Archives of Virology (2020) 165:557–570
https://doi.org/10.1007/s00705-020-04533-6

ORIGINAL ARTICLE

Codon usage pattern and its influencing factors in different genomes of hepadnaviruses

Bornali Deb[1] · Arif Uddin[2] · Supriyo Chakraborty[1]

Received: 31 August 2019 / Accepted: 7 December 2019 / Published online: 8 February 2020

Abstract

Codon usage bias (CUB) arises from the preference for a codon over codons for the same amino acid. The major factors contributing to CUB are evolutionary forces, compositional properties, gene expression, and protein properties. The present analysis was performed to investigate the compositional properties and the extent of CUB across the genomes of members of the family *Hepadnaviridae*, as previously no work using bioinformatic tools has been reported. The viral genes were found to be AT rich with low CUB. Analysis of relative synonymous codon usage (RSCU) was used to identify overrepresented and underrepresented codons for each amino acid. Correlation analysis of overall nucleotide composition and its composition at the third codon position suggested that mutation pressure might influence the CUB. A highly significant correlation was observed between GC12 and GC3 ($r = 0.910$, $p < 0.01$), indicating that directional mutation affected all three codon positions across the genome. Translational selection (P2) and mutational responsive index (MRI) values of genes suggested that mutation plays a more important role than translational selection in members of the family *Hepadnaviridae*.

Introduction

Amino acids are the building blocks of proteins, and the specific amino acids incorporated are determined by the genetic code. In the standard genetic code, a set of 61 codons encodes the 20 standard amino acids. Other than tryptophan and methionine, all amino acids are represented by more than one codon, resulting in codon redundancy. The condition of biased usage of some codons preferentially over other synonymous codons is known as codon usage bias (CUB), and it is specific for every genome [3, 29, 30, 53]. CUB differs among genomes as well as within the same genome, and studying these differences may help us to understand genome evolution among related species [66] as well as the relationship between host cells and viruses or immune reactions [62].

Various hypotheses have been proposed to explain the occurrence of CUB. In the neutral theory, mutational pressure at degenerate positions of a codon must be neutral, such that there is nonuniform usage of synonymous codons for a specific amino acid, indicating a lack of natural selection [48]. The level of gene expression has been shown to be associated with CUB [63, 64], whereas the selection-mutation-drift model postulates the importance of genetic drift, mutation pressure, and natural selection in the establishment of CUB [63, 64]. Natural selection of highly expressed genes can play an important role [31], and influences codon usage in various organisms [25]. Other notable determinants of innate CUB are base composition [5], skewness of bases [10], expression level of the gene [77], gene length [17], gene stability, replication [25, 39], translational selection [56], protein secondary structure [77] and hydrophobicity [13]. A previous study showed that variation in the tRNA pool and disparity in isochores of a cell are major determinants of CUB [4, 16].

Mutation is a major factor determining in configuring codon usage patterns in various viral genomes [52]. Investigation of constraints in codon usage provides information about molecular evolution of viruses and regulation of gene

Handling Editor: Carolina Scagnolari.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00705-020-04533-6) contains supplementary material, which is available to authorized users.

✉ Supriyo Chakraborty
supriyoch_2008@rediffmail.com

[1] Department of Biotechnology, Assam University, Silchar 788150, Assam, India

[2] Department of Zoology, Moinul Hoque Choudhury Memorial Science College, Algapur, Hailakandi 788150, Assam, India

expression and is useful for the design of vaccines [26]. Tao et al. reported mutational pressure to be a more important factor than selection constraints in CUB determination, as a significant correlation was observed between nucleotide content and CUB in 35 classical swine fever virus isolates [75]. Ma et al. have suggested that translational selection and mutational constraints are major evolutionary forces that govern CUB generation in hepatitis B virus [42]. Inspection of codon usage and compositional constraints of the DNA polymerase gene of herpes simplex virus type 1 has revealed higher usage of G or C bases over A or T at third codon position.

The family *Hepadnaviridae* includes enveloped viruses with a diameter of ~42 nm and an icosahedral core of ~34 nm [85]. A single capsid protein oligomerizes to form a capsid structure with icosahedral symmetry. Circular genomes with partially double-stranded DNA (~3.2 kbp) are synthesized by reverse transcription. Hepadnaviruses do not depend on host polymerase but instead encode their own polymerase [72] for reverse transcriptase activity, which converts RNA to DNA during genome replication.

CUB is a useful tool for understanding the factors responsible for governing viral evolution [32]. CUB in viral genes might be related to specific host selection, leading to a better grasp of viral evolution and the adaptive response of the host to infection [81]. A preliminary analysis of codon usage and base composition members of the genus *Flavivirus* has revealed a relationship between them [6].

In the current study, we investigated the compositional properties and pattern of codon usage in genes of members of the family *Hepadnaviridae* in order to identify their molecular characteristics and assess the role of evolutionary forces in shaping the CUB of genes. We identified overrepresented and underrepresented codons that could potentially be used in genetic engineering to develop better therapeutics. This analysis might help to elucidate host adaptive traits, mechanisms of viral evolution, and adaptive strategies of the host against infection.

Materials and methods

Retrieval of coding sequences

The complete coding sequences (cds) of genomes of members of the family *Hepadnaviridae* were retrieved from the GenBank database at the National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov). In the present analysis, we used only such cds that were exact multiples of three nucleotides and had a correct start and stop codon, eliminating all unknown bases from the cds. A list of genome sequences used in our analysis is shown in Supplementary File 1. We also compared the codon usage of these viruses with that of their hosts.

Effective number of codons (ENC)

The effective number of codons acts as a framework for quantifying the rate of CUB in cds independently of the length of the gene and the number of amino acids [84]. It indicates the degree of variation in codon usage in a gene from a completely uniform usage of synonymous codons. The ENC is calculated using the following formula:

$$ENC = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}$$

where F_k (k = 2, 3, 4 or 6) is the average of the F_k values for *k*-fold degenerate amino acids. The F value is the probability of two randomly chosen codons being identical for a particular amino acid.

Relative synonymous codon usage (RSCU)

To investigate the pattern of biased usage of synonymous codons, the relative synonymous codon usage (RSCU) value of each codon was determined. The RSCU value of a codon is the proportion of observed frequency to its predicted frequency within the synonymous codon family coding for a specific amino acid [65].

The RSCU value of a synonymous codon is estimated as

$$RSCUij = \frac{Xij}{\frac{1}{ni}\sum_{j=1}^{ni} Xij}$$

where, X_{ij} indicates for the frequency of the j^{th} codon for i^{th} amino acid and ni is the number of codons for the i^{th} amino acid (i^{th} codon family).

Base composition

The overall base content (A%, T%, G% and C%) and the base content at the third codon position (A3%, T3%, G3% and C3%) for the coding sequences of each genome were analysed. The overall GC content and its composition at the three codon positions (GC1%, GC2% and GC3%) were determined. Nucleotide skews, namely AT skew, GC skew, purine skew, pyrimidine skew, purine-pyrimidine skew, amino skew, and keto skew values of coding sequences over all genomes were computed.

PR2 bias plot analysis

A parity rule 2 (PR2) bias plot was made by plotting the GC bias on the abscissa [G3 /(G3 + C3)] and the AT bias [A3 /(A3

+ T3)] on the ordinate [44, 45]. In the PR2 plot, the midpoint is 0.5, which corresponds to the condition in which A = T and G = C and no bias is observed between mutation and selection pressure [46].

Neutrality plot

A neutrality plot was made by comparing GC3 (*x*-axis) and GC12 (*y*-axis) to account for the role of mutation-selection equilibrium in codon usage disparity. Each independent gene was represented by a dot in the plot. An effect of mutation pressure on the biased usage of codons is indicated by the slope of a regression line of GC12 vs. GC3, *i.e.*, if the value approaches 1 [70], whereas a scattered distribution of points indicates a significant role of natural selection in CUB generation.

