Genes & Genomics (2022) 44:773–791
https://doi.org/10.1007/s13258-022-01257-6

Online ISSN 2092-9293
Print ISSN 1976-9571

RESEARCH ARTICLE

Analysis of evolutionary and genetic patterns in structural genes of primate lentiviruses

Myeongji Cho[1,2] · **Xianglan Min**[1] · **Hyeon S. Son**[1,2,3]

Received: 2 February 2022 / Accepted: 9 April 2022 / Published online: 5 May 2022

Abstract

Background Primate lentiviruses (HIV1, HIV2, and *Simian immunodeficiency virus* [SIV]) cause immune deficiency, encephalitis, and infectious anemia in mammals such as cattle, cat, goat, sheep, horse, and puma.

Objective This study was designed and conducted with the main purpose of confirming the overall codon usage pattern of primate lentiviruses and exploring the evolutionary and genetic characteristics commonly or specifically expressed in HIV1, HIV2, and SIV.

Methods The *gag*, *pol*, and *env* gene sequences of HIV1, HIV2, and SIV were analyzed to determine their evolutionary relationships, nucleotide compositions, codon usage patterns, neutrality, selection pressure (influence of mutational pressure and natural selection), and viral adaptation to human codon usage.

Results A strong 'A' bias was confirmed in all three structural genes, consistent with previous findings regarding HIV. Notably, the ENC-GC3s plot and neutral evolution analysis showed that all primate lentiviruses were more affected by selection pressure than by mutation caused by the GC composition of the gene, consistent with prior reports regarding HIV1. The overall codon usage bias of *pol* was highest among the structural genes, while the codon usage bias of *env* was lowest. The virus groups showing high codon bias in all three genes were HIV1 and SIVcolobus. The codon adaptation index (CAI) and similarity D(A, B) values indicated that although there was a high degree of similarity to human codon usage in all three structural genes of HIV, this similarity was not caused by translation pressure. In addition, compared with HIV1, the codon usage of HIV2 is more similar to the human codon usage, but the overall codon usage bias is lower.

Conclusion The origin viruses of HIV (SIVcpz_gor and SIVsmm) exhibit greater similarity to human codon usage in the *gag* gene, confirming their robust adaptability to human codon usage. Therefore, HIV1 and HIV2 may have evolved to avoid human codon usage by selection pressure in the *gag* gene after interspecies transmission from SIV hosts to humans. By overcoming safety and stability issues, information from codon usage analysis will be useful for attenuated HIV1 vaccine development. A recoded HIV1 variant can be used as a vaccine vector or in immunotherapy to induce specific innate immune responses. Further research regarding HIV1 dinucleotide usage and codon pair usage will facilitate new approaches to the treatment of AIDS.

Keywords Human immunodeficiency virus 1 · Human immunodeficiency virus 2 · Simian immunodeficiency virus · Codon usage pattern · Relative synonymous codon usage · Mutational pressure · Natural selection

Abbreviations

HIV	Human immunodeficiency virus
AIDS	Acquired immunodeficiency syndrome
SIV	Simian immunodeficiency virus
A, C, G, and T	Adenine, cytosine, guanine and thymine
CD4+	Cluster of differentiation 4 +
dUTPase	DUTP diphosphatase

Myeongji Cho and Xianglan Min are co-first authors and contributed equally to this work.

✉ Hyeon S. Son
hss2003@snu.ac.kr

Extended author information available on the last page of the article

APOBEC3G	Apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like 3G
ZAP	Zinc finger antiviral protein
PR2	Parity rule 2
Val, Pro, Thr, Ala, Gly, Leu, Ser, Arg, Met, Trp, Tyr, Cys	Valine, Proline, Threonine, Alanine, Glycine, Leucine, Serine, Arginine, Methionine, Tryptophan, Tyrosine, Cysteine
RSCU	Relative synonymous codon usage
ENC	Effective number of codons
ICDI	Intrinsic codon deviation index
CAI	Codon adaptation index
eCAI	Expected codon adaptation index

Introduction

Human immunodeficiency virus 1 (HIV1) is the representative species of the lentivirus genus and the cause of the acquired immunodeficiency syndrome (AIDS) pandemic, an ongoing public health problem. Nearly 40 years after the first AIDS case was reported, neither vaccine nor cure has been developed. The *Lentivirus* genus belongs to the *Retroviridae* family and *Orthoretrovirinae* subfamily; this genus currently comprises 10 species. Lentiviruses other than primate lentiviruses (HIV1, HIV2, and *Simian immunodeficiency virus* [SIV]) cause immune deficiency, encephalitis, and infectious anemia in mammals such as cattle, cat, goat, sheep, horse, and puma. HIV species comprise the genetically distinct HIV1 and HIV2. Although they have the same transmission mode and similar immunological effects, HIV2 has lower pathogenicity and infectivity, longer incubation period, lower likelihood of AIDS development, and lower plasma viral load, compared with HIV1 (Vidyavijayan et al. 2017). In addition, unlike the global spread of HIV1, HIV2 is mainly confined to West Africa; however, it has recently been reported in several countries, including Angola, Mozambique, Portugal, France, the United States, and India (Visseaux et al. 2016). Genomic nucleotide sequence analysis of HIV1 and HIV2 has shown only 55% identity (Motomura et al. 2008); moreover, each species encodes distinct accessory proteins. In addition to the three structural proteins gag, pol, and env, HIV1 encodes tat, rev, vif, nef, vpr, and vpu regulatory/accessory proteins; HIV2 encodes the vpx protein, instead of vpu. According to genetic distance and phylogenetic analyses, HIV1 can be divided into M (main), O (outlier), N (not M-not O), and P groups. The M group is a global epidemic group, comprising 95% of HIV1; it is further divided into 10 subtypes (A-D, F-H, J-L) with the same genetic distance and circulating recombination forms formed by recombination between subtypes. HIV2 has eight lineages (A-G, U); HIV2 lineages are identical to HIV1 groups in terms of genetic distance or origin (Sharp et al. 1999; Robertson et al. 2000). According to epidemiological and phylogenetic studies regarding the *gag*, *pol*, and *env* genes of primate lentiviruses, groups M and N of HIV1 were transmitted from SIVcpz (Gao et al. 1999), while groups O and P were transmitted from SIVgor (Sharp and Hahn 2010; Hemelaar 2012), by at least four independent crossover events. At least eight independent transmissions of viruses from sooty mangabeys to humans gave rise to eight lineages of HIV2. Thus far, SIV infection has been confirmed in more than 40 non-human primates; although it does not cause disease in its native hosts (Rey-Cuille et al. 1998), old world monkeys, SIVmac (which originated from SIVsmm through an interspecies crossover event) causes simian AIDS in Asian macaques (Apetrei et al. 2005). Furthermore, it was initially presumed not to cause disease in chimpanzees, but Keele et al. confirmed that SIVcpz causes progressive cluster of differentiation 4+(CD4+) T cell loss and premature death in chimpanzees (in a manner similar to HIV1); these effects negatively impact health and reproduction. All lentiviruses have three structural genes in common (*gag*, *pol*, and *env*); however, unlike other lentiviruses, primate lentiviruses do not have a dUTP diphosphatase (dUTPase)-coding region in the *pol* open reading frame (Foley 2000).

Although HIV has a high mutation rate and undergoes frequent recombination, its genomic nucleotide composition has been robustly conserved for a long period of time, maintaining high adenine (A) content (up to 40%) and low cytosine (C) content (van der Kuyl and Berkhout 2012). Greater A-bias was observed in the *pol* gene and generally low A-bias was observed in the region overlapping with other regulatory/accessory protein-coding genes (van Hemert and Berkhout 1995). This high A content has been suggested as the cause of guanine(G) → A hypermutation of HIV1 RNA, mainly because of dNTP pool imbalance during reverse transcription (Vartanian et al. 2002). Apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like 3G (APOBEC3G), a human host restriction factor, also contributes to G → A hypermutation; however, this effect is weaker than the impact caused by dNTP pool imbalance (Yu et al. 2004). Therefore, although the vif of HIV binds to APOBEC3G and degrades it, the HIV genome has a high A-biased base composition pattern (Sheehy et al. 2002; Desimmie et al. 2014). The A-rich nucleotide composition of the HIV genome increases the selectivity for A-rich codons and tends to mainly substitute A for the third codon position, the most neutral part of the codon. An A-biased synonymous codon usage is observed in HIV; HIV1 shows a codon usage pattern lacking CpG dinucleotides, which are characteristic of eukaryotes. The zinc finger antiviral protein (ZAP), a host cellular antiviral protein, specifically targets a region with a high CpG dinucleotide content of HIV1 genomic RNA and repressed translation, thereby inhibiting viral replication (Ficarelli et al. 2020). This may promote the suppressed use of CpG dinucleotides in HIV1.

Most codon usage analyses of HIV have focused on describing the codon usage patterns of all nine genes (Pandit and Sinha 2011), comparing structural genes' codon usage patterns among geographical locations (Ahn and Son 2006), identifying codon usage pattern changes over time (year or early/late infection stages at the patient level) (Pandit and Sinha 2011; Meintjes and Rodrigo 2005), and contrasting the codon usages of HIV1 and HIV2 (Vidyavijayan et al. 2017). However, the indices used have been limited; codon usage analysis performed at the level of the primate lentivirus group, which includes more than 40 species of SIV, is insufficient. Therefore, we explored the evolutionary and genetic codon usage characteristics common and specific to the structural genes (*gag*, *pol*, and *env*) of HIV1, HIV2, and SIV.

Materials and methods

Data collection

Sequence data for phylogeny, compositional analysis, and codon usage pattern analysis were collected from NCBI Virus with a specific focus on the *gag*, *pol*, and *env* open reading frames of HIV1, HIV2, and SIV. To ensure the accuracy of sequence data, data cleaning including checking the length of each sequence (whether a multiple of 3), the presence of stop codons, and the inclusion of characters other than ATCG was performed in JAVA.

Phylogenetic analysis

To confirm the evolutionary relationships between HIV1, HIV2, and SIV, five HIV1 and HIV2 sequences were randomly selected, as well as two SIV sequences for each host species to construct a phylogenetic tree for each *gag*, *pol*, and *env* gene. As shown in Table 1, we collected the most available sequences of each *gag*, *pol*, and *env* gene to obtain sufficient data and manually selected accurate sequence information for codon usage indices analysis (Table 1). The phylogenetic analysis in our study had two main objectives. The first objective was to confirm phylogenetic relationships among the primate lentiviruses isolated from human hosts (HIV1 and HIV2) and non-human primate hosts (SIV) (Table 2). The second objective was to group sequences for codon usage indices analysis based on clustering patterns within the three gene trees including genus, tribe, and subfamily information of the SIV host. When grouping by host based on the *gag* and *pol* gene sequences, we had to consider the group SIVcer with the minimum number of available sequences (5) and, therefore, trees were repeatedly created by building datasets with sequences randomly extracted from the sequence pool for each group (Table 1). Afterwards, we confirmed that the clustering pattern was consistent among all reconstructed trees. Regarding SIV, some host species only had one sequence available; therefore, 60, 59, and 53 sequences for each *gag*, *pol*, and *env* gene were used to construct the final phylogenetic tree, respectively. After multiple sequence alignments had been performed using ClustalW (Thompson et al. 1994), phylogenetic trees were constructed using the maximum-likelihood method (Kishino and Hasegawa 1989) based on the Kimura-2 parameter model through MEGA-X (Kumar et al. 2018). The bootstrap parameter was set to 1000 to measure the reliability of each internal node. To minimize possible bias among the phylogenetic trees, the above process was performed five times for each *gag*, *pol*, and *env* gene; a representative tree with a high degree of agreement was generated. The evolutionary relationships of primate lentiviruses were confirmed. Regarding SIV, codon usage analysis grouping was performed based on the phylogenetic tree results and considering the genus, tribe, and subfamily of the host. The results of codon usage comparative analysis were interpreted considering the phylogenetic relationships among lentivirus species.

