The A-nucleotide preference of HIV-1 in the context of its structured RNA genome

Formijn J. van Hemert,* Antoinette C. van der Kuyl and Ben Berkhout*

[1]Laboratory of Experimental Virology; Department of Medical Microbiology; Center for Infection and Immunity Amsterdam (CINIMA); Academic Medical Center; University of Amsterdam; Amsterdam, The Netherlands

A bipartition of HIV-1 RNA genome sequences into single- and double-stranded nucleotides is possible based on the secondary structure model of a complete 9 kb genome. Subsequent analysis revealed that the well-known lentiviral property of A-accumulation is profoundly present in single-stranded domains, yet absent in double-stranded domains. Mutational rate analysis by means of an unrestricted model of nucleotide substitution suggests the presence of an evolutionary equilibrium to preserve this biased nucleotide distribution.

Introduction

The tendency of lentiviral open reading frames to become A-rich has been documented previously.[1] For instance, the single-stranded RNA genome of HIV-1 contains 36.2% A, 23.9% G, 22.2% U and 17.6% C. The increased A-content dictates the typical codon usage of this virus.[2,3] The apparent selection of A-rich codons in HIV genomes even contributes to a biased amino acid composition of the encoded proteins.[4] Also, HIV particles contain tRNAs that decode A-ending codons, suggesting a modulation of the cellular tRNA pool toward the typical codon preference of HIV genes.[5] These basic RNA properties are well conserved over time and among the different members of the Lentiviridae.[6,7] dCTP pool imbalance during reverse transcription has been proposed as a cause of G→A hypermutation of the HIV-1 genome.[8,9] dNTP pool imbalance appeared to contribute more to HIV evolution in vivo than sequence editing by the cellular restriction factors Apobec 3G/3F.[10] A reduction of the A-richness of HIV-1 polymerase sequences impaired viral DNA synthesis,[11] but a biological function for this typical lentiviral A-pressure has not yet been elucidated.[12]

Recently, a secondary structure model of the complete 9 kb HIV-1 RNA genome at single nucleotide resolution has been constructed by means of a combined chemical assay of nucleotide accessibility (SHAPE, see ref. 13) and RNA folding prediction (RNAstructure, see refs. 14 and 15). The biased nucleotide composition of the HIV-1 RNA genome will definitely have some implications for the distribution of the different nucleotides over the structured RNA genome. Even when assuming maximal base pairing across the genome, the character of the possible base pairs (G-C, A-U and G-U and the reverse set of three) dictates that not every A can be paired given the 14% surplus of A (36.2%) over its unique pairing partner U (22.2%). This would mean that the

single-stranded regions of HIV-1 RNA will statistically have a surplus of A and possibly G over U and particularly C. As these patterns could constitute a distinct molecular signature of the viral genome, we set out to further analyze the nucleotide distribution in the context of the HIV-1 RNA secondary structure model.[13]

Results

Nucleotide composition of the structured HIV-1 RNA genome. The nucleotide composition differs significantly between single- and double-stranded regions of the HIV-1 RNA structure model of the NL4-3 isolate (**Table 1**). Of the total 9,173 nucleotides, 59% and 41% are present in these ss and ds regions, respectively. As much as 79% of A nucleotides in this HIV-1 RNA genome participate in the ss parts. In other words, almost four of five A nucleotides are predicted to be unpaired in this highly structured RNA molecule. In contrast, 57% of U, 45% of G and only 38% of C are found in ss regions. These striking data indicate a differential nucleotide bias in the ss vs. ds domains of HIV-1 RNA. Apparently, the lentiviral property of A-pressure at the expense of C as described previously[2,4] is intensified in the ss regions (A-rich and C-poor with 79% and 38%, respectively) but absent in the ds parts that show a strikingly reversed pattern (C-rich and A-poor with 62% and 21%, respectively). Analysis of the HIV-1 sequence after partition into separate reading frames and codon positions (GAG, POL, ENV and NEF, excluding regions with gene overlap) confirmed these patterns (**Table S1**). The combined 5' and 3' non-coding regions displayed twice as much paired nucleotides than the genes and a concomitant decrease in A-content.[16,17]

Analysis of the base pairs in the HIV-1 RNA secondary structure model indicates that the most stable GC and CG base pairs are used more frequently than AU and UA pairs (**Fig. 1**). The

Table 1. Biased nucleotide composition of single- and double-stranded regions in HIV-1 NL4-3 RNA