Minimum free energy of mRNA

The energy released by mRNA secondary structure formation during the transcription process was estimated in kcal/mol. The minimum free energy was estimated as described by Ringnér and Krogh [57].Negative free energy values are indicated by a negative sign, and therefore, absolute values were used in our statistical analysis. A high absolute value thus indicates a greater loss of energy by the mRNA molecule when attaining a stable conformation. A highly stable mRNA conformation is expected to arise from a greater loss of energy as compared to the less stable mRNA [57].

mRNA stability index

The mRNA stability index of each coding sequence was estimated from the codon stabilization coefficient (csc) values of sense codons as described by Presnyak et al. [54]. The csc values of individual codons range from - 0.25 to + 0.25. The total stability of mRNA was estimated as the sum of the products of the individual codon csc values and their frequency in the coding sequence. The average mRNA stability index (per codon) was calculated and normalized for each coding sequence to range between - 1 (lowest stability) to + 1 (highest stability) [54].

Mutational responsive index (MRI)

The magnitude of mutational drift was measured using the mutational responsive index. A positive MRI value suggests a role of mutation on the coding sequence, while a negative MRI value indicates a role of translational selection across the gene. The MRI value was calculated as described previously [20] using the formula

$$\mathrm{MRI} = \mathrm{SCS} - \mathrm{CSCS}$$

where SCS is the scaled chi-square (SCS) value and CSCS is the corrected scaled chi-square (CSCS) value.

Translational selection (P2)

The P2 value indicates the efficiency of the interaction between the codon and anticodon and indicates the potential of the translational process in a gene. The P2 value was computed using the following formula:

$$\mathrm{P2} = \frac{(WWC + SSU)}{(WWC + SSY)}$$

where W = T or A, S = C or G and Y = T or C. A P2 value greater than 0.5 suggests an effect of translational selection on a coding sequence [22].

Software

A computer script written in the PERL language by the corresponding author (SC) was used to analyze the codon bias indices for different genes in hepadnavirus genomes. To measure the correlation among different parameters, namely the effective number of codons and compositional properties, we used the SPSS software package (Chicago, Illinois, USA). Paleontological statistics software (PAST) was used to perform correspondence analysis to study variations in codon usage, and cluster analysis to identify the most closely and distantly related genomes in the course of evolution.

Results

Codon usage bias analysis

In order to investigate the extent codon usage bias in the genomes of members of the family *Hepadnaviridae*, we determined the ENC values of the coding sequences for each virus listed in Table 1 and observed that the values varied from 42.40 to 56.33, with a mean of 52.49 (*i.e.* > 35). These values indicate that the codon usage bias in these genomes is low [9]. The RSCU values of the 59 sense codons indicated that almost half of them (28/59) were used frequently, indicating that more than one codon was used for several amino acids.

Pattern of codon usage

To examine the pattern of heterogeneous codon usage, we plotted the RSCU values of each codon as shown in Fig. 1. RSCU values greater than 1.6 indicated overrepresented codons, and RSCU values less than 0.6 indicated underrepresented codons. Three codons (TCT, AGA, GGA)

Table 1 Average ENC value of genes in hepadnavirus genomes

Virus	ENC
Woodchuck hepatitis virus	53.83
Ground squirrel hepatitis virus	52.88
Duck hepatitis B virus	53.23
Long-fingered bat hepatitis B virus isolate 776	55.05
inamou hepatitis B virus isolate 160050	54.22
Tibetan frog hepadnavirus isolate 243398	55.55
Woolly monkey hepatitis B virus clone WMHBV-2	54.66
White sucker hepadnavirus isolate RR173	53.97
Tent-making bat hepatitis B virus isolate TBHBV_Pan372_Uro_bil_PAN_2010	54.53
Horseshoe bat hepatitis B virus isolate HBHBV_GB09-403_Rhi_alc_GAB_2009	56.33
Roundleaf bat hepatitis B virus isolate RBHBV_GB09-256_Hip_rub_GAB_2009	54.48
Parrot hepatitis B virus	51.08
Snow goose hepatitis B virus	51.94
Ross's goose hepatitis B virus	49.10
Sheldgoose hepatitis B virus	46.30
Heron hepatitis B virus	52.78
Hepatitis B virus (strain ayw)	42.40

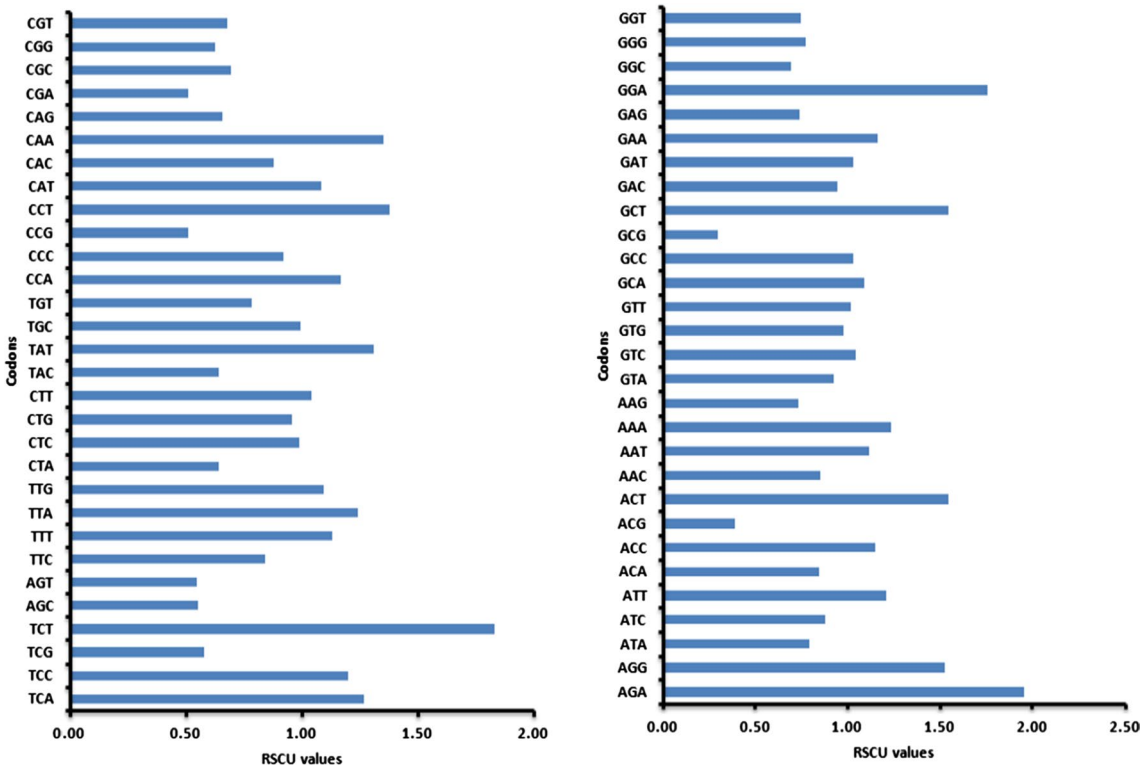

Fig. 1 Overall RSCU values of codons in hepadnavirus genomes

were found to be overrepresented, and seven (TCG, AGC, AGT, CCG, CGA, ACG, GCG) were underrepresented across the genome (Fig. 2). The preferred codon(s) for each amino acid are listed in Supplementary File 2.

