

# Amino acid sequence divergence of Tat protein (exon1) of subtype B and C HIV-1 strains: Does it have implications for vaccine development?

Abraham Joseph Kandathil<sup>1</sup>, Rajesh Kannangai<sup>1\*</sup>, Oriapadickal Cherian Abraham<sup>2</sup>, Susanne Alexander Pulimood<sup>3</sup> and Gopalan Sridharan<sup>1</sup>

<sup>1</sup>Department of Clinical Virology, <sup>2</sup>Department of Internal Medicine, <sup>3</sup>Department of Dermatology, Christian Medical College, Vellore, India; Rajesh Kannangai – Email: rajeshkannangai@hotmail.com; Phone: +914162282070; Fax: +914162232035; \*Corresponding author

Received November 01, 2009; Accepted November 15, 2009; Published December 12, 2009

## Abstract:

Functional genes of HIV-1 like the *tat* express proteins essential for viral survival and propagation. There are variations reported in levels of Tat transactivation among the different subtypes of HIV-1. This study looked at the amino acid differences in the different regions of Tat protein (exon 1) of subtype B and C strains of HIV-1 and tried to observe a molecular basis for protein function. HIV-1 sequences of subtype B (n=30) and C (n=60) strains were downloaded from HIV-1 Los Alamos data base. Among the 60 subtype C strain sequences, 30 each were from India and Africa. A HIV-1 Tat protein (exon 1) sequence, the consensus B and C sequence was obtained from the 'sequence search interface' in the Los Alamos HIV-1 sequence data. The sequences were visualized using Weblogo and the RNA binding regions of the three consensus sequences were also determined using BindN software program. Compared to subtype B, there was a high level of divergence in the auxiliary domain of *tat* exon 1 (amino acid positions 58- 69). The net charge of the subtype C (Indian) Tat protein (exon 1) auxiliary domain was -1.9 at pH 7 and it had an isoelectric point of 4.1. The net charge of the subtype C (African) auxiliary domain was -2.9 at pH 7 and it had an isoelectric point of 3.7 while the net charge of same region in subtype B was -0.9 at pH 7 with an isoelectric point of 4.9. The ratio of the hydrophilic residues to the total number of residues was 60% in the in both the Indian and African subtype C in the auxiliary domain while this was 50% in subtype B. The consensus subtype B sequence was found to have 36 RNA binding sites while subtype C (India) had 33 and subtype C (Africa) had 32 RNA binding sites. The HIV-1 Tat-TAR interaction is a potential target for inhibitors and being considered for its potential use in HIV-1 vaccines. Development of such inhibitor/vaccines would have to take into consideration the variation in amino acid sequence analyzed in this study as this could determine epitope presentation on MHC class I antigen for afferent immune response.

