BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 11(3)

Molecular characterization of full-length Tat in HIV-1 subtypes B and C

Chandra Nath Roy*, Irona Khandaker, Yuki Furuse & Hitoshi Oshitani*

Department of Virology, Tohoku University Graduate School of Medicine, 2-1 Seiryomachi, Aoba-ku, Sendai city, Miyagi, Japan-9808575; Hitoshi Oshitani – Email: oshitanih@med.tohoku.ac.jp; Chandra Nath Roy - chandranath@med.tohoku.ac.jp; Phone: 81-22-717-8211; Fax: 81-22-717-8212; *Corresponding authors

Received January 19, 2015; Accepted March 02, 2015; Published March 31, 2015

Abstract:
HIV-1Tat (trans-acting activator of transcription) plays essential roles in the replication through viral mRNA and genome transcription from the HIV-1 LTR promoter. However, Tat undergoes continuous amino acid substitutions. As a consequence, the virus escapes from host immunity indicating that genetic diversity of Tat protein in major HIV-1 subtypes is required to be continuously monitored. We analyzed available full-length HIV-1 sequences of subtypes B (n=493) and C (n=280) strains circulating worldwide. We observed 81% and 84% nucleotide sequence identities of HIV-1 Tat for subtypes B and C, respectively. Based on phylogenetic and mutation analyses, global diversity of subtype B was apparently higher compared to that of subtype C. Positively selected sites, such as positions Ser68 and Ser70 in both subtypes, were located in the Tat-transactivation responsive RNA (TAR) interaction domain. We also found positively selected sites in exon 2, such as positions Ser75, Pro77, Asp80, Pro81 and Ser87 for both subtypes. Our study provides useful information on the full-length HIV-1 Tat sequences in globally circulating strains.

Key words: full-length HIV-1 Tat, Tat, molecular evolution, Tat genetic diversity, Tat genetics

Background:
The human immune deficiency virus type 1(HIV-1) Tat (*trans*-acting activator of transcription), is one of the essential proteins, which directly enhances HIV-1 replication through interaction with HIV-1 long terminal repeat (LTR) promoter **[1].** Tat is therefore a promising target for developing HIV-1 vaccines and anti-HIV-1 drugs **[2-3].** Unlike viral essential enzymes, such as protease and reverse transcriptase, Tat undergoes continuous substitutions due to host selection pressure **[4],** leading to viral escape from Tat-specific CD8-positive T-lymphocyte responses **[5-6].** The sequence variation of target epitopes in Tat reduces antibody recognition and neutralization **[1, 7].** Molecular characterization of full-length Tat in globally circulating strains is therefore imperative.

Tat is a 101-amino acid protein encoded by two exons (exon 1: 1 to 72 residues, and the exon 2: 73 to 101 residues) in most of the clinical isolates **[8].** As shown in **Figure 1a**, Tat has been categorized into six different functional domains **[1, 8].** The first domain (residues 1 to 21) is the N-terminal acidic domain consisting of a Pro-rich tract and a conserved Trp residue at the position 20 (Trp20) **[1].** The second domain (residues 22 to 37) contains a highly conserved seven Cys tract at positions 22, 25, 27, 30, 31, 34, and 37 **[1, 8].** The third domain (residues 38 to 48) contains a hydrophobic core sequence: 43LeuGlyIleSerTry Gly48 **[4].** The fourth domain (residues 49 to 57) is a positively charged region composed of a well-conserved arginine-rich motif, 49-ArgLysLysArgArgGlnArgArgArg-57, and acts as a transactivation response element (TAR) binding domain **[9].** This domain has an extra ordinary property for nuclear localization **[10, 11]** and protein transduction, thus it has also been used to deliver various molecules inside the cells *in vitro* **[12-15].** The fifth domain (residues 58 to 72) is a Gln rich region **[4, 8],** and the fourth and fifth domains (residues 49 to 72) are

known as basic domains for transactivation **[1, 4, 8].** The sixth domain (residues 73 to 101) encoded by the exon 2 is known as RDG domain; this domain contains the highly conserved Glu Ser Lys Lys Lys Val Glu motif, which is related to optimal HIV-1 replication *in vivo* **[16],** thereby, the region may contribute to viral infectivity and binding to cell-surface integrins **[17, 18**]. It is therefore worthwhile to investigate the genetic diversity of HIV-1 full-length Tat in commonly circulating subtypes, which has an impact on clinical outcome of HIV disease as well as success of Tat-based vaccination and Tat-targeted antagonists. However, updated information regarding the evolution of HIV-1 focusing Tat protein at the global level is still unavailable. We therefore performed phylogenetic, selection pressure, and mutation analyses to understand diversity of Tat and its phylogenetic relationships between the subtypes using global data sets of sequences in subtypes B and C, the two major subtypes of HIV-1 circulating worldwide **[19].**

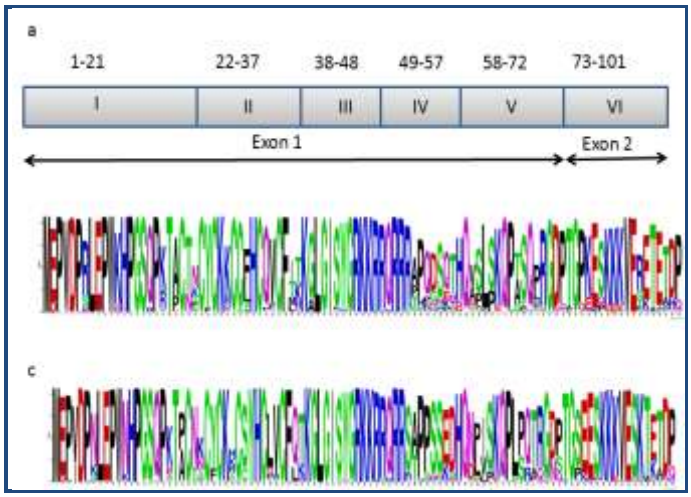

Figure 1: Functional domains of Tat and its genetic diversity. Schematic presentation of the domains of tat exon 1 and 2 were highlighted **(1a):** domain I (residues 1 to 21), an acidic/Pro-rich region; domain II (residues 22 to 37), a Cys-rich/Zn2 Finger domain; domain III (residues 38 to 48), containing conserved Phe (F); domain IV (residues 49–57, the basic domain); domain V (residues 58–72, a Glu rich domain); and domain VI (residues 3–101, encoded by the second exon). Sequence logo showing the Tat amino acid diversity observed at positions 1 to 100 in both subtype **B (1b)** and subtype **C (1c)**