Relationship between codon usage patterns of different hepadnaviruses and their hosts

Since viruses are obligate parasites, their codon usage patterns

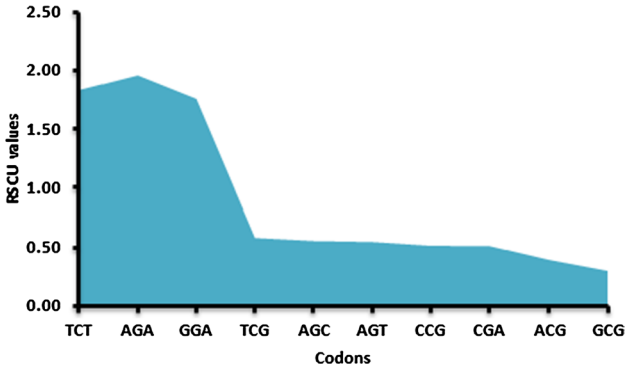

Fig. 2 Overrepresented codons (RSCU value > 1.6) and underrepresented codons (RSCU value < 0.6) in hepadnavirus genomes

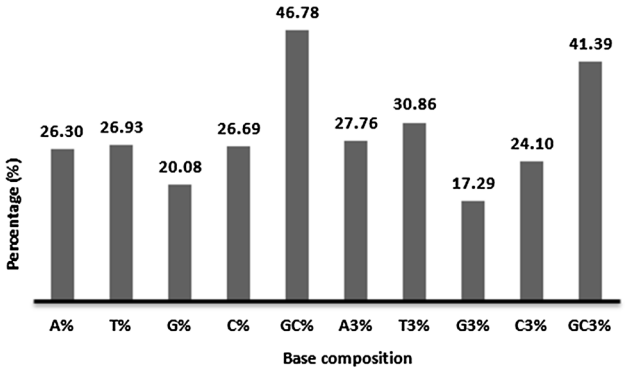

Fig. 3 Overall base content and base content at the third codon position of genes in hepadnavirus genomes

might be affected by those of their hosts [93]. Here, we analysed the codon usage patterns of a few hepadnaviruses and their respective hosts. As shown in Supplementary File 3, most of the viruses showed similarity to their hosts in their pattern of more and less frequently used codons and also had a few overrepresented and underrepresented codons in common, indicating a possible relationship between them. A similar pattern of relatedness has also been found with poliovirus [47], chikungunya virus [9], and coronaviruses [82] and their respective hosts.

Compositional properties

The biased choice of a codon preferably over other synonymous codons of the same family is strongly influenced by compositional characteristics of the genome [32]. In our analysis, we determined the overall base composition and the nucleotide frequency at the third codon position in all of the genomes in Fig. 3. An almost equal proportion of T, A and C (~27%) bases was observed, compared to G (~20%) base. The overall GC% and AT% were 46.78 and 53.22, respectively, indicating that hepadnavirus genes are AT rich. At the third codon position, T (30.86%) was most frequent, followed by A (27.76%), C (24.10%) and G (17.29%). The overall GC3% and AT3% content was 41.39 and 58.61, respectively, indicating AT richness at the third position across the genomes. GC composition has been reported to be a significant factor influencing codon usage bias across the genomes [79]. Here, GC content was highest in position 1, followed by positions 2 and 3. We found a significant correlation of ENC with the G3 and C3 content, ($p < 0.05$) (Table 2), indicating influence on the CUB. Our interpretation of nucleotide compositional properties and relative synonymous codon usage values suggests that mutational pressure might have a substantial effect on CUB. We also correlated codon usage with GC3 content and found a few positive and negative correlations between GC-ended codons and GC3 (Fig. 4). These results reveal the variation in codon usage in relation to GC constraints and provide a better understanding of the molecular architecture of genomes of hepadnaviruses [28].

Variation in codon usage

Correspondence analysis is a multivariate analysis that is used to explore the synonymous codon usage variation among genomes. In order to identify differences in the codon usage pattern across the genomes, we performed a correspondence analysis (CA) of the RSCU values of the 59 sense codons across the genomes (Fig. 5). In the figure, axis 1 and axis 2 are the two major contributors to the total variation. A green dot on the figure represents AT-ending codons, and a red dot indicates GC-ending codons. The figure shows a close distribution of the bases across the axes, suggesting that mutational pressure might have influenced the CUB of the genes, supporting the results reported by Wei et al.[80].

Cluster analysis was done with Past software, using the RSCU values of the 59 sense codons of each genome. These results, along with the findings of CA, revealed two major clusters (Fig. 6). Seven hepadnavirus genomes were found in one cluster, and 10 genomes were found in another cluster, indicating a close evolutionary relationship within each cluster.

Table 2 Correlation between ENC and base content of genes in hepadnavirus genomes

	A%	T%	G%	C%	GC%	A3%	T3%	G3%	C3%	GC3%
ENC	-0.178	-0.058	-0.167	0.267	0.144	-0.054	-0.102	-0.540*	-0.501*	0.094

*Significant at $p < 0.05$

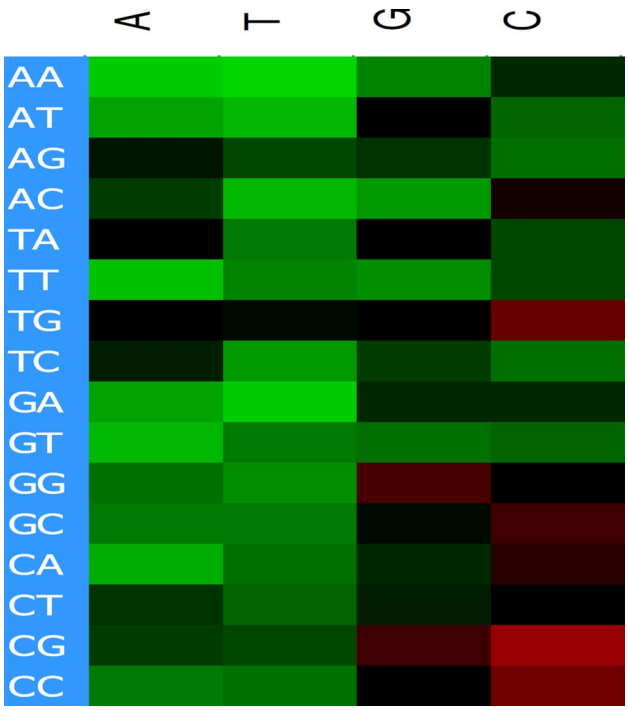

Fig. 4 Heat map with codon usage values of codons in hepadnavirus genomes. Green color indicates negative correlation, and red indicates a positive correlation of GC3 with A-, T-, G- and C-ending codons

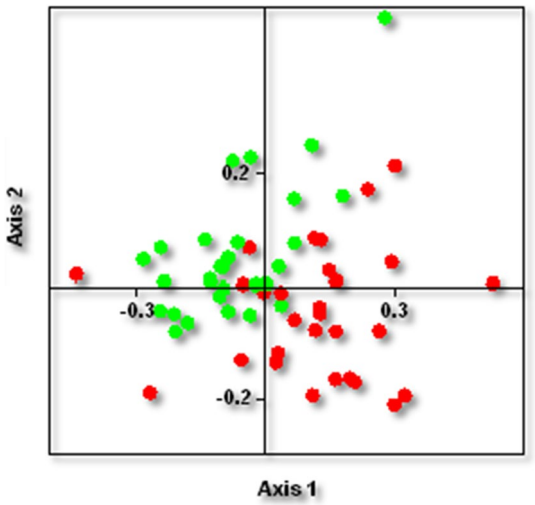

Fig. 5 Correspondence analysis of genomes of hepadnaviruses. Red color indicates GC-ending codons, and green indicates AT-ending codons

Parity rule 2 (PR2) bias plot analysis

A PR2 bias plot usually reveals the comparative magnitude of mutation and natural selection acting on genome composition [69]. A proportionate distribution of bases across the plot revealed that mutation might influence the CUB across the genomes, while a disproportionate distribution might point towards a role of both mutation and natural selection in determining the CUB [71]. We analysed the associations between the purine (A and G) and the pyrimidine (C and T) content, with A3/A3 + T3 on the ordinate and G3/G3+C3 on the abscissa in 2- fold, 4-fold and 6-fold PR2 bias plots (Fig. 7) to investigate the impact of evolutionary determinants on CUB. Notably, we observed an asymmetrical distribution of bases across the plot, indicating that both mutation and selection pressure might have affected the CUB [13].