Key words: HIV-1, Subtype C, India, *tat*

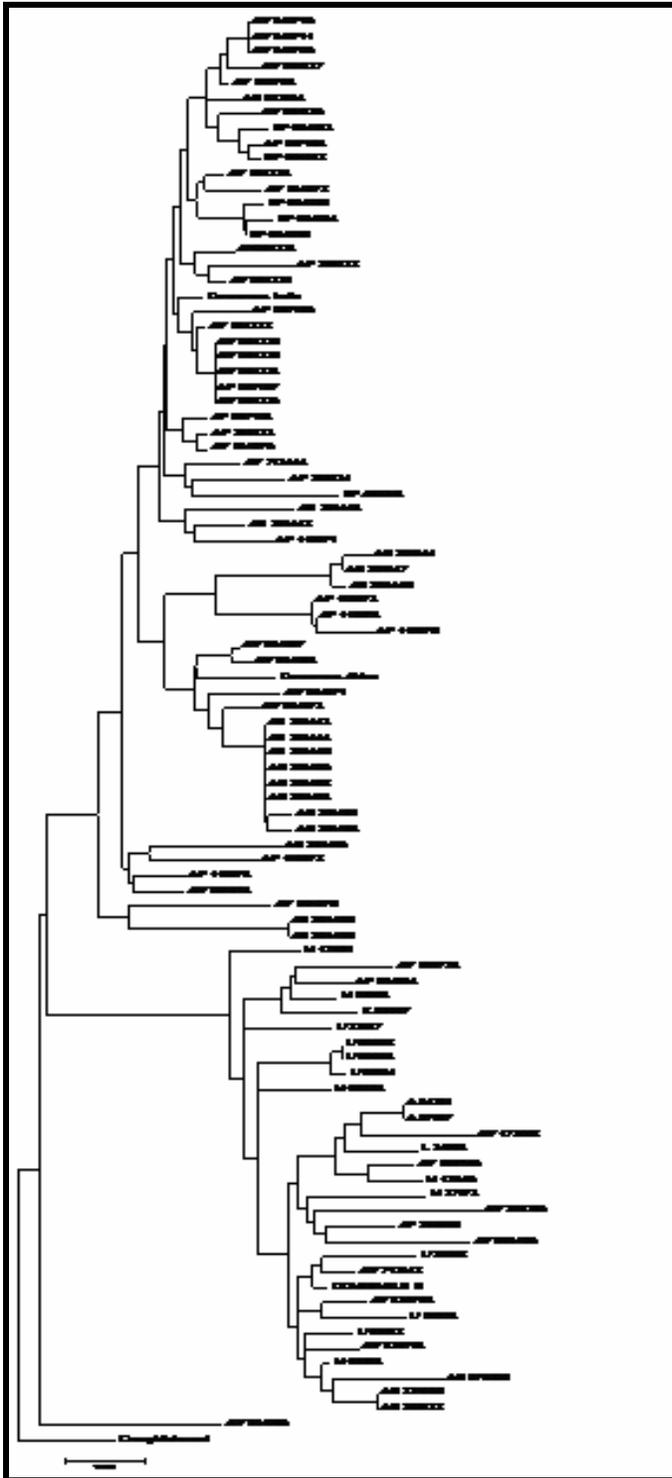
## Background:

The M group of human immunodeficiency virus-1 (HIV-1) is divided into nine non-recombinant subtypes (A-D, F-H, J, K) [1]. The amino acid variations in the *env* gene between the subtypes in the M group are 25-35 per cent, while in the *gag* gene it is about 15 per cent [1]. A recent study has shown that HIV disease progression can be influenced by the subtype [2]. In addition to structural genes HIV-1 has functional genes which express proteins essential for viral survival and propagation [3]. One such functional gene *tat* mediates an important role in transcription of the HIV-1 LTR [4]. Studies have also shown that there is a variation in levels of Tat transactivation among the different subtypes [5, 6]. The *tat* mRNA is a multiply spliced mRNA and consists of two coding (1 and 2) and one noncoding exon. Mutational analysis studies have shown that Tat protein can be functionally organized into different domains [7]. The study sequences encompassed the following domains: N-terminal domain (amino acid positions 1-20), Cys-rich domain (amino acid positions 21-40), Lys X Leu Gly Ile X Tyr motif (amino acid positions 41-48), basic domain (49-57) and the auxiliary domain (amino acid positions 58-67). Further studies on the Tat have shown that alteration of even one of the domains can affect the proper functioning of the Tat protein [7]. The Cys-rich domain has been suggested to be important for protein dimerization whereas the basic domain contains the arginine-rich RNA binding motif (ARM) and acts as a nuclear localization signal [2]. The auxiliary domain is believed to contribute to Tat activity by structural stabilization or by direct functional contribution [7]. We have used Tat (exon1) HIV-1 sequences of subtype B and subtype C strains obtained from GenBank and attempted to observe for amino acid differences in the different regions of Tat protein (exon 1) of subtype B and C strains to find a molecular basis for differences in protein function.