Table 1 Data collected for analysis of codon usage indices

Virus	Polyprotein	The number of data
HIV-1	Gag	22,699
	pol	11,021
	env	70,658
HIV-2	gag	76
	pol	45
	env	170
SIV	gag	750
	pol	609
	env	15,416

HIV-1 human immunodeficiency virus 1, *HIV-2* human immunodeficiency virus 2, *SIV* simian immunodeficiency virus

Compositional analysis

The nucleotide compositions of the *gag*, *pol*, and *env* coding sequences of HIV1, HIV2, and SIV were analyzed using CodonW (https://sourceforge.net/projects/codonw/) and CAIcal (http://genomes.urv.es/CAIcal/) programs. The relative frequencies of the overall (A%, T%, C%, and G%) and third codon (A3%, T3%, C3%, and G3%) positions of each

Table 2 SIV host information

Subfamily	Tribe	Genus	Species	Abbreviation
Cercopithecinae	Cercopithecini	Allochrocebus	lhoesti	SIVlst
			solatus	SIVsun
		Chlorocebus	aethiops	SIVagm-grv
			cynosuros	SIVagm-mal
			sabaeus	SIVagm-sab
			tantalus	SIVagm-tan
		Cercopithecus	albogularis	SIVsyk
			ascanius	SIVasc
			ascanius_whitesidei	SIVasw
			cephus	SIVmus
			mitis	SIVblu
			mona	SIVmon
			neglectus	SIVdeb
			nictitans	SIVgsn
		Miopithecus	ogouensis	SIVtal
	Papioini	Cercocebus	agilis	SIVagi
			atys	SIVsmm
			torquatus	SIVrcm
		Macaca	arctoides	SIVstm
			fascicularis	SIVcem
			mulatta	SIVmac
			nemestrina	SIVmne
Cercopithecinae	Papioini	Mandrillus	sphinx	SIVmnd-1
				SIVmnd-2
			leucophaeus	SIVdrl
Colobinae		Colobus	guereza	SIVcol
		Piliocolobus	badius_badius	SIVwrc
			badius_temminckii	SIVtrc
		Procolobus	verus	SIVolc
Homininae	Gorillini	Gorilla	gorilla gorilla	SIVgor
	Hominini	Pan	troglodytes_schweinfurthii	SIVcpz-eas
			troglodytes_troglodytes	SIVcpz-cen
			troglodytes_verus	SIVcpz-wes

Species are indicated by their Latin names, and the standard name of the species is preceded by the name of its genus

nucleotide, as well as the contents of GC and AT for the whole sequence and third codon position, were intuitively visualized using MS Excel 2016.

Parity rule 2 (PR2) analysis

For the *gag*, *pol*, and *env* genes of HIV1, HIV2, and SIV, PR2 plots were generated for the four amino acid codons (Valine [Val], Proline [Pro], Threonine [Thr], Alanine [Ala], Glycine [Gly]; Leucine [Leu]: CTT, CTC, CTA, CTG; Serine [Ser]: TCT, TCC, TCA, TCG; and Arginine [Arg]: CGT, CGC, CGA, CGG). In the PR2 plot, the x-axis is G3/(G3 + C3) and the y-axis is A3/(A3 + T3), indicating GC bias and AT bias, respectively. The G3/(G3 + C3) and A3/(A3 + T3) mean values for HIV1, HIV2, and SIV are shown as different icons; the relative positions of the lentivirus groups from the (0.5, 0.5) position, where mutation pressure and selection pressure are offset, were used to identify the direction and magnitude of the PR2 bias.

Neutrality analysis

A neutral evolutionary analysis was performed to quantify mutation and selection pressures by comparing the GC contents (GC12s and GC3s) of three codon positions. The effect of mutation usually has "directionality" toward higher or lower GC content; this directional mutation pressure often acts more on neutral parts (e.g., the third codon position) than on functionally important regions (Sueoka 1988). Mutations in the first and second codon positions, except six degenerate amino acids, are non-synonymous mutations that change the primary structure of a protein and eventually affect its function; these do not occur often. Hence, there are minimal differences in GC12s content for each sequence of the same virus under large selection pressure. Using the neutrality plot, the relative magnitude of the directional mutation pressure can be estimated by the degree of change in GC12s according to changes in GC3s (i.e., the slope of the regression line). For each of the *gag*, *pol*, and *env* genes, we created a scatter plot with GC12s as the ordinate and GC3s as the abscissa to determine whether the dots were correlated for each virus group. If a correlation existed, a regression line was generated to identify the major pressures on the codon usage pattern of the gene. A regression line slope of 1 (i.e., all points were distributed diagonally) was regarded as no or weak Darwinian selection pressure; in this context, codon usage bias is mediated solely by the mutation pressure. Conversely, a regression line slope of 0 (i.e., all points were located on a line parallel to the abscissa) indicated no difference in the GC content of the first and second codon positions of all gene sequences; in this context, evolution is driven solely by selection pressure.

Relative synonymous codon usage analysis

Relative synonymous codon usage (RSCU) is defined as the observed frequency of a codon divided by its expected frequency, which is the mean of all synonymous codons encoding that amino acid (Sharp and Li 1986). Excluding Methionine (Met), Tryptophan (Trp), and three stop codons, 59 RSCU values for each codon were allotted to a specific coding sequence. RSCU values are not affected by the sequence length and amino acid composition of the gene, enabling comparisons among genes of different lengths, as well as comparisons between amino acids (Sharp and Li 1986). A codon RSCU value of > 1 indicates that the corresponding codon is used frequently, rather than randomly. Moreover, a codon RSCU value of > 1.6 indicated over-representation, while a codon RSCU value of < 0.6 indicated under-representation in the coding sequence (Wong et al. 2010). RSCU values were calculated using CodonW and CAIcal softwares. The mean values of codons were obtained for each lentivirus group; a scatterplot was generated for each *gag*, *pol*, and *env* gene using XLSTAT for comparative analysis.

Effective number of codons and intrinsic codon deviation index analysis

The effective number of codons (ENC) is a simple measure of overall codon usage bias, which can be easily calculated from the codon frequency table; it is minimally influenced by gene length and amino acid composition. The ENC can assume values from 20 to 61; 20 indicates extreme codon usage bias (only one codon is used for each amino acid) and 61 indicates no bias (alternative synonymous codons are used equally for the corresponding amino acid) (Wright 1990). In general, an ENC value of < 35 indicates significant codon usage bias and an ENC value of > 50 indicates general random codon usage (Jiang et al. 2008). In general, it is considered that the severe codon usage bias is due to the high selection pressure, but there is a limit to suggesting such selection pressure only with ENC values. Since the GC3s % value of the sequence affects the ENC value, an ENC-GC3s plot (the X-axis is %GC3s and the Y-axis is ENC) is usually used to determine the magnitude of the selective pressure. To determine the mutation pressure affected by each gene in the virus, a curve of expected ENC values versus GC3s % values was generated assuming no selection pressure. If the ENC value of a gene is located near the expected curve, the gene is presumed to be under mutational pressure only. Conversely, as the distance from the expected curve increases, the codon usage pattern is mainly formed by natural selection or a pressure other than mutation pressure. The intrinsic codon deviation index (ICDI) is a useful measure of the codon bias of genes from species in which the optimal codons are unknown (Freirepicos et al. 1994); it is highly correlated with ENC and codon adaptation index (CAI). The ICDI ranges from 0 (for nonbiased genes) to 1 (for highly biased genes). Generally, an ICDI value of < 0.3 is considered low codon usage bias, while an ICDI value of > 0.5 is considered high codon usage bias (Freirepicos et al. 1994). Using the ICDI index, it is possible to evaluate the codon usage bias for each virus and check whether the pattern is consistent with the results of ENC analysis.

Calculation of the CAI and similarity D(A, B) analysis

The CAI represents the relative adaptiveness of the codon usage of a gene to the codon usage of highly expressed genes; it can be used to predict the level of gene expression, assess the adaptation of viral genes to a particular host, and compare codon usage among organisms (Sharp and Li 1987). CAI value ranges from 0 to 1 and is expected to have higher codon usage similarity and expression levels as the value increases (Jia et al. 2015). *Homo sapiens* was used as the reference set for HIV1, HIV2, and SIV because codon usage information was not available for some host species of SIV. To exclude the effects of the GC and amino acid compositions of the query sequence from the CAI value, 500 random sequences were generated by the Markov method for each HIV1, HIV2, and SIV group. Next, the Kolmogorov–Smirnov test was performed to evaluate whether the expected CAI (eCAI) values of the generated sequences followed a normal distribution; the critical value of the one-sided test was calculated at a significance level of 0.05. The mean and critical values of eCAI were calculated using the CAIcal web server. Next, the actual CAI values for each group were compared with 5% critical eCAI values; greater actual CAI values were presumed to indicate translation pressure. Then, using the Wilcoxon and Mann–Whitney *U* tests, values centered around the mean eCAI value for each group were compared (Puigbo et al. 2008a, b). If the query sequence was excessively large, it could not be successfully submitted to the CAIcal web server to calculate the eCAI value. Thus, for the three genes (*gag*, *pol*, and *env*) of HIV1 and the *env* gene of SIVmacaca, 500 sequences were randomly extracted and input as the query sequence. The similarity D(A, B) index proposed by Zhou et al. in 2013 quantified virus and host similarity by considering the RSCU values of 59 codons (except for Met and Trp) of the virus and the host as two space vectors in 59 dimensions, then calculating the cosine value (Zhou et al. 2013). The similarity D(A, B) index was calculated for all lentiviruses using *H. sapiens* as the host, and the RSCU values of 59 human codons were calculated from the codon usage table provided by the codon usage database. A correlation test between the similarity D(A, B) value and the CAI value was performed, and the mean similarity D(A, B) values of the HIV1 and HIV2 groups were compared. In addition, the HIV1 and HIV2 groups were compared with their origins, the SIVcpz_gor and SIVsmm groups.