Interrelationships among base compositions

The variation in the patterns of codon usage mainly stems from two evolutionary forces, *viz.*, mutational pressure and natural selection [44, 45]. Correlation analysis of compositional constraints can identify the primary forces determining the CUB [13]. We correlated overall base content (A, T, G and C %) with base content at the third codon position (A3, T3, G3 and C3%) using Karl Pearson's method (Table 3). A highly significant correlation was observed at $p < 0.01$ and $p < 0.05$, suggesting a strong influence of mutational pressure in determining the CUB of genes of hepadnaviruses, thereby supporting the results reported by Zhang et al. [88, 89].

Neutrality plot analysis

A neutrality plot is a useful tool to quantify the impact of two evolutionary forces on the genome. A highly significant positive correlation was observed between GC12 and GC3 ($r = 0.910$, $p < 0.01$), indicating the impact of directional mutation at all codon positions across the genomes, supporting the results reported by Sueoka [70]. Moreover, we plotted GC3 (abscissa) and GC12 (ordinate) to draw a linear regression line (Fig. 8). A regression coefficient (RC) value less than 0.5 suggests a greater role of natural selection, while an RC greater than 0.5 suggests a greater impact of mutational pressure. Our analysis yielded an RC value of 0.5118, indicating that the role of mutation pressure was slightly higher than that of natural selection in influencing the CUB of hepadnavirus genomes.

Nucleotide skewness

A nucleotide skewness analysis yielded negative values for mean GC skew (-0.14) and AT skew (-0.01), indicating a more frequent usage of C and T over G and A [80]. Previous studies have shown that the nucleotide skewness can influence the CUB of genes [13]. We therefore correlated base skew values with ENC using Karl Pearson's method

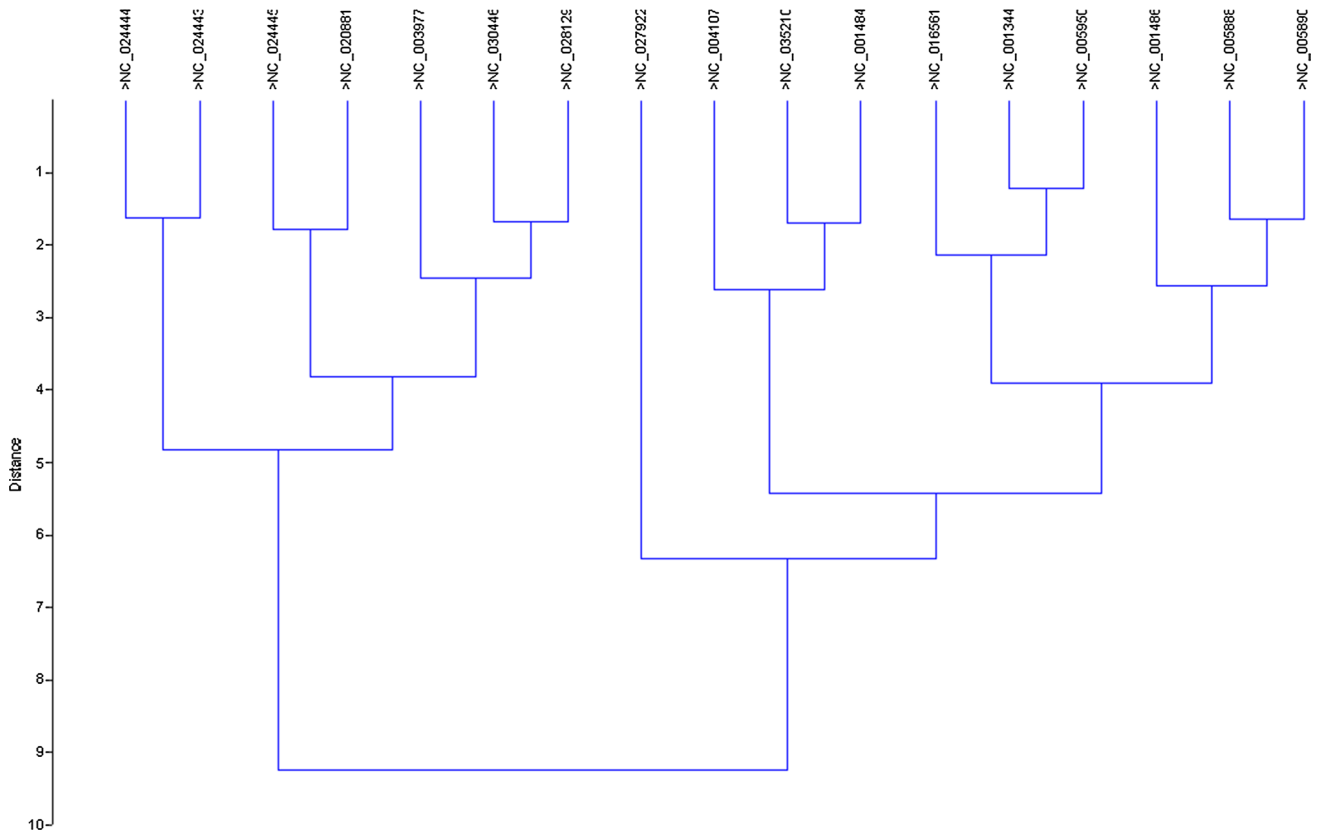

Fig. 6 Cluster analysis of genomes of hepadnaviruses

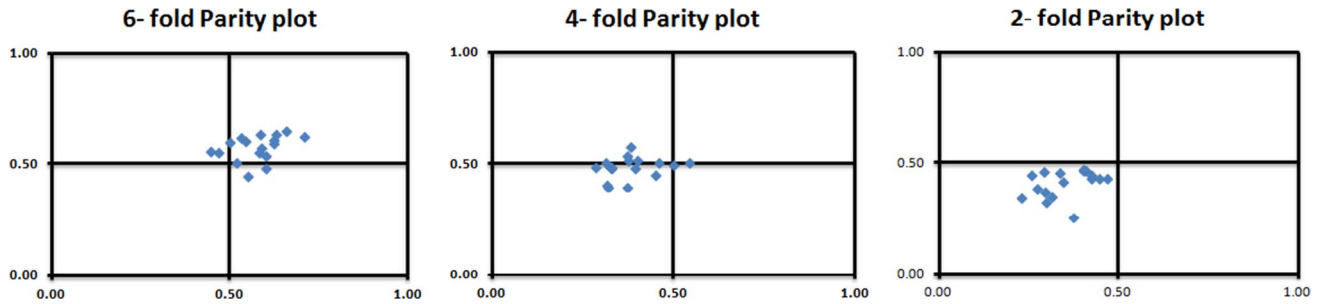

Fig. 7 Parity rule 2 bias plot of genomes of hepadnaviruses

Table 3 Interrelationships of overall base composition with base composition at the third codon position

	A3%	T3%	G3%	C3%	GC3%
A%	0.943**	0.510*	-0.463	-0.782**	-0.823**
T%	0.448	0.907**	-0.544*	-0.714**	0.813**
G%	-0.749**	-0.766**	0.797**	0.630**	0.883**
C%	-0.818**	-0.799**	0.466	0.937**	0.940**
GC%	-0.841**	-0.834**	0.597*	0.893**	0.975**

**, *Significant at $p < 0.01, 0.05$

and recorded a negative correlation of CUB with GC skew (-0.48), AT skew (-0.276), pu skew (-0.058), py skew (-0.096), amino skew (-0.236) and purine-pyrimidine skew (-0.374), but a positive correlation of ENC with keto skew (0.148), suggesting that all these nucleotide skews might influence the CUB of genes across the genome.

