Arch Virol
DOI 10.1007/s00705-017-3597-5

ORIGINAL ARTICLE

Codon usage of HIV regulatory genes is not determined by nucleotide composition

Supinya Phakaratsakul[1] · Thanyaporn Sirihongthong[1] · Chompunuch Boonarkart[1] · Ornpreya Suptawiwat[2] · Prasert Auewarakul[1]

Received: 18 April 2017 / Accepted: 30 August 2017
© Springer-Verlag GmbH Austria 2017

Abstract Codon usage bias can be a result of either mutational bias or selection for translational efficiency and/or accuracy. Previous data has suggested that nucleotide composition constraint was the main determinant of HIV codon usage, and that nucleotide composition and codon usage were different between the regulatory genes, *tat* and *rev*, and other viral genes. It is not clear whether translational selection contributed to the codon usage difference and how nucleotide composition and translational selection interact to determine HIV codon usage. In this study, a model of codon bias due to GC composition with modification for the A-rich third codon position was used to calculate predicted HIV codon frequencies based on its nucleotide composition. The predicted codon usage of each gene was compared with the actual codon frequency. The predicted codon usage based on GC composition matched well with the actual codon frequencies for the structural genes (*gag*, *pol* and *env*). However, the codon usage of the regulatory genes (*tat* and *rev*) could not be predicted. Codon usage of the regulatory genes was also relatively unbiased showing the highest effective number of codons (ENC). Moreover, the codon adaptation index (CAI) of the regulatory genes showed better adaptation to human codons when compared to other HIV genes. Therefore, the early expressed genes responsible for regulation of the replication cycle, *tat* and *rev*, were more similar to humans in terms of codon usage and GC content than other HIV genes. This may help these genes to be expressed efficiently during the early stages of infection.

List of abbreviations

RSCU	Relative synonymous codon usage
ENC	Effective number of codon
CAI	Codon adaptation index

Introduction

As 20 amino acids and three stop codons are encoded by a total of 64 codons in the genetic code, there is degeneracy wherein codons up to 6 different codons can be synonymous. Different species preferentially use some codons over others to encode the same amino acid resulting in codon usage bias. The frequency of different codon usage varies significantly between different organisms and sometimes varies within the same organism and even in the same operon [1]. Genes can be poorly expressed when introduced into a heterologous host species with a mismatched codon usage pattern. For example, human proteins with a human codon usage profile are poorly expressed in *E.coli* [2]. In order to increase the expression of heterologous proteins, codon optimization has therefore been widely used. A foreign gene is optimized to have a codon usage bias matched with that of a chosen host species. Several instances have shown dramatic increases in the expression levels of codon optimized mammalian proteins in bacteria and yeast. Also, many studies have shown that humanized viral genes express better than wild type

Handling editor: Li Wu.

Electronic supplementary material The online version of this article (doi:10.1007/s00705-017-3597-5) contains supplementary material, which is available to authorized users.

✉ Prasert Auewarakul
prasert.aue@mahidol.ac.th

[1] Department of Microbiology, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand

[2] Research and International Relations Division, HRH Princess Chulabhorn College of Medical Science, Chulabhorn Royal Academy, Bangkok 10210, Thailand

Published online: 25 October 2017

viral genes in human cells [3–6]. Codon optimization has also been widely applied in recombinant protein production and in DNA-based vaccine.

Theoretically, there are two major models proposed that explain the codon usage bias: the translation related model and the mutational model [7–11]. For translational efficiency, the codon usage bias correlates with the corresponding isoacceptor tRNA abundance. The isoacceptor tRNA concentration tends to co-evolve with high frequency codons in prokaryotes [2, 9]. Moreover, the rate of translation of mRNAs containing preferred codons was shown to be faster than mRNAs containing rare codons [10]. Most viruses have specific codon usage patterns, which do not match those of their hosts [9, 11]. Mutational pressure has been used to explain this phenomenon of codon usage bias of virus. Nucleotide composition was/is believed to be the major factor driving codon usage [8, 12, 13]. RNA editing mediated by the host immune response was found to be a factor influencing bias in the nucleotide content of viral genomes. For instance, APOBEC3G (apolipoprotein B mRNA editing enzyme, catalytic polypeptide 3G) is a cytidine (C) deaminase, which can catalyze deamination of dCTP to dUTP during retroviral DNA synthesis. Consequently, the deamination results in a guanosine to adenosine (G-to-A) hypermutation in the viral plus-strand DNA [14–16]. The guanosine (G)-to-adenosine (A) hypermutation is a major driving force in retroviral codon usage, which consequently prefers A at the third codon position. Therefore, A-richness in RNA genomes was found to be a natural property of the lentivirus family (e.g. it is 35-36% in HIV genomes). In contrast, human genomes are GC rich, and this GC content enhances gene expression in humans [17, 18].

The relationship between nucleotide composition and codon usage bias was previously described in a general model of codon bias due to GC content by Palidwor GA et al in 2010 [19]. In this model, the frequency of each codon can be calculated based on the GC content of the third codon position (GC3) and this approach was validated in prokaryotic, plant, and human genes. This provides a quantitative measurement for the influence of GC content on codon usage. If codon usage is only a result of nucleotide composition constraint, codon usage should be accurately predicted by this model. The codon usage bias of viruses tends to be affected by both genome nucleotide composition and functional selection for efficient infection in natural host species. However, it is not clear how much each factor influences codon usage bias within different genes of the same viral genome. This codon usage prediction model could be a useful tool to understand these factors.

Data suggests that viral codon usage could be influenced by the codon usage of the viruses' respective host. In the case of influenza viruses, which infect a broad range of hosts, viruses from different host species (e.g. from an avian or human host) showed different codon usage biases [7]. Moreover, after cross-species transmission, astrovirus codon usage exhibited evolution in a direction towards codon usage of the host species' genes [20]. Arboviruses such as dengue, chikungunya and Zika, which successfully replicate and transmit in multiple hosts and vectors, display various codon usage patterns that seem to balance in order to maintain efficient replication and survival in multiple hosts [21]. Therefore, the direction of codon usage adjustment of particular virus genes can respond to the codon usage of the host genome. Consequently, the evolution of both host and pathogen are influenced at an individual and population level, suggesting essentially that selective pressure can shape codon usage bias [22].

In the HIV genome, the codon usage is not uniform throughout the genome. A difference in codon usage between individual genes of HIV has been observed previously [17, 23]. Codon usage of the regulatory genes, *tat* and *rev*, is obviously distinct from other genes and this difference correlated with base composition, showing a low A content and high G and C content within these two genes [17, 18]. To understand the driving force behind this codon usage difference, we applied the previously published model of codon bias due to GC content to evaluate the influence of nucleotide composition on codon bias within these genes.

Materials and methods

Data acquisition

Complete genome sequences of 50 HIV-1, 13 HIV-2 and 23 SIV (lengths between 8719-10359 base pairs) were downloaded in FASTA format from the Los Alamos National Laboratory (LANL) HIV Sequence Database (www.hiv.lanl.gov, Mar 2017) (Table S1). The sequences were selected from various subtypes and CRFs (circulating recombinant forms) in different geographical regions and years. Nine HIV-1 genes (*gag, pol, env, vif, tat, rev, vpr, vpu, nef*) and nine HIV-2 and SIV genes (*gag, pol, env, vif, tat, rev, vpr, vpx, nef*) were obtained from the full length sequences using Bioedit program and Gene Cutter (www.hiv.lanl.gov).