Methodology:

Sequence data

Full-length HIV-1 Tat sequences were collected from the 'Web alignment' in the Los Alamos National Laboratory (LANL) HIV sequence database **[20].** The sequences were downloaded on February 2, 2014. Notably, the sequences 'up to 2012' were available on the accessed date. A total of 2156 sequences were initially downloaded. Sequence data of the other subtypes than subtype B or C, and of circulating recombinant forms (CRFs) were then excluded. Consequently, totals of 713 and 353 sequences for subtype B and subtype C, respectively, were obtained. The sequence data for full-length coding regions were only used after eliminating the problematic sequences. The reference strains for subtype B and C sequences were also obtained from the Los Alamos HIV-1 sequence data base. The reference sequences were also downloaded on February 2, 2014. As the reference sequences, accession numbers AY423387 (Europe), AY173951 (Asia), and AY331295 (North America) for subtype B and AF067155 (India), U52953 (South America), and AY772699 (Africa) for subtype C were used for alignment. Finally, we prepared a data set of totals of 493 and 280 sequences for subtype B and subtype C, respectively, which contained 100 amino acids encoded by 300 nucleotides (nt) from the positions 5831 to 6045 nt in exon 1 and 8379 to 8463 nt in exon 2, respectively of HXB2 genome (GenBank accession No. K03455), the details information of the sequences (the isolation year, isolated country etc), were mentioned in **Table 1 (Available with authors)**. A multiple-sequence alignment of the nucleotide sequences (without any gap) was made using the ClustalW [**21**]. The divergence of sequences was schematically visualized using Weblogo **[22].**

Estimating Phylogenetic tree

Maximum likelihood method was employed to built the phylogenetic trees. Notably, the method was selected as the best model by model test performed in MEGA 6 **[23].** A discrete gamma distribution was used to measure the evolutionary rate differences among sites (5 categories) and the analyses were done using 1,000 bootstrap replicates. The tree was rooted by using following reference strain: simian immune deficiency virus (SIV) sequence, CPZ.US.85.US_ Marilyn.AF103.

Selection pressure analysis

Global (ω) value of relative rates of non-synonymous (dN) and synonymous (dS) substitutions were calculated to measure the positive selection strength **[24].** All analyses were carried out using the online Datamonkey facility **[24-26]** after identifying the best fit model from every possible time-reversible model. Positive selection pressure analysis was performed at whole gene and site-by-site codon level using three likelihood methods: single-likelihood ancestor counting (SLAC), fixed effects likelihood (FEL), and interior branches Fixed Likelihood (iFEL). Briefly, in the SLAC method, the mean ratio of non-synonymous changes per non-synonymous site (dN) and the synonymous changes per synonymous site (dS) were measured using SLAC which considered inferred ancestral sequences for each internal node in a phylogeny using a codon model and then, calculated the synonymous and non-synonymous mutations by comparing each codon to its immediate ancestor. The FEL method is based on maximum-likelihood estimates. This method estimates the ratio of non-synonymous to synonymous substitutions on a site-by-site basis for the entire tree. iFEL is principally the same as FEL, except that selection is only tested along the internal branches of the phylogeny. To detect co-evolving sites from multiple alignments of amino acid sequence data and to identify significant associations among sites, we applied the Bayesian graphical models (BGM) method implemented in Spidermonkey through the Datamonkey web-based interface **[27].**

