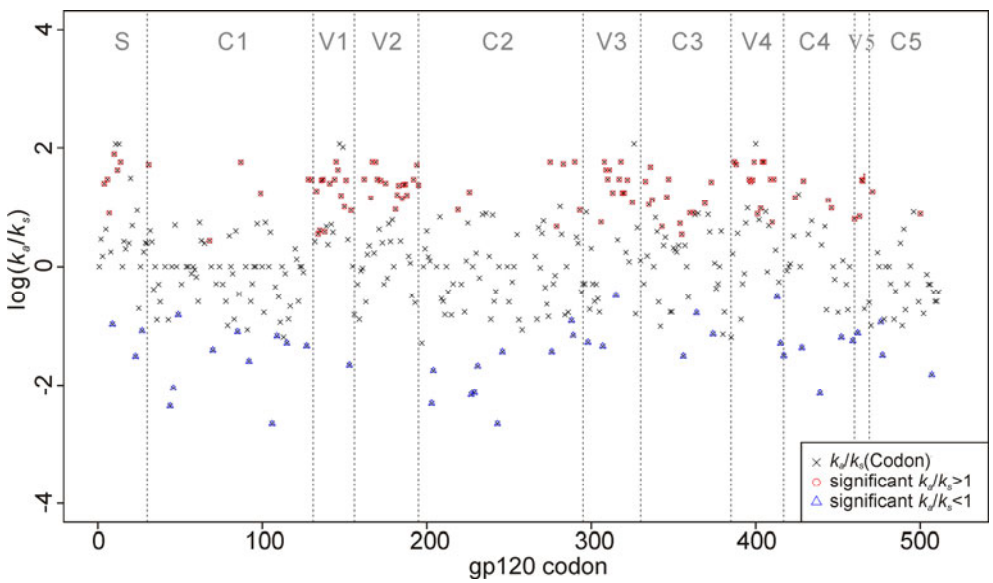

Figure 2. kaksCodon analysis along gp120 of East Asian subtype B'. For explanations see caption of Figure 1.

Do gp120 B and B' experience overall different selection?

While Figure 1 and 2 show that the selection pressure is distributed similarly over the substructures in B and B', there may still be a trend towards overall higher or lower selection pressure in B'. For instance, if B' is not yet well adapted to its host population, there could be overall a higher selection pressure in B', compared to, say, B in Europe or North-America.

If we have a global deviation between selection pressure on corresponding codons i in B and in B', this should show up in a plot of $(k_a/k_s)_{B,i}$ against $(k_a/k_s)_{B',i}$, or the logarithms thereof. Figure 3 indicates that there is no such systematic global deviation, as explained in the following. All points in the Figure are roughly normally distributed around the bisecting line (slope 1), which means that there is no strong global difference between k_a/k_s in B and B'. However, when we fit a linear model to the points we find a small but significant deviation from slope 1. For instance, if we use subtype B reference sequence HXB2 (GenBank K03455) as reference in kaksCodon (black circles), the least-square fit linear model (dashed black line) has a slope of 0.87±0.05 (± standard error). A naïve interpretation of this result is that the selection pressure in B' is more extreme, leading to a larger variation of the k_a/k_s -values in B', and thus to a slope <1 in Figure 3. But we have to keep in mind that kaksCodon uses a simplified estimation of k_a/k_s that assumes a star-like phylogeny rooted in a reference sequence. Accordingly, if we use the usual B' reference sequence RL42 as reference in k_a/k_s Codon, we have less variation along the $(k_a/k_s)_{B'}$ axis, and more along the $(k_a/k_s)_B$ axis, with a slope of the linear model of 1.11 ± 0.06. Hence, we can explain global deviations between selection pressures in B and B' obtained with kaksCodon as an artificial "reference sequence bias" due to the use of reference sequences in the computation of k_a/k_s. Thus we could not find evidence for globally different selection pressure between B and B'.

Codons under increased directional selection in East Asian B'

The lack of evidence for global differences of selection pressure between B and B' does not preclude differences between the two clades at specific codons. We therefore proceeded to a codon-wise analysis to identify codons with differences in selection pressure between B and B'.