Codon usage prediction

Predicted codon usage based on nucleotide composition was calculated using formulas for a general model of codon bias due to GC mutational bias, as previously described [19] as well as formulas with some modification. The modification reflected the A-rich third codon position of HIV genes leading to an imbalance between A and T. For codons ending with A or T, codon frequencies were adjusted by 2A/AT

(α) or 2T/AT (β), respectively. The formulas were modified accordingly, as follows (Table 1):

Codon frequency measurement

Each gene within the HIV and SIV sequences was submitted to a web-based program for codon usage analysis (http://www.bioinformatics.org/sms/codon_usage.html) [20]. The program generated codon tables containing the number and frequency of each codon and compared the frequencies of synonymous codons. The codon table for each gene was compared with the predicted codon usage generated from the modified formulas, as described above (Table 1). Furthermore, nucleotide composition, relative synonymous codon usage (RSCU) and effective number of codons (ENC) for each gene were also calculated by the CAIcal server, which is available at http://ppuigbo.me/programs/CAIcal/ [24]. Three stop codons (TAA, TAG and TGA), Met (ATG), and Trp (TGG) were excluded from the calculation since there are no synonymous codons. Codon usage for each HIV gene was compared to a reference human codon table and codon usage for each SIVcpz and SIVsmm gene was compared to reference chimpanzee and sooty mangabey codon tables, respectively, using the CAI calculator on the CAIcal server to generate a codon adaptation index (CAI) [25]. The reference codon usage tables for human, chimpanzee and sooty mangabey were obtained from the codon usage database (http://www.kazusa.or.jp/codon/) [26].

Table 1 Codon usage prediction using Palidwor's and adjusted formulas

Codons	Palidwor's formulas	Adjusted formulas
Two-codon		
- - A	1 − B*	(1 − B).α
- - T	1 − B	(1 − B).β
- - C or - - G	B	B
Four-codon		
- - A	1 − B/2	((1 − B)/2).α
- - T	1 − B/2	((1 − B)/2).β
- - C and - - G	B/2	B/2
Isoleucine		
ATA	1 − B/2 − B	((1 − B)/(2 − B)).α
ATT	1 − B/2 − B	((1 − B)/(2 − B)).β
ATC	B/2 − B	B/2 − B
Arginine		
AGA	$(1 - B)^2/1+B$	$((1 - B)^2/1+B).\alpha$
CGA	B(1 − B)/1+B	(B(1 − B)/1+B).α
CGT	B(1 − B)/1+B	(B(1 − B)/1+B).β
AGG	B(1 − B)/1+B	B(1 − B)/1+B
CGG, CGC	$B^2/1+B$	$B^2/1+B$
Leucine		
TTA	$(1 - B)^2/1+B$	$((1 - B)^2/1+B).\alpha$
CTA	B(1 − B)/1+B	(B(1 − B)/1+B).α
CTT	B(1 − B)/1+B	(B(1 − B)/1+B).β
TTG	B(1 − B)/1+B	B(1 − B)/1+B
CTG, CTC	$B^2/1+B$	$B^2/1+B$

B = GC3 (GC content of the 3rd codon position)

Evolutionary trends of HIV codons

Patterns of codon evolution and selection among HIV genes were investigated by counting codon changes at each dichotomy in a phylogenetic tree. Fifty HIV-1 gene sequences were aligned using the Bioedit program and then phylogenetic trees were generated in PHYLIP version 3.695 (April, 2013). The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolution of the 50 sequences. Subsequently, phylogenetic analysis by maximum likelihood was done using PAML version 4.9d (February, 2017). Model 0 from the Codeml menu was used to generate internal node sequences which represented the ancestral sequences of the branches. The numbers and types of codon changes that were predicted to occur at the dichotomy of each internal node were counted. As A-richness at the third codon position is the main characteristic of HIV codon usage, we classified codon changes into two groups as "toward A-richness" and "away from A-richness". Accordingly, the codon changes were separated into 4 groups based on changes to adenine (A) at the third codon position from their immediate ancestral sequences, specifically: (i) A to A, (ii) non-A to A, (iii) A to non-A and (iv) non-A to non-A. The selection trends of HIV-1 *gag*, *tat* and *rev* genes were analyzed using the Toward A-richness ratio defined by the following formula:

$$\sum(\text{A to A}) + (\text{non-A to A}) \,/\, \sum(\text{A to non-A}) + (\text{non-A to non-A})$$

Statistical analysis

Statistical analysis was done using Prism software version 7.01 (June, 2016). The nine genes of HIV and SIV were grouped into structural genes (*gag*, *pol* and *env*), regulatory genes (*tat* and *rev*), and accessory genes (*vif*, *vpr*, *vpu*/*vpx* and *nef*). Group comparison was tested by one-way ANOVA with Dunns' post-test to examine the absolute differences of predicted codon usage and actual codon frequency and CAI values and considered significant at p-value ≤ 0.05.

Results

An uneven effect of nucleotide composition on the codon usage bias of HIV genes

A-richness within the RNA genome was found to be a natural property of the lentivirus family. However, it is not uniform throughout the viral genome because of a difference in the nucleotide composition of each gene [23]. The lowest A content and the highest G and C content were recorded in the *tat* and *rev* genes [17]. It is not clear what caused this difference. In a previously published model, codon usage of organisms could be correctly predicted based on GC content at the third codon position [19]. However, the model does not take into account the AT imbalance and the A richness of lentivirus genomes, thus the formulas were adjusted.

To estimate the effect of nucleotide composition on codon usage bias, modified formulas corrected for the A richness were used to predict the codon usage of HIV genes. The predicted codon usage was then compared with the actual codon frequency. We analyzed all 9 HIV genes: structural genes (*gag, pol* and *env*), regulatory genes (*tat* and *rev*), and accessory genes (*vif, vpr, vpu/vpx* and *nef*). While the predicted codon usage of the structural and accessory genes for both HIV-1 and HIV-2, with both Palidwor's original formulas and the formulas modified for A richness, matched reasonably well with the actual codon usage, the codon usage calculation from nucleotide composition failed to predict the actual codon usage of the regulatory genes of both viruses (Fig. 1). While this general trend is similar between HIV-1 and HIV-2, there are some differences between equivalent genes from the two viruses. The nucleotide composition and codon usage of HIV-2 *env* is relatively unbiased, whereas HIV-1 *env* shows typical A-richness and a codon usage pattern similar to those of other structural genes. In contrary, the nucleotide composition of HIV-1 *tat* and *rev* is very neutral resulting in flat lines for predicted codon usage, whereas that of HIV-2 *tat* and *rev* is less neutral (Fig. 1A, B).