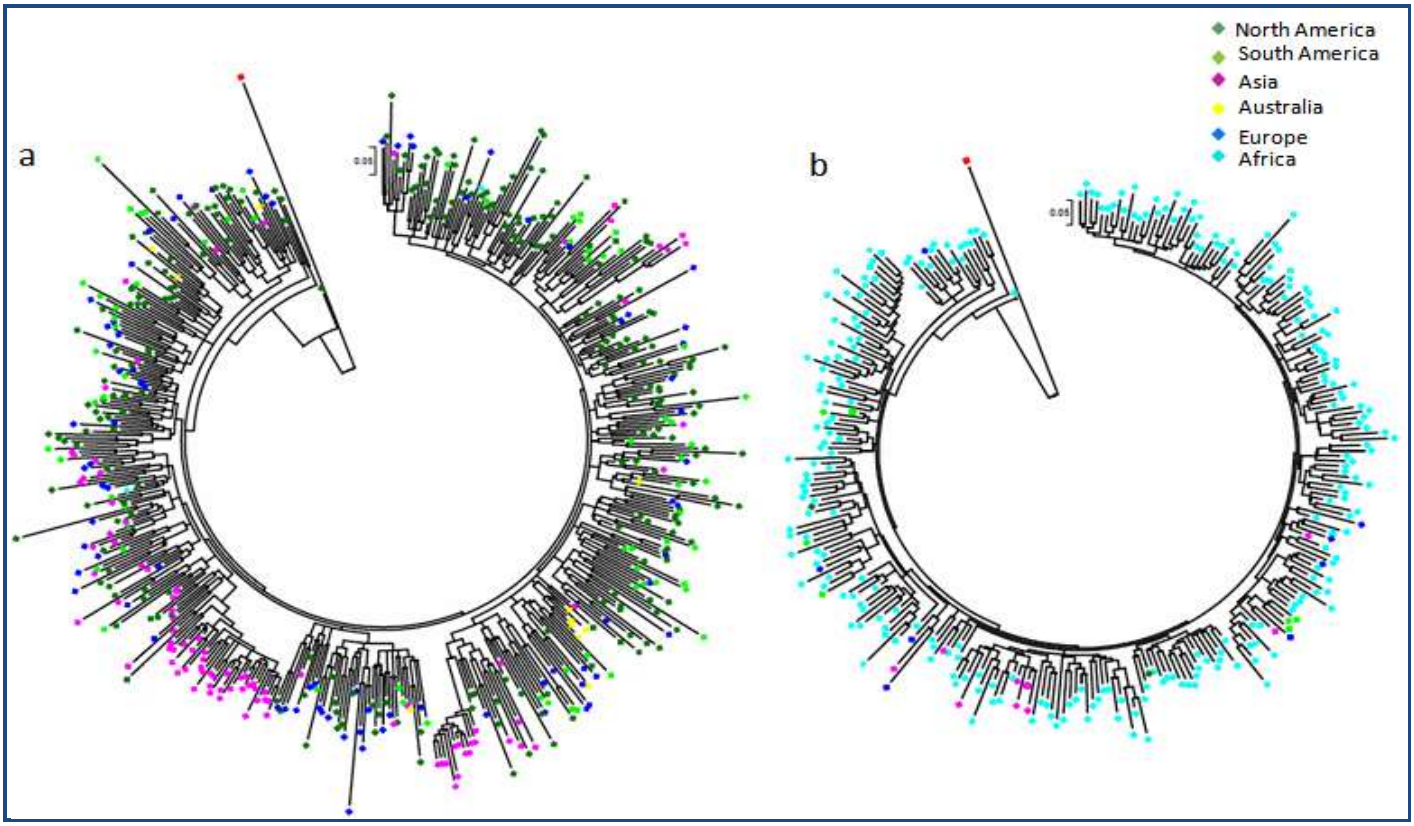

Figure 2: Phylogenetic trees of HIV-1 subtypes B and C. Maximum likelihood (ML) phylogenetic tree of HIV-1 subtypes B **(a)** and C **(b)** sequences based on 300 nucleotide sites of Tat gene sequence generated through the Los Alamos database. GTR+I+Γ5nucleotide substitution model was employed with 1000 bootstrapped data by MEGA 6. The reference Tat sequences were downloaded from the Los Alamos database showed in round bullet. SIV sequence (Ref.CPZ.US.85.US_Marilyn.AF103) was used to root the tree showed in red square bullet

Results:

Sequence homology and Mutation analysis

Nucleotide sequence identities of HIV-1 Tat subtypes B and C were 81% and 84%, respectively. In comparison to subtype C, amino acid substitutions were more frequently observed in subtype B, as shown in **Table 2 (see supplementary material).** Amino acid sequences diversity in both subtypes B and C were illustrated by sequence logo **(Figure 1b & c**). The detailed substitution positions in each domain were described below following HXB2 numbering in both subtypes (the position of amino acid, aa, which had major changes).

In domain I, we found that positions 1, 11, 14, 15, and 16were completely conserved in both subtypes; in addition, Asp5 in subtype C was also completely conserved. Notably, Lys12Asn (90%) and Ala21Pro (53%) were frequently observed only in subtype C.

In domain II, among the conserved Cys positions, Cys27 in subtype B and Cys22 and Cys34 in subtype C were completely conserved.Lys28 in subtype B and His33 in subtype C were also found as completely conserved in our analysis. Cys31Ser was observed in subtype C (83%), and all sequences of subtype C in our data sets were substituted as Phe32Tyr.

In domain III, we found mutations in all positions in this domain except Lys41 in subtype C and Gly48 in subtype B. In subtype C, 80% sequences were substituted as Ile39Gln. In both subtypes, most of the sequences contained Ala42Gly (99%).

In domain IV, the consensus Arg-rich motif provided two of the key functions of Tat, nuclear localization and membrane transduction **[10-15],** a total of 6 Arg at the positions 49, 52, 53, 55, 56, and 57. We did not find any conserved Arg position in subtype B. Only positions 49 and 55 were found conserved in subtype C accompanied with Arg57Ser substitution predominantly (88%). In domain V, only Gln66 was conserved in both subtypes; besides, Gln72 was almost completely conserved in subtype Band subtype C. For other sites, His59Pro substitution was apparent in both subtypes (81% and 95% in subtype B and C, respectively). In addition, Asn61Asp (64%) and Asn67Val (44%) in subtype B, and also Gln60Pro (89%), Asn61Ser (92%), Gln63Glu (75%), Thr64Asp (87%), Ala67Asn (73%), and Lys69Ile (61%) substitutions were frequently found in subtype C.

The domain VI also known as the RGD (Arg-Gly-Asp) domain, and RGD is a ligand for several integrins, which play important roles in HIV replication and cell surface binding **[17, 28, 29].** Even though representative conserved domain in Tat was well maintained in both subtypes, we observed genetic variation in the sixth domain of our data set. In subtype B, 87Ser (77%) was most frequently found with many different substitutions in small percentages atother positions of exon 2. In subtype C, high frequencies of substitutions were observed, includingThr74Leu (92%), Pro77Thr (79%), Pro84Ser (80%),

Lys85Glu (89%) and 87Ser (95%) indicating that this domain is also variable. In subtype C Tat exon 2, we mostly observed substitutions, 93Ser (98%) and 94Lys (97%).Overall, we found genetic variation in domains IV, V, and VI, which may have an impact on the functional properties of Tat, such as Tat-TAR interactions, protein transductions as well as cell surface binding and replication. The summary of substitutions was shown in **Table 2 (see supplementary material).**

Phylogenetic inference

Analysis of the phylogenetic relationship was performed using the maximum likelihood (ML) tree based on the nucleotide sequences of subtypes B and C **(Figure 2a & b).** The phylogeny of Tat in subtype B featured with well intermixing of sequences among the different continents. The wide genetic diversity and several poorly defined clusters were observed in the sequences particularly in the strains from USA, South America and Europe. However, Asian strains, especially Thai and Korean strains clustered together distinctly, may result from different single introduction. In subtype C, it was hard to define clear clustering **(Figure 2b**).In the tree, the majority of subtype C strains were found in Africa and those African stains did not show any monophyletic distribution, and there was a interspersing of Asian, South American and European strains with African stains. Overall, we observed a slow and continuous introduction of new HIV-1 strains in different parts of the world with repeated cross-border transmission as reflected by diffuse distributions and intermixing of the different HIV-1 variants in both subtypes B and C.

