Entropy and codon bias in HIV-1

Aakash Pandey
Department of Biotechnology, Kathmandu University
aakash.biophys@gmail.com

Abstract

HIV is rapidly evolving virus with a high mutation rate. For heterologous gene expression system, the codon bias has to be optimized according to the host for efficient expression. Although DNA viruses show a correlation on codon bias with their hosts, HIV genes show low correlation for both nucleotide composition and codon usage bias with its human host which limits the efficient expression of HIV genes. Despite this variation, HIV is efficient at infecting hosts and multiplying in large number. In this study, I have performed information theoretic analysis of nine genes of HIV-1 based on codon statistics of the whole HIV genome, individual genes and codon usage of human genes. Despite being poorly adapted to the codon usage bias of human hosts, I have found that the Shannon entropies of nine genes based on overall codon statistics of HIV-1 genome are very similar to the entropies calculated from codon usage of human genes probably suggesting co-evolution of HIV-1 along with human genes. Similarly, for the HIV-1 whole genome sequence analyzed, the codon statistics of the third reading frame has the highest bias representing minimum entropy and hence maximum information.

Keywords: HIV-1 Genome, Shannon Entropy, Codon Usage Bias, Codon Adaptation Index, Expected Codon Adaptation Index

Introduction

Every organism has its own pattern of codon usage. All the synonymous codons for a particular amino acid are not used equally. Some synonymous codons are highly expressed, whereas the uses of others are limited. The use depends on the species [1] [2]. The difference in codon usage has also been observed among genes of the same organism [3]. Codon bias has been linked to specific tRNA levels that are mainly determined by the number of tRNA genes that code for a particular tRNA [4]. The choice of codon affects the expression level of genes. This can be seen in the expression pattern of transgenes. Gustafsson *et. al.* showed that use of particular codons can increase the expression of the transgene by over 1,000 fold [5]. In bacteria, the gene expressivity correlates with codon usage [6]. Although bacteriophages have been shown to have codons that are preferred by their hosts [7], however, the codon usage pattern of RNA virus seems to differ with its host [8]. Despite this variation, HIV virus can effectively multiply in human T cells. Although codon usage of early genes (*tat*, *rev*, *nef*) shows higher correlations with human codon usage [9], however, late genes show little correlation. It raises a question of how such variation in codon usage still allows for efficient viral gene expression. *van Weringh et. al.* showed that there is a difference in the tRNA pool of HIV-1 infected and uninfected cells. Although they speculated that HIV-1 modulates the tRNA pool of the host making it suitable for its efficient genome translation, however, the extent to which such modulation helps in efficient translation is still unknown.

After Shannon published his groundbreaking paper "A Mathematical theory of Communication" [10], there have been several attempts in using information theory in the context of living systems. Shannon used the term *information* differently than classical information theorists have used. Here, I have tried to use the information as mentioned by Shannon to see whether information theoretic analysis leads to some novel insights into the problem. According to Shannon, for a possible set of events with probability distribution given by {p_1, p_2, p_3, …, p_n } the entropy or uncertainty is given by,

$$H = -\sum_{i=1}^{n} p_i * log(p_i)$$

This is in fact the observed entropy of a sequence with the given probability distribution. H is the maximum when all p_i are equally likely. In this condition the information content is zero. The amount of information or 'negentropy' in a sequence can then be given as,

$$I = H_{max} - H_{obs}$$

where, H_{obs} is the entropy obtained from given probability distribution [11]. DNA comprises 4 nucleotides A, G, C, T whose distribution pattern varies among different species. Gatlin deduced information content based on this distribution pattern [12] and using transition probability values obtained from the neighbor data [13]. Lately, the information theoretic value of a given DNA sequence was obtained using the Shannon formula as double sum [14],

$$H = \sum_{i=1}^{i=n_{aa}} \left(-\sum_{j=1}^{j=n_{syncod(i)}} P_{(i,j)} log_2\left(P_{(i,j)}\right)\right)$$