Results

Evolutionary relationships among primate lentiviruses

Phylogenetic analyses of *gag*, *pol*, and *env* confirmed that HIV1 and HIV2 clustered together with the SIV species identified as their respective origins (Fig. 1). The taxonomies of SIV hosts and the abbreviations of SIV according to host species are shown in Table 2. Separate phylogenetic analyses of each gene showed that HIV1 formed the same clades with SIVcpz and SIVgor, while HIV2 formed the same clades with SIVsmm and SIVmacaca. SIVcpz has been proposed as a recombinant of SIVgsn(3') and SIVrcm(5') (Sharp et al. 2005), and the gene tree constructed in this study also showed the evolutionary distance distribution pattern supporting this assumption; SIVcpz was clustered close to SIVgsn and SIVrcm in the *env* and *pol* gene trees, respectively. Furthermore, SIVlst and SIVsun, whose hosts belong to the genus *Allochrocebus*, were closely related to SIVcolobus. Among the three phylogenetic trees, the *env* gene exhibited greater variation than the *gag* and *pol* genes. Given that the *env* gene encodes the HIV envelope glycoprotein, a known target for cytotoxic T lymphocytes and neutralizing antibodies, the distinctive level and pattern of *env* gene variations may be due to selection for antigenic diversity. The selective pressure on the *env* gene by cell-mediated immunity might result in escape mutations. Thus, the host's antibody response could continuously drive viral evolution with an initial response to the transmitted virus and with sequential responses to escape variants (Richman et al. 2003). The fact that the *env* gene can evolve to encode envelope glycoproteins for interactions with altered receptors or coreceptors, which allow the virus to enter alternative host cells, further supports the distinctive level and pattern of *env* gene variations compared to that of other genes. In most instances where the host monkey species were closely related (i.e., they belonged to the same genus), the corresponding SIV species were also closely related. Therefore, groupings for codon analysis of SIV were based on the same genus; SIV species were classified into groups after checking phylogenetic relationships considering tribe or subfamily information. In the grouping process, the specificities of SIVmnd1 and SIVsmm in the phylogenetic tree topology were considered. SIVmnd1 was closely related to SIVmandrillus (comprising SIVmnd2 and SIVdrl) only in the *env* gene tree. Although sooty mangabeys belong to the genus *Cercocebus*, SIVsmm was more closely related to SIVmacaca group than to SIVagi and SIVrcm, which infect members of the genus *Cercocebus*. Only one complete SIVmnd1 sequence (M27470) was available, and there were large differences between SIVmnd1 and SIVmandrillus in terms of *gag* and *pol* genes; thus, this complete sequence was excluded from the SIVmandrillus group. SIVsmm does not cause AIDS in sooty mangabey; therefore, it was isolated from the genus *Cercocebus* group and analyzed separately. The remaining SIVagi and SIVrcm were analyzed as SIVcerco. The *Colobus*, *Piliocolobus*, and *Procolobus* genera belong to the African group of the subfamily Colobinae and formed the same phylogenetic clade in all three genes; accordingly, they were grouped as SIVcolobus. Chimpanzees and gorillas are members of the same subfamily, Hominidae; analyses of the three genes showed that SIVcpz and SIVgor formed a clade with HIV1. Therefore, these were included in the SIVcpz_gor group. The genera *Cercopithecus* and *Miopithecus* belong to the Cercopithecini tribe and were closely located in the phylogenetic tree; they were grouped as the SIVpithecus group. The nucleotide composition and codon usage patterns were analyzed by naming the genus *Chlorocebus* as the SIVagm group and the genus *Allochrocebus* as the SIVallo group. The number of gene sequences of each virus used for the final index calculation is shown in Table 1.

Compositional properties of lentivirus structural genes

Regarding the overall nucleotide compositions of the *gag*, *pol*, and *env* genes, the A content was most abundant in all three genes, while the C content was lowest in *pol* and

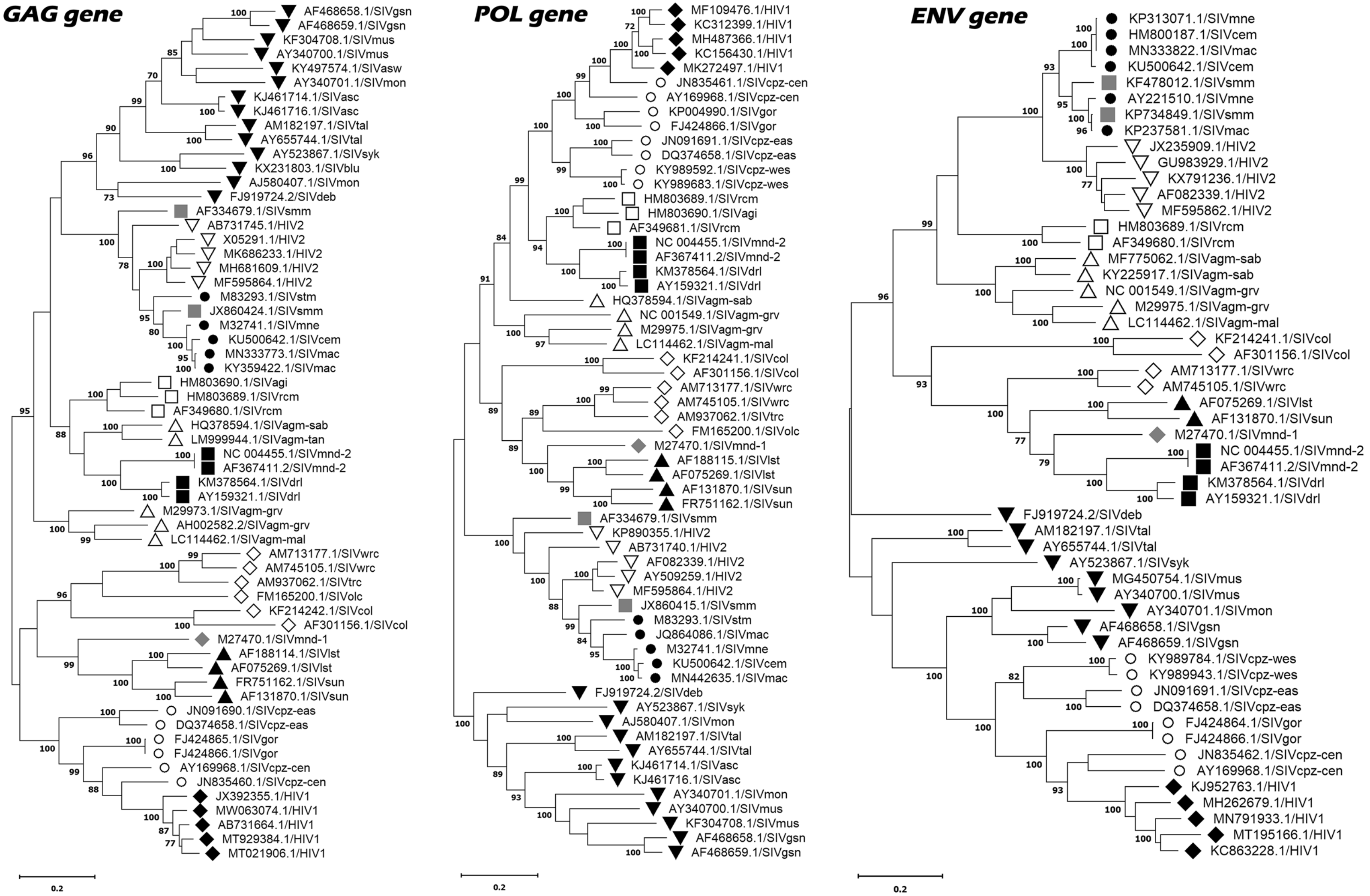

Fig. 1 Phylogenetic trees constructed with nucleotide sequences of lentivirus structural genes. Maximum likelihood trees using nucleotide sequences show evolutionary relationships of primate lentiviruses. Bootstrap values > 70% based on 1000 replications are shown at each node, and branch lengths represent evolutionary distances. [◆] HIV1; [▽] HIV2; [△] SIVagm; [▲] SIValló; [□] SIVcer; [◇] SIVcol; [○] SIVcpz_gor; [●] SIVmac; [■]SIVmnd; [▼] SIVpit; [▨] SIVsmm

env (Fig. 2). For *gag* gene, the SIVallo, SIVcerco, and SIVcolobus groups had the lowest C content, but the T content was the lowest in other 8 groups. Analysis of the nucleotide composition at the third codon position among the three genes showed that the A3 content was highest in all genes; the mean A3 contents of *pol*, *gag*, and *env* were 45.03%, 38.76%, and 33.9%, respectively. *Gag* exhibited the highest A3 content in the HIV1 group, while *pol* and *env* exhibited the highest A3 contents in the SIVcpz_gor group. In all three genes, A3 was more frequent than T3, while G3 was more frequent than C3 in all groups except the HIV2 group of the *env* gene. Figure 2 shows that, with respect to gene-specific AT/GC bias, all three genes were rich in AT. Generally, AT3 bias was more severe than overall AT bias, but the HIV2 and SIVpithecus groups of the *env* gene uniquely showed the opposite pattern. Therefore, the AT bias in these two groups mainly originated from the AT compositions of the first and second codon positions. The top 4 AT3 biases in *pol* and *env* genes were exhibited by HIV1, SIVcerco, SIVcpz_gor, and SIVcolobus. Regarding the *gag* gene, SIVallo showed the greatest AT3 bias, followed by HIV1, SIVcerco, and SIVcpz_gor. In contrast, the groups with the lowest AT3 biases were HIV2 and SIVpithecus for *pol* and *env* genes. The *gag* gene showed the lowest AT3 bias in the SIVpithecus group, followed by SIVmandrillus and HIV2. Overall, HIV1 and SIVcpz_gor showed greater AT3 biases in all three genes, compared with HIV2.

Influences of mutational pressure and natural selection on lentivirus codon usage bias

PR2 analysis showed that, for six amino acids (excluding Leu and Arg), all virus groups generally exhibited similar bias patterns (Fig. 3). Val and Gly generally use codons ending with A or G, whereas Pro, Thr, Ala, and Ser use codons with A or C in the third position. Codons ending in C in Pro, Thr, Ala, and Ser are preferred to codons ending in G; this minimizes the use of CpG dinucleotides, thereby circumventing the restrictive effect of host cell antiviral proteins and permitting viral replication. The host cell zinc finger antiviral protein binds to the high CpG dinucleotide region and inhibits viral replication (Ficarelli et al. 2020); this finding suggests that the HIV1 evasion mechanism may have played a role in the formation of this genetic pattern. Regarding the *gag* gene, the HIV1 group almost completely avoided the use of ACG codons for Thr; only the SIVsmm group was biased towards the codon end with C instead of

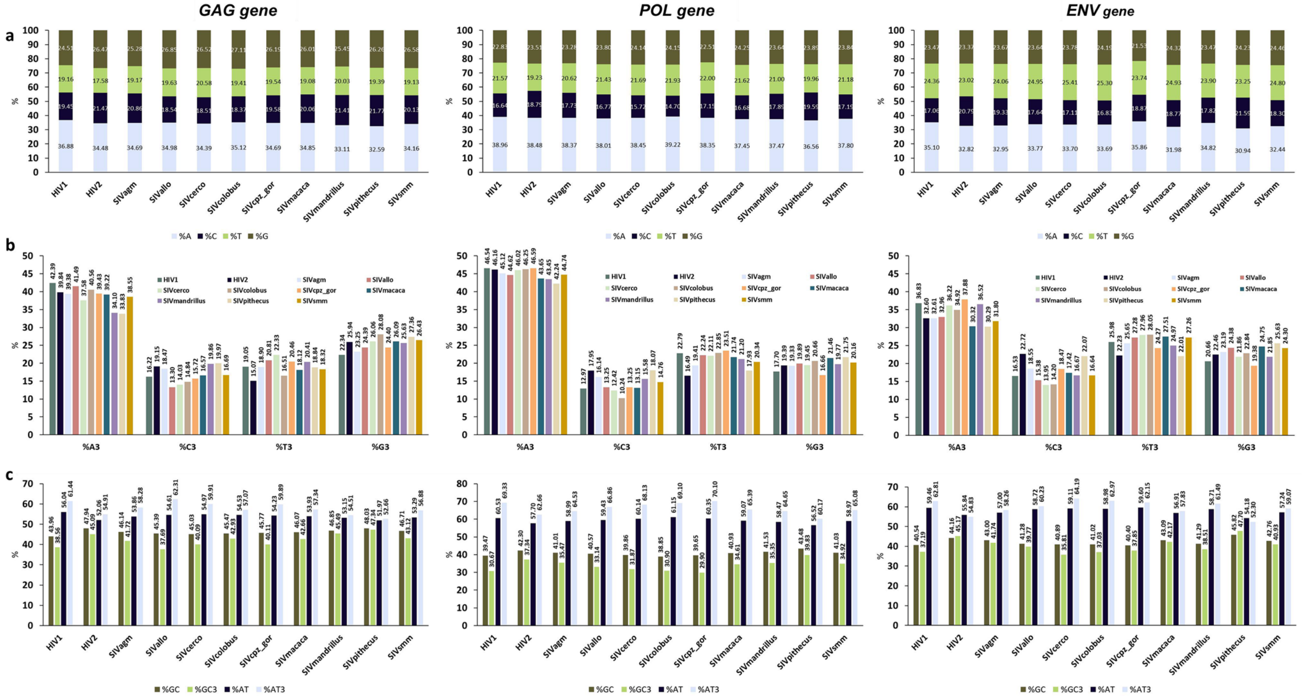

Fig. 2 Compositional features of lentiviral *gag*, *pol*, and *env* genes. **a** Distribution of A, T, C, and G in lentiviral *gag* genes. **b** Distribution of A, T, C, and G at the third codon position. **c** Total GC/AT content and GC/AT content at the third codon position. High AT contents and much higher AT3 contents were observed, indicating the predominant use of A-end codons in primate lentiviral genes

