Distinct Sequence Patterns Characterize the V3 Region of HIV Type 1 gp120 from Subtypes A and C

KLÁRA FELSÖVÁLYI[1], **ARTHUR NÁDAS**[2], **SUSAN ZOLLA-PAZNER**[2], and **TIMOTHY CARDOZO**[1]

1 *Department of Pharmacology and*

2 *Department of Pathology, New York University School of Medicine and the New York Veterans Affairs Medical Center, New York, New York 10016.*

Abstract

The known sequences of HIV-1 viruses have been categorized into subtypes based on the phylogenetic partitioning of their *env* and *gag* gene sequences. The *env* gene encodes the protein gp120, which contains five sequence-variable regions (V1 to V5), of which the V3 loop is of central importance to viral infectivity. The V3 loop consensus sequences of HIV-1 subtype A and C viruses are similar, and more similar to one another than the V3 consensus sequences of any other two HIV-1 subtypes. However, using a position-specific statistical comparison, we found that the V3 region of these two subtypes is statistically distinct ($p = ~0.0$). (The *p*-value calculated to the lowest limit of representation on the computer used to run the calculation. This lowest limit was 10^{-16}. Although theoretically a *p*-value cannot be equal to 0.0, the *p*-value for the comparisons in question can be intuitively considered to be extremely small, or ~0.0.)

The protein gp120 encoded by *env* requires interaction with CD4 and chemokine receptors in the membrane of target cells for HIV-1 to infect cells. Deletion of the V3 loop of gp120 abolishes virus infectivity.[1] The V3 loop interacts with the chemokine receptors, and specific V3 residues dictate the choice of coreceptor usage.[2] Coreceptor utilization is often predicted using HIV-1 envelope sequence data, with viral tropism, or "phenotype," being assigned on the basis of the sequence of the V3 loop.[3] These well-documented findings provide evidence that the V3 loop plays a critical role, through direct chemokine receptor interaction, in virus infectivity. Several studies have also demonstrated that the V3 loop of gp120 contains epitopes susceptible to neutralization by broadly reactive antibodies.[2]

Distinct subtype definitions for HIV-1, based on full length *env* sequences of worldwide strains, have previously been established.[4] The consensus V3 loop sequences of most subtypes are distinct, but the consensus V3 sequences of subtypes A and C were originally thought to be identical. More recently, expanded data indicate that the consensus sequences of these two HIV-1 subtypes are indeed quite similar, differing at only one position (as compared to the consensus sequences of subtypes A and B, which differ at five positions, or B and D, which differ at four).[5]

Significant similarity in the V3 sequences of subtypes A and C, the two most prevalent subtypes causing the HIV pandemic, might imply that these two subtypes could be treated as one with respect to structural studies of the V3 loop and with respect to the design of vaccines that will induce neutralizing antibodies to the V3 loop. However, inspection of the sequences of many isolates from these two subtypes reveals that the V3 region of subtypes A and C do, in fact,

Address reprint requests to: *Timothy Cardozo, 550 First Avenue, MSB 497A, New York, New York 10016, E-mail:* timc@saturn.med.nyu.edu.

differ at several key positions (Table 1). This study was undertaken to determine whether a statistically significant difference exists in the sequences of the V3 loop as it appears in HIV-1 subtype A and C viruses.

Differences between subtypes is usually quantified by bootstrap value,[6] but there is no established significance threshold for this procedure, and it does not allow identification of the key positions that may underlie the statistical difference between two globally related sequence groups. We used a classical statistical approach to compare the amino acid distributions at each position in the V3 loop between subtype A and subtype C to quantify the difference, if any, between these two groups.

All of the subtype A and subtype C V3 loop sequences were downloaded from the Los Alamos HIV sequence database (http://www.hiv.lanl.gov).[5] Duplicates were removed and 3562 unique sequences from subtype A and 2704 unique sequences from subtype C (as defined by LANL annotation) were organized into separate multiple sequence alignments using the multiple sequence alignment assignments of the LANL database. An equivalence was established between the two multiple sequence alignments at each of the 35 common V3 loop positions such that the first cysteine was labeled position 1, the terminal cysteine was labeled position 35, and each homologous position in between was numbered accordingly. (The GPG of the crown is thus numbered 15–17.) Gaps in the individual sequences were removed, leaving only the aligned 35 V3 loop positions. The number of occurrences of each amino acid was recorded at each of these 35 common positions in the V3 loop for each alignment (not shown). These counts were also expressed as the percentage of the total number of viruses in the subtype for display purposes [e.g., threonine (T) occurs in position 2 of the V3 loop in 59% of subtype A viruses; Table 1]. A comparison of the amino acid distributions at each of the 35 positions was then made independently using the chi-squared test in a two-by-two table.