Here, n_{aa} is the number of distinct amino acids, $n_{syncod(i)}$ is the number of synonymous codons (or micro states) for amino acid *i* (or macro state) whose value range from 1 to 6, and $P_{(i,j)}$ is the probability of synonymous codon *j* for amino acid *i*. Also, we know that information is not absolute. It depends on the environment. This means that the same sequence of DNA may represent different amounts of information depending on what environment it is in or on the machinery that interprets the sequence. We exploit this to calculate Shannon entropy for nine genes of HIV-1 based on codon distribution of the viral genome, individual genes and that of its host – human codon usage frequency. Information is calculated based on the codon distribution for three possible reading frames. To the best of my knowledge, I believe that such study has not been carried out yet. Viruses show overlapping genes and are speculated to be present to increase the density of genetic information [15]. These genes are read by ribosomal frameshifting [16]. For those nine genes, I have also calculated the intrinsic entropy of the sequence which can be defined as the entropy based on its own codon usage (i.e. codon usage within the same gene) to compare with other entropy values.

Heterologous expression systems, such as viruses, use host translational machinery for their replication. They are under evolutionary pressure to adapt to the host tRNA pool. To estimate a degree of evolutionary adaptiveness of host and viral codon usage, Codon Usage Index (CAI) can be used [17]. But, for sequences with a high biased nucleotide composition, interpretation of CAI can be tricky [18]. So, to know whether the value of CAI is statistically significant and has arisen from codon preferences or is merely artifacts of nucleotide composition bias, expected CAI (eCAI) can be a threshold value for comparison [19].

Methodology:

The DNA sequences were obtained from NCBI database in FASTA format. For each sequence, codon statistics was obtained by entering the sequences on online Sequence Manipulation Suite [20] and using the standard genetic code as the parameter. Number and fractions of each possible codons were noted. First nucleotide was deleted to shift the reading frame by +1 to include other possible codon patterns and again the number and frequency were noted. The process was repeated for +2 reading frame. Now, as any of the reading frames can contain the gene of interest, all three reading frame statistics were used to calculate the Shannon entropy separately. The assumption made in the calculation is that reading the message occurs in a linear fashion without slippage of the reading frame (RF). The fraction was normalized for a particular amino acid, but not with the total number of codons. Thus,

$$\sum_{j=1}^{j=n_{syncod(i)}} P_{(i,j)} = 1$$

First, a row matrix was constructed with fractions of synonymous codons used.

$$codon0 = [p_1, p_2, \dots, p_{64}]$$

The fractions in the matrix were treated as microstate to calculate Shannon entropy and thus another matrix was constructed consisting of Shannon entropy of each fraction distribution.

$$H = [-codon0(1) * log_2 codon0(1), \cdots, -codon0(64) * log_2 codon0(64)]$$

The total Shannon entropy of the sequence is then calculated as:

$$H_{gene} = \sum_{i=1}^{64} N_i * H_i$$

Here N_i is the total number of a particular codon present in the gene of interest. Such calculation was performed for all three RFs.

The correlation coefficient for each gene's codon statistics was calculated with human codon usage statistics. Correlation coefficients for two genes *vpr* and *vpu* were calculated again removing the codon data for which no amino acid is present in that gene. Then, Shannon entropy was calculated for all nine genes using the human codon usage statistics. Intrinsic entropy, which is the entropy based on own codon statistics of each gene was also calculated. Again, the assumption is that there is no slippage of reading frame during translation of the message. Thus, codon statistics for single reading frame starting with start codon was used to calculate intrinsic entropy. Similarly, average entropy was calculated by averaging the fractions of synonymous codons for all three RFs. For the calculation of percentage overall GC content and position specific GC content of codons of nine genes and CAI values, http://genomes.urv.cat/CAIcal/ [21] online site was used again using the standard genetic code as the parameter. For calculation of the expected codon adaptation index was performed in E-CAI server (http://genomes.urv.es/CAIcal/) using Markov chain and standard genetic code as the parameters. Human codon usage statistics was obtained from the online site (http://genomes.urv.cat/CAIcal/CU_human_nature.html). Computations were performed in R.

Result and Discussion:

We see that there is a high correlation between codon present in HIV-1 early genes (*tat*, *rev*, *nef*) and human codon usage (fig 1), but correlation is low for other genes: *env*, *gag*, *pol*, *vpr*, *vif*, *vpr* (table 1); *vpu* and *vif* genes have the lowest correlation with human codon usage. The degree of correlation also differs for 3 RFs codon statistics and nine genes: high correlation probably suggesting the location of the particular gene at that RF.

Table 1. Correlation coefficients calculated among human codon usage, codon usage of HIV-1 genes and codon statistics for all three reading frames of HIV-1 genome

Genes	Correlation coefficients			
	Human codon	RF 0	RF +1	RF +2
env	0.48	0.78	0.74	0.87
gag	0.55	0.84	0.76	0.83
tat	0.72	0.88	0.96	0.91
rev	0.62	0.91	0.87	0.94
vpu	0.26 / 0.41	0.61	0.54	0.68
vpr	0.48 / 0.55	0.61	0.66	0.65
vif	0.44	0.79	0.74	0.76
nef	0.73	0.78	0.76	0.73
pol	0.57	0.84	0.82	0.89
Human	-	0.74	0.78	0.66

Table showing the correlation between codon fractions of nine genes with human codon usage fraction (column 2) and also with three different reading frames. RF 0 represents the codon fractions of the initial reading frame of HIV-1 sequence, RF +1 represents a codon fractions of single frame shift and RF +2 represents codon fractions of +2 frameshifts of the HIV-1 genome sequence. Human codon usage statistics is also compared with the three possible reading frames of HIV-1 genome where +1 shifted reading frame shows the highest correlation. For *vpu* and *vpr* gene, the number after backlash is calculated removing the data for which no codons are present for certain amino acids.

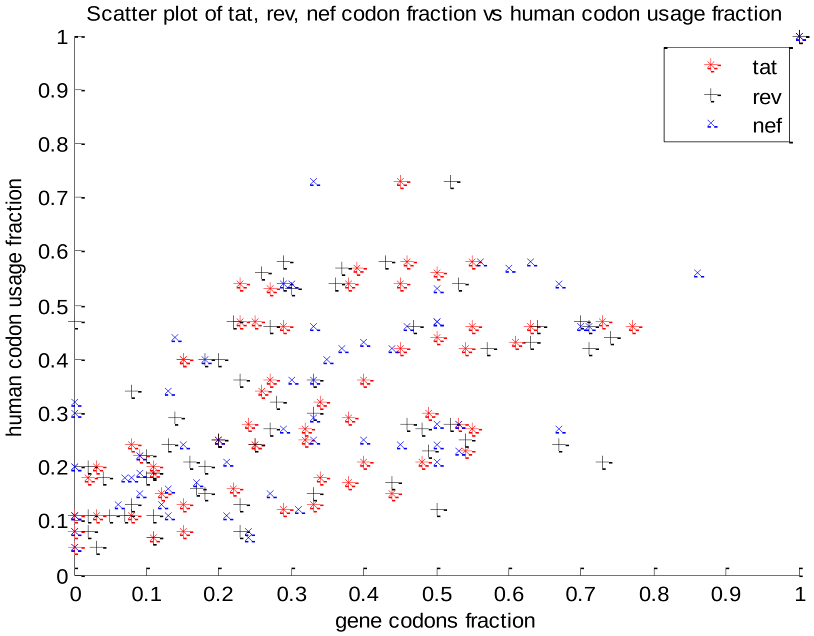

Fig 1: Scatter plot between codons fractions of *tat*, *rev* and *nef* genes of HIV-1 and human codon usage fraction with the correlation coefficient of 0.72, 0.62 and 0.73 respectively. These three genes have the higher correlation with human codon usage fraction and are also the early genes.