To compare the degree of concordance between the predicted and actual codon usage among groups of the HIV genes, absolute differences between predicted and actual RSCU were analyzed. Structural genes exhibited the lowest difference meaning that the codon usage of these genes was the most accurately predicted, based on GC content. In contrast, the absolute differences in the regulatory genes was the highest among HIV-1 genes, significantly higher than that of the structural genes and accessory genes. This implies that the codon usage of regulatory genes is less dependent on nucleotide composition (Fig. 2A). The codon usage of regulatory genes might be driven by functional selection related to the efficiency of translation in the host. Apart from the *nef* gene, accessory genes also showed significantly higher differences in predicted and actual RSCU when compared to structural genes, albeit slightly lower than that of the regulatory genes. The RSCU difference value of *nef* was comparable to the group of structural genes and significantly lower than that of *vif* and *vpr* (Fig. 2A).

Generally, HIV-2 also demonstrates a similar trend of absolute differences between predicted and actual RSCU among its nine genes. Codon usage of the structural genes are also the most accurately predicted while the RSCU differences among the regulatory genes are also the highest; however, differences in some genes were found. While the RSCU difference of *gag* was the highest among structural genes in HIV-1, it is slightly lower than *pol* in HIV-2. Codon usage of the *env* gene was the most accurately predicted among the HIV-2 genes and was more predictable than that of HIV-1, showing a lower difference in RSCU. In the regulatory genes, the RSCU differences of *tat* and *rev* showed the same trend as that of HIV-1 but with lower levels of differences. Nevertheless, accessory genes showed a different pattern from that of HIV-1. The RSCU differences for *vpx* and *nef* were the first and the second highest among the accessory genes and were significantly higher than all the structural genes. These two genes are comparable to the regulatory genes (Fig. 2B). Overall, the RSCU differences for the HIV-2 genes were slightly lower than that of the HIV-1 genes indicating that the codon usage of HIV-2 genes could be better predicted than for HIV-1 (Fig. 2A, B).

Although codon usage of the three structural genes could be properly predicted by nucleotide composition, some codons with A at the third codon position are more frequently used than their predicted frequencies, especially the AGA codon for Arginine. This over-representation suggests a selective pressure favoring this codon. In contrast, codon usage within the regulatory genes failed to be predicted based on nucleotide composition. Over-represented codons in the structural genes with A at the third codon position were also preferred by the regulatory genes, albeit with lower frequencies. In addition AAC, ATC and CCT codons for asparagine, isoleucine and proline, respectively, were preferred in the regulatory genes but not in the structural genes (Fig. 1A). These codons are also preferred within normal human codon usage patterns. Thus regulatory gene codon usage might be linked to efficient replication in the context of the virus within the host. The codon usages for the accessory genes were partially predictable and this pattern varies within this group.

As the second exons of *tat* and *rev* genes overlap with the *env* ORF, only the first exons of both genes were analyzed, in order to clarify whether the higher GC content and the different pattern of codon usage within these two genes are actual properties or an effect of the overlapping frame with *env*. The first exon of *tat* and *rev* are 215 and 75 base pairs, respectively. Although, the GC3 of *tat* exon 1 (43.32%) is much lower than the full length (50.12%), the actual codon

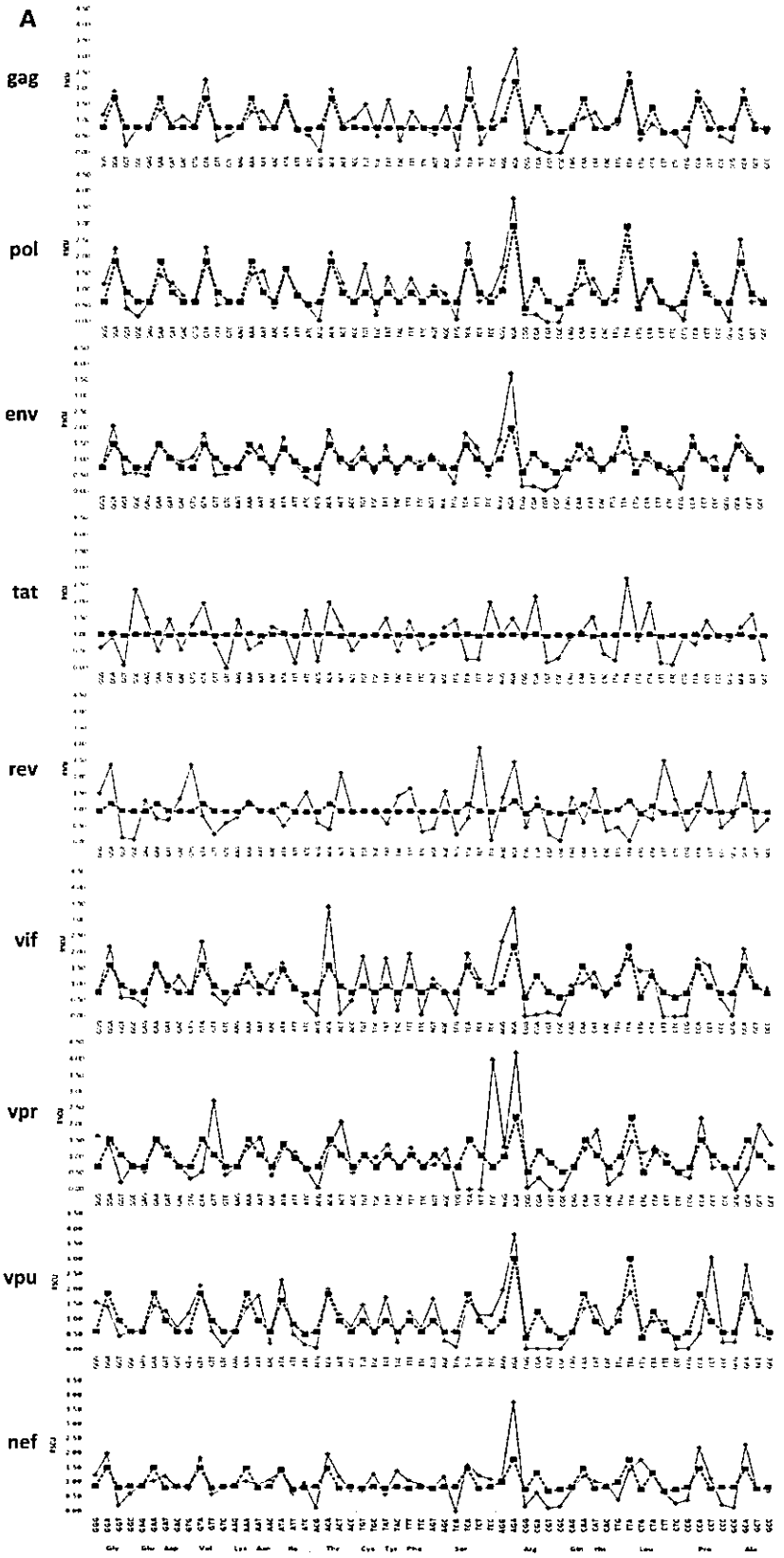
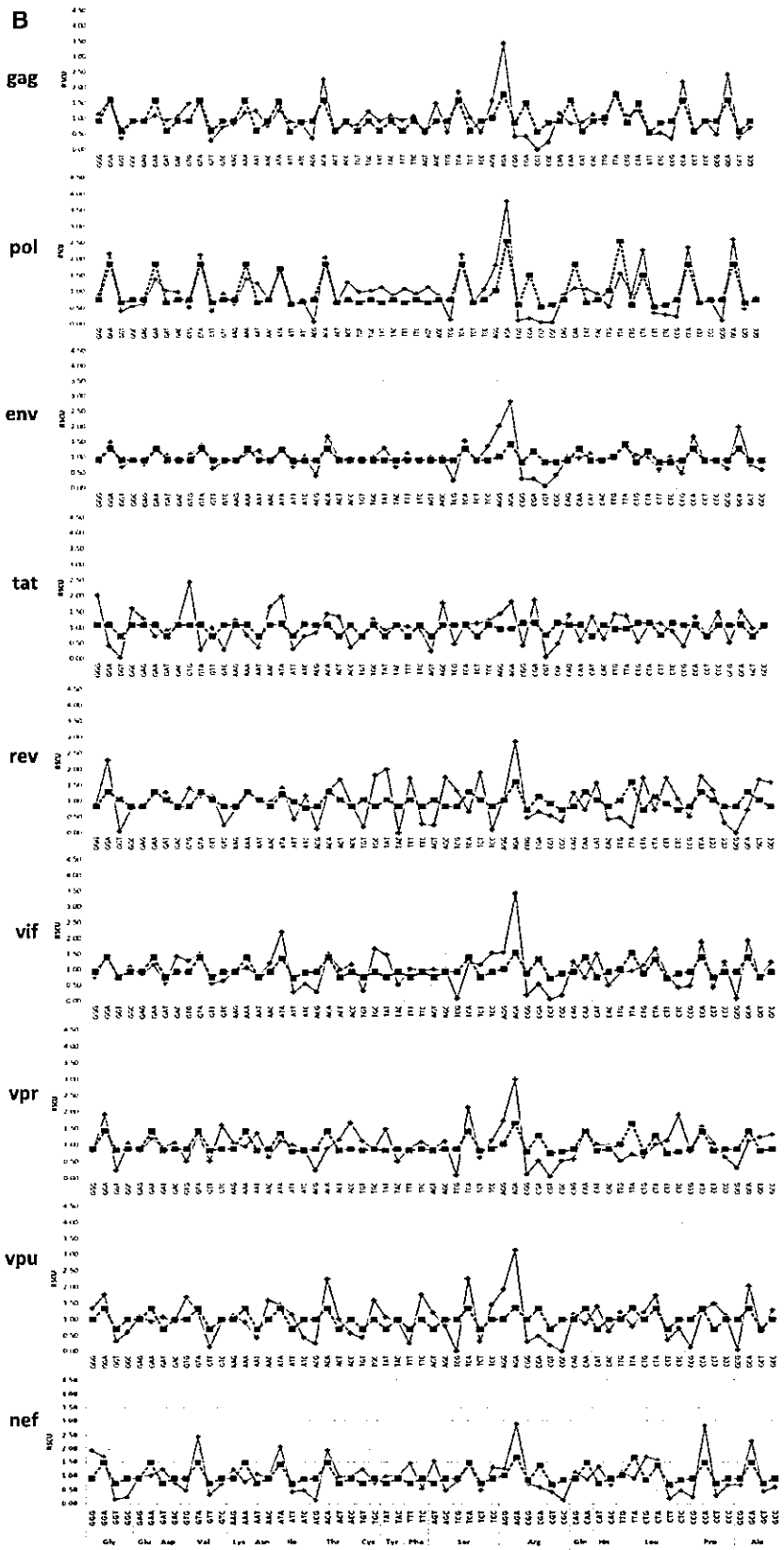