1. To assess whether the two probability distributions of amino acids at each position in the two subtypes are different, it is not necessary to compare the distributions in every respect; rather it suffices to compare some interesting or important features. The comparison of amino acid counts at equivalent positions in the two different sets (subtypes) can be made by comparing the two categories of consensus and nonconsensus sequences, instead of using 20 categories for the 20 amino acids. Thus, for example, for position 2 of the V3 loop, instead of comparing the two boxed columns of 20 elements seen in Table IA (which would be an invalid chi-square comparison since the occurrence of many amino acids would be zero), we compared just the number of occurrences of the consensus amino acid at position 2 in subtype A (threonine, 59%), the nonconsensus counts of subtype A (the other 19 amino acids added up = 41%) and the corresponding counts at that position in subtype C. Thus, the two-by-two table for the chi-square comparison at position 2 is as follows:

	Consensus	*Non-consensus*
Subtype A	2072 threonines (59%)	1489 other amino acids (41%)
Subtype C	2202 threonines (76%)	502 other amino acids (24%)

This simplified comparison matches more specifically the question we are asking in the first place: do the two *consensus* sequences differ.

2. This simplification results in the comparison of two Bernoulli (two-dimensional multinomial) distributions, one for each subtype, with *the null hypothesis being that the two distributions of consensus V3 loop amino acids do not differ*.

3. The interdependence between positions in the V3 loop is significant,[7] requiring a Bonferroni-like correction of the *p*-value for this lack of independence between nearby positions along the sequence. Thus, the significance threshold for this comparison is much less than $p < 0.05$.

The chi-squared tests were computed in S-PLUS 7.0 (Insightful Corp.) using the program "chisq.test" provided therein. Note that the calculation was done on amino acid counts and not the percentages shown in Table 1. This was done in order to satisfy the minimum expected count threshold of valid chi-square tests. The minimum expected count for amino acids in the two-by-two tables ranged between 17 and 1244 over the 35 positions, thus amply satisfying the traditional requirement that such counts should exceed five.

If a *p*-value shows a statistically significant difference at any one of the 35 positions of the V3 loop, it may be concluded quantitatively that the consensus sequences, and therefore the amino acid makeup, of the V3 loops of subtype A and C are different. The result of our position-specific analysis is that the null hypothesis is rejected for 9 of the 35 positions with a *p*-value of essentially 0.0 (the *p*-value calculated below the limit of representation on the particular computer used, which read out as "$< 10^{-16}$"), well below any conceivable threshold of statistical significance (Table 2). Since the null hypothesis was established with no difference between subtype A and C V3 loop consensus sequences, the results show that there *is* a statistically significant difference between the two sets. In addition, the specific amino acid positions responsible for the difference are distributed throughout the V3 loop, indicating that the differences observed are not an anomaly related to a single site or to boundary effects.

The differences in the amino acid distributions between subtypes A and C at these positions do not represent differences in the distribution of electrostatic charges between the subtypes. For example, there are no positions at which a positively charged lysine, arginine, or histidine is prevalent in subtype A while a negatively charged aspartate or glutamate is prevalent at that same position in subtype C. Instead, the differences derive from relatively charge-conservative alterations in the amino acid distribution at a particular position, such as T to A at position 19 or H to R at position 13 (Table 1).

A large portion of the crown of the V3 loop has been resolved in three dimensions for several subtype B sequences and is a β-hairpin structure regardless of the sequence variation.[2] At least one conformation of the whole V3 loop including this β-hairpin has also been published,[8] allowing us to map the specific amino acid positions responsible for the differences we see statistically. These amino acid positions are clustered at the extreme stems and at the crown or tip of the V3 loop (Fig. 1). The majority of the positions between the stem and crown do not appear to vary between the subtypes. Interestingly, the specific amino acid positions responsible for the difference occur at symmetric positions in the parallel β-strands of the stem and crown.