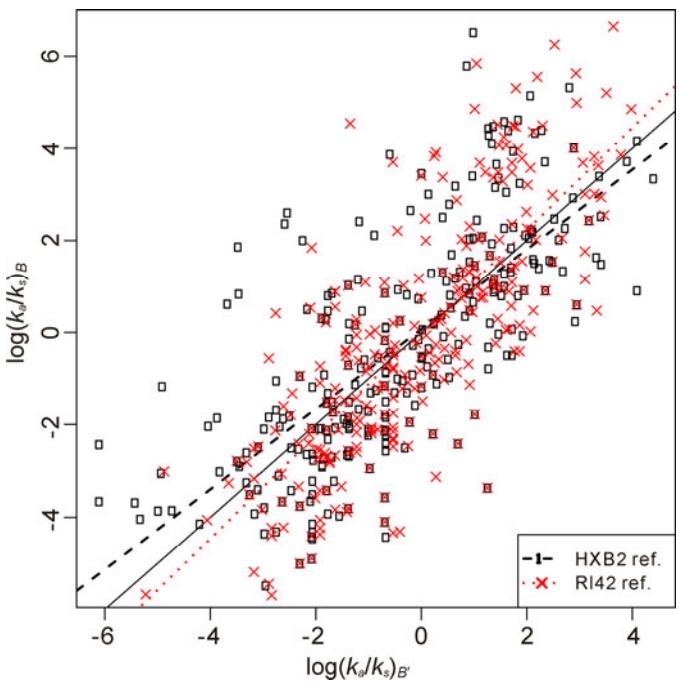

Figure 3. Relation between overall selection pressures on B and B'. For all codons common to the two reference sequences HXB2 and RL42 the $\log(k_a/k_s)$ values for these codons in B' (x-axis) and B (y-axis) are plotted. Black symbols: $\log(k_a/k_s)$ values with reference sequence HXB2 in kaksCodon method; red symbols: $\log(k_a/k_s)$ values for reference sequence RL42. The black dashed and red dotted lines are the least-square fits to the red and black symbols, respectively. The solid black line corresponds to $\log(k_a/k_s)_B = \log(k_a/k_s)_{B'}$.

Our initial hypothesis was that the encounter of HIV-1 with a new host population that has a different composition of HLA-types etc. may exert differential selection pressure at specific sites of gp120. We therefore focused in our analysis on a subset of codons *i* satisfying the following three conditions: (1) Positions *i* had to have non-gap counterparts in both reference sequences HXB2 and RL42; this allowed us to check for reference sequence bias. (2) k_a/k_s for the respective codon *i* for the B' sequences had to be higher than the estimate for the corresponding codon *i* in the B sequences:

$$\Delta_i = (\frac{k_a}{k_s})_{B',i} - (\frac{k_a}{k_s})_{B,i} > 0 \quad (5)$$

We required Eq (5) to hold irrespective of whether reference sequence HXB2 or RL42 was used in the kaksCodon calculation. (3) Only positions were considered with *LOD* > 3 for the subtype B ensemble of sequences; in this way we made sure that the k_a/k_s-shift Δ_i was operating on a position under significant selection pressure.

Figure 4 shows the distribution of the 83 codons *i* fulfilling all three conditions (see Appendix for complete list). At each position, two values of Δ_i are given computed with kaksCodon, one for reference sequence HXB2 (crosses), another for reference sequence RL42 (circles). The distance between cross and circle is a measure of the precision of Δ_i. In most cases, this distance is small, i.e. the precision is good. Interestingly, most of the variable loops do not carry many positions with $\Delta_i > 0$, with the exception of V2. However, high-Δ_i positions seem to cluster at the boundaries of most variable loops, and in some C-regions between variable loops. This observation and Figure 1 and 2 are in agreement with the layered structural model of gp120 proposed by Pancera M (2010). In this model, gp120 is organized in several layers, from a conserved layer around the core interface to the gp41 fusion machinery, over an adapter layer, to a highly variable outer layer facing the host immune system. According to Figure 1 and 2, codons under directional selection pressure cluster most strongly in the variable loops. The changes in these regions may be buffered by the adapter layer, probably including the regions flanking the variable loops. This buffering could necessitate a softening of this adapter layer, leading to a decreased stabilizing selection that may even turn into a directional selection.

Another remarkable feature of Figure 4 is that Δ_i is small for many of these codons. This is consistent with the observation that many of these codons lie in C-regions where stabilizing selection dominates ($k_a/k_s < 1$), see also Figure 1. Since our codon subset contains only positions with $\Delta_i > 0$ this means that for positions under stabilizing selection in subtype B, this stabilizing selection could be weakened in B', so that k_a/k_s is shifted towards the neutral value of 1. We discuss this point in the following.