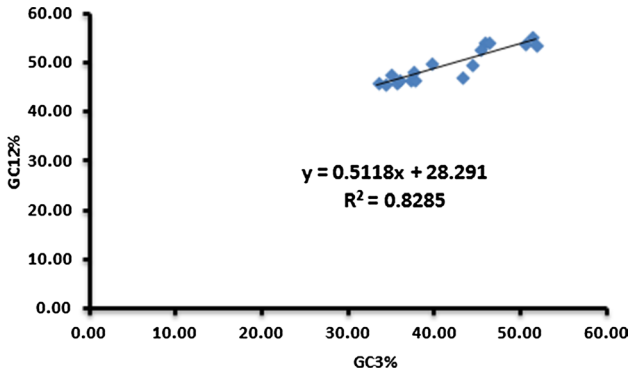

Fig. 8 Neutrality plot of genomes of hepadnaviruses

Role of minimum free energy

Minimum free energy refers to the quantum of energy released by an mRNA molecule during the process of transcription [57]. The minimum free energy of genes in each genome in the family *Hepadnaviridae* is presented in Supplementary File S4. The minimum free energy of genes over all genomes in the family *Hepadnaviridae* was found to be -314.07 kcal/mol (ranging from -150.78 to -576.79) with the negative sign indicating a loss of energy. This suggests that energy is released during the transcription process, resulting in a more stable conformation. Higher energy release might lead to the formation of a less stable structure and thus affect the translation process. Here, we correlated the ENC value with the minimum free energy of each gene and found a non-significant relationship between them, suggesting that the loss of minimum free energy was not related to codon usage bias. Further, we correlated minimum free energy with GC composition (overall GC, GC1, GC2, GC3 and GC12%) using Karl Pearson's method and observed a highly significant negative correlation (Table 4), indicating that minimum free energy might be associated with GC compositions and that the conformation/stability of mRNA transcripts could be related to GC constraints.

Role of mRNA stability

The degradation rate of mRNA is a major factor in gene expression; unstable mRNA molecules usually contain non-optimal codons, leading to substantial destabilization in protein expression, while stable mRNA molecules contain optimal codons [54]. We determined the mRNA stability index of each coding sequence, and the mean for each genome in shown in Supplementary File S5. Seven genomes were found to have a positive value, while nine genomes had a negative value. A positive mRNA stability index indicates higher stability of the mRNA and vice versa. We correlated the ENC value with the mRNA stability index using Karl Pearson's product moment method and found a nonsignificant relationship between mRNA stability and CUB, indicating that codon bias was not associated with mRNA stability. However, on correlating the mRNA stability index with base composition (Table 5), a highly significant negative correlation was observed with C% and a significant positive correlation was observed with G1% and A3% ($p < 0.05$), suggesting that the stability of mRNA might be associated with its nucleotide composition.

Role of translational selection (P2)

To gain insights into the role of CUB on mRNA translation, we determined the mean P2 value across the genomes of hepadnaviruses and found the value to be 0.15. A P2 value less than 0.5 suggests a lesser role of translational selection in CUB determination [10]. On further correlation analysis between ENC and P2 values, we found a highly significant negative correlation (-0.653**), suggesting, an inverse relationship between translational selection and CUB. This indicated that a coding sequence with a low ENC value

Table 4 Correlation between minimum free energy (mFE) of mRNA and GC composition of genes

Correlation	GC%	GC1%	GC2%	GC3%	GC12%
mFE	0.800**	0.600**	0.744**	0.784**	0.780**

**Significant at $p < 0.01$

Table 5 Correlation between mRNA stability index and base composition of genes in hepadnavirus genomes

Correlation	mRNA stability index
A%	0.433
T%	0.234
G%	-0.132
C%	-0.489*
A1%	0.416
T1%	-0.128
G1%	0.562*
C1%	-0.481
A2%	0.190
T2%	0.340
G2%	-0.385
C2%	-0.262
A3%	0.491*
T3%	0.194
G3%	-0.115
C3%	-0.434

*Significant at $p < 0.05$

(implying high codon bias) might have been subjected to a high degree of translational selection during evolution.

Role of the mutation responsive index (MRI)

The mutation responsive index is a useful parameter for quantifying the effect of mutation and translational selection on CUB [10]. A positive MRI value suggests directional mutation pressure, while a negative MRI value indicates a role of translational selection on the CUB. The mean MRI value in our analysis was 0.47, *i.e.,* positive, suggesting an influence of directional mutation pressure across the genomes of hepadnaviruses.

Discussion

A point mutation at the third nucleotide position of a codon usually leads to a synonymous substitution that does not alter the encoded amino acid, and thus the stability of the organism is not affected. However, nonsynonymous substitutions can result in phenotypic changes that allow natural selection to act upon genes [74]. Mutation and natural selection are the two major evolutionary forces that contribute to the CUB of genes. Other factors affecting CUB include base composition, gene expression, genetic drift, nonsense mutation, missense mutation, and mRNA stability.

Hepatitis B is a global health concern [86], with approximately 2 billion individuals infected and 0.6 million deaths each year [83]. Infected patients develop acute or chronic hepatitis, which can cause liver cirrhosis or primary hepatocellular carcinoma [61]. Woodchuck hepatitis virus acts on neonates and can cause acute hepatitis in woodchucks, which can eventually become chronic carriers of the virus [12]. Chronically infected woodchucks have a high probability of developing hepatocellular carcinoma, although cirrhosis may not occur [2]. Ground squirrel hepatitis virus can also cause hepatitis, and chronically infected animals sometimes develop hepatocellular carcinoma [43]. Infected pekin ducks show few disease symptoms, with noncytopathic replication occurring in hepatocytes [1]. Snow goose hepatitis B virus forms large numbers of virions with single-stranded DNA in its host [23].

In this study, we examined the degree of CUB, overall compositional properties, overrepresented and underrepresented codons, role of evolutionary forces, impact of nucleotide skewness, and the role of minimum free energy and mRNA stability in the genomes of members of the family *Hepadnaviridae*. The results provide an in-depth understanding of gene expression, the role of mutation and selection pressure on genes, and identification of the preferred codons for each amino acid. These results were compared with those obtained with other organisms to identify similarities and differences.

The effective number of codons (ENC) indicates the magnitude of codon bias across the genome. In the present study, the mean ENC value was low, indicating that there is little bias in codon usage in members of the family *Hepadnaviridae* [10]. A lack of strong CUB is expected to promote efficient usage of more codons and thereby speed up the translation process [32]. Fu reported a lower ENC value in five members of the *Herpesviridae* family of DNA viruses than in other members of the family [19]. Similarly, the ENC value of 11 human bocavirus isolates was in the range of 40.87 to 48.42, with a mean value 44.45, indicating low CUB [91]. However, low ENC values implying high bias have been reported in several viruses, including Orgyia pseudotsugata nucleopolyhedrovirus and Lymantria dispar nucleopolyhedrovirus [33]. Lower CUB might relate to efficient replication in different cell types with different codon preferences [32].

We observed similar usage of three nucleotides, T, A and C, across the genomes (Fig. 1) and identified three overrepresented and seven underrepresented codons in the mRNA molecules of hepadnaviruses (Fig. 2). RSCU analysis of classical swine fever virus has demonstrated a preferential use of G-, C-, and A- ending codons, with no T-ending codon across the genome [75]. Jiang et al. analysed the codon usage pattern in baculovirus genomes and found nine overrepresented codons namely, TAC, TTT, TTG, CAA, CAC, ATT, AAA, GAA, and GTG [33]. RSCU analysis of mimivirus elucidated higher usage of A/T-ending codons over G/C-ending codons [60].