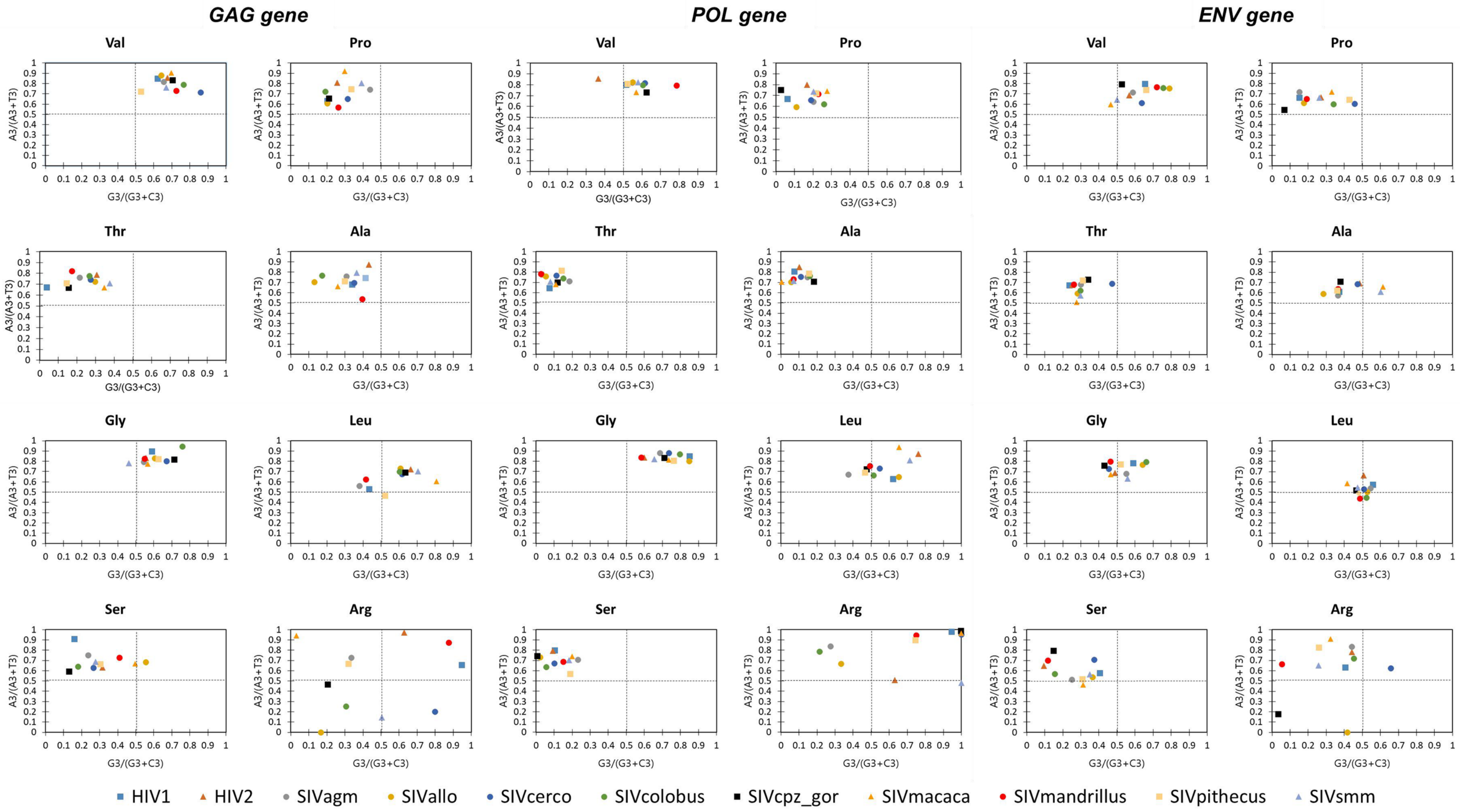

Fig. 3 Mean PR2-bias plot of *gag*, *pol*, and *env* genes of 11 lentivirus groups. AT-bias (A3/(A3 + T3)) and GC-bias (G3/(G3 + C3)) were calculated separately for each four-fold degenerate amino acid (Val, Pro, Thr, Ala, and Gly) and four-codon family of Leu, Ser, and Arg. For all three structural genes of primate lentiviruses, consistent G-end codon avoidance was observed in the amino acids Pro(CC-), Thr(AC-), Ala(GC-), and Ser(TC-), and this CpG dinucleotide avoidance pattern was particularly evident in *pol* genes

G for Gly. SIVpithecus and SIVmacaca favored a codon end with G (rather than C) for Ser, unlike the other groups. In the *pol* gene, the G3/(G3 + C3) values for the nine groups ranged from 0.52 to 0.62, whereas those values for the HIV2 and SIVmandrillus groups were 0.36 and 0.79, respectively. In contrast to the other groups, HIV2 was found to prefer the GTC codon (rather than GTG) for Val in the *pol* gene. For Ala in the *env* gene, the SIVsmm and SIVmacaca groups had a G3/(G3 + C3) value of 0.6 or higher, indicating that a codon ending in G was used more frequently than a codon ending in C. Furthermore, these two groups did not avoid the use of CpG dinucleotides; the HIV2 and SIVcerco groups, which formed the same clade with SIVsmm and SIVmacaca in phylogenetic analysis of *env*, also used the codons GCC and GCG evenly. The GC biases of Leu amino acids were in the ranges 0.38–0.80 and 0.37–0.76 in *gag* and *pol*, respectively; in *env*, the A3/(A3 + C3) and G3/(G3 + C3) values did not deviate significantly from point (0.5, 0.5).

Neutral evolution analysis (Fig. 3) enables evaluation of the relationship between the content of GC3s and the content of GC12s, which can be used to quantify the relative magnitudes of selection and mutation pressures (Fig. 4). Theoretically, the absolute value of the slope of the regression line is within the range 0 to 1; values closer to 0 indicate greater selection pressure. Thus, if the GC12s value changes in a manner similar to the GC3s value, the virus is under weak selection pressure; if changes in the GC12s value are limited to a small range, the virus is under high selection pressure. A correlation analysis of each virus group was performed for each of the three genes. Regarding the *gag* gene, HIV1 ($r = -0.053$, $p < 0.0001$), HIV2 ($r = 0.37$, $p = 0.001$), SIVcerco ($r = -0.9746$, $p = 0.0048$), SIVmacaca ($r = -0.2895$, $p < 0.0001$), SIVmandrillus ($r = -0.9302$, $p = 0.0008$), and SIVpithecus ($r = 0.5435$, $p = 0.0363$) showed significant correlations between GC12s and GC3s values. Regarding the *pol* gene, HIV1 ($r = 0.2014$, $p < 0.0001$), HIV2 ($r = 0.4803$, $p = 0.0008$), SIVagm ($r = 0.3819$, $p = 0.015$), SIVcolobus ($r = 0.7664$, $p = 0.0445$), SIVcpz_gor ($r = -0.6292$, $p < 0.0001$), and SIVpithecus ($r = 0.589$, $p = 0.034$) showed significant correlations between GC12s and GC3s values. Regarding the *env* gene, HIV1 ($r = 0.0653$, $p < 0.0001$), SIVagm ($r = 0.2926$, $p = 0.0038$), SIVmacaca ($r = -0.6714$, $p < 0.0001$), SIVpithecus ($r = 0.6066$, $p = 0.0076$), and SIVsmm ($r = -0.6667$, $p < 0.0001$) showed significant correlations between GC12s and GC3s values. All three genes showed a distribution generally parallel to the x-axis, suggesting that selection pressure plays a major role in the formation of codon usage in primate lentiviruses. Regarding the *gag* gene, all groups except HIV2 had a negative slope. The slope of the regression line for the virus group indicates the relative magnitude of the directional mutation pressure experienced by the virus. Regarding the *gag* gene, the relative contributions of directional mutation pressure in codon usage pattern formation in HIV1, HIV2, SIVcerco, SIVcolobus, SIVmacaca, SIVmandrillus, and SIVpithecus groups were 2.55%, 17.25%, 41.07%, 4.68%, 33.28%, and 13.94%, respectively. Thus, in the *gag* gene, the HIV1 group showed the highest selection pressure (97.45%), followed by the SIVmacaca group (95.32%), the SIVpithecus group (86.06%), the HIV2 group (82.75%), the SIVmandrillus group (66.72%), and the SIVcerco group (58.93%). Regarding the *pol* gene, the slopes of the HIV1, HIV2, SIVagm, SIVcolobus, and SIVpithecus groups, but not the SIVcpz_gor group, were > 0, indicating positive correlations. The directional mutation pressures of the HIV1, HIV2, SIVagm, SIVcolobus, SIVcpz_gor, and SIVpithecus groups in the *pol* gene were 7.26%, 41.36%, 0.01%, 17.51%, 19.38%, and 7.83%, respectively. Among them, SIVagm (99.99%) and HIV2 (58.64%) were under the maximum and minimum selection pressures in *pol*; the HIV1, SIVpithecus, SIVcolobus, and SIVcpz_gor groups were affected by 92.74%, 92.17%, 82.49%, and 80.62% selection pressures, respectively. Regarding the *env* gene, the GC12s and GC3s values showed positive correlations in the HIV1, SIVagm, and SIVpithecus groups, while they showed negative correlations in the SIVmacaca and SIVsmm groups. Regarding the *env* gene, the magnitudes of directional mutation pressure in the HIV1, SIVagm, SIVmacaca, SIVpithecus, and SIVsmm groups were 3.06%, 9.18%, 15.24%, 7.97%, and 40.67%, respectively; the groups receiving the most and least selection pressure were HIV1 (96.94%) and SIVsmm (59.33%), respectively. In addition, selection pressures in the SIVpithecus, SIVagm, and SIVmacaca groups were 92.03%, 90.82%, 84.76%, and 59.33%, respectively.