For a codon *i* fulfilling Eq (5) there are in general three possibilities: (a) We can have stabilizing selection in B and a weaker stabilizing selection in B' if $(k_a/k_s)_{B',I} = (k_a/k_s)_{B'i} + \Delta_i < 1$; (b) Δ_i can shift selection from stabilizing to directional; (c) Δ_i can turn directional selection in B into a stronger directional selection in B'. Figure 5 shows which of these possibilities are realized amongst the 83 positions. The Figure confirms that most of these codons are under stabilizing selection in subtype B ($\log(k_a/k_s)_{B'i} < 0$), and that therefore most codons fall into cases (a) or (b), i.e. the stabilizing selection is weakened or turned into directional selection. However, one has to be cautious with the interpretation of the differences, since for the smaller B' set the power of kaksCodon is much lower than for the larger B set, and hence for some of the codons $\Delta_i > 0$ may due to this smaller power. Still, some of the codons have $\log(k_a/k_s)_{B'I} > 0$ and thus fall into category (c) (upper right corner in Figure 5).

A detailed, codon-by-codon discussion of these 83 codons lies beyond the scope of this work. However, a preliminary analysis gave first insights into possible

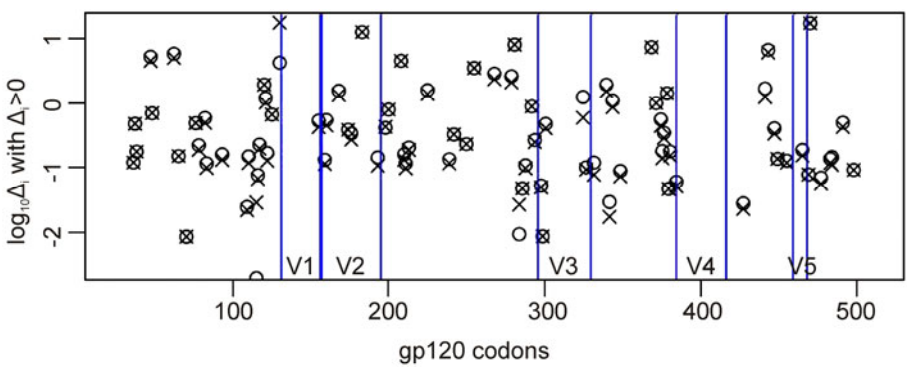

Figure 4. Distribution of codons *i* with $\Delta_i > 0$ along gp120 sequence. At each codon *i*, cross and circle gives Δ_i obtained from kaksCodon with reference sequences HXB2 and RL42, respectively. Blue vertical lines indicate boundaries of variable loops V1 to V5.

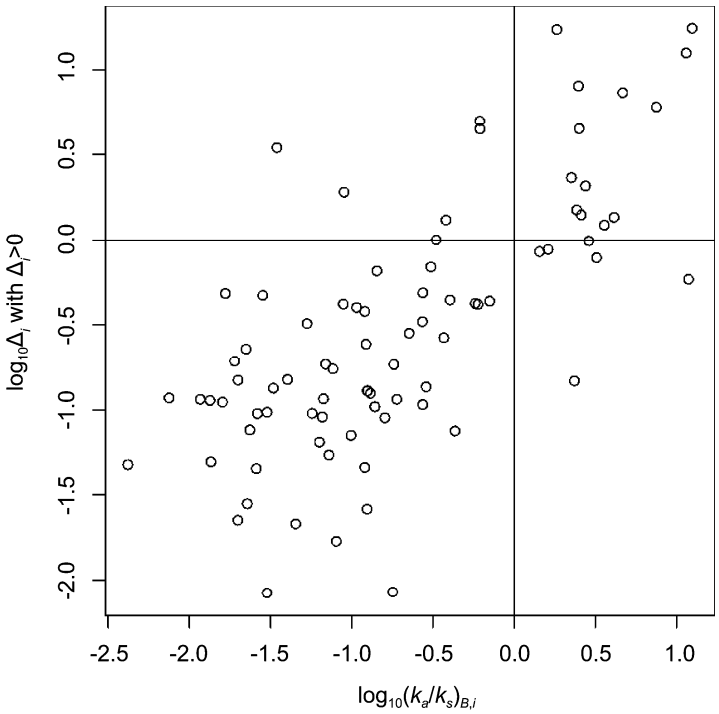

Figure 5. Selection pressure $(k_a/k_s)_{B,i}$ in subtype B vs. change Δ_i of selection pressure at the same 83 codons *i* in East Asian B', both quantities scaled logarithmically.