A related pattern of codon usage was observed between a few hepadnavirus genomes and their respective hosts, which had the majority of the more and less frequently used codons as well as a few overrepresented and underrepresented codons in common. Similar patterns have also been reported for foot-and-mouth disease virus [94], papillomavirus [95], astroviruses [78], and equine influenza virus [38] and their hosts. These patterns of relatedness suggest that selection pressure from the host might affect the CUB in the viral genome, allowing the virus to adjust to its cellular environment [9]. Previously, it was reported that similar CUB patterns in viral and hosts genomes increases the efficiency of translation. [21].

Codon usage has been reported to be significantly influenced by the base composition of the gene [10]. In the present analysis, the base frequency of A, T, C was nearly equal, but the frequency of G was different. The hepadnaviruses were found to have AT-rich genomes (Fig. 3). Mutation pressure is assumed to play an important role in shaping CUB in some genes if these have a very high content of A and T or G and C [35, 67, 90, 92]. The AT content leads to low

thermodynamic stability, which plays a significant role in initiation of replication [55].

The most frequent nucleotide in the third codon position was T, followed by A, C, and G. The frequency of G and C was highest in the first position and lowest in the third position. The GC content has been reported previously to be proportionally related to CUB [77]. Zhang et al. [89], analysed the base content of torque teno sus virus 1 and reported A% > G% > C% > T%, suggesting the preferred use of A- over T-ending codons [88, 89]. Bouquet et al., analysed genetic and codon characteristics of hepatitis E virus and reported the distribution of G and T bases to be ~25% each, while A was highly preferred over C [7]. Sheng et al. reported the GC composition of the porcine circovirus genome to be 48.61%, playing a preferential role in synonymous codon usage [40].

Correlation analysis of GC3 bias and codon usage revealed a few positive and negative relationships with some GC-ending codons, while a negative relationship was observed with AT-ending codons across the genomes (Fig. 4). Palidwor et al. reported a similar pattern of codon usage across bacteria, yeast and humans [49]. Choudhuryet al. reported a positive association of GC3 with GC-ending codons in the human SPANX gene [11]. Similarly, Uddin et al. reported a negative relationship of AT-ending codons and a positive relationship of GC-ending codons across birds and mammals [76]. The relationship of codon usage to GC bias suggests a significant impact of GC content on CUB and the molecular makeup of genes.

The essence of variation in codon usage is multifactorial; therefore, a multivariate statistical approach, correspondence analysis, was performed to estimate the rate of variation in codon usage. A closer distribution of bases was observed across the axes, depicting the impact of mutation (Fig. 5). The major trends of variation in codon preference of baculoviruses were clearly depicted with COA [33]. Zhang et al. [89], performed COA with RSCU values of codons in torque teno sus virus 1 and reported geographical diversity as a limiting factor for CUB in whole viral genome [88].

Cluster analysis was performed to investigate the evolutionary relatedness of hepadnaviruses, and it revealed two major branches with a close association between them (Fig. 6). Deka et al. reported three major clusters of influenza A virus genes [14]. A feasible relationship of porcine circovirus 3 to other circoviruses was found with hierarchical clustering among them [18]. Samuel et al. reported that foot-and-mouth disease viruses obtained from the Middle East and North Africa were not closely related to classical European vaccine strains [58].

The role of two major evolutionary forces driving CUB was investigated using parity rule 2 bias plots for 2-fold, 4-fold and 6-fold degenerate codon families (Fig. 7). Disproportional distributions of bases were found across the scatter plot, suggesting a role of both mutation pressure and natural selection in CUB [77]. Gun et al. used a PR2 bias plot to analyze the PB2 genes of influenza A H7N9 virus isolates from several host species and reported no constraint in selection and mutation between two complementary strands of DNA [24]. A strong preference for A and G bases over T and C in 4-fold degenerate codon families was found in the coding sequences of Zika virus [8].

Discrepancy in codon usage associated with mutational pressure is generally related to deamination and demethylation of DNA, errors in non-random replication process, and chemical decomposition of nucleotides [36]. The mutation biases are usually neutral, act on DNA sequences of an organism, and do not influence the properties of the encoded protein. Some mutations are caused by replication errors and methylation. Different fidelities in replication of the leading and lagging strands can also cause strand-specific mutational bias in bacteria [41] and eukaryotes [50]. However, in numerous non-chordate species with a large population size, natural selection acts on the synonymous codon usage pattern [37]. This could perhaps influence the translation rate due to matching of transfer RNA abundance and codon usage [96].

In the present study, a significant correlation was found between the overall base content and the base contents at the third codon position, indicating that mutation is a driving force in the establishment of CUB. Similarly, a significant relationship to base content, namely A, T, G, C, G + G and A3, T3, G3, C3, (G + C)3, has been reported in polioviruses [87]. A significant positive or negative correlation with compositional constraints has been identified in the torque teno sus virus 1 genome, indicating an effect of mutational pressure [88, 89]. Correlation analysis of porcine circovirus also revealed a significant relationship among T%, A%, G%,C%, GC%, and T3%, A3%, G3%, C3%, GC3%, showing that base content had a major impact in codon preference [40].

An analytical method based on GC12 and GC3 content could be used to determine the magnitude of evolutionary forces. We therefore used a neutrality plot in this study to analyze hepadnavirus genomes (Fig. 8). The regression coefficient from the plot was found to be 0.5118, indicating that mutation plays a more important role than natural selection. Furthermore, a significant correlation was observed between GC12 and GC3 ($r = 0.910$, $p < 0.01$), suggesting a directional role of mutation. High mutation pressure leads to evolution. Most viruses have a high evolutionary rate due to large population size, high mutation rate, and short generation time. Viral mutations arise from errors made during replication of the viral genome. The mutation rate is used to determine the amount of genetic variation within a population, and this allows natural selection to operate [51]. A high mutation rate can lead to a higher degree of genetic diversity [59].

A neutrality plot of the PB2 gene of influenza A H7N9 virus suggested that natural selection is more important than mutational pressure for determining the CUB of viral genes [24]. A significant positive correlation was observed between G12 and GC3 in coding sequences of Zika virus, and the slope of the regression line was found to be 0.032, indicating a major influence of selection over mutational pressure in determining the CUB [8].

An asymmetrical distribution of nucleotides in coding sequences is measured using the nucleotide skewness parameter x-y/x+y, where x and y represent two different nucleotides. A positive skewness value indicates a preponderance of x over y and a negative value indicates the opposite. Nucleotide skewness has been reported to influence the CUB of genes [10]. Our analysis revealed higher usage of C over T (pyrimidine) and G over A (purine) nucleotides. Correlation analysis of ENC with nucleotide skews showed an inverse relationship between CUB and GC skew, AT skew, PU skew, PY skew, amino skew and PU-PY skew and a positive correlation between CUB and keto skew. Skew analysis across retroviral genomes revealed higher usage of A over G and C over T [8]. Nucleotide skew values of Nipah virus genes showed significant correlation with codon usage across the genome [10].

Since mRNA is required for translation, factors affecting mRNA either directly or indirectly influence the translation process [57].The folding of an mRNA molecule is thought to affect the stability of codons in protein expression. In the present study, the mFE values of coding sequences ranged from - 150.78 to - 576.79 kcal/mol, with a mean value of - 314.07 kcal/mol. Coding sequences with a low mFE value are weakly folded, and coding sequences with high mFE value are strongly folded. Coding sequences with low mFE values tend to produce more protein than those with high mFE values because strong secondary structure in the mRNA molecule disfavors translation. We therefore speculate that the coding sequences in hepadnavirus genomes with a low absolute ΔG value of have a higher translation rate than those with a low absolute ΔG value [57].

A significant correlation was observed between the mFE value and GC content, suggesting that the stability of the mRNA molecule is affected by GC content. Since the GC content is associated with the stability of mRNA secondary structure, this might play a role in determining the extent of gene expression. Similarly, the mFE value of mitochondrial ATP6 gene across the phylum Platyhelminthes also shows a significant correlation with GC content [44, 45]. The mRNA stability index is another parameter that was used in our analysis to estimate mRNA stability. Our results showed that seven genomes had a positive value (highly stable mRNA), while nine genomes had a negative value (less stable mRNA). A significant correlation was found between the stability index and base composition in hepadnavirus genomes. As the stability of the mRNA molecule has been found to be related to the stability of the protein [27], the stability of gene products can be predicted using the mRNA stability index for the expressed genes. The mRNA stability index can also be used to predict the level of expression of a gene.