## Methodology:

HIV-1 sequences of subtype B (n=30) and C (n=60) strains were downloaded from HIV-1 Los Alamos data base ([www.hiv.lanl.gov/content/sequence/HIV/](http://www.hiv.lanl.gov/content/sequence/HIV/)). Among the 60 subtype C strain sequences downloaded, 30 were from India and the other 30 were from Africa. The accession numbers and subtypes of the strains are mentioned in **Table 1**. Among the 30 subtype B strains downloaded, information on co-receptor usage was available for 10 strains. The accession numbers of 5 strains that utilized CCR5 co-receptors were M93258, U23487, U04908, M65024 and M68893 and accession numbers for the 5 that utilized CXCR4 were U39362, M17449, M17451, K02007 and L31963. The CCR5 strains were considered as NSI and the CXCR4 were considered as SI for this study. The accession numbers of the CXCR4 and CCR5 utilizing strain were obtained from the dataset used to construct classifiers for Wetcat, which allows determination of HIV-1 co-receptor usage (<http://genomiac2.ucsd.edu:8080/wetcat/v3.html>).

A HIV-1 Tat protein (exon 1) sequence was obtained from the 'sequence search interface' in the Los Alamos HIV-1 sequence data <http://www.hiv.lanl.gov/components/sequence/HIV/search/search.html>. The consensus B and C sequence was also obtained from the Los Alamos HIV-1 sequence data using the following options "Alignment type: Subtype reference, Year: 2007, Organism: HIV1, Region: TAT, Subtype: All, DNA/Protein: DNA, Format: FASTA" (<http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html>). The sequences were downloaded on 10<sup>th</sup> June 2009. The subtype C reference sequence, accession number AF067155 (India) and AY772699 (Africa), was used for alignment of the 60 sequences obtained from GenBank. The nucleotide sequences were translated using the ExPasy translate tool ([http://sosnick.uchicago.edu/translate\\_dna.html](http://sosnick.uchicago.edu/translate_dna.html)). The alignment of the nucleotide and amino acid sequences obtained was done using clustalW (<http://www.ebi.ac.uk/Tools/clustalw/>).



**Figure 1:** The phylogenetic tree constructed using HIV-1 Tat protein (exon 1) amino acid sequences of subtype B and clade C strains. The evolutionary history was inferred using the Minimum Evolution method. The bootstrap consensus tree inferred from 500 replicates is taken to represent the evolutionary history of the taxa analyzed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) is shown next to the branch. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site. The ME tree was searched using the Close-Neighbor-Interchange (CNI) algorithm at a search level of 1. The Neighbor-joining algorithm was used to generate the initial tree. All positions containing gaps and missing data were eliminated from the dataset.

The sequences were visualized using Weblogo (<http://weblogo.berkeley.edu/logo.cgi>). The net charges of the amino acid sequence of the consensus subtype C and B sequences were also calculated (<http://www.innovagen.se/custom-peptide-synthesis/peptide-property-calculator/peptide-property-calculator.asp>). The RNA binding regions of the three consensus sequence (subtype C from India and Africa and subtype B) were also determined using BindN

(<http://bioinfo.ggc.org/bindn/>) [8]. A specificity of 80% was used, by default, for determining the RNA binding regions. Phylogenetic analysis was performed for the amino acid sequences with Mega 4 software using a minimum evolution method and bootstrap value of 500 replicates [9]. The evolutionary distances were computed using the Poisson correction method.