functions of these codons. When comparing the 83 positions with a high Δ_i with functionally annotated positions of the HXB2 reference sequence as available from the HIV database at LANL (http://hiv.lanl.gov/), we discovered that a conspicuously high number of the codons were situated close to one of the 23 annotated glycosylation sites in gp120.9 of these sites were a maximum of one codon away from one of the 83 positions. A Fisher exact test yields a p-value of 0.006 for an association between glycosylation and membership in the set of the 83 codons with $\Delta_i > 0$, i.e. glycosylation sites are significantly over-represented in this set. This could mean that tuning of glycosylation sites could be an adaptation mechanism in the transition from subtype B to its East Asian variant B'.

DISCUSSION

The importance of glycosylation is underlined by a very recent systematic study by Wang W (2013a), showing firstly, a significant effect of specific Env glycosylations on infectivity, and secondly, that this effect depends strongly on HIV-1 subtype. Our findings are also consistent with an earlier study (Choisy M, et al., 2004) that had found with a maximum-likelihood method a strong association of N-glycosylation sites with sites under directional selection in *env* in various subtypes of HIV-1 and HIV-2. However, at the moment we cannot exclude that there are confounding factors, such as surface exposure. Glycosylation sites are naturally exposed to the solvent, and adapter function may also be easier to accommodate at the protein surface. We hope that the ongoing efforts to determine the structure and thus surface exposure of the functional Env spike (Liu J, et al., 2008; Mao Y, et al., 2012; Tran E E, et al., 2012) will hopefully allow to decide this question. Moreover, we expect that in the future the increasing availability of sequences will allow for higher statistical power in the analysis of selection pressure, and also for more detailed analyses of the evolutionary dynamics of HIV-1 and its changing selection patterns over time.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge funding by Deutsche Forschungsgemeinschaft (http://www.dfg.de), grant TRR60/A6; the University of Duisburg-Essen (http://www.uni-due.de); and the Chinese Key National Science and Technology Program in the 12th Five-Year Period, grant 2012ZX10001006-002.

AUTHOR CONTRIBUTIONS

Designed research: DH; contributed data: YW, RY; performed research: SD, DH; interpreted data: SD, YW,

BB, JV, RY, DH; wrote the paper: DH.

APPENDIX

This is the list of the 83 codons fulfilling the criteria 1-3 described in the text (numbering refers to gp120 in reference genome HXB2 as defined in the HIV database at LANL, http://hiv.lanl.gov):

25, 36, 37, 38, 43, 47, 48, 51, 59, 62, 65, 68, 70, 72, 76, 78, 79, 82 83, 91, 93, 94, 102, 109, 110, 111, 115, 116, 117, 120, 121, 122, 125, 130, 155, 156 159, 160, 168, 171, 174, 176, 178, 183, 193, 197, 198, 200, 202, 208, 210, 211, 213, 216 221, 224, 225, 239, 242, 250, 255, 258, 268, 269, 278, 279, 281, 284, 286, 288, 292, 294 296, 298, 299, 300, 301, 302, 307, 321, 325, 327, 330, 332, 334, 339, 340, 342, 344, 349 369, 372, 375, 376, 377, 379, 380, 381, 385, 386, 389, 419, 421, 428, 432, 442, 444, 448 450, 453, 456, 466, 469, 470, 471, 478, 482, 484, 485, 492, 499, 504, 508, 509, 511

REFERENCES

Chen L, Lee C. 2006. Distinguishing hiv-1 drug resistance, accessory, and viral fitness mutations using conditional selection pressure analysis of treated versus untreated patient samples. Biol Direct, 1:14.

Chen L, Perlina A, Lee C J. 2004. Positive selection detection in 40,000 human immunodeficiency virus (hiv) type 1 sequences automatically identifies drug resistance and positive fitness mutations in hiv protease and reverse transcriptase. J Virol, 78: 3722-3732.

Choisy M, Woelk C H, Guégan J F, and Robertson D L. 2004. Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. J Virol, 78:1962-1970.