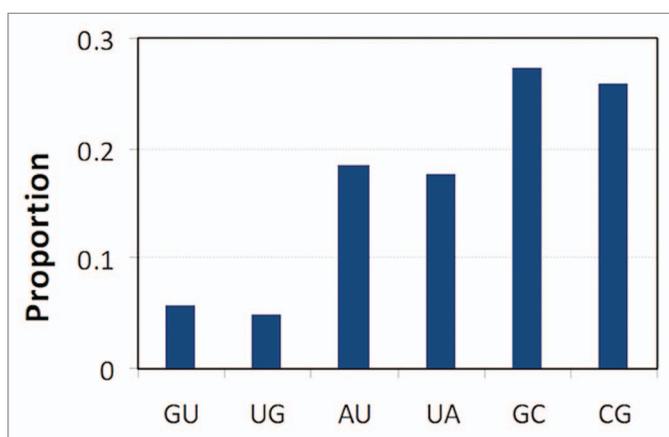| Nucl	ss ds	Number (proportion)		ss/ds
All	ss	5377	(0.59)	1.416
	ds	3796	(0.41)	
A	ss	2596	(0.79)	3.779
	ds	687	(0.21)	
U	ss	1155	(0.57)	1.304
	ds	886	(0.43)	
C	ss	623	(0.38)	0.616
	ds	1012	(0.62)	
G	ss	1003	(0.45)	0.828
	ds	1211	(0.55)	

Figure 1. Base pair composition of the double-stranded portion of the HIV-1 NL4-3 RNA structure.

least stable GU and UG pairs are present at an even lower frequency. This unequal base pair composition correlates with the slightly preferred occurrence of G and C in the ds parts of the HIV-1 genome (**Table 1**: 55% and 62%, respectively).

The structure of the NL4-3 RNA genome has been deduced from a combination of experimental RNA structure probing data and computational RNA structure prediction. In short, the SHAPE reactivity assay monitored the accessibility of nucleotides in the RNA structure by chemical base modification. These experimental data were fed into the RNA-folding software[14] to obtain a pairing probability value for each individual nucleotide.[13] We analyzed the distribution of the four nucleotides for SHAPE reactivity (**Fig. 2**). The A-nucleotides show a peak in SHAPE reactivity around 0.8, which contrasts with the much lower values calculated for the other three nucleotides. The SHAPE reactivity of the C-nucleotides is most restricted and largely confined to the 0.2–0.4 window. These results indicate that A nucleotides are in general more exposed to chemical modification than the other nucleotides because the most As are single-stranded, whereas Cs are best protected against the modifying agent by base pairing. These structure-probing results are in agreement with the biased

nucleotide composition based on the predicted HIV-1 RNA structure model: A is overrepresented in ss regions and C is found preferentially in ds domains (**Table 1**). This points to an intimate relationship between the nucleotide composition of the HIV-1 RNA genome and its structure.

The connection between base composition and secondary structure in the RNA genome of the HIV-1 strain NL4-3 may be exemplary for other virus isolates. To test this, the ss/ds designation of NL4-3 was projected onto the corresponding nucleotides of 448 aligned HIV-1 subtype B sequences taken from the Los Alamos database (year 2010, no recombinants). An ss/ds bipartition was created without affecting the individual base-to-base alignments. Indeed, nucleotide frequencies differ between these two data sets quite similarly as described above for NL4-3 RNA (**Table 2**). The small values for standard deviation (StD) indicate considerable conservation of the typical nucleotide composition in the ss and ds compartments. Apparently, the property of A-pressure at the expense of C is prominent in portions of HIV-1 subtype B RNAs that represent unpaired regions in NL4-3 RNA.

Different nucleotide substitution pattern in ss and ds domains of HIV-1 RNA. The strikingly different nucleotide composition of ss and ds RNA regions may point to different evolutionary rates of the nucleotides in these two domains. Maximum likelihood estimates of relative evolutionary rates for A, U, C and G nucleotides in ss and ds alignments confirmed this expectation (**Table 3**). A positive value indicates the substitution probability of a row nucleotide by one of the column nucleotides in the same row. A negative value on the matrix's diagonal represents the quantity to reduce the summarized values of the substitution probabilities in the same row to zero. From inspection of the Qss matrix, it is obvious that the A-nucleotide shows the lowest probability and the C-nucleotide the highest probability of being substituted (-0.699428 and -1.776556, respectively). The single-stranded As alter most frequently into G (0.461079), followed by C (0.156830) and U (0.081518). G-nucleotides, in turn, rapidly change into A (1.148242), while G→C and G→U are relatively rare mutational events (0.114724 and 0.084045, respectively). Likewise, the C→U substitution is more prominent than U→C (1.007012 vs. 0.568940), C→A outscores A→C (0.574634 vs. 0.156830) and U→A exceeds A→U (0.138854 vs. 0.081518). This nucleotide substitution pattern will lead to an accumulation of A at the expense of C, G and U until an equilibrium is reached, which is exactly the nucleotide distribution that has been observed in the ss regions of HIV-1 subtype B RNAs (**Table 2**).

The Qds matrix contrasts strongly with the Qss matrix. The A-nucleotide is most prone to substitution (-1.499043) and G→A is slightly less probable than A→G (0.736551 and 0.904678, respectively), which is in line with the enhanced proportion of G and the equivalent diminishment of A in ds domains of HIV-1 RNA genomes (**Table 2**). It should be noted that these matrices have been constructed by means of an unrestricted model of nucleotide substitution without any constraining condition like reversibility, (partial) rate equality or fixed transition/transversion ratios. In addition, the RNA genomes of different HIV-1 isolates generated nearly identical Q matrices (**Table S2**).

We report that ss and ds regions in HIV-1 RNA employ different mutational patterns/signatures to maintain their distinct nucleotide composition. This may relate to experimental findings that indicate that local RNA structure can influence pausing of the reverse transcriptase enzyme,[18,19] which may increase the probability of misincorporation.[20] Overall, the secondary RNA structure seems to pose serious constraints on the nucleotide composition and evolution of the HIV-1 RNA genome.

Discussion

It is known that the RNA genomes of retroviruses do not use an equal portion of the four possible nucleotides, the HIV-1 genome being particularly A-rich (36.2%) and C-poor (17.6%).[2] We now evaluated these biases with respect to the ss and ds nature of the nucleotides in the viral RNA genome.[13] We document a strikingly different nucleotide signature for the ss and ds regions. The bias is put to the extreme for the ss regions (47.5% A, 21.3% U, 19.2% G and 11.9% C) and approaches a more neutral nucleotide composition for the ds regions (19.9% A, 23.6% U, 30.7% G and 25.8% C). We subsequently show that distinct mutational patterns can be observed in these two regions that will result in the maintenance of the typical nucleotide composition of the ss and ds regions.

The paired/unpaired status of a nucleotide in a viral RNA structure can have several biological effects. For instance, chemical and Apobec 3G-mediated nucleotide modification affects ss RNA more than ds RNA.[21,22] Error rates of the HIV-1 reverse transcriptase differ by template structure, being higher for ss than ds RNA.[23,24] The biology of an RNA molecule is obviously determined by properties other than the ss/ds nature. The protein-coding capacity dictates the selection of certain strings of nucleotides to form the required codons. In protein-coding sequences, which concern nearly the entire HIV genome, shifts in codon bias are restricted by the availability of cellular aminoacyl-tRNAs, overlapping reading frames (*tat*, *rev* and *env*) and overlapping regulatory sequences (e.g., *nef* overlaps with the 3' long-terminal repeat). Indeed, the viral genome is riddled with specific sequence elements that control RNA splicing and many other processes such as RNA packaging into virion particles. Despite these multiple constraints, we disclosed a relatively simple pattern of biased nucleotide composition that is highly related to the base-paired structure of the RNA molecule: excessive A-usage and C-restriction in the ss domains.

The molecular mechanism responsible for the creation of this typical A-rich genome configuration remains unknown, but the new findings do specify our thoughts on the possible evolutionary events. A priori, two possible scenarios can be envisaged that relate to the two independent steps of evolution: mutation and selection. The A-bias might arise through a preferred mutational activity and/or evolutionary selection. According to the first scenario, the generation of an A-rich genome may be caused by an enzymatic property of the error-prone reverse transcriptase enzyme or cellular editing activities encoded by the Apobec functions, which may induce G→A hypermutation in HIV
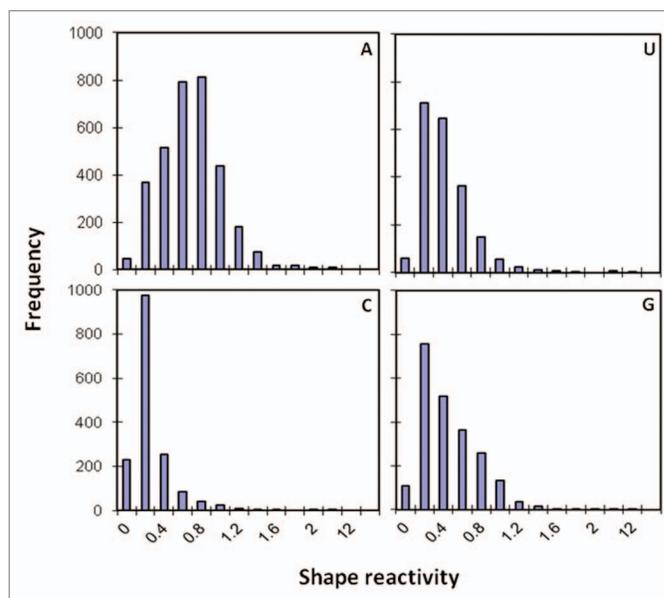

Figure 2. SHAPE reactivity of each nucleotide (A, U, C and G) in HIV-1 NL4-3 RNA. The histograms show increasing SHAPE reactivity in windows of 0.2 (X-axis, relative units). Frequency refers to the number of nucleotides per SHAPE window. Note the deviant SHAPE reactivity of the A-nucleotide.

Table 2. Nucleotide frequencies in single- and double-stranded regions of HIV-1 subtype B RNA genomes

448 HIV-1 isolates		A	U	C	G
All	AVG	36.20	22.23	17.64	23.94
	StD	0.55	0.17	0.31	0.33
ss	AVG	47.50	21.30	11.90	19.20
	StD	0.40	0.21	0.26	0.30
ds	AVG	19.90	23.60	25.80	30.70
	StD	0.44	0.23	0.29	0.30

The alignment of 448 sequences (All) was divided in two parts by the ss or ds designation of bases in the NL4-3 RNA structure. The individual nucleotide compositions were used for the calculation of average and standard deviation.

sequences.[4,9,10,25] The new finding of a clustering of A nucleotides in ss regions of the HIV-1 RNA genome does not support these mutational scenarios as a driving force for the acquisition of A-richness, as this would create an ubiquitously A-rich genome. Of course, we cannot exclude a mutational activity that is selective for ss regions.

According to the second scenario, HIV-1 and other lentiviruses have become A-rich (and C-poor) over evolutionary times by selective pressure. It is currently unknown what purpose is served by the strikingly differential base content of the HIV-1 RNA genome, but the new finding that excessive A usage is restricted to the ss domains does support this scenario and further specifies the typical lentiviral genome requirements. Our favorite suggestion would be that an RNA genome with A-rich ss domains provides a molecular signature that is recognized

Table 3. Different patterns of nucleotide substitution for ss and ds nucleotides in HIV-1 subtype B RNA genomes

Qss	A	U	C	G	Qds	A	U	C	G
A	-0.699428	0.081518	0.156830	0.461079	A	-1.499043	0.254533	0.339833	0.904678
U	0.138854	-0.899975	0.568940	0.192180	U	0.175181	-0.868417	0.532371	0.160865
C	0.574634	1.007012	-1.776556	0.194910	C	0.300107	0.471187	-0.837239	0.065945
G	1.148242	0.084045	0.114724	-1.347011	G	0.736551	0.068547	0.064782	-0.869880

Patterns of nucleotide substitution are presented as rate matrices (Qss and Qds). A positive value of a row represents the rate of substitution of the row nucleotide into one of the column nucleotides. A negative value on the matrix diagonal is the quantity by which the sum of the positive row becomes reduced to zero (meaning a zero rate of substitution). An unrestricted model of nucleotide substitution was used. The two alignments of ss and ds nucleotides were analyzed in five batches of 80 sequences. The resulting matrices (**Table S2**) were arithmetically averaged to obtain the two "consensus" matrices (Qss and Qds).

during virus replication. This recognition could occur in the context of the virus replication cycle, e.g., in selective packaging of this RNA molecule into virion particles amidst an excess of other transcripts. Alternatively, this recognition could occur in the context of the virus-host interplay, e.g., in recognition of the invading RNA by cellular factors of the innate immune system. The virus may have adopted a particular genome architecture to adapt to cellular defense mechanisms. Interestingly, *gag*, *pol* and *env* transcripts lose the ability to induce type 1 interferon responses upon "translational optimization" of the codons.[26] A more accurate description of this lentiviral RNA structure by biophysical means, 3D-modeling and functional studies, e.g., binding studies with candidate viral or cellular proteins, should help to unravel the underlying biological meaning of this particular RNA genome architecture.

Materials and Methods

The RNA sequence of HIV-1 isolate NL4-3, belonging to subtype B, its structure, SHAPE reactivity data and base-pairing probabilities were taken from Watts et al.[13] The Los Alamos HIV database (www.hiv.lanl.gov) provided aligned genomes of HIV-1 subtype B isolates (year 2010, 448 genomes, no recombinants). The NL4-3 RNA sequence, including its single-stranded (ss) or double-stranded (ds) designation for each nucleotide position, was manually made part of this alignment. All nucleotides involved in base pairing (regular Watson-Crick and G-U/U-G pairs) were scored as ds, and unpaired nucleotides (interhelical segments, hairpin loops and internal loops, bulges) as ss. Subsequently, a bipartition was created guided by the ss or ds designation under stringent

preservation of aligned nucleotide positions. Nucleotide frequency estimates and tree building was performed using MEGA5.[27] The unrestricted model 8 of the BASEML module of PAML V4.4[28] was used to estimate the mutational rate parameters of U→C, U→A, U→G, C→U, C→A, C→G, A→U, A→C, A→G, G→U and G→C relative to the fixed value of 1 for G→A by means of maximum likelihood (ML) iteration. The pattern of nucleotide substitution is presented as a matrix (Q) specifying relative rates multiplied by a constant so that the average rate is made equal to 1 when the process is in equilibrium (see the PAML manual for details). In view of the parameter richness of this model 8, both ss and ds alignments were analyzed in five portions of 80 isolates without the final 48 isolates of the original alignment. Arithmetic averaging was applied to generate two "consensus" Q matrices showing different nucleotide substitution patterns between paired and unpaired nucleotides in 401 HIV-1 subtype B RNAs (the source EXCEL sheet is available as **Table S2**).

References

1. Kypr J, Mrázek J. Unusual codon usage of HIV. Nature 1987; 327:20; PMID:3646480; http://dx.doi.org/10.1038/327020a0.

2. van Hemert FJ, Berkhout B. The tendency of lentiviral open reading frames to become A-rich: constraints imposed by viral genome organization and cellular tRNA availability. J Mol Evol 1995; 41:132-40; PMID:7666442.

3. Jenkins GM, Holmes EC. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res 2003; 92:1-7; PMID:12606071; http://dx.doi.org/10.1016/S0168-1702(02)00309-X.

4. Berkhout B, van Hemert FJ. The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. Nucleic Acids Res 1994; 22:1705-11; PMID:8202375; http://dx.doi.org/10.1093/nar/22.9.1705.

5. van Weringh A, Ragonnet-Cronin M, Pranckeviciene E, Pavon-Eternod M, Kleiman L, Xia X. HIV-1 modulates the tRNA pool to improve translation efficiency. Mol Biol Evol 2011; 28:1827-34; PMID:21216840; http://dx.doi.org/10.1093/molbev/msr005.

6. Berkhout B, Grigoriev A, Bakker M, Lukashov VV. Codon and amino acid usage in retroviral genomes is consistent with virus-specific nucleotide pressure. AIDS Res Hum Retroviruses 2002; 18:133-41; PMID:11839146; http://dx.doi.org/10.1089/08892220252779674.

7. Bronson EC, Anderson JN. Nucleotide composition as a driving force in the evolution of retroviruses. J Mol Evol 1994; 38:506-32; PMID:8028030; http://dx.doi.org/10.1007/BF00178851.

8. Cohen A, Barankiewicz J, Lederman HM, Gelfand EW. Purine and pyrimidine metabolism in human T lymphocytes. Regulation of deoxyribonucleotide metabolism. J Biol Chem 1983; 258:12334-40; PMID:6605343.

9. Vartanian J-P, Meyerhans A, Sala M, Wain-Hobson SG. G-->A hypermutation of the human immunodeficiency virus type 1 genome: evidence for dCTP pool imbalance during reverse transcription. Proc Natl Acad Sci U S A 1994; 91:3092-6; PMID:7512722; http://dx.doi.org/10.1073/pnas.91.8.3092.

10. Deforche K, Camacho R, Laethem KV, Shapiro B, Moreau Y, Rambaut A, et al. Estimating the relative contribution of dNTP pool imbalance and APOBEC3G/3F editing to HIV evolution in vivo. J Comput Biol 2007; 14:1105-14; PMID:17985990; http://dx.doi.org/10.1089/cmb.2007.0073.

11. Keating CP, Hill MK, Hawkes DJ, Smyth RP, Isel C, Le SY, et al. The A-rich RNA sequences of HIV-1 pol are important for the synthesis of viral cDNA. Nucleic Acids Res 2009; 37:945-56; PMID:19106143; http://dx.doi.org/10.1093/nar/gkn1015.

12. van der Kuyl AC, Berkhout B. The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus. Retrovirology 2012; 9:92; PMID:23131071; http://dx.doi.org/10.1186/1742-4690-9-92.

13. Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW Jr., Swanstrom R, et al. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 2009; 460:711-6; PMID:19661910; http://dx.doi.org/10.1038/nature08237.

14. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci U S A 2004; 101:7287-92; PMID:15123812; http://dx.doi.org/10.1073/pnas.0401799101.

15. Mathews DH. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA 2004; 10:1178-90; PMID:15272118; http://dx.doi.org/10.1261/rna.7650904.

16. Berkhout B, van Wamel JL. The leader of the HIV-1 RNA genome forms a compactly folded tertiary structure. RNA 2000; 6:282-95; PMID:10688366; http://dx.doi.org/10.1017/S1355838200991684.

17. Klasens BIF, Huthoff HT, Das AT, Jeeninga RE, Berkhout B. The effect of template RNA structure on elongation by HIV-1 reverse transcriptase. Biochim Biophys Acta 1999; 1444:355-70; PMID:10095059; http://dx.doi.org/10.1016/S0167-4781(99)00011-1.

18. Harrison GP, Mayo MS, Hunter E, Lever AML. Pausing of reverse transcriptase on retroviral RNA templates is influenced by secondary structures both 5' and 3' of the catalytic site. Nucleic Acids Res 1998; 26:3433-42; PMID:9649630; http://dx.doi.org/10.1093/nar/26.14.3433.

19. Wu W, Blumberg BM, Fay PJ, Bambara RA. Strand transfer mediated by human immunodeficiency virus reverse transcriptase in vitro is promoted by pausing and results in misincorporation. J Biol Chem 1995; 270:325-32; PMID:7529228; http://dx.doi.org/10.1074/jbc.270.1.325.

20. Parthasarathi S, Varela-Echavarría A, Ron Y, Preston BD, Dougherty JP. Genetic rearrangements occurring during a single cycle of murine leukemia virus vector replication: characterization and implications. J Virol 1995; 69:7991-8000; PMID:7494312.

21. Yu Q, König R, Pillai S, Chiles K, Kearney M, Palmer S, et al. Single-strand specificity of APOBEC3G accounts for minus-strand deamination of the HIV genome. Nat Struct Mol Biol 2004; 11:435-42; PMID:15098018; http://dx.doi.org/10.1038/nsmb758.

22. Ji J, Hoffmann JS, Loeb L. Mutagenicity and pausing of HIV reverse transcriptase during HIV plus-strand DNA synthesis. Nucleic Acids Res 1994; 22:47-52; PMID:7510388; http://dx.doi.org/10.1093/nar/22.1.47.

23. Yoshida K, Nakamura M, Ohno T. Mutations of the HIV type 1 V3 loop under selection pressure with neutralizing monoclonal antibody NM-01. AIDS Res Hum Retroviruses 1997; 13:1283-90; PMID:9339845; http://dx.doi.org/10.1089/aid.1997.13.1283.

24. Armitage AE, Deforche K, Chang CH, Wee E, Kramer B, Welch JJ, et al. APOBEC3G-induced hypermutation of human immunodeficiency virus type-1 is typically a discrete "all or nothing" phenomenon. PLoS Genet 2012; 8:e1002550; PMID:22457633; http://dx.doi.org/10.1371/journal.pgen.1002550.

25. Vabret N, Bailly-Bechet M, Najburg V, Müller-Trutwin M, Verrier B, Tangy F. The biased nucleotide composition of HIV-1 triggers type I interferon response and correlates with subtype D increased pathogenicity. PLoS One 2012; 7:e33502; PMID:22529893; http://dx.doi.org/10.1371/journal.pone.0033502.

26. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 2011; 28:2731-9; PMID:21546353; http://dx.doi.org/10.1093/molbev/msr121.

27. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 1997; 13:555-6; PMID:9367129.