Entropic calculation shows a general trend for the sequences analyzed. First, +2 frameshifted reading frame shows lower entropy as compared to two other reading frames. This marked distinction of the third reading frame among three possible reading frames of the sequence analyzed is surprising. As it has the lowest entropy among three reading frames, the sequence with the codon usage pattern of third RF represents the highest information (in Shannon's sense). This probably suggests that there is a genome-wide conservation of codon usage for that reading frame, but the reason is unclear. *env*, *rev*, *pol and vpu* genes have the highest correlation with the third reading frame as compared to other two reading frames. Similarly, *gag* and *vpr* also have a high correlation coefficient. If we use codon statistics of third RF to calculate the Shannon entropy, we get the minimum entropy and hence maximum information. But, then again we run into a problem as this third RFs shows the lowest correlation coefficient with human codon usage pattern. So there has to be a balance between these contrasts: maximizing information or maximizing correlation. Take *gag* for example, it shows high correlation with RF0 and RF2 both of which have lower correlation with human codon usage. This means that the expression of *gag* is affected by this choice of codons. In fact, the ratio of native and optimized codons determine the HIV-1 *gag* expression [22]. This also supports the speculation that codon bias leads to sub-optimal expression in infected cells. There is, in fact, good evidence that HIV-1 gene expression is not the maximum but, is fine-tuned to allow regulation of diverse processes [23]. More evidence of sub-optimal expression is shown by the fact that when codon optimized genes that are better adapted to the host tRNA pool were introduced, it led to higher expression [24][25][26]. From the entropic calculation based on human codon usage, we can see that *vpu* and *vif* have the lowest entropy, but as they have a low correlation with human codon usage, their expression is limited. Codon optimization of these genes results in the increase of expression level [27]. However, high correlation does not imply that the gene is in that reading frame. It is possible that such bias may or may not affect biological function, but it is likely that such distinction of lower entropy has some evolutionary importance.

Table 2: Entropies of HIV-1 genes based on various codon distributions

Genes	***pol***	***env***	***nef***	***vif***	***vpr***	***vpu***	***rev***	***tat***	***gag***
Entropy H	259.60	129.72	93.84	84.72	43.00	36.17	398.83	400.92	225.52
Entropy 0	264.11	131.32	94.79	87.81	42.86	38.53	399.91	399.71	229.41
Entropy +1	262.34	129.77	94.26	87.79	42.76	38.83	398.85	403.42	227.55
Entropy +2	256.70	125.54	92.12	85.16	41.27	37.15	388.22	390.52	222.10
Entropy I	229.54	110.10	91.23	75.25	38.21	27.32	378.79	397.98	216.69
Average entropy	264.48	130.58	94.79	87.95	42.77	38.69	400.59	402.66	229.09

Shannon Entropic values in bits for nine genes based on different codon statistics. Entropy H denotes entropic value based on human codon usage statistics. Entropy 0, Entropy +1 and Entropy +2 represent entropic values based on first reading frame, +1 shifted reading frame and +2 shifted reading frame of HIV-1 genome sequence respectively. Entropy I represents intrinsic entropy. Average entropy is calculated by averaging the codon statistics for three possible reading frames.

Intrinsic entropy differs greatly with other entropic values as it shows lowest values. Such low intrinsic entropies may have significance for free-living organisms as lower entropies suggest higher bias. But, for heterologous expression systems such as HIV-1, entropy H probably represents the best entropic values for the genes analyzed as host (human) gene usage codon statistics was used for the calculation. Average entropy, which is closer to entropy H, rather than intrinsic entropy gives a better representation for entropic value and hence for the amount of information a gene contains inside a human host. Although there is great variation in the synonymous codon usage statistics between HIV-1 genes and human genes, the entropic values for the HIV-1 genes based on the overall code distribution of the HIV genome shows almost similar values as compared to the calculation based on human codon usage statistics (Table 2). Even for *a vpu* gene, which has a very low correlation coefficient (0.26), the entropic values based on overall codon statistics of HIV-1 genome and human codon usage statistics show similarity: 36.17 and 38.69 bits respectively. Even if we remove the data for which there is no single codon for certain amino acids in that gene, the correlation coefficient is still low. In *vpu* gene, codons for Cysteine, Threonine and Phenylalanine are absent. If we remove that data, we get a correlation coefficient of 0.41, which is still low. However, this removal does not affect the entropic calculation. Similarly, for *vpr* gene, codons for Cysteine are absent. Removing that data new correlation coefficient obtained is 0.55 and average entropic values and entropy H are close: 42.77 and 43.00 bits respectively. HIV is a highly variable virus which undergoes rapid mutation. Although HIV cannot match its codon bias with that of the host, but it can have a stable codon usage pattern. To maintain the overall codon statistics, it has to maintain the nucleotide composition, which is the determinant of codon bias [28]. It has been shown that, despite its high mutation rate, the biased nucleotide composition of HIV is constant over time [29]. If the genes have same codon biases as that of the host then it might lead to their highest expression. But this is not desired as it would not allow for efficient tuning of its complex processes. If codon bias is completely different from that of host, then it might result in very low expression putting its ability to survive in the host into question. So, HIV has to find a solution which results in sub-optimal expression of genes. So from the calculation of table 2 (Entropy H and Average Entropy), it seems that HIV has found a solution in which its codon bias is different from that of host to allow sub-optimal expression, but at the same time represent the same level of information as can be obtained from the codon bias of its host. This might suggest that, despite having a different nucleotide composition with human, HIV-1 viruses have co-evolved with human genes to represent the same level of information.

Besides having similar entropy, we can note that all the genes have high CAI values (Table 3) although with a varying degree of GC content. All the CAI values are greater than 0.6 with *vpu* having the lowest value of 0.62 whereas *tat* has the highest value of 0.77. CAI well above 0.5 is usually considered to be showing a good level of adaptation towards the host, however, care should be given while interpreting these values as they may not reveal the level of adaptation just by themselves. Such values may be due to the bias in nucleotide composition. So, to know whether these values actually represent the adaptation we need to set a threshold set by the bias in nucleotide composition so we may say that CAI does represent the level of adaptation. For that expected CAI (eCAI) is calculated and compared [19]. From these comparisons, none of the genes seems to be well adapted to the human codons usage pattern. We can note that the GC content of all the genes is below average. Also, GC content of second and third nucleotides of codons shows the greatest variability. Thus, we can conclude that nucleotides of these positions are the determinant of codon bias in HIV genes rather than the selection pressure for codons.

Table 3: Codon adaptation index (CAI) and GC content of HIV-1 genes.

Gene	Length	CAI	%GC	%GC1	%GC2	%GC3	eCAI (p<0.05)
gag	1503	0.73	44.0	50.3	43.5	38.3	0.74
nef	621	0.76	49.4	58.0	44.4	45.9	0.76
tat	2592	0.77	39.8	38.8	41.0	39.7	0.78
pol	1746	0.71	38.3	48.8	36.4	29.6	0.72
rev	2682	0.73	40.5	41.5	39.7	40.2	0.74
vif	579	0.72	42.0	46.6	41.5	37.8	0.74
vpr	291	0.73	45.0	54.6	34.0	46.4	0.74
vpu	249	0.62	37.8	54.2	31.3	27.7	0.66

CAI and eCAI values with overall GC percentage and position specific GC percentage for nine HIV-1 genes is reported. %GC 1, %GC2 and %GC3 represent GC percentage of the first, second and third position of the codons respectively. Second and third position of the codon shows greatest bias in GC content as compared to the first position (except for *tat* and *rev* gene which shows almost no bias for all three positions)

Conclusion:

Despite many studies, HIV viral genome still possesses several mysteries. HIV is evolving along with its human host. However, it is not clear why its nucleotide composition and synonymous codon usage bias differ greatly from its host. From the comparison of CAI with eCAI, we can conclude that HIV genes are poorly adapted to the tRNA pools of humans. So it can be inferred that selection pressure on HIV to adapt to tRNA pools is minimal as compared to the rapid mutation it has on its genome [8]. Because of this, HIV evolves as a separate entity, although there is selection pressure on different levels. It is not clear whether nucleotide composition bias can give rise to the asymmetry in the observed information content along three possible reading frames. However, despite having large differences in nucleotide composition and synonymous codon usage bias, HIV genes are seem to have evolved to represent the same level of information as obtained by the codon bias of human genes. How HIV is able to attain such uniformity, despite differing from its host, is yet another mystery this study has surfaced. Further work is needed, which can bring together the differences in one place to give a clear picture of the evolution of HIV viral genome.

Conflict of interest: The authors declare no conflict of interest.

References:

1. Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. Codon catalog usage and the genome hypothesis. Nucl Acids Res. 1980;8(1):197-197.

2. Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucl Acids Res. 1981;9(1):213-213.

3. Sharp P, Cowe E, Higgins D, Shields D, Wolfe K, Wright F. Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens ; a review of the considerable within-species diversity. Nucl Acids Res. 1988;16(17):8207-8211.