Delport W, Poon A F, Frost S D, and Kosakovsky Pond S L. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics, 26:2455-2457.

Deng X, Liu H, Shao Y, Rayner S, and Yang R. 2008. The epidemic origin and molecular properties of b': a founder strain of the hiv-1 transmission in asia. AIDS, 22: 1851-1858.

Graf M, Shao Y, Zhao Q, Seidl T, Köstler J, Wolf H, and Wagner R. 1998. Cloning and characterization of a virtually full-length hiv type 1 genome from a subtype b'-thai strain representing the most prevalent b-clade isolate in china. AIDS Res Hum Retroviruses, 14: 285-288.

Huelsenbeck J P, Jain S, Frost S W, and Pond S L. 2006. A dirichlet process model for detecting positive selection in protein-coding dna sequences. Proc Natl Acad Sci U S A, 103: 6263-6268.

Katoh K, Misawa K, Kuma K, and Miyata T. 2002. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res, 30:3059-3066.

Kimura M. 1968. Evolutionary rate at the molecular level. Nature, 217:624-626.

Li W H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol, 36:96-99.

Li Z, He X, Li F, Yang Y, Wang Q, Wang Z, Xing H, Takebe Y, and Shao Y. 2012a. Tracing the origin and history of hiv-1 subtype b' epidemic in china by near full-length genome analyses. AIDS, 26: 877-884.

Li Z, Huang Y, Ouyang Y, Shao Y, and Ma L. 2012b. CorMut: Detect the correlated mutations based on selection pressure. R package version 1.2.0.

Liu J, Bartesaghi A, Borgnia M J, Sapiro G, and Subramaniam S. 2008. Molecular architecture of native hiv-1 gp120 trimers. Nature. 455:109-113.

Mao Y, Wang L, Gu C, Herschhorn A, Xiang S H, Haim H, Yang X, and Sodroski J. 2012. Subunit organization of the membrane-bound hiv-1 envelope glycoprotein trimer. Nat Struct Mol Biol. 19:893–899.

Murrell B, Wertheim J O, Moola S, Weighill T, Scheffler K, and Kosakovsky Pond S L. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8: e1002764.

Nei M, and Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.

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

Pancera M, Majeed S, Ban Y E, Chen L, Huang C C, Kong L, Kwon Y D, Stuckey J, Zhou T, Robinson J E, Schief W R, Sodroski J, Wyatt R, and Kwong P D. 2010. Structure of hiv-1 gp120 with gp41-interactive region reveals layered envelope architecture and basis of conformational mobility. Proc Natl Acad Sci U S A. 107:1166–1171.

Pond S L, Frost S D, Grossman Z, Gravenor M B, Richman D D, and Brown A J. 2006. Adaptation to different human populations by hiv-1 revealed by codon-based analyses. PLoS Comput Biol. 2:e62.

R Development Core Team. 2006. R: A Language and Environment for Statistical Computing. R version 3.0.0; http://www.R-project.org.

Ratner L, Haseltine W, Patarca R, Livak K J, Starcich B, Josephs S F, Doran E R, Rafalski J A, Whitehorn E A, and Baumeister K. 1985. Complete nucleotide sequence of the aids virus, htlv-iii. Nature. 313:277-284.

Rice P, Longden I, and Bleasby A. 2000. Emboss: the european molecular biology open software suite. Trends Genet. 16: 276-277.

Tran E E, Borgnia M J, Kuybeda O, Schauder D M, Bartesaghi A, Frank G A, Sapiro G, Milne J L, and Subramaniam S. 2012. Structural mechanism of trimeric hiv-1 envelope glycoprotein activation. PLoS Pathog. 8:e1002797.

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

Wang W, Nie J, Prochnow C, Truong C, Jia Z, Wang S, Chen X S, and Wang Y. 2013a. A systematic study of the n-glycosylation sites of hiv-1 envelope protein on infectivity and antibody-mediated neutralization. Retrovirology. 10:14.

Wang Y, Rawi R, Wilms C, Heider D, Yang R, and Hoffmann D. 2013b. A small set of succinct signature patterns distinguishes chinese and non-chinese hiv-1 genomes. PLoS One. 8: e58804.

Wernersson R, and Pedersen A G. 2003. Revtrans: Multiple alignment of coding dna from aligned amino acid sequences. Nucleic Acids Res. 31:3537-3539.