Gouy and Gautier have suggested that codon usage bias of highly expressed genes is determined by translational selection [22], since preferred codons of highly expressed genes are recognized by the most abundant tRNA molecules in cells [29, 30]. According to Gouy and Gautier, a P2 value less than 0.5 indicates bias in favour of translational efficiency [22], but in the current study, we found that the average P2 value was 0.15, indicating that selection for translational efficiency had little influence on the codon bias in the studied genomes.

The average MRI value was 0.47 in our current study, suggesting a strong influence of mutational pressure on CUB. Consequently, the coding sequences of hepadnavirus genomes probably over-respond to mutational pressure [73]. The results of P2 and MRI analysis were further supported by the neutrality plot analysis, *i.e.*, mutation pressure was found to play a predominant role in hepadnavirus genomes.

Analysis of MRI and P2 indices in Nipah virus also indicated an important role of mutation pressure and translational selection [10]. Deka et al. reported lower degree of translational efficiency in the M1 and M2 matrix protein genes of influenza A virus [15].Viruses usually exploit the host cell's transcription and translational machinery to replicate, and thus the infected host might affect viral evolution [68]. The latent-stage genes in Epstein-Barr virus have been reported to deoptimize codon usage to lessen competition with the host's cell translation machinery [34].

Conclusion

In the present study, we investigated the nucleotide composition and biases in codon usage pattern of genes across the genomes of members of the family *Hepadnaviridae*. The overall codon usage bias across the genomes was low, indicating high variability in synonymous codon usage in viral genes. Almost equal usage of T, A and C was found, which differed from that of G, and AT richness was found. Three overrepresented codons (TCT, AGA, GGA) and seven underrepresented codons (TCG, AGC, AGT, CCG, CGA, ACG, GCG) were identified. Both mutational pressure and natural selection appear to have shaped the codon usage pattern of genes in hepadnavirus genomes during evolution.

Acknowledgements The authors are grateful to Assam University, Silchar, Assam India, for providing necessary facilities to carry out the research work.

Compliance with ethical standards

The study is based on DNA sequence analysis accessed from publicly available database. Ethical clearance is therefore not required.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Bajunaid HA (2013) Genetic variability of Hepatitis B virus, University of Nottingham
2. Beasley RP (1988) Hepatitis B virus. The major etiology of hepatocellular carcinoma. Cancer 61(10):1942–1956
3. Bennetzen JL, Hall BD (1982) Codon selection in yeast. J Biol Chem 257(6):3026–3031
4. Bernardi G, Olofsson B et al (1985) The mosaic genome of warm-blooded vertebrates. Science 228(4702):953–958
5. Bibb M, Findlay P et al (1984) The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. Gene 30(1–3):157–166
6. Blitvich B, Firth A (2015) Insect-specific flaviviruses: a systematic review of their discovery, host range, mode of transmission, superinfection exclusion potential and genomic organization. Viruses 7(4):1927–1959
7. Bouquet J, Cherel P et al (2012) Genetic characterization and codon usage bias of full-length Hepatitis E virus sequences shed new lights on genotypic distribution, host restriction and genome evolution. Infection, Genetics and Evolution 12(8):1842–1853
8. Butt AM, Nasrullah I et al (2016) Evolution of codon usage in Zika virus genomes is host and vector specific. Emerg Microbes Infect 5(1):1–14
9. Butt AM, Nasrullah I et al (2014) "Genome-wide analysis of codon usage and influencing factors in chikungunya viruses. PloS One 9(3):e90905
10. Chakraborty S, Deb B et al (2019) Analysis of codon usage patterns and influencing factors in Nipah virus. Virus Res 263:129–138
11. Choudhury MN, Chakraborty S (2015) Codon usage pattern in human SPANX genes. Bioinformation 11(10):454
12. Cote PJ, Toshkov I et al (2000) Temporal pathogenesis of experimental neonatal woodchuck hepatitis virus infection: increased initial viral load and decreased severity of acute hepatitis during the development of chronic viral infection. Hepatology 32(4):807–817
13. Deb B, Uddin A et al (2018) Analysis of codon usage pattern of mitochondrial protein-coding genes in different hookworms. Mol Biochem Parasitol 219:24–32
14. Deka H, Chakraborty S (2016) Insights into the usage of nucleobase triplets and codon context pattern in five influenza A virus subtypes. J Microbiol Biotechnol 26(11):1972–1982
15. Deka H, Nath D et al (2019) DNA compositional dynamics and codon usage patterns of M1 and M2 matrix protein genes in influenza A virus. Infect Genet Evol 67:7–16
16. Dittmar KA, Goodenbour JM et al (2006) Tissue-specific differences in human transfer RNA expression. PLoS Genet 2(12):e221
17. Eyre-Walker A (1996) Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? Mol Biol Evol 13(6):864–872
18. Franzo G, Segales J et al (2018) The analysis of genome composition and codon bias reveals distinctive patterns between avian and mammalian circoviruses which suggest a potential recombinant origin for *Porcine circovirus* 3. PloS One 13(6):e0199950
19. Fu M (2010) Codon usage bias in herpesvirus. Arch Virol 155(3):391–396
20. Gatherer D, McEwan NR (1997) Small regions of preferential codon usage and their effect on overall codon bias-The case of the plp gene. IUBMB Life 43(1):107–114
21. Goldhirsch A, Wood WC et al (2011) Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen international expert consensus on the primary therapy of early breast cancer 2011. Ann Oncol 22(8):1736–1747
22. Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10(22):7055–7074
23. Greco N, Hayes MH et al (2014) Snow goose hepatitis B virus (SGHBV) envelope and capsid proteins independently contribute to the ability of SGHBV to package capsids containing single-stranded DNA in virions. J Virol 88(18):10705–10713
24. Gun L, Haixian P et al (2018) Codon usage characteristics of PB2 gene in influenza A H7N9 virus from different host species. Infect Genet Evol 65:430–435
25. Gupta S, Ghosh T (2001) Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. Gene 273(1):63–70
26. Gustafsson C, Govindarajan S et al (2004) Codon bias and heterologous protein expression. Trends Biotechnol 22(7):346–353
27. Hargrove JL, Schmidt FH (1989) The role of mRNA and protein stability in gene expression. FASEB J 3(12):2360–2370
28. Hassan H, Mohamed M et al (2010) Effect of using organic acids to substitute antibiotic growth promoters on performance and intestinal microflora of broilers. Asian Austr J Anim Sci 23(10):1348–1353
29. Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 146(1):1–21
30. Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 151(3):389–409
31. Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2(1):13–34
32. Jenkins GM, Holmes EC (2003) The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res 92(1):1–7
33. Jiang Y, Deng F et al (2008) An extensive analysis on the global codon usage pattern of baculoviruses. Arch Virol 153(12):2273
34. Karlin S, Blaisdell BE et al (1990) Contrasts in codon usage of latent versus productive genes of Epstein-Barr virus: data and hypotheses. J Virol 64(9):4264–4273
35. Karlin S, Mrázek J (1996) What drives codon choices in human genes? J Mol Biol 262(4):459–472
36. Kaufmann WK, Paules RS (1996) DNA damage and cell cycle checkpoints. FASEB J 10(2):238–247
37. Kober KM, Pogson GH (2013) Genome-wide patterns of codon bias are shaped by natural selection in the purple sea urchin, *Strongylocentrotus purpuratus*. G3: Genes|Genomes|Genetics 3(7):1069–1083
38. Kumar N, Bera BC et al (2016) Revelation of influencing factors in overall codon usage bias of equine influenza viruses. PloS One 11(4):e0154376
39. Liu Q (2006) Analysis of codon usage pattern in the radioresistant bacterium *Deinococcus radiodurans*. Biosystems 85(2):99–106
40. Liu X-S, Zhang Y-G et al (2012) Patterns and influencing factor of synonymous codon usage in porcine circovirus. Virol J 9(1):68
41. Lobry J (1996) Origin of replication of *Mycoplasma genitalium*. Science 272:745–746