This study was prompted by observations that the human antibody response to V3 is different in individuals infected with subtypes A and B.[9] Indeed, the amino acid distribution between the V3 regions of various HIV subtypes is well recognized. For example, the database of HIV sequences establishes that while the vast majority of subtype B V3 sequences contain the GPGR motif at the tip of the loop, GPGQ is the more common motif at the tip of V3 in all of the other HIV-1 subtypes.[5] This latter R/Q change constitutes a major antigenic distinction.[10] Another example includes the finding that in contrast to the V3 region of subtype B, the V3 of subtype C is far less variable.[11]

Subtypes A and C were originally thought to have identical consensus V3 loops and their consensus sequences are now known to be extremely similar. This similarity could indicate that the immune response to this region of these two subtypes might also be similar, implying

that immune responses induced by a vaccine to the V3 of subtype A might induce neutralizing antibodies against subtype C as well. However, on the basis of the results presented above, the patterns of amino acid usage in the V3 loops of these two virus subtypes differ significantly, implying that it is important to determine if there are differences in the type of anti-V3 antibodies induced by viruses carrying the subtype A vs. the subtype C envelopes. Such studies should reveal if there are anti-V3 antibodies that cross-react with and cross-neutralize subtype A and C primary isolates (as there are between subtypes A and B), or if subtype-specific antibodies predominate. This information could have significant implications for vaccine design.

These data may also illuminate the structural basis of the biological differences between viruses of subtypes A and C. In the β-strand structural fold of the V3 loop (Fig. 1), symmetric positions tend to be part of a single continuous protein surface despite their separation in the primary sequence. This suggests that the symmetrically located side chains of the crown and stem that vary significantly between the subtypes form continuous exposed surfaces that differ in shape between the subtypes. As the amino acid differences we found are charge conservative, the underlying β-strand structure and the electrostatic nature of these surfaces probably do not, on the other hand, alter significantly between the subtypes.[5] The shape difference might be sufficient to confer escape of one subtype from antibodies that bind this surface in the other subtype. The V3 loop is in fact essential for infectivity,[1] and therefore variably shaped, exposed residues in V3 might be selected to confer better viral fitness. Therefore, one possible interpretation of these results is that the crown and extreme stem of the V3 loop are exposed fragments in the unliganded or liganded forms (or both) of gp120 and therefore subjected to the evolutionary pressures that have produced the various virus subtypes.

References

1. Chiou SH, Freed EO, Panganiban AT, Kenealy WR. Studies on the role of the V3 loop in human immunodeficiency virus type 1 envelope glycoprotein function. AIDS Res Hum Retroviruses 1992;8:1611–1618. [PubMed: 1457207]

2. Zolla-Pazner S. Identifying epitopes of HIV-1 that induce protective antibodies. Nat Rev Immunol 2004;4:199–210. [PubMed: 15039757]

3. Hoffman NG, Seillier-Moiseiwitsch F, Ahn J, Walker JM, Swanstrom R. Variability in the human immunodeficiency virus type 1 gp120 Env protein linked to phenotype-associated changes in the V3 loop. J Virol 2002;76:3852–3864. [PubMed: 11907225]

4. Louwagie J, Janssens W, Mascola J, et al. Genetic diversity of the envelope glycoprotein from human immunodeficiency virus type 1 isolates of African origin. J Virol 1995;69:263–271. [PubMed: 7983718]

5. HIV sequence compendium. Theoretical Biology and Biophysics Group. Los Alamos National Laboratory; Los Alamos, NM: 2005. LA-UR 04-7420

6. Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci USA 1996;93:7085–7090. [PubMed: 8692949]

7. Korber BT, Farber RM, Wolpert DH, Lapedes AS. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. Proc Natl Acad Sci USA 1993;90:7176–7180. [PubMed: 8346232]

8. Huang CC, Tang M, Zhang MY, et al. Structure of a V3-containing HIV-1 gp120 core. Science 2005;310:1025–1028. [PubMed: 16284180]

9. Krachmarov C, Pinter A, Honnen WJ, et al. Antibodies that are cross-reactive for human immunodeficiency virus type 1 clade A and clade B v3 domains are common in patient sera from