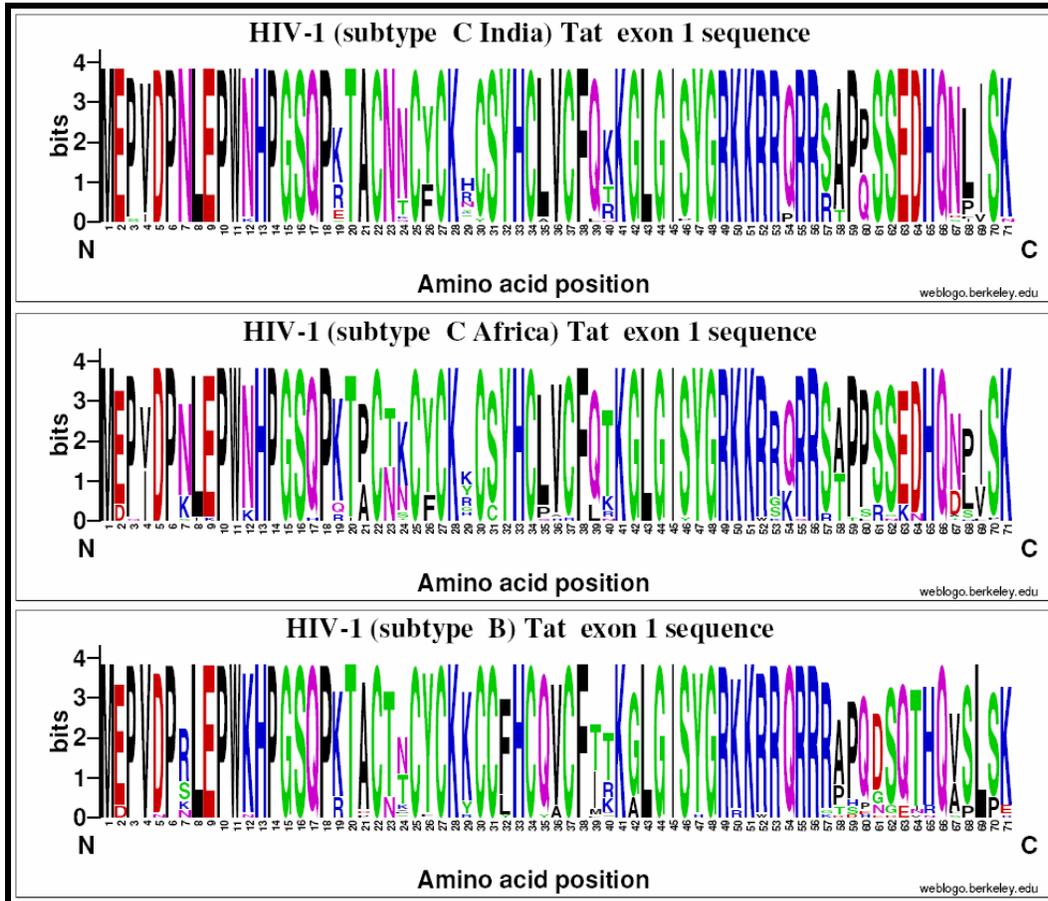


Figure 2: The amino acid characters for all the subtype C from India (n=30) **fig 2a**, Africa (n=30) **fig 2b** and clade B (n=30) **fig 2c**. The height depicts the relative proportion of the amino acid at a site. The taller the logo the lesser the variability at the site. The x axis depicts the amino acid position. The seq logo was developed at <http://weblogo.berkeley.edu/logo.cgi>.



Figure 3: The RNA binding sites as predicted by BindN (<http://bioinfo.ggc.org/bindn/>). The binding residues are labeled with '+' and non-binding residues labeled with '-' and in green. The confidence is denoted from level 0 (lowest) to level 9 (highest).

**Results:**

When compared to subtype B, subtype C strains from India showed variation at amino acid positions 7, 19, 29,35, 39, 57,60,61,63,64,67,68 and 69. In the phylogenetic tree the study sequences were shown to form two divergent clusters which were composed exclusively of either subtype B or subtype C strains (Figure 1). When compared to subtype B, subtype C strains from Africa showed variation at amino acid positions 7, 12, 19, 24,29,31,35, 39, 53, 54 57,60,61,63,64,67,68 and 69. When compared to subtype C strains from Africa, subtype C strains from India showed variation at amino acid positions 2,7, 21, 23, 29, 31, 36, 37, 53, 60, 61 – 64 and 71. In this study, there was a high level of divergence in the auxiliary domain of subtype B and C sequences (amino acid positions 58- 69). Among the Indian subtype C, 30 strains (100%) had Asn at position 7 as compared to Arg, Ser, Lys and Asn in the consensus subtype B and C (African) sequence. The other amino acid changes between Indian subtype C and subtype B are as follows (in parenthesis indicates the amino acid present in the consensus subtype B strain): 90% showed Asn (Lys) at position 12, 90% showed Gln (Thr,Ile) at position 39, 93% showed Ser (Arg) at position 57, 97% showed Pro (Gln) at position 60, 87% showed Ser (Asp) at position 61, 90% showed Glu (Gln) at position 63, 100% showed Asp (Thr) at position 64, 97% showed Asn (Val) at position 67, 80% showed Leu (Ser) at position 68 and 93% showed Ile (Leu) at position 69. The amino acid sequences of subtype B and C can be visualized as a sequence logo in Figure 2. When comparing the amino acid sequences of SI and NSI subtype B strains there were no changes that could be classified as being distinct for each group.