Analysis for selection pressure

Global ω values for the nucleotide sequences of Tat in subtypes B and C were 0.883 and 0.760, respectively (less than 1), indicating that there is no detectable positive selection on the entire Tat genes. We found that a total of 23 and 18 amino acid positions in subtypes B and C, respectively, were under significant positive selection based on the FEL, iFEL and SLAC methods. The detailed positively selected codons with statistical significance were calculated with SLAC, FEL, and iFEL methods as shown in **Table 3, 4, 5, & 6 (see supplementary material).** Position Thr40 in the third domain was positively selected in both subtypes by all three methods. We found several positively selected sites in the basic region of Tat, such as Ser68 and Ser70 in both subtypes; His59, Asn61, Ser62, and Thr64, His65 in subtype B; and Ala58, Ala67, and Leu69 in subtype C. Remarkably, in both subtypes of exon 2, Ser75, Pro77, Asp80, Pro81, and Ser87 were positively selected by all 3 methods.

Discussion:

We observed high genetic diversity in HIV-1 full length Tat in both subtypes B and C irrespective of their region of origin as revealed by the phylogenetic tree and mutation analysis. Genetic diversity reflected by relative branch lengths in the phylogenetic tree, particularly in subtype B suggests that the clustering occurs due to the transmission network at individual or at local level. Previous study also showed high genetic divergence of HIV-1 Tat exon-1 in both subtypes **[30].** Furthermore, we found several positively selected sites located in the sixth domain of Tat encoded in exon 2. We found substitutions even in highly conserved Cys-rich and49Arg Lys Lys Arg Arg Gln Arg Arg Arg57 domains. In addition, we found a Ser31Cys substitution (both belong to nucleophilic amino acid group) in HIV-1 subtype C as described previously **[31**]. Absence of a critical Cys31 in the Cys-rich domain has been reported only in subtype C **[31];** this position may play a role in evolution of subtype C. In fact, we observed substitutions such asArg57Gly or Thr in subtype B, Arg57Ser in subtype C, and Gln63Asn in both subtypes B and C, which are within and close to the basic domains, respectively. Interestingly, we found position 63 was under significant diversifying selection for subtype B. However, position 57 was not positively selected in either subtype. Rather, it showed purifying selection in subtype B (data not shown). Previous study showed that mutated Tat in HIV-1 subtype C in those sites exhibited greater transcriptional activity in Jurk at cells compared with subtypes B and E, without LTR sequence dependency **[32, 33].** Overall, the amino acid diversity that we found in well conserved positions is likely to have an impact on Tat mediated viral transcription as described previously **[31, 33, 34].**

We have found that Ser68 and Ser70 positions in the basic domains were positively selected in both subtypes which were previously reported as genetically variable region **[4],** and we also observed similar results. This result implies that changes in amino acids in basic region may have a functional impact, and those changes may fix in the virus population. However, further *in vitro* experiment is needed to validate this hypothesis. We found that positions encoded in exon 2 such as Ser75, Pro77, Asp80, Pro81, and Ser87 were positively selected. As reported previously, exon 2 plays a role in the kappa-light-chain enhancer of activated B cell-(NF-κB) dependent control of HIV-1 transcription in T cells **[8, 35].** It has been previously reported that unlike laboratory-passaged strains, such as HIV-1_{HXB2} with premature stop codon at position 87, majority of HIV-1 strains encode 101 amino acids without any truncation beyond the position 86 **[1].** We found Ser87 in subtypes B and C which was positively selected. As previously noted, the existence of two exons is essential to maintain stability of Tat *in vivo* **[36];** therefore, this position may be crucial to maintain the functional stability of Tat. Again, mutations of the exon 2 were found particularly at intimately networked coevolving sites with exon1 in the fourth, fifth, and sixth domains (data not shown). This may also have some impact on HIV mRNA transcription through Tat-TAR interaction and initiation of reverse transcription, which were previously reported as influenced by genetic variation of Tat **[8, 37].**

Developing antivirals targeting the interaction site between HIV-1 Tat and TAR has been under process **[38].** In addition, Tat based vaccine development is also underway **[39].** Examining the molecular diversity of full-length Tat gene in globally circulating strains is therefore imperative. Thus, our study findings accomplish to understand the genetic diversity of full-length Tat in common HIV-1 subtypes like B and C.

Acknowledgement:

We would like to thank Ms. Karen Lewis for her careful reading the manuscript. We also thank Drs. Yasuhiro Suzuki Junji Imamura and Eiichi N. Kodama for their guidance, inspiration and critical comments regarding writing the manuscript.