Fig. 1 Comparison of predicted codon usage and actual codon frequency in the nine genes of HIV-1 (**a**) and HIV-2 (**b**). Predicted codon usage was calculated by modified formulas from the model developed by Palidwor GA et al., and is showed with dashed lines while actual codon frequency is shown with solid lines. The codon usage of each codon (X-axis) is demonstrated as RSCU (Y-axis)

Fig. 1 (continued)



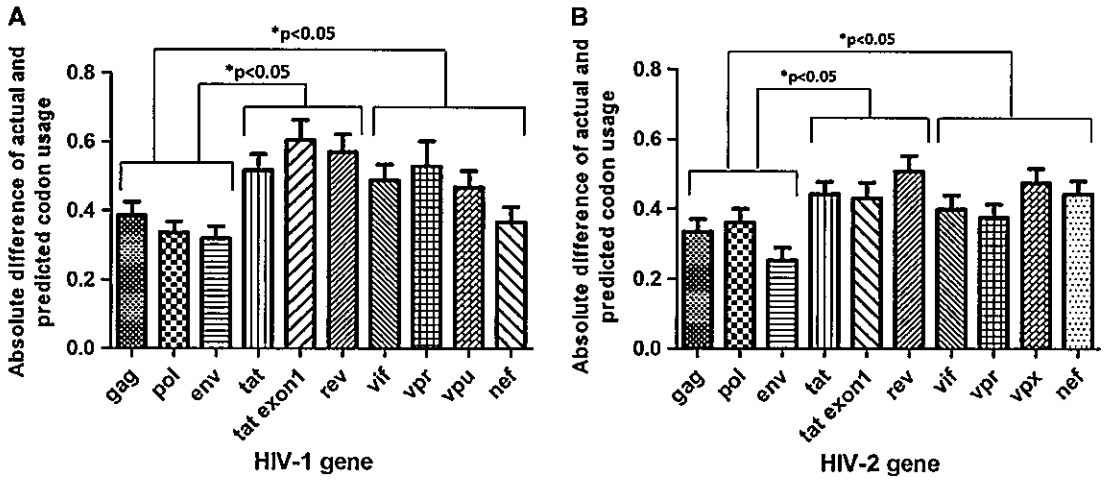


Fig. 2 Absolute differences in predicted codon usage and actual codon frequency within the nine HIV-1 (a) and HIV-2 genes (b). Three groups of nine genes (structural, regulatory and accessory) were analyzed using one-way ANOVA with Dunns' post-test for individual comparisons and was considered significant at $p \leq 0.05$

usage of *tat* exon 1 is further from significant prediction than that of full length HIV-1 and more comparable to full length HIV-2, suggesting that codon usage of the *tat* gene is truly unpredictable by base composition (Fig. 2). Because the *rev* exon 1 is very short and the result may be unreliable, we did not analyze its codon usage. In addition, the third codon position of *tat* exon 2 overlaps with the first codon position of *env* and the third codon position of *rev* exon 2 overlaps with the second codon position of *env*. The GC3 of *tat* and *rev* are dramatically higher than that of GC1 and GC2 of *env*, respectively, so the high GC3 of these genes is probably not a result of the overlapping codon position (1 and 2) within the *env* open reading frame.