Variation in RSCU value and codon usage preference

Through analysis of the mean RSCU values for all virus groups using the *gag*, *pol*, and *env* genes, overrepresented codons (RSCU > 1.6) were identified for each group (Fig. 5). Among all codons of the three genes, the CTG (Leu) codon in the *gag* gene region had the largest RSCU difference between virus groups, notably overrepresented (SIVmacaca: 1.68) and underrepresented (HIV1: 0.51) virus groups coexisted. The AGA codon of the *gag* gene was overexpressed most frequently, and the AGC codon showed the largest variation. Regarding the mean RSCU value of AGC, HIV1 and HIV2 had the highest mean value (2.2), followed by SIVmacaca and SIVsmm (1.7), and then SIVallo and SIVpithecus (0.7). Figure 5 shows that the RSCU values of the TTA, ATA, and GTA codons of *gag* were significantly higher in the HIV1 group than in the other groups. Concerning the 2-codon amino acids Tyrosine (Tyr) and Cysteine (Cys), the HIV1 group was biased toward the use of the TAT and TGT codons, in contrast to the average use of two synonymous

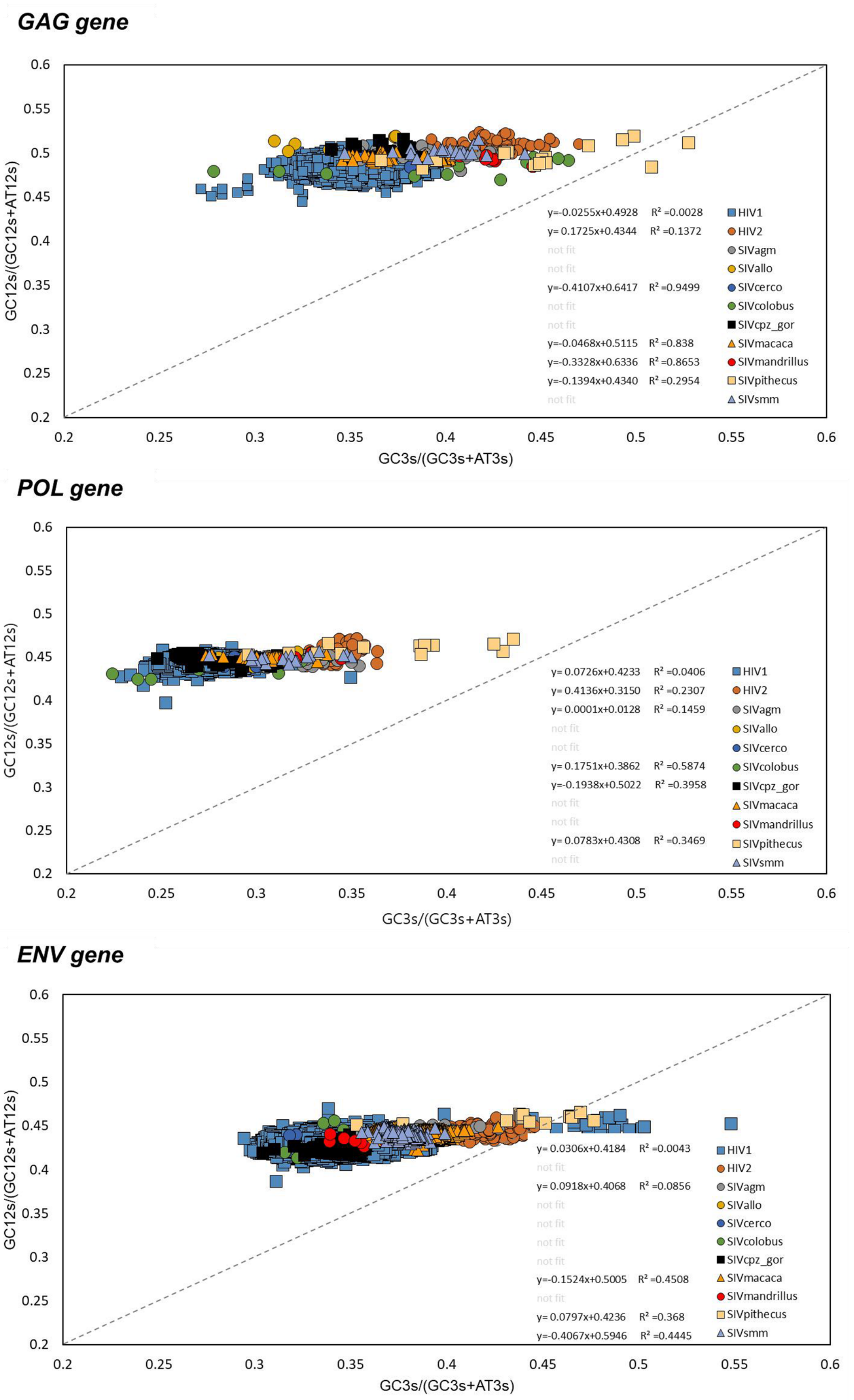

Fig. 4 Neutrality plot for lentiviral *gag*, *pol*, and *env* genes. Different virus groups are represented by different icons, and each icon represents a sequence. If a correlation exists, the linear regression equation is presented; "not fit" indicates that no correlation was present between GC12s and GC3s values. The extent of change in GC12s was very limited compared to that in GC3s, implying that selection pressure plays a major role in shaping codon usage patterns in primate lentiviruses compared to mutational pressure

codons in the HIV2 group. In terms of Leu codon usage in *pol*, the HIV2 group was considerably different. Among the 10 groups, with the exception of HIV2, the mean RSCU value of the TTA codon was high (2.18); in HIV2, it was 1.52. Concerning the CTA codon, the mean RSCU value of the other groups was 1.27, whereas HIV2 showed a value of 2.27; HIV2 thus exhibited a greater preference for the CTA codon than for the TTA codon. In terms of Ser codon usage in *pol*, the TCA and AGT codons were preferred; the HIV2, SIVallo, SIVmacaca, and SIVsmm groups had a greater preference for the TCA codon, while the remaining groups had a greater preference for the AGT codon. In terms of Cys

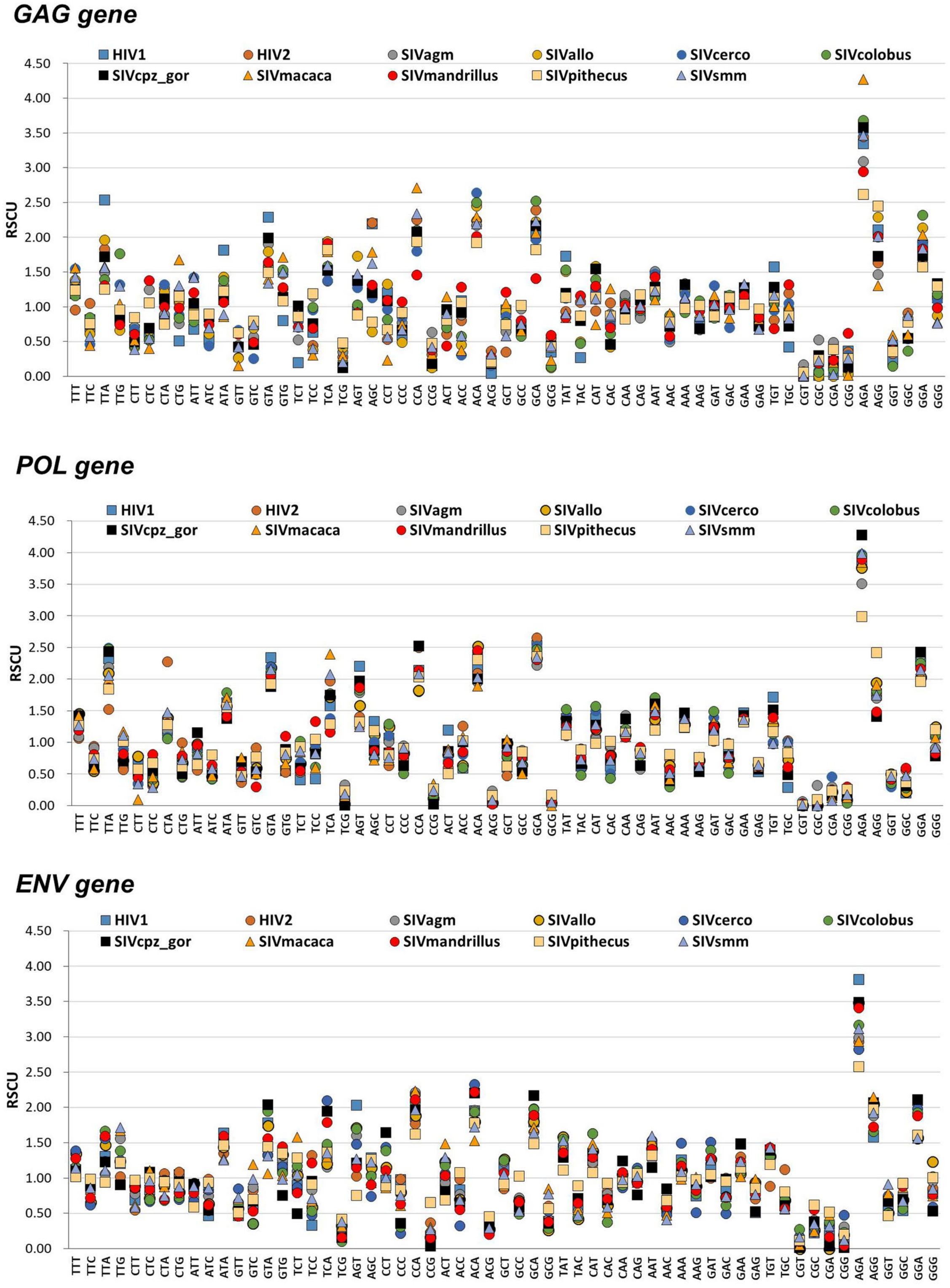

Fig. 5 RSCU plot of lentiviral *gag*, *pol*, and *env* genes. The degree of variation among the 11 groups was the smallest in the *pol* gene. Generally, for RSCU values > 1.6, codons were considered over-expressed; for RSCU values < 0.6, codons were considered under-expressed. CCA, ACA, GCA, AGA, AGG, and GGA codons were over-expressed (RSCU > 1.6) in all structural genes of lentiviruses and, based on RSCU values, there was a dominant preference for the A-end codon

codon usage in *pol*, synonymous codons were used evenly in most groups; a pattern of TGT codon overexpression was observed in the HIV1 and SIVcpz_gor groups. Regarding the *pol* gene, a codon ending in A or T was preferred in all two-codon amino acids except for Cys; Ser codon usage in *env* differed significantly among virus groups. All groups except SIVmacaca (most frequently used TCT) preferred TCA or AGT; the HIV1, SIVagm, SIVallo, and SIVcolobus groups preferred AGT. In terms of Cys codon usage in *env*, all groups preferred the TGT codon except for HIV2 (most frequently used TGC). Concerning Ile codon usage, *pol* and *env* were biased toward the ATA codon; in *gag*, the SIVcerco, SIVmacaca, SIVmandrillus, and SIVsmm groups preferred ATT. Thr amino acids showed a pattern of overexpression and underexpression of the ACA and ACG codons, respectively, in all three genes. The CCA, ACA, GCA, AGA, AGG, and GGA codons were overexpressed; the corresponding amino acids Pro, Thr, Ala, Arg, and Gly showed a similar pattern in all three structural genes.