42. Ma M-R, Ha X-Q et al (2011) The characteristics of the synonymous codon usage in hepatitis B virus and the effects of host on the virus in codon usage pattern. Virol J 8(1):544
43. Marion PL, Knight SS et al (1983) Ground squirrel hepatitis virus infection. Hepatology 3(4):519–527
44. Mazumder GA, Uddin A et al (2018) Codon usage pattern of complex III gene of respiratory chain among platyhelminths. Infect Genet Evol 57:128–137
45. Mazumder GA, Uddin A et al (2018) Preference of A/T ending codons in mitochondrial ATP6 gene under phylum Platyhelminthes: codon usage of ATP6 gene in Platyhelminthes. Mol Biochem Parasitol 225:15–26
46. Moritz C, Dowling T et al (1987) Evolution of animal mitochondrial DNA: relevance for population biology and systematics. Annu Rev Ecol Syst 18(1):269–292
47. Mueller S, Papamichail D et al (2006) "Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. J Virol 80(19):9687–9696
48. Nakamura Y, Gojobori T et al (1997) Codon usage tabulated from the international DNA sequence databases. Nucleic Acids Res 25(1):244–245
49. Palidwor GA, Perkins TJ et al (2010) A general model of codon bias due to GC mutational bias. PloS One 5(10):e13431
50. Pavlov IP, Anrep GV (2003) Conditioned reflexes. Courier Corporation, Mineola
51. Peck KM, Lauring AS (2018) Complexities of viral mutation rates. J Virol 92(14):e01031-01017
52. Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nature Rev Genetics 12(1):32
53. Plotkin JB, Robins H et al (2004) Tissue-specific codon usage and the expression of human genes. Proc Natl Acad Sci 101(34):12588–12591
54. Presnyak V, Alhusaini N et al (2015) Codon optimality is a major determinant of mRNA stability. Cell 160(6):1111–1124
55. Rajewska M, Wegrzyn K et al (2012) AT-rich region and repeated sequences–the essential elements of replication origins of bacterial replicons. FEMS Microbiol Rev 36(2):408–434
56. Reis MD, Savva R et al (2004) Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 32(17):5036–5044
57. Ringnér M, Krogh M (2005) Folding free energies of 5′-UTRs impact post-transcriptional regulation on a genomic scale in yeast. PLoS Comput Biol 1(7):e72
58. Samuel A, Knowles N et al (1999) Genetic analysis of type O viruses responsible for epidemics of foot-and-mouth disease in North Africa. Epidemiol Infect 122(3):529–538
59. Sanjuán R, Nebot MR et al (2010) Viral mutation rates. J Virol 84(19):9733–9748
60. Sau K, Gupta S et al (2006) Factors influencing synonymous codon and amino acid usage biases in Mimivirus. Biosystems 85(2):107–113
61. Seeger C, Mason WS (2000) Hepatitis B virus biology. Microbiol Mol Biol Rev 64(1):51–68
62. Shackelton LA, Parrish CR et al (2006) Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. J Mol Evol 62(5):551–563
63. Sharp PM, Li W-H (1986) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucleic Acids Res 14(19):7737–7749
64. Sharp PM, Li W-H (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24(1–2):28–38
65. Sharp PM, Li W-H (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15(3):1281–1295
66. Sharp PM, Matassi G (1994) Codon usage and genome evolution. Curr Opin Genet Dev 4(6):851–860
67. Sharp PM, Stenico M et al (1993) Codon usage: mutational bias, translational selection, or both? Biochem Soc Trans 21(4):835
68. Su M-W, Lin H-M et al (2009) Categorizing host-dependent RNA viruses by principal component analysis of their codon usage preferences. J Comput Biol 16(11):1539–1547
69. Sueoka N (1961) Compositional correlation between deoxyribonucleic acid and protein. Cold Spring Harbor symposia on quantitative biology. Cold Spring Harbor Laboratory Press, New York
70. Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci 85(8):2653–2657
71. Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40(3):318–325
72. Sun D, Zhu L et al (2018) Recent progress in potential anti-hepatitis B virus agents: structural and pharmacological perspectives. Eur J Med Chem 147:205–217
73. Sur S, Sen A et al (2007) Mutational drift prevails over translational efficiency in Frankia nif operons. Indian J Biotechnol 6(3):321–328
74. Tamura K, Nei M et al (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci 101(30):11030–11035
75. Tao P, Dai L et al (2009) Analysis of synonymous codon usage in classical swine fever virus. Virus Genes 38(1):104–112
76. Uddin A, Chakraborty S (2016) Codon usage trend in mitochondrial CYB gene. Gene 586(1):105–114
77. Uddin A, Chakraborty S (2018) Codon usage pattern of genes involved in central nervous system. Mol Neurobiol 2018:1–12
78. van Hemert FJ, Berkhout B et al (2007) Host-related nucleotide composition and codon usage as driving forces in the recent evolution of the Astroviridae. Virology 361(2):447–454
79. Wan X-F, Xu D et al (2004) Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. BMC Evol Biol 4(1):19
80. Wei L, He J et al (2014) Analysis of codon usage bias of mitochondrial genome in *Bombyx mori* and its relation to evolution. BMC Evol Biol 14(1):262
81. Wong EH, Smith DK et al (2010) Codon usage bias and the evolution of influenza A viruses. Codon usage biases of influenza virus. BMC Evol Biol 10(1):253
82. Woo PC, Wong BH et al (2007) Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses. Virology 369(2):431–442
83. World Health Organization (2017) Hepatitis B factsheet. World Health Organization, Geneva
84. Wright F (1990) The 'effective number of codons' used in a gene. Gene 87(1):23–29
85. Xu C, Guo H et al (2010) "Interferons accelerate decay of replication-competent nucleocapsids of hepatitis B virus. J Virol 84(18):9332–9340
86. Zanetti AR, Van Damme P et al (2008) The global impact of vaccination against hepatitis B: a historical overview. Vaccine 26(49):6266–6273
87. Zhang J, Wang M et al (2011) Analysis of codon usage and nucleotide composition bias in polioviruses. Virol J 8(1):146
88. Zhang Z, Dai W et al (2013) Synonymous codon usage in TTSuV2: analysis and comparison with TTSuV1. PloS One 8(11):e81469

89. Zhang Z, Dai W, Wang Y, Lu C, Fan H (2013) Analysis of synonymous codon usage patterns in torque teno sus virus 1 (TTSuV1). Arch Virol 158:145–154
90. Zhao S, Zhang Q et al (2007) The factors shaping synonymous codon usage in the genome of *Burkholderia mallei*. J Genet Genom 34(4):362–372
91. Zhao S, Zhang Q et al (2008) Analysis of synonymous codon usage in 11 Human Bocavirus isolates. Biosystems 92(3):207–214
92. Zhong J, Li Y et al (2007) Mutation pressure shapes codon usage in the GC-Rich genome of foot-and-mouth disease virus. Virus Genes 35(3):767–776
93. Zhou H, Wang H et al (2005) "Heterogeneity in codon usages of sobemovirus genes. Arch Virol 150(8):1591–1605
94. Zhou J-H, Gao Z-L et al (2013) The analysis of codon bias of foot-and-mouth disease virus and the adaptation of this virus to the hosts. Infect Genet Evol 14:105–110
95. Zhou J, Liu WJ et al (1999) Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. J Virol 73(6):4972–4982
96. Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. Evol Genes Proteins 97:97–166