The codon usage bias of each gene was also analyzed for 'effective number of codons' (ENC), which was introduced by Frank Wright in 1990 [27]. The ENC ranges from between 20 and 61. A value of 20 means an extreme bias as just one codon is used for each amino acid, whereas a value of 61 means no bias, as all codons are used equally [27, 28]. ENCs were plotted against GC3 for each HIV gene to analyze the effects of mutational pressure and natural selection (Fig. 3). ENC and GC3 plots for the regulatory and accessory genes are more variable than those of the structural genes. In HIV-1, the ENC-GC3 plot for the regulatory genes clusters separately from structural genes with high ENC and GC3, and in fact overlaps with the accessory genes (Fig. 3A). The high ENC and GC3 for the regulatory genes (mean ENC and GC3: 51.9 and 48.46, respectively) infer that these genes are relatively unbiased. On the other hand, the codon usage of the structural genes has a higher bias with a lower GC3 (mean ENC and GC3: 44.54 and 35.75, respectively) (Fig. 3A). The ENC and GC3 of the accessory genes are variable. *Vpu* shows the most bias and the lowest GC3, whereas *nef* is the least bias and contains the highest GC3 among the accessory genes. In HIV-2, the ENC and GC3 of the structural and accessory genes of HIV-2 are generally higher than those of HIV-1, indicating lower levels of codon bias, while HIV-2 *tat* and *rev* show comparable levels of ENC and GC3 when compared to HIV-1 *tat* and *rev* (Fig. 3B). Therefore the regulatory genes, *tat* and *rev*, contain the highest GC3 and are the least bias, in terms of codon usage, when compared to other HIV genes.

Furthermore the pattern of codon selection in evolution of the regulatory genes was compared to *gag*, a representative structural gene, using a 'Toward A-richness' ratio. The number of codon changes at each internal node for these genes was counted and grouped according to whether they resulted in A-ending or non A-ending codons, and a Toward A-richness ratio was calculated (Table 2). Codon changes within the *gag* gene distributed evenly among the four groups of codon changes. Since codons within the structural genes are generally A-rich, this even distribution suggested a lower rate of changes to A-ending codons and a relative purifying selection for A-ending codons. On the other hand, the regulatory genes *tat* and *rev* had much lower frequencies of mutations that resulted in A-ending codons and most of the changes were non A-ending codons. This resulted in markedly lower Toward A-richness ratios and suggested a purifying selection for non A-ending codons. These data further supported a difference in the evolution and selection of codon usage among the different classes of HIV genes.

The level of codon adaptation to specific hosts

The similarity of codon usage between foreign genes and the host genome is used to predict translation efficiency in host

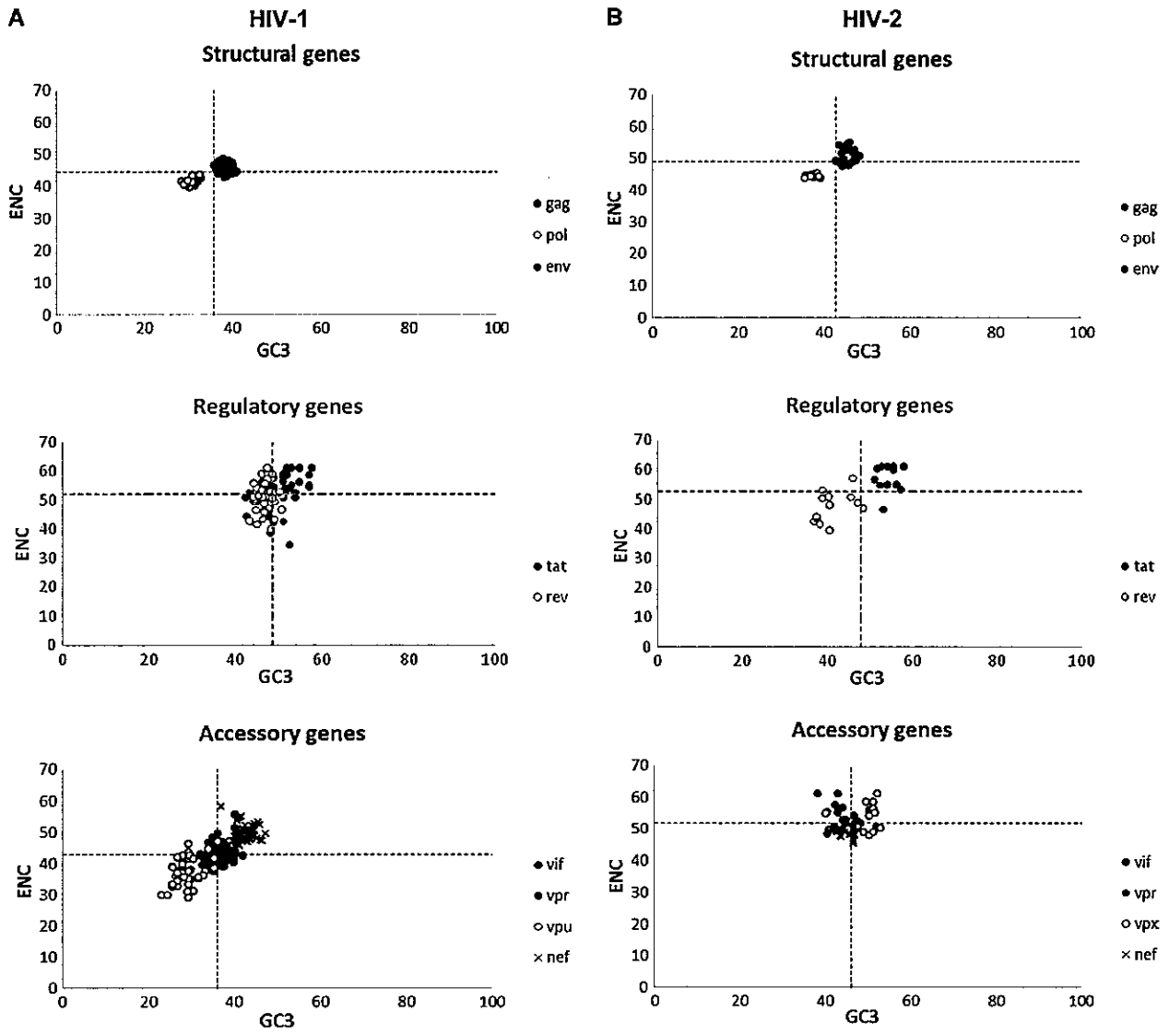


Fig. 3 Effective number of codons (ENC) and GC3 plots. The ENC-values for the nine HIV-1 (**a**) and HIV-2 (**b**) genes was plotted against the proportion of GC at the third codon position (GC3). Three groups of nine genes (structural, regulatory and accessory) were analyzed. The dashed lines indicate the average ENC and GC3 of each group

cells. A widely used tool to quantify codon usage similarity and predict translation efficiency is the codon adaptation index (CAI). This is performed by measuring the similarity between synonymous codon usage within a gene and the synonymous codon frequencies of a reference database. CAI ranges from zero to one, closer to one meaning frequencies that match the reference database [29]. A CAI analysis of HIV genes was performed, based on a reference human codon table derived from 93,487 coding sequences [26]. Analysis of the nine HIV-1 genes demonstrated that the highest CAI was in the regulatory genes. This means that codon usage of *tat* and *rev* genes is closer to humans than the other HIV-1 genes (Fig. 4A). The codon usage of the regulatory genes may need to be more compatible with the human host in order to be expressed efficiently during the early stages of infection. The CAI of *vif* and *vpr* were comparable to that of the structural genes. *Vpu* showed the lowest CAI among the nine genes and *nef* exhibited the highest CAI within the accessory genes. In HIV-2, the CAI of most genes was similar to HIV-1; however, *vpx* exhibited the highest CAI (Fig. 4B). The CAI of SIVcpz and SIVsmm was also investigated. The reference chimpanzee and sooty mangabey codon tables were derived from 118 and 30 coding sequences, respectively, and showed a similar trend for the CAI values as per HIV. The CAI of the regulatory genes was significantly higher than that of the structural genes.