The net charge of the consensus subtype C (Indian) Tat protein (exon 1) was 8 at pH 7 and it had an isoelectric point of 9.6. The net charge of the consensus subtype C (African) Tat protein (exon 1) was 5 at pH 7 and it had an isoelectric point of 8.8. While the net charge of the consensus subtype B Tat protein (exon 1) was 11 at pH 7 and it had an isoelectric point of 10.1. The molecular weight of the Tat protein (exon 1) consensus subtype B was 8235.6. The molecular weight of the Tat protein (exon 1) consensus subtype C, Indian and African based on the amino acid sequences was 8162.4 and 8194.3, respectively. The ratio of the hydrophilic residues to the total number of residues was 49% in subtype C (Indian) which was higher than subtype B(44%) and African subtype C (45%). The net charge of the subtype C (Indian) Tat protein (exon 1) auxiliary domain (amino acid positions 58-67) was -1.9 at pH 7 and it had an isoelectric point of 4.1. The net charge of the subtype C (African) Tat protein (exon 1) auxiliary domain (amino acid positions 58-67) was -2.9 at pH 7 and it had an isoelectric point of 3.7. The ratio of the hydrophilic residues to the total number of residues was 60% in the in both the Indian and African subtype C in the auxiliary domain. The net charge of same region in subtype B was -0.9 at pH 7 and it had an isoelectric point of 4.9. The ratio of the hydrophilic residues to the total number of residues was 50% in subtype B. The molecular weight of the auxiliary domain of the Tat protein (exon 1) consensus subtype B was 1110.1. The molecular weight of the Tat protein (exon 1) subtype C based on the amino acid sequences at amino acid positions 58-67 for the Indian and African sequences were 1081.1 and 1112.1, respectively. The predicted RNA binding regions are shown in Figure 3. The consensus subtype B sequence was found to have 36 RNA binding sites while subtype C (India) had 33 and subtype C (Africa) had 32 RNA binding sites. The predominant predicted RNA binding sites for subtype B was from amino acid positions 46-64, while for the subtype C both from Africa and India it was in amino acid positions 46-63.

**Discussion:**

The amino acid sequences of Tat protein (exon 1) from subtype B and C strains showed distinct differences. This was observed in the phylogenetic tree where subtype B and C diverged and formed distinct clusters, showing a certain level of distinct evolution. There was also a difference in the ratio of the hydrophilic residues to the total number of residues between the subtypes. There was one amino acid change at position 57 in 13 subtype C strains on comparison to subtype B strains. Ninety three percent of clade C strains from India showed Ser at position 57 instead of Arg which is seen at position 57 among 90% subtype B strains. Serine has uncharged and small side chains as compared to Arg. The basic domain in which these changes were seen is responsible for the transportation of the Tat to the nucleus and nucleolus<sup>7</sup>. The implication of this change in the basic domain of Tat possibly affects relocation to the nucleus but will have to be further investigated. Most cytosol proteins function optimally when their isoelectric point is close to the ambient pH which in a resting T cells is 7.2. The isoelectric point of subtype C (India) Tat protein (exon 1) protein (9.6) is closer to 7.2 as compared to subtype B (10.1) and may be hence more efficient in transactivation of LTR element. The subtype C (African) Tat protein (exon 1) had an isoelectric point of 8.8 which is the closest to 7.2 compared to the values of the other consensus sequences uses in this study. The HIV-1 Tat-TAR (transactivation responsive) interaction is a potential target for inhibitors [10]. The Tat protein is also being considered for its potential use in HIV-1 vaccines [11]. This could be potentially important especially with the recent failure of the Merck HIV-1 vaccine trials [12]. The Merck vaccine made use of sequences from the *gag*, *env* and the *nef* region. Although Tat based vaccines would not be able to prevent an individual acquiring infection it could block viral replication and disease onset (post exposure vaccine) [11]. *In-vitro* studies have shown subtype C Tat protein to have a higher transactivational potential than subtype B [5, 6, 13]. Development of such inhibitor/vaccines would have to take into consideration the variation of the Tat proteins not only in terms of their activity but also the variation in amino acid sequence analyzed in this study which could determine epitope presentation on MHC class I antigen for afferent immune response.