Codon usage patterns in *gag*, *pol* and *env*

The mean ENC values of *gag*, *pol*, and *env* were 47.80, 43.47, and 49.18, and the ranges were 44.27–52.52, 40.62–46.62, and 46.53–53.42, respectively. The mean ENC value of *pol* was lowest, and it showed the least variance among virus groups (Fig. 6). Therefore, the overall codon usage bias of *pol* was highest among the structural genes, while the codon usage bias of *env* was lowest. The virus groups showing high codon bias in all three genes were HIV1 and SIVcolobus; the SIVpithecus group consistently showed low codon bias. The HIV2 and SIVsmm groups showed low codon usage bias only in *env*. As shown in Fig. 7, the SIVagm, SIVcpz_gor, and SIVmandrillus groups demonstrated a vertical distribution with small GC3s % variation and large ENC value variation in the *gag* gene (Fig. 7). Therefore, smaller ENC value is indicative of greater selection pressure. Regarding the *gag* gene, the SIVpithecus group had a distinct distribution of sequences on the ENC-GC3s plot. The SIVasc group was under greater selective pressure,

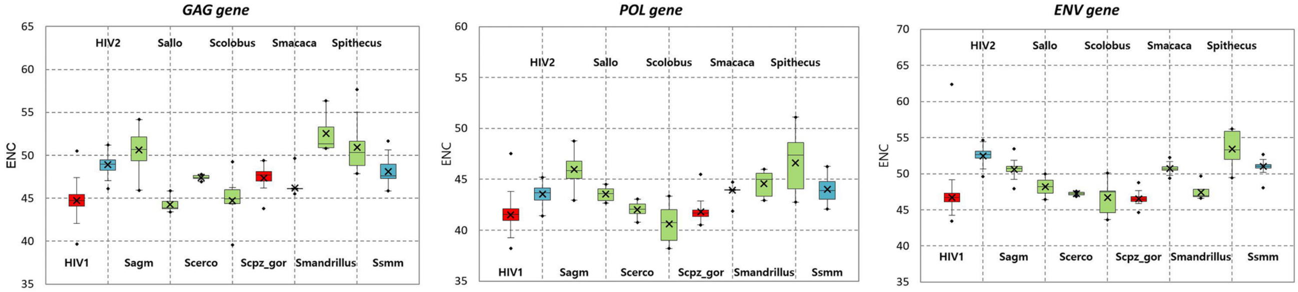

Fig. 6 ENC box plot of lentiviral *gag*, *pol*, and *env* genes. Cross: mean ENC value; dots: minimum and maximum ENC values. The HIV1 and SIVcpz_gor groups are highlighted by red boxes and the HIV2 and SIVsmm groups are highlighted by blue boxes. The mean ENC values for *gag*, *pol*, and *env* were 47.8, 43.47, and 49.18, respectively, indicating that the codon diversity was greater in *gag* and *env* genes than in the *pol* gene. HIV1 showed a strong preference for using synonymous codons in all three genes

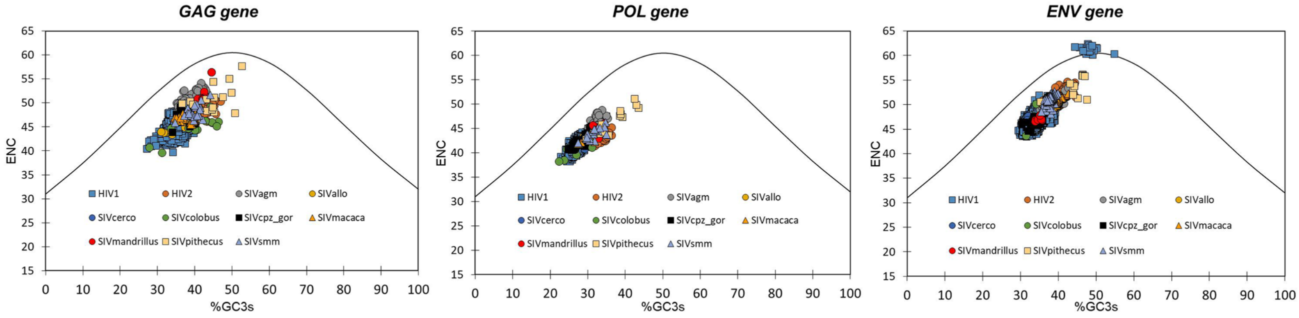

Fig. 7 ENC-GC3s plot of lentiviral *gag*, *pol*, and *env* genes. The smooth line is the expected value of ENC calculated based on GC3 content. In the *gag* gene, sequences within the same group are subject to greater differences in selection pressure than in the *pol* and *env* genes. Sequences within the same group of the *gag* gene show a greater difference in selection pressure than those of the *pol* and *env* genes. A small portion of HIV1 sequences located on the expected curve within the *env* ENC-GC3s plot are clones of highly neurovirulent HIV-1 isolates

compared with the SIVdeb group. Regarding the *pol* gene, the degrees of selection pressure were similar among virus groups. Contrary to the overall pattern, the SIVagm group was under less selection pressure; the smaller ENC value in the SIVagm group indicated greater selection pressure. Regarding the *pol* gene, the HIV2 group was under slightly greater selection pressure, compared with the HIV1 group; SIVlhoest in the SIVallo group was less affected by selection pressure than SIVsun. Regarding the *env* gene, the SIVsmm and SIVcolobus groups showed little variation in GC3s % and a large difference in ENC values; therefore, the smaller ENC values in the two groups were indicative of greater selection pressure. Regarding the *env* gene, SIVcol was less affected by selection pressure than SIVwrc. Regarding the *pol* gene, SIVlhoest was under greater selection pressure, compared with SIVsun, in the *env* gene region. Regarding the *env* gene, the SIVmacaca group showed a distribution parallel to the abscissa with a large GC3s % variation and a small ENC value variation, suggesting that the larger GC3s % value was indicative of greater selection pressure. Although the sequences in the SIVpithecus group of *env* were located throughout the ENC-GC3s plot, SIVtal was under greater selection pressure, compared with SIV that infects other species of the genus *Cercopithecus*. Consistent with the ENC results, the ICDI values of all three genes were < 0.3, indicating low codon usage bias (Table 3). In terms of ICDI results, the groups with comparatively high codon usage bias in the three genes were HIV1 and SIVcolobus. ENC and ICDI index analyses showed that HIV1 had a greater codon usage bias than HIV2, consistent with a prior report (Vidyavijayan et al. 2017). Based on Pearson correlation analysis of ENC, ICDI, CAI, and similarity D(A, B) findings for each gene, pairwise negative correlations were observed between ENC and ICDI, and between CAI and similarity D(A, B). The ICDI and CAI values also showed negative correlations (Table 4). Therefore, greater overall codon usage bias for a particular gene was associated with lower similarity to the *H. sapiens* codon usage pattern.

Differences in viral adaptation to codon usage in humans

The CAI value of *gag* ranged from 0.715 to 0.746 among the 11 groups; it was 0.746 ± 0.005 in the SIVmacaca group, indicating the greatest adaptability to the reference set, *H. sapiens* (Table 5). The CAI values of *pol* and *env* were highest in the SIVmandrillus group (0.708 ± 0.004) and HIV2 group (0.730 ± 0.005), respectively; they were lowest in the HIV1 group (0.691 ± 0.004) and SIVcpz_gor group (0.700 ± 0.006), respectively. Among the three genes, *pol* and *gag* showed the smallest and largest differences, respectively, in CAI values among virus groups. Because the CAI value is affected by the nucleotide and amino acid composition of the sequence, the eCAI value at a significance level of 0.05 was calculated using the method established by Puigbò et al. Five hundred sequences were generated by the Markov method to reflect the nucleotide and amino acid compositions of the three genes. The Kolmogorov–Smirnov normality test showed that the 500 nucleotide sequences generated in all groups exhibited a normal distribution (< 0.061). Next, the actual CAI value for each group and the eCAI value at the 5% significance level were compared. There were no significant differences between the CAI and mean eCAI values in the 11 groups in terms of *pol* and *env* genes. Therefore, CAI values of *pol* and *env* genes can be fully achieved due to the GC content of the sequence itself, suggesting that they experienced less translation pressure. Regarding the *gag* gene, the SIVallo, SIVcolobus, SIVcpz_gor, SIVmacaca, and SIVsmm groups were adapted to the codon usage pattern of *H. sapiens*; the other groups did not exhibit a significant

Table 3 ICDI values of lentiviral *gag*, *pol*, and *env* genes

Group	*gag*		*pol*		*env*	
	N	ICDI	*N*	ICDI	*N*	ICDI
HIV1	22,699	0.169 ± 0.022	11,021	0.199 ± 0.019	70,658	0.130 ± 0.015
HIV2	76	0.097 ± 0.013	45	0.134 ± 0.013	170	0.065 ± 0.012
SIVagm	43	0.113 ± 0.023	40	0.130 ± 0.016	96	0.088 ± 0.012
SIValo	6	0.164 ± 0.012	6	0.153 ± 0.016	2	0.114 ± 0.034
SIVcer	5	0.132 ± 0.017	5	0.187 ± 0.017	2	0.153 ± 0.012
SIVcol	12	0.161 ± 0.053	7	0.223 ± 0.051	5	0.148 ± 0.018
SIVcpz_gor	16	0.135 ± 0.019	172	0.189 ± 0.015	192	0.117 ± 0.007
SIVmac	624	0.117 ± 0.002	341	0.149 ± 0.005	14,611	0.101 ± 0.008
SIVmnd	8	0.081 ± 0.013	6	0.140 ± 0.016	8	0.119 ± 0.010
SIVpit	15	0.091 ± 0.018	13	0.116 ± 0.034	18	0.061 ± 0.016
SIVsmm	21	0.114 ± 0.013	19	0.145 ± 0.017	482	0.105 ± 0.011

ICDI values are means ± standard deviations

Table 4 Correlations of ENC, ICDI, CAI, and D(A, B)

gag gene				
Pearson's correlation (N = 23525, α = 0.05, two-tailed)	ENC	ICDI	CAI	D(A, B)
ENC	1	− 0.783*	0.129*	− 0.794*
ICDI		1	− 0.289*	0.593*
CAI			1	− 0.509*
D(A, B)				1
pol gene				
Pearsons correlation (*N* = 11675, α = 0.05, two-tailed)	ENC	ICDI	CAI	D(A, B)
ENC	1	− 0.896*	0.453*	− 0.893*
ICDI		1	− 0.453*	0.700*
CAI			1	− 0.581*
D(A, B)				1
env gene				
Pearson's correlation (N = 86244, α = 0.05, two-tailed)	ENC	ICDI	CAI	D(A, B)
ENC	1	− 0.826*	0.594*	− 0.928*
ICDI		1	− 0.559*	0.762*
CAI			1	− 0.714*
D(A, B)				1

*Significant at the 0.05 level (two-tailed)

Table 5 CAI and 5% critical eCAI values of lentiviral genes

Group	*gag*		*pol*		*env*	
	CAI	eCAI	CAI	eCAI	CAI	eCAI
HIV1	0.722 ± 0.006	0.733	0.691 ± 0.004	0.708	0.711 ± 0.007	0.716
HIV2	0.735 ± 0.009	0.738	0.694 ± 0.005	0.719	0.730 ± 0.005	0.739
SIVagm	0.715 ± 0.009	0.734	0.695 ± 0.007	0.717	0.728 ± 0.009	0.733
SIValo	**0.727 ± 0.013**	0.720	0.695 ± 0.009	0.707	0.726 ± 0.006	0.734
SIVcer	0.725 ± 0.008	0.732	0.701 ± 0.007	0.713	0.716 ± 0.004	0.717
SIVcol	**0.735 ± 0.016**	0.731	0.698 ± 0.013	0.713	0.719 ± 0.004	0.723
SIVcpz_gor	**0.726 ± 0.009**	0.724	0.691 ± 0.005	0.699	0.700 ± 0.006	0.725
SIVmac	**0.746 ± 0.005**	0.730	0.703 ± 0.002	0.717	0.724 ± 0.006	0.734
SIVmnd	0.729 ± 0.008	0.738	0.708 ± 0.004	0.717	0.730 ± 0.009	0.734
SIVpit	0.737 ± 0.017	0.747	0.703 ± 0.01	0.720	0.720 ± 0.009	0.738
SIVsmm	**0.738 ± 0.008**	0.737	0.702 ± 0.005	0.717	0.724 ± 0.011	0.730

CAI value are means ± standard deviations; the reference set was Homo sapiens

eCAI is represented as 5% critical value

Bold type indicates a CAI value significantly greater than the mean eCAI value

difference from the calculated eCAI values. Therefore, HIV1 and HIV2 have high codon usage similarity with *H. sapiens* in all genes, but this similarity was not mainly derived from translation pressure. Regarding the *gag* gene, the CAI value was centered on the mean value of the eCAI for the SIVallo, SIVcolobus, SIVcpz_gor, SIVmacaca, and SIVsmm groups. At a significance level of 0.05, the Mann–Whitney U test indicated that the SIVmacaca group had significantly greater adaptability to *H. sapiens* than the other four groups, while the SIVsmm group had greater adaptability to *H. sapiens* than the SIVcpz_gor group.