Table 2 Number of codon changes (grouped according to A- and non A-ending) in the *gag*, *tat* and *rev* genes along all the branches of a phylogenetic tree

Gene	Σ (A to A)	Σ (non-A to A)	Σ (A to non-A)	Σ (non-A to non-A)	Toward A-richness ratio [Σ (A to A) + Σ (non-A to A)/Σ (A to non-A) + Σ (non-A to non-A)]
gag	237	210	245	237	0.927
tat	40	35	23	115	0.543
rev	34	31	42	131	0.376

In accessory genes, the *vpu* of SIVcpz showed the lowest CAI, whereas the *vpx* of SIVsmm showed the highest CAI (Fig. 4C, D). This pattern was similar to HIV-1 and HIV-2, respectively.

The adaptation of viral codon usage to the host was also compared between different hosts. Reference human, chimpanzee or sooty mangabey codon tables were used to calculate the CAI of HIV-1/SIVcpz and HIV-2/SIVsmm, respectively (Fig. 5). Unexpectedly, using the reference chimpanzee codon table, both HIV-1 and SIVcpz demonstrated higher CAIs than when using the reference human codon table, meaning that the codon usage of HIV-1 and SIVcpz may be more adapted to chimpanzees than human (Fig. 5A,B). Similarly, using a reference sooty mangabey codon table, both HIV-2 and SIVsmm demonstrated higher CAIs than the reference human codon table, inferring that HIV-2 and SIVsmm may be more adapted to sooty mangabeys than humans (Fig. 5C, D). These data suggest that the codon usage of HIV-1and HIV-2 have not moved closer to humans than chimpanzees or sooty mangabeys, and that SIVcpz and SIVsmm prefer the codon usage of their original hosts. The evolutionary time delineating SIV from HIV may not be long enough for a relationship between HIV codon usage to have been established.

Overall, the codon usage pattern did vary within the HIV genome. The early expressed genes, *tat* and *rev*, responsible for regulation of the replication cycle were more similar to humans in terms of codon usage and GC content, than the other genes. This may help these genes to be expressed more efficiently during the early stages of infection.

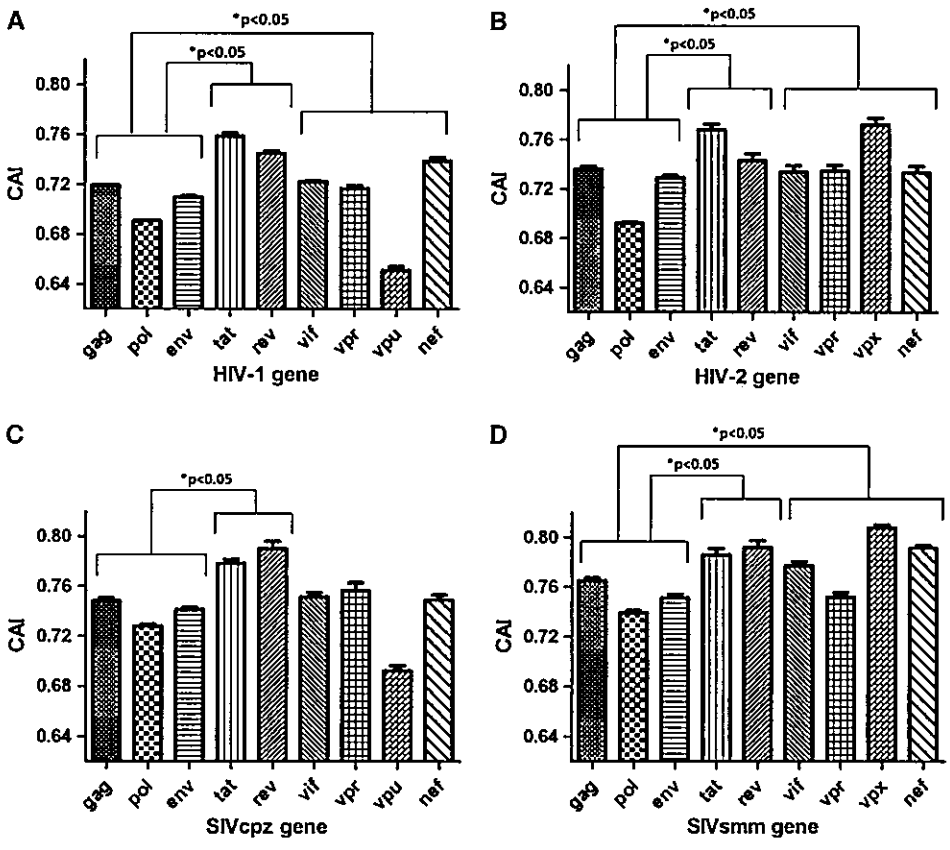


Fig. 4 The codon adaptation index (CAI) of HIV-1 (**a**) and HIV-2 (**b**) genes was compared to a reference human codon usage database. The CAI of SIVcpz (**c**) and SIVsmm (**d**) genes was compared to a reference chimpanzee and sooty mangabey codon usage database. Three groups of nine genes (structural, regulatory and accessory) were analyzed using one-way ANOVA using Dunns' post-test for individual comparisons and was considered significant at $p \leq 0.05$

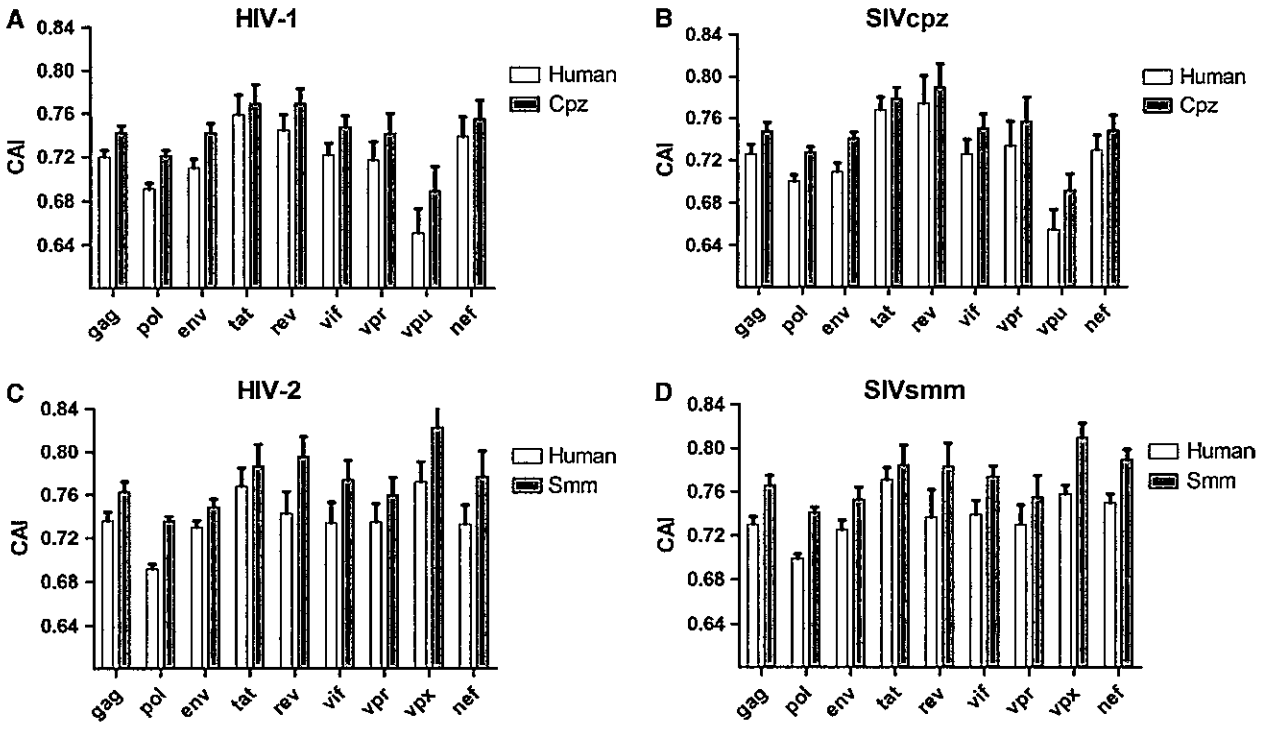


Fig. 5 Comparison of CAI calculated using the different reference host codon tables. The CAI of HIV-1 (a) and SIVcpz (b) genes was compared to reference human and chimpanzee codon tables. The CAI of HIV-2 (c) and SIVsmm (d) genes was compared to reference human and sooty mangabey codon tables

Discussion

HIV has caused a global public health problem; one of the main difficulties in coping with the epidemic is the extremely high viral genetic variability resulting from rapid replication and an extremely high mutation rate [30]. Although the main selective force for HIV evolution is host immunity, functional selection and other constraints also contribute. Understanding the biological components of viral genome evolution will help us to better understand viral evolution and diversity. Functionally, nucleotide composition and codon usage are important components exhibiting biological constraint on the viral genome. Nucleotide composition and codon usage are interlinked and both can affect each other. Theoretically, the difference in nucleotide composition and codon usage between the regulatory and structural genes of HIV could be caused by either uneven mutational bias along the genome or functional selection on either nucleotide composition or codon translation efficient. It was previously described in the selection-mutation-drift theory that the use of synonymous codons in a gene is a balance between the forces of mutation and selection [31]. A-richness is one of the characteristics of nucleotide composition within HIV genomes, a feature which is shared with most members of the genus *Lentivirus* [18, 23]. This is a result of G-to-A hypermutation caused by APOBEC cytidine deaminase enzymes, innate cellular anti-retroviral proteins [14–16]. On the other hand, A-to-G hypermutation can be caused by RNA editing through the adenosine deaminase enzymes, ADAR. RNA editing by ADAR has been shown in various viruses including HIV-1 [32–35]. These editing mechanisms, as well as the mutational bias of the viral reverse transcriptase itself, dictate the viral genome's nucleotide composition. RNA editing by ADAR was shown to depend on the target RNA sequence and secondary structure [36]. This suggests that hypermutation could be uneven along the viral genome, which would result in different nucleotide compositions in different genes or regions.

The nucleotide composition of viral RNA can have functional consequences. A previous study showed that local variation in A-richness could affect the stability of the local RNA structure. Reduction of A content in the *gag-pol* ORF resulted in excessive stability of the viral RNA and had a negative effect on reverse transcription resulting in a reduction in cDNA synthesis [37]. Another study showed that suppression of CpG dinucleotide content, to avoid silencing through methylation, might influence HIV-1 genome composition [38]. The biased nucleotide composition of HIV-1 was also shown to be responsible for the induction of a type I interferon response, whereas human codon optimized *gag*,

pol and *env* RNA transcripts lost the ability to induce IFN-α/β production [39]. These functional effects may act as selective pressures to maintain optimal viral genome nucleotide composition, consequently affecting viral codon usage.

Considering the effect of translational selection on codon usage, optimal codons are more frequently found in highly expressed genes. These are believed to be translated more rapidly and accurately than other synonymous codons [40, 41]. As HIV requires the host's translational machinery to translate their mRNA, host tRNA pools must be important for viral replication, at least at the early steps. The distinct codon usage of the regulatory genes might support the level, rate and accuracy of translation. *Tat* and *rev* are early viral gene products which contain more optimal codons for expression in human than the other HIV genes. Thus this might provide a benefit for HIV as these transcripts may be rapidly translated at the early steps of genome replication, using tRNA pools in a way similar to the normal host mRNAs. In addition, another study has suggested that there could be alteration in the host's tRNA pool at the later stages of infection to suit late gene transcript's translation [42].

The nucleotide composition of the lentiviral genome is strikingly stable over time, even in a highly variable virus such as HIV. Between 1983 and 2009 the base composition varied less than 1%, per base position per isolate [18]. On the other hand, evidence of gradual adaptation of HIV-1 codon usage over a period of two decades since the early phases of the epidemic have also been observed [17]. This suggests adaptation of HIV-1 codon usage to the new human host. The stability in the nucleotide composition suggests that codon adaptation may have evolved under specific constraints. Although SIV has spent much longer in its natural host, its adaptation to chimpanzee codon usage was found to be comparable to that of HIV, with a similar pattern of high adaptation in regulatory genes. This suggests that viral adaptation of codon usage is counteracted by mutational bias and other constraints acting on genome composition. In addition, the codon usage of HIV-2 is less biased than that of HIV-1 in the structural and accessory genes as indicated by higher GC3s and ENCs, which resulted in a higher CAI. This implies that HIV-2 may be more adapted to humans. This also applies to SIVcpz and SIVsmm, which are the predecessors of HIV-1 and HIV-2, respectively. SIVsmm also demonstrated higher CAI than SIVcpz in most of the structural and accessory genes. It is possible that the SIVsmm and HIV-2 lineage may be better adapted naturally to their hosts than the SIVcpz and HIV-1 lineage. Alternatively, SIVcpz and HIV-1 may be more efficient in adjusting the tRNA pool and translational machinery to suit their codon usage at the later stages of infection, and thus are capable of having their codon usage more deviated from that of their hosts. Nevertheless, the CAIs of the regulatory genes of HIV-1 and HIV-2/ SIVcpz and SIVsmm are comparable, which suggests that there is a necessity to efficiently express the regulatory genes at the early stages of infection in the context of a normal cellular environment. Our data adds to the understanding of interactions and driving forces that shape HIV genome composition, codon usage and ultimately their evolution.