**Reference:**

- [1] AJ Kandathil *et al.*, *Indian J Med Res.* **121**: 333 (2005) [PMID: 15817947]
- [2] N Kiwanuka *et al.*, *J Infect Dis.* **197**: 707 (2008) [PMID:18266607]
- [3] WK Wang *et al.*, *J Microbiol Immunol Infect.* **33**:131 (2000) [PMID: 11045374]
- [4] J Brady, F Kashanchi, *Retrovirology* **2**: 69 (2005) [PMID: 16280076]
- [5] T Kurosu *et al.*, *Microbiol Immunol.* **46**: 787 (2002) [PMID: 12516777]
- [6] Y Desfosses *et al.*, *J Virol.* **79**: 9180 (2005) [PMID: 15994812]
- [7] M Kuppuswamy *et al.*, *Nucleic Acids Res.* **17**: 3551 (1989) [PMID: 2542902]
- [8] L Wang, SJ Brown, *Nucleic Acids Res.* **34**:W243 (2006) [PMID: 16845003]
- [9] K Tamura *et al.*, *Mol Biol Evol.* **24**: 1596 (2007) [PMID: 17488738]
- [10] M Yang *Curr Drug Targets Infect Disord.* **5**: 433 (2005) [PMID: 16535863]
- [11] A Caputo *et al.*, *Curr HIV Res.* **2**: 357 (2004) [PMID: 15544457]
- [12] RP Sekaly *J Exp Med.* **205**: 7 (2008) [PMID: 18195078]
- [13] P Roof *et al.*, *Virology* **296**: 77 (2002) [PMID: 12036319]

Edited by P. Kanguane

Citation: Kandathil *et al.*, Bioinformation 4(6): 237-241 (2009)

**License statement:** This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

## Supplementary material:

**Table 1:** The Genbank Ids and the subtype of the HIV-1 Tat (exon1) sequences downloaded from the HIV-1 Los Alamos data base

| Subtype             | GenBank Ids  |
|---------------------|--|
| Subtype C<br>India  | AF067158, AF286223, AB023804, AF286232, AF067154, EF694033, AY064229, AY064224, AY049709, AY049710, AY049711, AY064227, AY049708, AY064223, AY064226, AY064220, AF067157, AY064225, AY064222, AY064221, AY064228, AF067159, AF067155, AY713414, EF694034, EF694036, EF694035, AF286231, EF469243, EF694032 |
| Subtype C<br>Africa | AB254142, AY901972, AB254148, AF110971, AB254149, AY901974, AF110972, AF110978, AY901970, AB254141, AB254147, AB254146, AF110973, AF110974, AF110975, AB254151, AB254153, AB254152, AB254150, AB254145, AB254144, AB254143, AB254154, AY901971, AY901973, AY901967, AY901968, AY901969, AB254155, AB254156 |
| Subtype B           | A04321, A07867, M17451, U23487, M65024, U69592, U69593, U69591, U63632, U39362, AY713412, AB221005, AB253432, M68893, AY835758, AB078005, AY835763, U04908, AY835749, M17449, M93258, K02007, AF004394, AY835774, AY173952, AY560109, L31963, M37573, AF286365, AY308760                                   |