References:

[1] Jeang KT *et al. J Biol Chem* 1999 **274**: 28837 [PMID: 10506122]
[2] Ensoli B *et al. AIDS* 2006 **20**: 2245 [PMID: 17117011]
[3] Hamy F *et al. Chem Biol* 2000 **7**: 669 [PMID: 10980447]
[4] Allen TM *et al. Nature* 2000 **407**: 386 [PMID: 11014195]
[5] Mason RD *et al. Virology* 2009 **388**: 315 [PMID: 19394064]
[6] Goldstein G *et al. Vaccine* 2001 **19**: 1738 [PMID: 11166899]
[7] Ruckwardt TJ *et al. J Virol.* 2004 **78**: 13190 [PMID: 15542671]
[8] Li L *et al. Adv Virol* 2012 **2012:** 123605 [PMID: 22899925]
[9] Rana TM & Jeang KT, *Arch Biochem Biophys.* 1999 **365**: 175 [PMID: 10328810]
[10] Truant R & Cullen BR, *Mol Cell Biol*. 1999 **19**: 1210 [PMID: 9891055]
[11] de la Fuente JM & Berry CC, *Bioconjug Chem* 2005 **16**: 1176 [PMID: 16173795]
[12] Ziegler A & Seelig J, *Biophys* J. 2004 **86:** 254 [PMID: 14695267]
[13] Roy S *et al. Genes Dev*. 1990 **4**: 1365 [PMID: 2227414]
[14] Fawell S *et al. Proc Natl Acad Sci U S A* 1994 **91**: 664 [PMID: 8290579]
[15] Hidema S *et al. J Biosci Bioeng* 2012 **113**: 5 [PMID: 22019405]
[16] Smith SM *et al. J Biol Chem* 2003 **278**: 44816 [PMID: 12947089]
[17] Brake DA *et al. J Cell Biol* 1990 **111:** 1275 [PMID: 2202737]
[18] Orsini MJ *et al. J Neurosci* 1996 **16:** 2546 [PMID: 8786430]
[19] Buonaguro L *et al. J Virol* 2007 **81:** 10209 [PMID: 17634242]
[20] Los alamos National Laboratory (LANL) HIV sequence database http://www.Hiv.Lanl.Gov/content/sequence /newalign /align.Html.2014 Accessed 2 February, 2014
[21] http://www.genome.jp/tools/clustalw/ Acccessed February 8, 2014
[22] http://weblogo.berkeley.edu/logo.cgi;2008 Accessed on February 15, 2014
[23] Tamura K *et al. Mol Biol Evol* 2013 **30**: 2725 [PMID: 24132122]
[24] Pond SL & Frost SD, *Bioinformatics* 2005 **21**: 2531 [PMID: 15713735]
[25] Kosakovsky Pond SL & Frost SD, *Mol Biol Evol* 2005 **22:** 1208 [PMID: 15703242]
[26] Delport W *et al. Bioinformatics* 2010 **26**: 2455 [PMID: 20671151]
[27] Poon AF *et al. Bioinformatics* 2008 **24**: 1949 [PMID: 18562270]
[28] El-Sayed A & Futaki S *et al.* AAPS J, 2009 **11:**13 [PMID: 19125334]
[29] Sood V & Ranjan R *et al* .AIDS 2008 **22:** 1683 [PMID: 18670233]
[30] Kandathil AJ *et al. Bioinformation* 2009 **4:** 237 [PMID: 20975916]
[31] Kurosu T *et al. Microbiol Immunol* 2002; **46**: 787 [PMID: 12516777]
[32] Rossenkhan R *et al. J Virol* 2013 **87**: 5732 PMID: 23487450]
[33] Desfosses Y *et al. J Virol* 2005 **79**: 9180 [PMID: 15994812]
[34] Opi S *et al. J Biol Chem.* 2002 **277**: 35915 [PMID: 12080071]
[35] Mahlknecht U *et al. J Leukoc Biol*. 2008 **83**: 718 [PMID: 18070983]
[36] Campbell GR & Loret EP**,** *Retrovirology* 2009 **6**: 50 [PMID: 19467159]
[37] Harrich D *et al. EMBO J*, 1997 **16:** 1224 [PMID: 9135139]
[38] Hamasaki T *et al. Antimicrob Agents Chemother* 2013 **57**: 1323 [PMID: 23274668]
[39] Goldstein G & Chicca JJ, *Hum Vaccin Immunother*. 2012 **8**: 479 [PMID: 22336878]

Edited by P Kangueane
Citation: **Roy *et al.*** Bioinformation 11(3): 151-160 (2015)

BIOINFORMATION

open access

Supplementary material:

Table 1: Lists of sequences (accession numbers) used in this study **(Available with authors)**

Table 2: Comparison of the numbers of amino acid residue changes

Sub type B (n=493)	Subtype C (n=280)
M(493)	M(280)
D(63)	D(22)
S(1)	L(10)Q(4)
I(13)A(1)	I(74)
E(1)N(2)	0
Y(1)H(4)S(1)	H(4)
K(21)S(67)N(24)	D(1)K(46)S(20)N(109)
I(3)	I(4)
D(3)K(1)A(6)Q(1)	D(4)A(1)
A(1)	S(2)
0	0
N(16)E(10)Q(27)R(5)	N(253)E(3)S(1)H(1)
Q(2)	R(2)
O	0
P	0
Q	0
R(22)K(5)	R(3)
H(1)	H(1)
R(122)Q(13)A(2)S(7)E(7)N(1)V(1)T(2)G(1)L(1)	R(14)Q(15)A(3)S(8)E(13)N(10)D(1)T(13)G(1)L(1)I(1)
A(1)	A(1)I(1)
P(95)D(6)E(3)S(2)R(1)T(2)	P(150)V(1)S(2)T(1)
S(1)	0
N(119)P(1)S(6)	N(129)P(1)S(7)
A(8)D(1)G(2)K(88)P(50)Q(4)R(2)S(43)T(57)	A(4)G(6)H(1)K(136)P(7)Q(12)R(7)S(16)T(26)
R(1)	R(1)
F(25)H(2)	C(1)F(44)H(2)M(1)
0	S(1)
0	Q(1)R(1)
A(10)E(3)H(1)I(3)M(1)N(2)Q(28)R(60)S(6)	A(7)C(7)F(3)G(4)H(80)I(1)L(3)N(1)Q(6)R(55)S(13)V(2)Y(49)
G(1)R(1)	R(3)Y(1)
S(17)	S(232)T(2)
L(95)M(4)W(13)Y(67)	Y(280)
P(1)	0
G(1)	0
A(1)I(2)L(3)M(2)P(17)V(1)Y(1)	A(1)I(5)Q(223)P(20)R(1)T(2)V(4)
A(22)C(2)D(1)F(1)H(2)I(2)K(18)L(13)M(2)S(1)R(3)T(3)W(1)Y(1)	A(21)D(1)K(1)F(1)R(2)L(1)S(6)T(1)
0	0
L(1)S(1)	L(3)
A(1)L(62)M(76)S(1)T(140)V(6)	H(3)L(51)R(1)Q(225)
A(1)E(2)H(1)K(180)N(4)Q(20)R(42)S(15)	A(4)K(57)N(4)R(8)S(2)
Q(1)	0
T(1)G(492)	D(1)G(279)
Q(1)	S(1)
S(2)	A(1)S(1)
L(2)V(2)	T(3)
F(19)I(1)P(1)V(1)Y(10)	H(1)P(1)Y(17)
F(1)H(14)N(5)	H(3)N(4)
0	D(2)
G(1_)K(1)S(1)	0
R(1)	R(1)
N(1)	E(1)R(1)
G_(1)W(10)	
G(8)K(7)N(1)S(8)	W(8)
K(1)P(8)R(7)	K(4)P(3)R(7)
G(1)	0
H(1)P(1)Q(2)	H(1)
A(3)G(12)K(4)S(2)T(6)	G(2)H(1)K(1)N(9)S(246)T(2)
D(1)H(2)L(1)N(2)P(100)S(46)T(92)	G(1)P(1)S(2)T(104)
A(1)D(5)N(3)P(400)S(24)R(1)T(4)Y(1)	S(6)T(2)P(267)
D(1)E(27)G(11)H(11)K(14)L(5)N(5)P(47)R(9)S(2)T(1)Y(2)	A(5)H(1)L(1)R(1)S(19)T(1)P(250)
A(2)D(316)E(3)G(90)H(13)S(31)Y(2)	D(2)G(5)R(8)S(257)

C(10)D(4)G(20)H(6)N(42)R(15)T(1)	C(1)G(8)N(16)
A(2)E(71)K(62)p(27)S(7)W(1)	A(5)G(1)K(63)T(2)E(209)
A(29)D(21)G(1)H(1)I(22)N(53)P(9)S(15)V(2)	A(6)D(245)E(2)G(12)N(6)S(8)
D(32)N(43)P(1)R(6)Y(5)	N(4)R(3)S(1)Y(1)
0	0
D(47)E(13)G(25)I(12)K(1)L(1)N(3)S(8)T(10)V(219)	D(62)H(2)I(1)N(206)S(4)T(5)
A(14)D(9)F(1)H(9)N(4)P(97)T(4)Y(8)	F(5)H(8)I(4)L(124)N(1)P(127)
A(1)I(8)P(7)V(7)	I(171)T(1)V(80)
P(142)Q(3)	P(34)L(1)
D(1)E(33)N(6)Q(2)R(1)T(1)	E(11)N(2)
0	R(1)
S(20)	L(1)S(3)
A(106)D(3)G(1)I(2)P(6)N(1)S(52)V(3)	F(5)I(1)S(13)L(259)
A(8)P(30)T(31)	P(176)T(5)
K(1)L(1)P(2)R(7)W(2)	H(4)K(1)P(3)R(47)
R(3)S(25)T(18)V(1)Y(1)	A(35)H(1)I(1)S(1)T(223)V(1)
A(1)G(23)H(2)P(3)Q(12)R(2)	P(5)Q(50)
E(1)K(1)R(3)	E(1)R(4)
E(3)N(25)V(1)	E(1)I(2)N(83)V(3)
Q(38)S(9)T(8)	Q(41)S(72)W(1)
A(26)D(1)E(7)K(6)	A(4)K(3)P(1)
D(1)S(2)	S(2)
Q(48)S(3)T(4)	A(1)E(1)L(4)Q(2)S(226)T(3)
E(87)N(1)Q(3)T(10)	E(250)I(1)Q(1)
K(23)V(1)	K(7)Q(1)
K(1)P(48)Q(60)R(1)S381)W(2)	P(9)Q(4)S(266)T(1)
E(35)Q(4)R(1)T(6)	E(1)Q(2)R(1)T(1)
E(52)S(1)T(1)	T(1)E(1)
A(1)E(29)T(10)	T(1)E(18)
M(3)	A(2)M(2)A(1)G(2)
A(4)K(1)	A(1)G(2)
A(1)E(1)I(1)G(15)K(31)N(6)S84)T(46)	G(2)N(3)S(273)
A(6)K(67)Q(6)T(5)	T(1)Q(3)R(3)K(271)
A(25)E(1)R(2)	A(21)K(76)
A(11)D84)G(5)K(7)Q(5)T(13)V(1)	A(4)D(2)K(76)R(1)S(1)T(5)
A(63)D(1)E(1)G(1)K(8)I(1)L(1)M(10)N(5)P(6)R(2)S(3)V(12)	A(35)E(1)K(1)P(5)R(8)S(4)
A(6)G(5)H(72)N(8)P(1)R(1)S(1)V(1)Y(2)	A(2)G(1)H(1)N(12)T(1)Y()1)
H(2)L(1)Q(68)R86)S(3)T(2)	L(2)Q(49)R(10)
A(12)C(8)D(66)E(21)G(47)H(12)I(18)K(8)L(29)N(11)P(1)Q(7)R(41)S(18)T(5)V(142)Y(12)	C(34)G(1)H(1)L(7)S(13)R(3)V(1)Y(14)

Table 3: Positively selected codon positions HIV-1 Tat

Subtype	Codon[a]	Domain	p-value		
B			SLAC[b]	FEL[c]	iFEL[d]
	7	First (Acidic)	<0.05	<0.05	<0.05
	24	Second (Cysteine-rich)	<0.001	<0.001	<0.001
	32	Second (Cysteine-rich)	<0.001	<0.001	<0.001
	40	Third	<0.05	<0.001	<0.05
	59	Fifth (Basic)	<0.001	<0.001	<0.001
	61	Fifth (Basic)	<0.05	<0.001	<0.001
	62	Fifth (Basic)	<0.05	<0.05	<0.05
	63	Fifth (Basic)	<0.001	<0.05	<0.05
	64	Fifth (Basic)	<0.001	<0.001	<0.001
	65	Fifth (Basic)	<0.001	<0.001	<0.001
	68	Fifth (Basic)	<0.001	0	<0.001
	70	Fifth (Basic)	<0.001	<0.001	<0.001
	75	Sixth (Exon 2)	<0.001	<0.001	<0.001
	77	Sixth (Exon 2)	<0.001	<0.001	<0.05
	80	Sixth (Exon 2)	<0.05	<0.001	<0.05
	81	Sixth (Exon 2)	<0.001	<0.001	<0.001
	84	Sixth (Exon 2)	<0.001	<0.001	<0.05
	85	Sixth (Exon 2)	<0.001	<0.001	<0.001
	87	Sixth (Exon 2)	<0.001	0	<0.001
	88	Sixth (Exon 2)	<0.001	<0.001	<0.05
	90	Sixth (Exon 2)	<0.05	<0.05	<0.05
	93	Sixth (Exon 2)	<0.001	<0.001	<0.001
	98	Sixth (Exon 2)	<0.001	<0.001	<0.001
C	4	First (Acidic)	<0.001	<0.001	<0.001
	21	First (Acidic)	<0.05	<0.05	<0.05

29	Second (Cysteine-rich)	<0.05	<0.001	<0.001
39	Third	<0.05	<0.05	<0.001
40	Third	<0.05	<0.001	<0.05
58	Fifth (Basic)	<0.001	<0.001	<0.001
67	Fifth (Basic)	<0.001	<0.001	<0.001
68	Fifth (Basic)	<0.001	<0.001	<0.001
69	Fifth (Basic)	<0.001	<0.001	<0.001
70	Fifth (Basic)	<0.001	<0.001	<0.05
75	Second Exon	<0.001	<0.001	<0.001
77	Second Exon	<0.05	<0.001	<0.05
80	Second Exon	<0.001	<0.001	<0.001
81	Second Exon	<0.001	<0.001	<0.001
87	Second Exon	<0.05	<0.05	Not selected
95	Second Exon	<0.001	<0.001	<0.05
97	Second Exon	<0.001	<0.001	<0.05
100	Second Exon	<0.05	<0.001	<0.001

Foot note: [a] according to HIV-1HXB2 numbering
[b]single-likelihood ancestor counting (SLAC)
[c]fixed effects likelihood (FEL), and
[d]interior branches likelihood (iFEL) approach
Grey colored rows indicate commonly selected sites in both subtypes. p-values were shown as <0.05 and <0.001.

Table 4: Site under positive selection by SLAC method in subtype B (4a) and C (4b)

Codon	dN-dS	Normalized dN-dS	p-value
4	8.75216	0.366369	0.00365278
7	20.6135	0.862891	0.0341446
24	72.0506	3.01607	1.57E-07
32	66.9937	2.80438	2.78E-09
39	42.1178	1.76307	0.00184905
40	34.158	1.42987	0.00256497
47	11.2377	0.470414	0.0100499
58	111.025	4.64755	8.44E-32
59	50.5059	2.1142	1.82E-14
61	36.6593	1.53457	0.00477469
62	25.929	1.0854	0.00262506
63	44.9872	1.88319	0.00029172
64	69.8476	2.92385	1.21E-13
65	35.9778	1.50605	2.77E-07
68	74.1966	3.1059	3.33E-21
70	66.7315	2.79341	1.69E-13
73	11.3045	0.473213	0.00067664
75	36.4257	1.5248	6.12E-11
77	62.2126	2.60425	1.71E-06
80	11.7369	0.491312	0.0241565
81	25.9187	1.08497	4.94E-06
84	24.0092	1.00504	2.31E-07
85	38.1916	1.59872	4.46E-07
86	11.6579	0.488005	0.0135706
87	77.9872	3.26458	4.40E-19
88	24.0204	1.0055	5.34E-06
89	14.837	0.621085	0.0150515
90	15.3424	0.64224	0.00333789
93	55.4161	2.31974	1.28E-08
95	13.1843	0.551899	0.00331566
97	31.1695	1.30477	0.00060578
98	37.06	1.55135	1.01E-06
100	96.7752	4.05105	3.76E-09
Codon	**dN-dS**	**Normalized dN-dS**	**p-value**
4	13.0933	1.11386	2.77E-06
19	12.0428	1.02449	0.0194111
21	8.76653	0.745778	0.00640897
29	20.5078	1.74462	0.00422291
36	6.03257	0.513197	0.0434762
39	8.55281	0.727596	0.00391241
40	12.3986	1.05476	0.00113586
58	14.1613	1.20472	2.80E-07
59	3.70586	0.315261	0.0116663
67	13.915	1.18376	0.00020476

68	28.5869	2.43192	1.34E-15
69	19.095	1.62443	1.71E-06
70	10.8347	0.921717	2.06E-05
71	3.54406	0.301497	0.048156
75	19.2189	1.63497	9.18E-11
77	9.09342	0.773587	0.00779295
80	15.6258	1.3293	4.49E-05
81	22.5102	1.91496	2.93E-12
87	6.36906	0.541822	0.00217784
90	4.54426	0.386585	0.0194814
95	6.06914	0.516308	0.00067664
97	15.1829	1.29163	3.69E-07
98	5.60021	0.476415	0.00582018
100	11.747	0.999329	0.00784085

Table 5: Site under positive selection by FEL method in Subtype B(5a) and C (5b) 4a)Subtype B