4. Hershberg RPetrov D. Selection on Codon Bias. Annu Rev Genet. 2008;42(1):287-299.

5. Gustafsson C, Govindarajan S, Minshull J. Codon bias and heterologous protein expression. Trends in Biotechnology. 2004;22(7):346-353.

6. Gouy MGautier C. Codon usage in bacteria: correlation with gene expressivity. Nucl Acids Res. 1982;10(22):7055-7074.

7. Lucks J, Nelson D, Kudla G, Plotkin J. Genome Landscapes and Bacteriophage Codon Usage. PLoS Computational Biology. 2008;4(2):e1000001.

8. Jenkins G, Holmes E. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Research. 2003;92(1):1-7.

9. 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. Molecular Biology and Evolution. 2011;28(6):1827-1834.

10. Shannon, C.. A Mathematical Theory of Communication. Bell Systems Technical Journal, 1948;27: 279-423, 623-656.

11. Brillouin L. Science and information theory. New York: Academic Press; 1962.

12. L.L. Gatlin. The information content of DNA. Journal of Theoretical Biology, 1966;10,281-300.

13. J. Josse, A.D. Kaiser, A. Kornberg. Enzymatic synthesis of deoxyribonucleic acid: VIII. frequencies of nearest neighbor base sequences in deoxyribonucleic acid. Journal of Biological Chemistry, 1961;236, 864-875.

14. Zeeberg B. Shannon Information Theoretic Computation of Synonymous Codon Usage Biases in Coding Regions of Human and Mouse Genomes. Genome Research. 2002;12(6):944-955.

15. Lamb RHorvath C. Diversity of coding strategies in influenza viruses. Trends in Genetics. 1991;7(8):261-266.

16. Wilson W, Braddock M, Adams S, Rathjen P, Kingsman S, Kingsman A. HIV expression strategies: Ribosomal frameshifting is directed by a short sequence in both mammalian and yeast systems. Cell. 1988;55(6):1159-1169.

17. Sharp P,Li W. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucl Acids Res. 1987;15(3):1281-1295.

18. Grocock RSharp P. Synonymous codon usage in Pseudomonas aeruginosa PA01. Gene. 2002;289(1-2):131-139.

19. Puigbò P, Bravo I, Garcia-Vallvé S. E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI). BMC Bioinformatics. 2008;9(1):65.

20. Codon Usage [Internet]. Bioinformatics.org. 2016 [cited 4 April 2016]. Available from: http://www.bioinformatics.org/sms2/codon_usage.html

21. Puigbò P, Bravo I, Garcia-Vallve S. CAIcal: A combined set of tools to assess codon usage adaptation. Biology Direct. 2008;3(1):38.

22. Kofman A, Graf M, Bojak A, Deml L, Bieler K, et al. HIV-1 gag expression is quantitatively dependent on the ratio of native and optimized codons. Tsitologiia 2003;45: 86–93.

23. Marzio G, Vink M, Verhoef K, de Ronde A, Berkhout B. Efficient Human Immunodeficiency Virus Replication Requires a Fine-Tuned Level of Transcription. Journal of Virology. 2002;76(6):3084-3088.

24. Haas J, Park E, Seed B. Codon usage limitation in the expression of HIV-1 envelope glycoprotein. Current Biology. 1996;6(3):315-324.

25. Anson Dunning K. Codon-Optimized Reading Frames Facilitate High-Level Expression of the HIV-1 Minor Proteins. Molecular Biotechnology. 2005;31(1):085-088.

26. Ngumbela K, Ryan K, Sivamurthy R, Brockman M, Gandhi R, Bhardwaj N et al. Quantitative Effect of Suboptimal Codon Usage on Translational Efficiency of mRNA Encoding HIV-1 gag in Intact T Cells. PLoS ONE. 2008;3(6):e2356.

27. Nguyen K, Llano M, Akari H, Miyagi E, Poeschla E, Strebel K et al. Codon optimization of the HIV-1 vpu and vif genes stabilizes their mRNA and allows for highly efficient Rev-independent expression. Virology. 2004;319(2):163-175.

28. Bronson EAnderson J. Nucleotide composition as a driving force in the evolution of retroviruses. J Mol Evol. 1994;38(5):506-532.

29. van der Kuyl ABerkhout B. The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus. Retrovirology. 2012;9(1):92.