Cameroon, but their neutralization activity is usually restricted by epitope masking. J Virol 2005;79:780–790. [PubMed: 15613306]

10. Zolla-Pazner S, Zhong P, Revesz K, et al. The cross-clade neutralizing activity of a human monoclonal antibody is determined by the GPGR V3 motif of HIV type 1. AIDS Res Hum Retroviruses 2004;20:1254–1258. [PubMed: 15588347]

11. Gaschen B, Taylor J, Yusim K, et al. Diversity considerations in HIV-1 vaccine selection. Science 2002;296:2354–2360. [PubMed: 12089434]
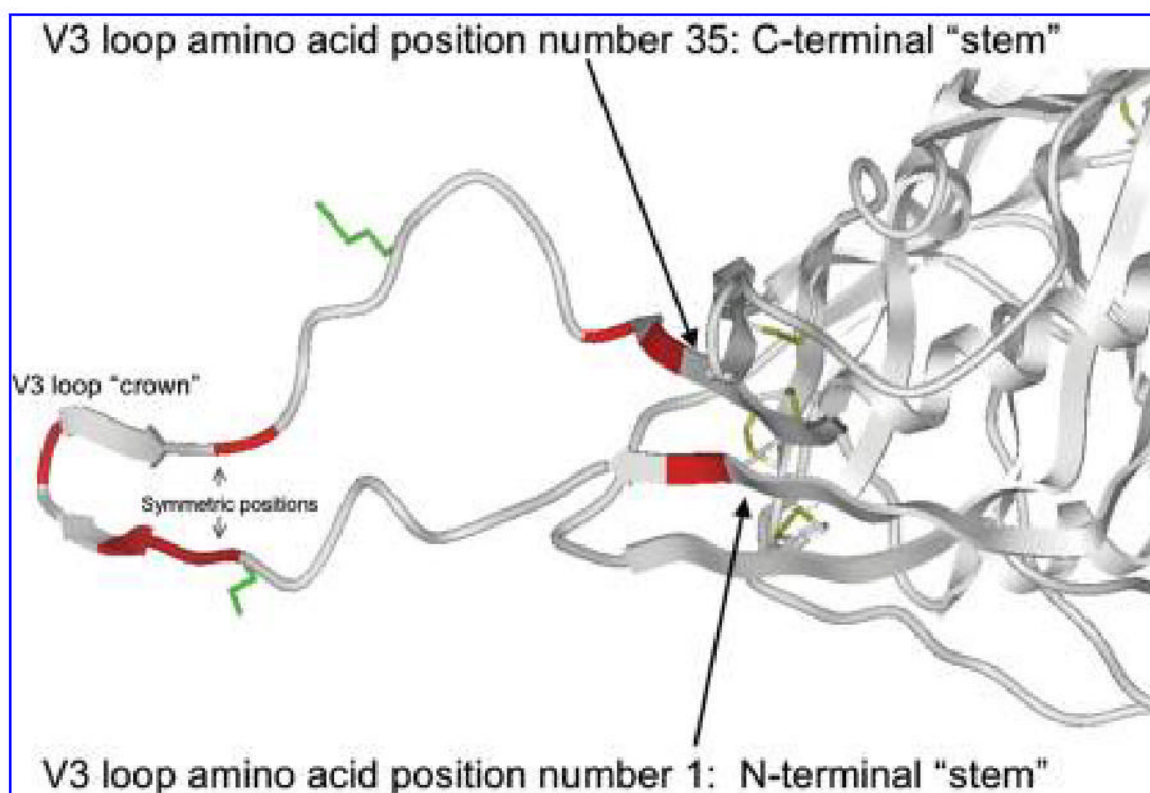
FIG. 1.
Location of statistically significant differences between subtype A and subtype C on the 3D structure of gp120 and the V3 loop. The crystallographically derived 3D model of gp120 including the V3 loop from Huang *et al*.[8] is displayed in ribbon depiction. Amino acid positions in the V3 loop found in this study to differ significantly between subtypes A and C are colored red. The coreceptor interacting 11 and 25 positions of the V3 loop are displayed in green stick depiction. Disulfide bridges in the gp120 core are also displayed in stick depiction.