Quantification of codon usage similarity between virus and host showed that the similarity D(A, B) values of HIV1 and HIV2 in *gag* were 0.117 ± 0.006 and 0.086 ± 0.008, respectively (Table 6). The similarity D(A, B) values of the SIVcpz_gor and the SIVsmm groups (the presumed origins of HIV1 and HIV2) were 0.095 ± 0.008 and 0.087 ± 0.007, respectively. The Mann–Whitney *U* test showed that the similarity D(A, B) value of HIV1 was significantly higher than the corresponding value of HIV2 at a significance level of 0.05; the value of HIV1 was significantly higher than the value of SIVcpz_gor. However, the similarity D(A, B) values did not significantly differ between the HIV2 and SIVsmm groups. The similarity D(A, B) values of HIV1, HIV2, SIVcpz_gor, and SIVsmm to *H. sapiens* in *pol* were 0.132 ± 0.005, 0.121 ± 0.006, 0.132 ± 0.006, and 0.119 ± 0.008, respectively. The Mann–Whitney *U* test showed results identical to the findings in *gag* analysis, using a significance level of 0.05. The similarity D(A, B) values of HIV1, HIV2, SIVcpz_gor, and SIVsmm to *H. sapiens* in *env* were 0.096 ± 0.005, 0.063 ± 0.005, 0.099 ± 0.004, and 0.077 ± 0.005, respectively. Using the Mann–Whitney *U* test for assessment of *env*, the similarity D(A, B) value of the HIV1 group was significantly higher than the corresponding value of the HIV2 group; however, the similarity D(A, B) values of HIV1 and HIV2 were significantly lower than the corresponding values of SIVcpz_gor and SIVsmm, respectively. In general, in terms of codon usage, lentiviruses showed high similarity to *H. sapiens*; this similarity was highest in *env* and lowest in *pol*. In all three structural genes, HIV2 exhibited comparatively greater similarity to the human codon usage pattern than HIV1. In the case of *gag* gene, HIV1 showed low similarity with human codon usage compared to SIVcpz_gor. In the *env* gene, both HIV1 and HIV2 showed a codon usage pattern more similar to that of humans than SIVcpz_gor and SIVsmm, the origin viruses.

Discussion

Because of codon degeneracy, 18 amino acids (excluding Met and Trp) are encoded by two or more codons. Synonymous codons encoding the same amino acid are not used randomly, and varying degrees of codon usage bias have been confirmed in almost all species. The use of different synonymous codons directly affects translation, including its initiation (Kudla et al. 2009; Goodman et al. 2013), efficiency, and accuracy (Drummond and Wilke 2008), as well as RNA structure and folding (Shabalina et al. 2006), thus affecting gene function. Optimizing the codon usage of the HIV1 *gag* gene increases the expression level of the virus (Deml et al. 2001; Gao et al. 2003; Smith et al. 2004), confirming that improvements in mRNA stability and nuclear escape are critical factors, rather than increased translational efficiency (Kofman et al. 2003). In addition, HIV1 replication and *env* expression were aborted by substitution of the synonymous codon (AGG → CGU) encoding Arg in the gp41 gene region, due to disruption of the secondary structure of intronic splicing silencer RNA (Jordan-Paiz et al. 2020). Analysis of the synonymous codon usage pattern enables prediction of the evolutionary direction of the viral genome, and an adequate understanding of viral codon usage facilitates codon editing. These aspects allow identification of gene function and elucidation of novel antiviral mechanisms in the innate immune system, as well as unknown areas of the viral life cycle.

In the codon usage analysis of primate lentiviruses in this study, a strong A bias was confirmed in all three structural genes, consistent with previous findings regarding HIV (van Hemert and Berkhout 1995; Pandit and Sinha 2011). Unlike the nucleotide composition that minimizes C nucleotide in *pol* and *env*, the use of T and C nucleotides is diminished in *gag*; we found that the use of T

Table 6 Similarity D(A, B) values of lentiviral *gag*, *pol*, and *env* genes

Group	*gag*		*pol*		*env*	
	N	D(A,B)	*N*	D(A,B)	*N*	D(A,B)
HIV1	22,699	0.117 ± 0.006	11,021	0.132 ± 0.005	70,658	0.096 ± 0.005
HIV2	76	0.086 ± 0.008	45	0.121 ± 0.006	170	0.063 ± 0.005
SIVagm	43	0.085 ± 0.010	40	0.110 ± 0.009	96	0.073 ± 0.006
SIValo	6	0.116 ± 0.016	6	0.120 ± 0.005	2	0.087 ± 0.011
SIVcer	5	0.096 ± 0.004	5	0.129 ± 0.005	2	0.099 ± 0.006
SIVcol	12	0.108 ± 0.018	7	0.139 ± 0.016	5	0.101 ± 0.012
SIVcpz_gor	16	0.095 ± 0.008	172	0.132 ± 0.006	192	0.099 ± 0.004
SIVmac	624	0.092 ± 0.002	341	0.120 ± 0.002	14,611	0.078 ± 0.003
SIVmnd	8	0.068 ± 0.007	6	0.109 ± 0.006	8	0.089 ± 0.005
SIVpit	15	0.074 ± 0.019	13	0.100 ± 0.020	18	0.058 ± 0.011
SIVsmm	21	0.087 ± 0.007	19	0.119 ± 0.008	482	0.077 ± 0.005

Similarity D(A, B) values are means ± standard deviations; the reference set was *Homo sapiens* nucleotides was lowest in all primate lentivirus groups except SIVallo, SIVcerco, and SIVcolobus. Furthermore, in *env*, only the HIV2 and SIVpithecus groups had a lower AT3 bias, compared with the overall AT bias; the AT bias in these two groups was mainly formed by the AT composition of the first and second codon sites. The codon usage pattern of Arg shows extreme A bias in the HIV1 genome and lacks CpG dinucleotides; thus, the AGA codon is used most frequently and all CGNs are underexpressed (van Hemert and Berkhout 1995). In this study, we found that the synonymous codon usage pattern of the Arg amino acid was very conservative in all primate lentivirus groups. Moreover, the use of synonymous codons in the two-codon amino acid Cys of HIV1 was heavily biased toward the TGT codon, unlike the typical use of two synonymous codons in humans (van Hemert and Berkhout 1995). An RSCU analysis of HIV2 and SIVs showed that only HIV1 in *gag* and HIV1 and SIVcpz_gor in *pol* exhibit these codon usage biases. In the *gag* and *pol* genes of the other primate lentiviruses, the TGT and TGC codons are used evenly, similar to the findings in humans. In a PR2 analysis, an A > T bias was observed in all four codon amino acids of *gag* and *pol*, and a G > C bias was observed in Val and Gly. However, a C > G bias was observed in Pro, Thr, and Ala; this is presumed to prevent the use of CpG dinucleotides at Pro, Thr, and Ala. Furthermore, for Val in *pol*, only the HIV2 group used GTC more frequently than GTG. Regarding the *env* gene, the SIVsmm and SIVmacaca groups used GCG more frequently than GCC, unlike the strong GCC codon usage bias in *gag* and *pol*; the HIV2 and SIVcerco groups used the two codons in a similar manner. Phylogenetic analysis of *env* indicated that these four groups (SIVsmm, SIVmacaca, HIV2, and SIVcerco) showed the closest evolutionary relationship, consistent with the synonymous codon usage pattern of Ala. Therefore, editing the corresponding codon would confirm a biologically important function. Notably, the ENC-GC3s plot and neutral evolution analysis showed that all primate lentiviruses were more affected by selection pressure than by mutation caused by the GC composition of the gene, consistent with prior reports regarding HIV1. The CAI and similarity D(A, B) values indicated that although there was a high degree of similarity to human codon usage in all three structural genes of HIV, this similarity was not caused by translation pressure. In addition, compared with HIV1, the codon usage of HIV2 is more similar to the human codon usage, but the overall codon usage bias is lower. Finally, the origin viruses of HIV (SIVcpz_gor and SIVsmm) exhibit greater similarity to human codon usage in the *gag* gene, confirming their robust adaptability to human codon usage. Therefore, HIV1 and HIV2 may have evolved to avoid human codon use by selection pressure in the *gag* gene after interspecies transmission from SIV hosts to humans.

Conclusion

We confirmed the overall codon usage patterns of primate lentiviruses, then explored the evolutionary and genetic characteristics of HIV1, HIV2, and SIV. Because the grouping of sequence data for codon usage pattern analysis is based on the phylogenetic topology of *gag*, *pol*, and *env*, as well as the classification systems of SIV hosts, differences in patterns within groups and differences due to HIV lineages or subtypes and geographic distribution were not considered. Information such as codon deoptimization, dinucleotide usage, and codon pair usage can be applied to multiple RNA viral genomes to generate novel attenuated vaccines. By overcoming safety and stability issues, information from codon usage analysis will be useful for attenuated HIV1 vaccine development. A recoded HIV1 variant can be used as a vaccine vector or in immunotherapy to induce specific innate immune responses. Further research regarding HIV1 dinucleotide usage and codon pair usage will facilitate new approaches to the treatment of AIDS.

Acknowledgements Portions of this work were presented and published in thesis form in fulfillment of the requirements for the M.Ph. for Student XM from Seoul National University.

Author contributions MC and XM analyzed and interpreted the results, MC and XM wrote the manuscript, HS designed and supervised the study. All authors read and approved the final manuscript.

Funding This work was supported by grants from the National Research Foundation (NRF) of Korea funded by the Ministry of Education (MOE, No. 2020R1F1A1076161). This research was also funded by the MOE under the BK21-Four project supervised by the NRF.