Acknowledgements This work was financial supported by the Thailand Research Fund through the Royal Golden Jubilee Ph.D. Program (Grant No. PHD/0030/2556) and a research grant (Grant No. IRN60W0002) from Thailand Research Fund.

Compliance with ethical standards

Conflict of interest The authors declare no conflicts of interest.

Research involving human participants and/or animals No part of this study was performed with human participants or animals.

References

1. Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10:7055–7074
2. Gustafsson C, Govindarajan S, Minshull J (2004) Codon bias and heterologous protein expression. Trends Biotechnol 22:346–353
3. Hu X, Shi Q, Yang T, Jackowski G (1996) Specific replacement of consecutive AGG codons results in high-level expression of human cardiac troponin T in Escherichia coli. Protein Expr Purif 7:289–293
4. Deng T (1997) Bacterial expression and purification of biologically active mouse c-Fos proteins by selective codon optimization. FEBS Lett 409:269–272
5. Kotula L, Curtis PJ (1991) Evaluation of foreign gene codon optimization in yeast: expression of a mouse IG kappa chain. Biotechnology (NY) 9:1386–1389
6. Burgess-Brown NA, Sharma S, Sobott F, Loenarz C, Oppermann U, Gileadi O (2008) Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study. Protein Expr Purif 59:94–102
7. Wong EH, Smith DK, Rabadan R, Peiris M, Poon LL (2010) Codon usage bias and the evolution of influenza A viruses. Codon usage biases of influenza virus. BMC Evol Biol 10:253
8. Belalov IS, Lukashev AN (2013) Causes and implications of codon usage bias in RNA viruses. PLoS One 8:e56642. doi:10.1371/journal.pone.0056642
9. Bulmer M (1987) Coevolution of codon usage and transfer RNA abundance. Nature 325:728–730
10. Ermolaeva MD (2001) Synonymous codon usage in bacteria. Curr Issues Mol Biol 3:91–97
11. Haas J, Park EC, Seed B (1996) Codon usage limitation in the expression of HIV-1 envelope glycoprotein. Curr Biol 6:315–324
12. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH (2004) Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci USA 101:3480–3485
13. Sharp PM, Stenico M, Peden JF, Lloyd AT (1993) Codon usage: mutational bias, translational selection, or both? BiochemSoc Trans 21:835–841
14. Knoepfel SA, Di Giallonardo F, Däumer M, Thielen A, Metzner KJ (2011) In-depth analysis of G-to-A hypermutation rate in HIV-1 env DNA induced by endogenous APOBEC3 proteins using massively parallel sequencing. J Virol Methods 171:329–338

15. Imahashi M, Nakashima M, Iwatani Y (2012) Antiviral mechanism and biochemical basis of the human APOBEC3 family. Front Microbiol 3:250
16. Romani B, Engelbrecht S, Glashoff RH (2009) Antiviral roles of APOBEC proteins against HIV-1 and suppression by Vif. Arch Virol 154:1579–1588
17. Pandit A, Sinha S (2011) Differential trends in the codon usage patterns in HIV-1 genes. PLoS One 6:e28889. doi:10.1371/journal.pone.0028889
18. van der Kuyl AC, Berkhout B (2012) The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus. Retrovirology 9:92. doi:10.1186/1742-4690-9-92
19. Palidwor GA, Perkins TJ, Xia X (2010) A general model of codon bias due to GC mutational bias. PLoS One 5:e13431. doi:10.1371/journal.pone.0013431
20. van Hemert FJ, Berkhout B, Lukashov VV (2007) Host-related nucleotide composition and codon usage as driving forces in the recent evolution of the Astroviridae. Virology 361:447–454
21. Butt AM, Nasrullah I, Qamar R, Tong Y (2016) Evolution of codon usage in Zika virus genomes is host and vector specific. Emerg Microbes Infect 5:e107. doi:10.1038/emi.2016.106
22. Gilchrist MA, Coombs D (2006) Evolution of virulence: interdependence, constraints, and selection using nested models. Theor Popul Biol 69:145–153
23. Kypr J, Mrazek J (1987) Unusual codon usage of HIV. Nature 327:20
24. Stothard P (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 28:1102–1104
25. Puigbò P, Bravo IG, Garcia-Vallve S (2008) CAIcal: a combined set of tools to assess codon usage adaptation. Biol Direct 3:38. doi:10.1186/1745-6150-3-38
26. Nakamura Y, Gojobori T, Ikemura T (2000) Codon usage tabulated from international DNA sequence databases. Nucleic Acids Res 28:292
27. Wright F (1990) The 'effective number of codons' used in a gene. Gene 87:23–29
28. Fuglsang A (2006) Estimating the "effective number of codons": the Wright way of determining codon homozygosity leads to superior estimates. Genetics 172:1301–1307
29. Sharp PM, Li WH (1987) The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295
30. Turner BG, Summers MF (1999) Structural biology of HIV. J Mol Biol 285:1–32
31. Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907
32. Cattaneo R, Schmid A, Eschle D, Baczko K, ter Meulen V, Billeter MA (1988) Biased hypermutation and other genetic changes in defective measles viruses in human brain infections. Cell 55:255–265
33. Samuel CE (2012) ADARs: viruses and innate immunity. Curr Top Microbiol Immunol 353:163–195. doi:10.1007/82_2011_148
34. Jayan GC, Casey JL (2002) Increased RNA editing and inhibition of hepatitis delta virus replication by high-level expression of ADAR1 and ADAR2. J Virol 76:3819–3827
35. Doria M, Neri F, Gallo A, Farace MG, Michienzi A (2009) Editing of HIV-1 RNA by the double-stranded RNA deaminase ADAR1 stimulates viral infection. Nucleic Acids Res 37:5848–5858
36. Sapiro AL, Deng P, Zhang R, Li JB (2015) Cis regulatory effects on A-to-I RNA editing in related Drosophila species. Cell Rep 11:697–703
37. Keating CP, Hill MK, Hawkes DJ, Smyth RP, Isel C, Le SY et al (2009) The A-rich RNA sequences of HIV-1 pol are important for the synthesis of viral cDNA. Nucleic Acids Res 37:945–956
38. Meintjes PL, Rodrigo AG (2005) Evolution of relative synonymous codon usage in Human Immunodeficiency Virus type-1. J Bioinform Comput Biol 3:157–168
39. Vabret N, Bailly-Bechet M, Najburg V, Müller-Trutwin M, Verrier B, Tangy F (2012) The biased nucleotide composition of HIV-1 triggers type I interferon response and correlates with subtype D increased pathogenicity. PLoS One 7:e33502. doi:10.1371/journal.pone.0033502
40. Ray SK, Baruah VJ, Satapathy SS, Banerjee R (2014) Cotranslational protein folding reveals the selective use of synonymous codons along the coding sequence of a low expression gene. J Genet 93:613–617
41. Yu CH, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS et al (2015) Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. Mol Cell 59:744–754
42. van Weringh A, Ragonnet-Cronin M, Pranckeviciene E, Pavon-Eternod M, Kleiman L, Xia X (2011) HIV-1 modulates the tRNA pool to improve translation efficiency. Mol Biol Evol 28:1827–1834