Table 1

Distribution of Amino Acids Found at Each Position in the V3 Loop in All Unique Clade A and Clade C Sequences from the LANL Database[a]

A. Distribution of amino acids at each position in the V3 loop

Subtype A — V3 position

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
G	0	0	0	0	19	0	0	0	0	0	7	0	0	0	97	0	99	0	0	0	0	0	0	83	5	0	0	98	0	0	0	0	0	0	0
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	62	0	0	90	4	0	6	0	0	0	0	0	0	0	98	0	0
V	0	0	0	0	0	0	0	0	0	0	0	45	0	0	0	0	0	0	0	5	0	0	0	0	0	12	2	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96	3	0	0	0	0	0	0	0	0	0	0	0	0	2	0
P	0	0	0	97	0	0	0	0	0	0	0	0	2	0	0	98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
I	0	29	0	0	0	0	0	0	0	0	0	39	0	89	0	0	0	0	0	0	0	0	0	0	0	85	92	0	0	98	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	6	0	0	5	0	0	0	0	0	89	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
T	0	59	0	0	0	0	0	97	0	18	0	0	0	0	0	0	0	0	28	0	0	7	81	0	0	0	3	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	94	0	0	0	0	0	0	0	0	0	0	0	0	32	0
H	0	0	0	0	0	0	0	0	0	0	0	0	32	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	61	0
C	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99
N	0	0	0	0	72	97	98	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	6	1	2	0	0	0	19	0	0	0	0	1	0
Q	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	93	0	0	0	0	0	0	0	0	0	0	0	0	0	71	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	58	0	0	0	76	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	22	0	0	0
R	0	0	99	0	0	0	0	0	97	11	0	0	52	0	0	0	0	4	0	0	0	0	2	0	0	0	0	0	0	0	95	3	0	0	0

Subtype C — V3 position

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
G	0	0	0	0	24	0	0	0	0	0	0	3	0	4	82	0	99	0	0	0	0	0	0	71	11	0	0	96	0	0	0	0	0	0	0
A	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	93	1	0	9	0	0	0	0	0	0	1	97	0	0
V	0	3	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	3	0	0	0	0	6	1	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	1	96	0	0	0	0	0	0	0	0	0	0	0	81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	1	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	11	0	0	0	0	0	0	0	0	0	60	0	79	0	17	0	0	0	0	0	0	0	0	0	92	92	0	0	96	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	7	0	0	0	0	0	93	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
T	0	76	0	0	0	0	2	94	0	4	0	0	0	0	0	0	0	0	66	0	0	5	83	5	0	0	3	0	0	1	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0	0	0	21	0
H	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	74	0
C	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99
N	0	0	0	0	62	98	96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	14	4	0	0	0	18	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	96	0	0	0	0	0	0	1	0	0	0	0	0	0	78	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	49	0	0	0	77	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	1	17	0	0	0	1	0	0	6	0	0	0
K	0	0	0	0	0	0	1	0	0	68	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	1	8	0	0	0
R	0	1	98	0	0	0	0	2	94	10	0	0	91	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	1	0	0	97	2	0	0

B. Consensus sequences and percentages of subtype A and subtype C derived from (A)

Subtype A

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
C99	T59 I29	R99	P97	N72 G19	N97	N98	T97	R97	K64 T18 R11	S89 G7	V45 I39	R52 H32	I89	G97	P98	G99	Q93	A62 T28	F96	Y94	A90	T81	G83	D58 E12	I85 V12	I92	G98	D76 N19	I98	R95	Q71 K22	A98	H61 Y32	C99

Subtype C

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
C99	T76 I11	R98	P96	N62 G24 S7	N98	N96	T94	R94	K68 R10 E10	S93	I60 V22 M14	R91	I79	G82	P81 I17	G99	Q96	T66 A25	F95	Y86 F11	A93	T83	G71 N14	D49 E17 G11 A9	I92	I92	G96	D77 N18	I96	R97	Q78 K8 E6	A97	H74 Y21	C99

[a] Only the amino acids occurring more than 6% of the time are displayed for clarity.

Table 2

Statistically Significant Differences between Amino Acid Distributions at Indicated Amino Acid Positions in the V3 Loop

Amino acid position in the V3 loop	p-value
2	0.0
12	0.0
13	0.0
14	0.0
15	0.0
19	0.0
23	0.0
32	0.0
34	0.0