Codon	dS	dN	dN/dS	Normalized dN-dS	p-value
4	0	0.26109	Infinite	0.01093	0.00072
6	0	0.11999	Infinite	0.00502	0.02921
7	1.08199	2.13551	1.974	0.0441	0.00259
24	1.3935	4.13474	2.967	0.11474	9.02E-07
32	0.37976	2.50879	6.606	0.08912	2.92E-08
39	1.86091	3.12232	1.678	0.0528	0.01533
40	1.19826	2.64097	2.204	0.06039	0.00025
42	0.44096	0.93895	2.129	0.02084	0.00821
47	0	0.32407	Infinite	0.01356	0.00391
58	3.19E-15	3.95358	1.2E+15	0.16549	0
59	0	1.4852	Infinite	0.06217	6.76E-14
61	0.79398	2.63138	3.314	0.07691	4.85E-07
62	0.46742	1.32344	2.831	0.03583	0.00294
63	1.77934	3.00339	1.688	0.05124	0.01364
64	0.34998	2.71356	7.753	0.09893	1.61E-13
65	0	1.13187	Infinite	0.04738	6.63E-08
68	0	2.28148	Infinite	0.0955	0
70	0.25966	2.49759	9.619	0.09367	5.55E-15
73	0	0.31138	Infinite	0.01303	4.05E-05
74	2.24711	3.31207	1.474	0.04458	0.00872
75	0	1.06386	Infinite	0.04453	5.40E-14
77	1.28366	3.88778	3.029	0.109	4.08E-11
80	0	0.36209	Infinite	0.01516	0.00073
81	0.13846	0.84292	6.088	0.02949	7.70E-05
84	0	0.67209	Infinite	0.02813	6.82E-07
85	0.19831	1.38755	6.997	0.04978	1.00E-05
86	0	0.37184	Infinite	0.01556	0.00157
87	0.09124	2.55542	28.008	0.10314	0
88	0	0.73766	Infinite	0.03088	1.07E-05
90	0.13217	0.57497	4.35	0.01853	0.01284
93	0.34624	2.46781	7.127	0.0888	3.46E-12
95	0.11018	0.49713	4.512	0.0162	0.00278
97	0.72291	1.91335	2.647	0.04983	0.0001
98	0.06485	1.34701	20.772	0.05367	1.97E-07
100	1.56697	7.53994	4.812	0.25001	0

Table 6: Site under positive selection by iFEL method in subtype B (6a) and C (6b)

Codon	dS	dN	dN Leaves	dN/dS	Normalized dN-dS	p-value
7	1.08201	2.12779	2.13764	1.967	0.04377	0.04202
24	1.42378	5.66242	3.65908	3.977	0.17742	1.30E-07
32	0.38051	2.98721	2.36496	7.851	0.10911	2.11E-07
39	1.87096	3.47108	3.01451	1.855	0.06698	0.01897
40	1.19903	2.76516	2.60479	2.306	0.06555	0.00577
42	0.44079	1.14548	0.88401	2.599	0.0295	0.02491
58	0	5.38848	3.53435	Infinite	0.22555	0
59	0	0.94903	1.63235	Infinite	0.03972	2.77E-06
61	0.79523	2.98052	2.52858	3.748	0.09147	3.44E-05
62	0.46735	1.29419	1.33154	2.769	0.03461	0.02804
63	1.7805	3.45885	2.8745	1.943	0.07025	0.02071
64	0.34908	2.28587	2.8417	6.548	0.08107	8.04E-06
65	0	1.26826	1.0944	Infinite	0.05309	1.16E-06
68	0	2.57985	2.19288	Infinite	0.10799	6.85E-14

70	0.2597	3.03509	2.33262	11.687	0.11617	2.85E-08
73	0	0.16321	0.35041	Infinite	0.00683	0.0249
74	2.247	4.87607	2.87592	2.17	0.11005	0.00026
75	4.10E-17	1.06864	1.06249	2.6E+16	0.04473	4.11E-08
77	1.28296	2.9234	4.16761	2.279	0.06866	0.00513
80	0	0.31367	0.3748	Infinite	0.01313	0.00856
81	0.13841	0.54947	0.92368	3.97	0.01721	0.04938
84	0	0.35146	0.75671	Infinite	0.01471	0.00379
85	0.19829	1.34688	1.39876	6.792	0.04808	0.00068
87	0.09095	2.24841	2.64267	24.721	0.09031	1.65E-08
88	2.90E-16	0.23197	0.87465	8E+14	0.00971	0.04359
90	0.13241	0.79305	0.51645	5.989	0.02765	0.00903
93	0.33512	3.28117	2.24416	9.791	0.12331	2.19E-09
98	0.06443	1.12169	1.40951	17.411	0.04425	0.00012
100	1.58732	10.6766	6.541	6.726	0.38046	0

5b)Subtype B

Codon	dS	dN	dN Leaves	dN/dS	Normalized dN-dS	p-value
4	0.07826	2.34093	1.27402	29.912	0.1925	1.67E-06
21	0.69618	2.64321	1.51427	3.797	0.16565	0.00292
29	2.25573	10.6737	6.69246	4.732	0.71617	2.26E-07
39	0.44023	2.77875	0.93268	6.312	0.19895	0.00037
40	0.73899	2.59738	2.35517	3.515	0.1581	0.00696
58	0.07549	2.29101	1.52913	30.348	0.18849	4.70E-06
67	0.31341	2.01221	2.02799	6.42	0.14453	0.00245
68	0	3.80073	2.94594	Infinite	0.32335	3.09E-11
69	0.1082	3.06838	2.03639	28.357	0.25184	5.67E-07
70	0.06815	1.17413	1.24805	17.23	0.09409	0.0029
75	0	3.52053	1.67304	Infinite	0.29951	1.22E-12
77	0.59912	1.71616	1.89864	2.864	0.09503	0.04187
80	0.07268	2.46722	1.81151	33.948	0.20372	7.45E-07
81	1.00E-06	3.90974	2.0962	3909740	0.33263	3.21E-11
95	0	0.67051	0.65981	Infinite	0.05704	0.01594
97	0.07123	1.0356	2.21379	14.538	0.08204	0.00796
100	0.06539	2.17681	2.16339	33.288	0.17963	4.09E-05