Data availability All data and materials described in the manuscript are available.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Ahn I, Son HS (2006) Epidemiological comparisons of codon usage patterns among HIV-1 isolates from Asia, Europe, Africa and the Americas. Exp Mol Med 38(6):643–651. https://doi.org/10.1038/emm.2006.76

Apetrei C, Kaur A, Lerche NW, Metzger M, Pandrea I, Hardcastle J, Falkenstein S, Bohm R, Koehler J, Traina-Dorge V, Williams T, Staprans S, Plauche G, Veazey RS, McClure H, Lackner AA, Gormus B, Robertson DL, Marx PA (2005) Molecular epidemiology of simian immunodeficiency virus SIVsm in US primate centers unravels the origin of SIVmac and SIVstm. J Virol 79(14):8991–9005. https://doi.org/10.1128/Jvi.79.14.8991-9005.2005

Deml L, Bojak A, Steck S, Graf M, Wild J, Schirmbeck R, Wolf H, Wagner R (2001) Multiple effects of codon usage optimization on expression and immunogenicity of DNA candidate vaccines encoding the human immunodeficiency virus type 1 Gag protein. J Virol 75(22):10991–11001. https://doi.org/10.1128/JVI.75.22.10991-11001.2001

Desimmie BA, Delviks-Frankenberrry KA, Burdick RC, Qi DF, Izumi T, Pathak VK (2014) Multiple APOBEC3 restriction factors for HIV-1 and one Vif to rule them all. J Mol Biol 426(6):1220–1245. https://doi.org/10.1016/j.jmb.2013.10.033

Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134(2):341–352. https://doi.org/10.1016/j.cell.2008.05.042

Ficarelli M, Antzin-Anduetza I, Hugh-White R, Firth AE, Sertkaya H, Wilson H, Neil SJD, Schulz R, Swanson CM (2020) CpG dinucleotides inhibit HIV-1 replication through zinc finger antiviral protein (ZAP)-Dependent and -independent mechanisms. J Virol 94(6):e01337-e1419. https://doi.org/10.1128/JVI.01337-19

Foley BT (2000) An overview of the molecular phylogeny of lentiviruses. HIV sequence compendium. pp 35–43

Freirepicos MA, Gonzalezsiso MI, Rodriguezbelmonte E, Rodrigueztorres AM, Ramil E, Cerdan ME (1994) Codon usage in Kluyveromyces-Lactis and in yeast cytochrome C-Encoding genes. Gene 139(1):43–49. https://doi.org/10.1016/0378-1119(94)90521-5

Gao F, Bailes E, Robertson DL, Chen YL, Rodenburg CM, Michael SF, Cummins LB, Arthur LO, Peeters M, Shaw GM, Sharp PM, Hahn BH (1999) Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature 397(6718):436–441. https://doi.org/10.1038/17130

Gao F, Li Y, Decker JM, Peyerl FW, Bibollet-Ruche F, Rodenburg CM, Chen Y, Shaw DR, Allen S, Musonda R, Shaw GM, Zajac AJ, Letvin N, Hahn BH (2003) Codon usage optimization of HIV type 1 subtype C gag, pol, env, and nef genes: in vitro expression and immune responses in DNA-vaccinated mice. AIDS Res Hum Retroviruses 19(9):817–823. https://doi.org/10.1089/088922203769232610

Goodman DB, Church GM, Kosuri S (2013) Causes and effects of N-terminal codon bias in bacterial genes. Science 342(6157):475–479. https://doi.org/10.1126/science.1241934

Hemelaar J (2012) The origin and diversity of the HIV-1 pandemic. Trends Mol Med 18(3):182–192. https://doi.org/10.1016/j.molmed.2011.12.001

Jia X, Liu SY, Zheng H, Li B, Qi Q, Wei L, Zhao T, He J, Sun J (2015) Non-uniqueness of factors constraint on the codon usage in *Bombyx mori*. BMC Genomics 16(1):1–12. https://doi.org/10.1186/s12864-015-1596-z

Jiang Y, Deng F, Wang HL, Hu ZH (2008) An extensive analysis on the global codon usage pattern of baculoviruses. Arch Virol 153(12):2273–2282. https://doi.org/10.1007/s00705-008-0260-1

Jordan-Paiz A, Nevot M, Lamkiewicz K, Lataretu M, Franco S, Marz M, Martinez MA (2020) HIV-1 lethality and loss of Env protein expression induced by single synonymous substitutions in the virus genome intronic-splicing silencer. J Virol. https://doi.org/10.1128/JVI.01108-20

Kishino H, Hasegawa M (1989) Evaluation of the maximum-likelihood estimate of the evolutionary tree topologies from DNA-sequence data, and the branching order in Hominoidea. J Mol Evol 29(2):170–179. https://doi.org/10.1007/Bf02100115

Kofman A, Graf M, Deml L, Wolf H, Wagner R (2003) Codon usage-mediated inhibition of HIV-1 gag expression in mammalian cells occurs independently of translation. Tsitologiia 45(1):94–100

Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in Escherichia coli. Science 324(5924):255–258. https://doi.org/10.1126/science.1170160

Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35(6):1547–1549. https://doi.org/10.1093/molbev/msy096

Meintjes PL, Rodrigo AG (2005) Evolution of relative synonymous codon usage in Human Immunodeficiency Virus type-1. J Bioinform Comput Biol 3(1):157–168. https://doi.org/10.1142/s0219720005000953

Motomura K, Chen JB, Hu WS (2008) Genetic recombination between human immunodeficiency virus type 1 (HIV-1) and HIV-2, two distinct human lentiviruses. J Virol 82(4):1923–1933. https://doi.org/10.1128/Jvi.01937-07

Pandit A, Sinha S (2011) Differential trends in the codon usage patterns in HIV-1 genes. PLoS ONE 6(12):e28889. https://doi.org/10.1371/journal.pone.0028889

Puigbo P, Bravo IG, Garcia-Vallve S (2008a) CAIcal: a combined set of tools to assess codon usage adaptation. Biol Direct 3(1):1–8. https://doi.org/10.1186/1745-6150-3-38

Puigbo P, Bravo IG, Garcia-Vallve S (2008b) E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI). BMC Bioinformatics 9(1):1–7. https://doi.org/10.1186/1471-2105-9-65

Rey-Cuille MA, Berthier JL, Bomsel-Demontoy MC, Chaduc Y, Montagnier L, Hovanessian AG, Chakrabarti LA (1998) Simian immunodeficiency virus replicates to high levels in sooty mangabeys without inducing disease. J Virol 72(5):3872–3886. https://doi.org/10.1128/JVI.72.5.3872-3886.1998

Richman DD, Wrin T, Little SJ, Petropoulos CJ (2003) Rapid evolution of the neutralizing antibody response to HIV type 1 infection. Proc Natl Acad Sci 100(7):4144–4149. https://doi.org/10.1073/pnas.0630530100

Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, Gao F, Hahn BH, Kalishc ML, Kuiken C, Learn GH, Leitner T, Mccutchan F, Osmanov S, Peeters M, Pieniazek D, Salminen M, Sharp PM, Wolinsky S, Korber B (2000) HIV-1 nomenclature proposal. Science 288(5463):55–56. https://doi.org/10.1126/science.288.5463.55d

Shabalina SA, Ogurtsov AY, Spiridonov NA (2006) A periodic pattern of mRNA secondary structure created by the genetic code. Nucleic Acids Res 34(8):2428–2437. https://doi.org/10.1093/nar/gkl287

Sharp PM, Bailes E, Robertson DL, Gao F, Hahn BH (1999) Origins and evolution of AIDS viruses. Biol Bull 196(3):338–342. https://doi.org/10.2307/1542965

Sharp PM, Hahn BH (2010) The evolution of HIV-1 and the origin of AIDS. Philos Trans R Soc Lond B Biol Sci 365(1552):2487–2494. https://doi.org/10.1098/rstb.2010.0031

Sharp PM, Li WH (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24(1–2):28–38. https://doi.org/10.1007/Bf02099948

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(3):1281–1295. https://doi.org/10.1093/nar/15.3.1281

Sharp PM, Shaw GM, Hahn BH (2005) Simian immunodeficiency virus infection of chimpanzees. J Virol 79(7):3891–3902. https://doi.org/10.1128/Jvi.79.7.3891-3902.2005

Sheehy AM, Gaddis NC, Choi JD, Malim MH (2002) Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature 418(6898):646–650. https://doi.org/10.1038/nature00939

Smith JM, Amara RR, Campbell D, Xu Y, Patel M, Sharma S, Butera ST, Ellenberger DL, Yi H, Chennareddi L, Herndon JG, Wyatt LS, Montefiori D, Moss B, McClure HM, Robinson HL (2004) DNA/MVA vaccine for HIV type 1: effects of codon-optimization and the expression of aggregates or virus-like particles on the immunogenicity of the DNA prime. AIDS Res Hum Retroviruses 20(12):1335–1347. https://doi.org/10.1089/aid.2004.20.1335

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85(8):2653–2657. https://doi.org/10.1073/pnas.85.8.2653

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. Nucleic Acids Res 22(22):4673–4680. https://doi.org/10.1093/nar/22.22.4673

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(1):1–14. https://doi.org/10.1186/1742-4690-9-92

van Hemert FJ, Berkhout B (1995) The tendency of lentiviral open reading frames to become A-rich: constraints imposed by viral genome organization and cellular tRNA availability. J Mol Evol 41(2):132–140. https://doi.org/10.1007/BF00170664

Vartanian JP, Henry M, Wain-Hobson S (2002) Sustained G→A hypermutation during reverse transcription of an entire human immunodeficiency virus type 1 strain Vau group O genome. J Gen Virol 83(4):801–805. https://doi.org/10.1099/0022-1317-83-4-801

Vidyavijayan KK, Hassan S, Precilla LK, Ashokkumar M, Chandrasekeran P, Swaminathan S, Luke Elizabeth Hanna LE (2017) Biased nucleotide composition and differential codon usage pattern in HIV-1 and HIV-2. AIDS Res Hum Retroviruses 33(3):298–307. https://doi.org/10.1089/AID.2015.0320

Visseaux B, Damond F, Matheron S, Descamps D, Charpentier C (2016) Hiv-2 molecular epidemiology. Infect Genet Evol 46:233–240. https://doi.org/10.1016/j.meegid.2016.08.010

Wong EHM, Smith DK, Rabadan R, Peiris M, Poon LLM (2010) Codon usage bias and the evolution of influenza A. viruses Codon Usage Biases of Influenza Virus. BMC Evol Biol 10(1):1–14. https://doi.org/10.1186/1471-2148-10-253

Wright F (1990) The effective number of codons used in a gene. Gene 87(1):23–29. https://doi.org/10.1016/0378-1119(90)90491-9

Yu Q, Chen D, Konig R, Mariani R, Unutmaz D, Landau NR (2004) APOBEC3B and APOBEC3C are potent inhibitors of simian immunodeficiency virus replication. J Biol Chem 279(51):53379–53386. https://doi.org/10.1074/jbc.M408802200

Zhou JH, Zhang J, Sun DJ, Ma Q, Chen HT, Ma LN, Ding YZ, Liu YS (2013) The distribution of synonymous codon choice in the translation initiation region of dengue virus. PLoS ONE 8(10):e77239. https://doi.org/10.1371/journal.pone.0077239

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Myeongji Cho[1,2] · **Xianglan Min**[1] · **Hyeon S. Son**[1,2,3]

[1] Laboratory of Computational Virology & Viroinformatics, Graduate School of Public Health, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea

[2] Institute of Health and Environment, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea

[3] Interdisciplinary Graduate Program in Bioinformatics, College of Natural Science, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea