

Article

Insights into Synonymous Codon Usage Bias in Hepatitis C Virus and Its Adaptation to Hosts

Rekha Khandia ^{1,*} , Azmat Ali Khan ² , Noushad Karuvantevida ³ , Pankaj Gurjar ⁴,
Igor Vladimirovich Rzhepakovsky ⁵  and Isabel Legaz ^{6,*} 

¹ Department of Biochemistry and Genetics, Barkatullah University, Bhopal 462026, India

² Pharmaceutical Biotechnology Laboratory, Department of Pharmaceutical Chemistry, College of Pharmacy, King Saud University, Riyadh 11451, Saudi Arabia

³ College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai P.O. Box 505055, United Arab Emirates

⁴ Department of Science and Engineering, Novel Global Community Educational Foundation, Hebersham, NSW 2770, Australia

⁵ Medical and Biological Faculty, North Caucasus Federal University, 355017 Stavropol, Russia

⁶ Department of Legal and Forensic Medicine, Biomedical Research Institute (IMIB), Regional Campus of International Excellence “Campus Mare Nostrum”, Faculty of Medicine, University of Murcia, 30120 Murcia, Spain

* Correspondence: rekha.khandia@bubhopal.ac.in (R.K.); isalegaz@um.es (I.L.)

Abstract: Hepatitis C virus (HCV) is enveloped RNA virus, encoding for a polyprotein that is processed by cellular proteases. The virus is responsible for liver cirrhosis, allograft rejection, and human hepatocellular carcinoma. Based on studies including compositional analysis, odds ratio analysis, parity analysis, skew analysis, relative synonymous codon usage, codon bias, and protein properties, it was evident that codon usage bias in HCV is dependent upon the nucleotide composition. Codon context analysis revealed CTC-CTG as a preferred codon pair. While CGA and CGT codons were rare, none of the codons were rare in HCV-like viruses envisaged in the present study. Many of the preferred codon pairs were valine amino acid-initiated, which possibly infers viral infectivity; hence the role of selection forces appears to act on the HCV genome, which was further validated by neutrality analysis where selection accounted for 87.28%, while mutation accounted for 12.72% force shaping codon usage. Furthermore, codon usage was correlated with the length of the genome. HCV viruses prefer valine-initiated codon pairs, while HCV-like viruses prefer alanine-initiated codon pairs. The HCV host range is very narrow and is confined to only humans and chimpanzees. Based on indices including codon usage correlation analysis, similarity index, and relative codon deoptimization index, it is evident in the study that the chimpanzee is the primary host of the virus. The present study helped elucidate the preferred host for HCV. The information presented in the study paved the way for generating an attenuated vaccine candidate through viral recoding, with finely tuned nucleotide composition and a perfect balance of preferred and rare codons.

Keywords: hepatitis C virus (HCV); codon usage; liver cirrhosis; hepatocellular carcinoma; similarity index; relative codon deoptimization index; allograft rejection; liver transplantation



Citation: Khandia, R.; Khan, A.A.; Karuvantevida, N.; Gurjar, P.; Rzhepakovsky, I.V.; Legaz, I. Insights into Synonymous Codon Usage Bias in Hepatitis C Virus and Its Adaptation to Hosts. *Pathogens* **2023**, *12*, 325. <https://doi.org/10.3390/pathogens12020325>

Academic Editor: David J Allen

Received: 1 December 2022

Revised: 25 January 2023

Accepted: 2 February 2023

Published: 15 February 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The Hepatitis C virus (HCV) is an enveloped, single-stranded RNA virus of 9.6 kb genome size flanked by 5' and 3' untranslated regions. A single polyprotein is transcribed of over 3000 amino acids cleaved into structural and nonstructural proteins [1]. HCV is a major cause of liver cirrhosis and hepatocellular carcinoma, with increased mortality and dismal survival worldwide [2,3]. Liver cancer may be prevented using new curative hepatitis C antivirals [4]. Liver transplantation in HCV-positive recipients displays higher mortality rates, and HCV infection leads to allograft rejection [5]. Understanding the

virus-host interactions can help prevent and control HCV infection. The HCV displays a limited host range where it robustly infects only two hosts, humans and chimpanzees [6].

Degeneracy in the genetic code allows the encoding 20 amino acids from 61 codons. Except for TAA, TAG, and TGA, which encode for stop codon, and Met and Trp, which are encoded by only a single codon, all other amino acids are encoded by multiple codons, known as synonymous codons. However, synonymous codons are not randomly used, and specific codons are preferred over others, called codon bias. Codon bias can be tissue-specific [7], organ-specific [8], species-specific [9], and environment-dependent [10].

Since the viruses are intracellular pathogens, they rely on the host machinery to replicate and exhibit various levels of selection while infecting different hosts. Co-evolution and adaptation of the viruses to the hosts are the widely studied parameters that are analyzed using synonymous codon usage bias [11]. Codon bias is the result of non-random mutational patterns, selectional forces, and genome composition. It is related to the gene expression level [12], gene length [13], selective transcription [14], presence of rare codons [15], preferred codons [16], preferred codon pairs [17], protein properties [18], mRNA structure [19] roles of translation efficacy [20] and accuracy [21]; accordingly, virus fitness in any host [22] may be explained. Unfit viruses are attenuated in their infectivity and may serve as a vaccine candidate.

In the present study, we envisaged 54 genomes of the HCV virus (complete polyprotein) and various parameters like compositional parameters, nucleotide disproportion, dinucleotide frequency, codon usage, codon bias, relative synonymous codon usage, presence of rare codons, preferred codon pairs, the effect of major evolutionary forces, and gene expression were studied. The relative codon deoptimization index (RCDI) and similarity index analysis was carried out for its hosts, chimpanzee and human, to determine which host the virus is more adapted to. The analysis helps in understanding molecular signatures and the extent of mutational and selectional forces associated with HCV, and also provide information that can be useful in designing a vaccine candidate against it.

2. Materials and Methods

2.1. Sequence Collection

The complete coding sequences for the hepatitis C virus (HCV) were retrieved from National Centre for Biotechnology Information (<http://www.ncbi.nlm.nih.gov>, Accessed on 10 November 2022). Based on the criteria for selecting the sequence, the polyprotein sequence must be complete, devoid of ambiguous nucleotides, divisible by three, and start with the start codon and terminate with the stop codon; a total of 64 sequences were qualified. Out of these 64 sequences, 54 belonged to 1a, 1b, 2a, 2b, 2c, 2k, (non-recombinants) while 3a, 5a, and 6a genotypes were present as recombinant with 2a genotype in ten recombinant isolates. The information on accession numbers, genotypes, and resistance status towards antivirals are given in the supplementary file S1. To rationally compare the data from HCV with data from other HCV-like viruses, we selected 03 other HCV-like viruses Bovine hepatitis virus (BovHepV), Equine hepatitis virus (EqHV) and rodent hepatitis virus (RHV). For BovHepV, EqHV, and RHV, 27, 23, and 41 sequences were used to compare with HCV. These HCV-like sequences were also retrieved from the NCBI database, and the selection criteria were kept the same as for HCV.

2.2. Base Composition

The overall base composition (A%, T%, G%, and C%) and the base composition at the third codon position (A3%, T3%, G3%, and C3%) were analyzed. Average %GC content with GC content at all three codon positions (GC1%, GC2%, and GC3%) was also determined.

2.3. Dinucleotide Odds Ratio

There may be 16 dinucleotides derived from 4 bases, influencing codon bias and amino acid composition [23]. The odds ratio is the likelihood of observing a dinucleotide and

is calculated using the formula online available link <https://www.bioinformatics.nl/cgi-bin/emboss/compseq>, accessed on 15 November 2022). Here, the values below 0.78 and above 1.23 are considered underrepresentation and overrepresentation of dinucleotide, respectively [24].

2.4. Nucleotide Skew

A disproportionate usage of nucleotide is termed nucleotide skew, and it originates from the asymmetric replication in replication fork, mutational and selectional biases [25]. Nucleotide skews, namely, AT skew, GC skew, purine skew, pyrimidine skew, amino skew, and keto skew, were calculated using the formula $\text{Skew} = (A + B) / (A - B)$ where A and B are former and later nucleotides [26]. A positive skew value indicates the abundance of the first nucleotide over the second one, and vice versa [27].

2.5. Codon Usage

Codon usage analysis for HCV and its hosts, humans (*Homo sapiens*) and chimpanzees (*Pan troglodytes*), and HCV like viruses and their respective hosts *Bos taurus* (bovine), *Equus caballus* (equine), and *Peromyscus maniculatus* (rodent), for BovHepV, EqHV, and RHV was done using the Kazusa database (<https://www.kazusa.or.jp/codon/>, accessed on 10 November 2022). The codon usage by each organism is given in the supplementary file S2. The codon usage here describes as the usage of codon per thousand codons. The codon usage values (per thousand) obtained from the Kazusa database was based on 93487 and 857 CDSs for human and chimpanzee genomes, respectively. HCV, BovHepV, EqHV and RHVHV codon usage per thousand was calculated using CAIcal software [28].

2.6. Relative Synonymous Codon Usage (RSCU)

RSCU is a statistical tool to determine the codon usage bias of a single gene or entire genome. It is observed to the expected frequency of codon out of many synonymous codons available for a single amino acid in a given gene, set of genes, or genome [29]. RSCU values above 1.6 suggest overrepresentation, while values below 0.6 show underrepresentation [30,31]. RSCU values were obtained by the CAIcal server (<https://www.kazusa.or.jp/codon/>, accessed on 10 November 2022) developed by Puigbo P et al. (2008) [28].

2.7. Neutrality Plot Analysis

A neutrality plot is a method to determine the influence of two major evolutionary forces, selection and mutation, on the gene. The neutrality plot is constructed by regressing %GC12 at the Y axis and %GC3 at the X axis, and it accounts for the mutation-selection equilibrium during codon bias. Each of the genes is represented as a dot in the plot. A regression coefficient value less than 0.5 suggests a more significant role of natural selection, while greater than 0.5 suggests a greater impact of mutational pressure [32].

2.8. The Effective Number of Codons (ENc)

The effective number of codons (ENc) is one of the common measures to explain the usage bias of synonymous codons [33]. It ranges between 20 to 61. When only one codon is used out of many available synonymous codons, a value of 20 is achieved and indicates the highest bias. On the other hand, when all the synonymous codons are used equally, a value of 61 is obtained, which is suggestive of no bias. Generally, values above 35 are considered low bias [13]. The ENc values were calculated using the COUSIN tool (<https://cousin.ird.fr/>, Accessed on 15 November 2022) [34].

2.9. Codon Adaptation Index (CAI)

The CAI is the similarity measure between the synonymous codon usage of a gene and that of a reference set. It may be used as a predictive of the protein expression level of genes [35]. Higher CAI values indicate higher adaptability and a higher expression level,

and indicate the codons with higher RSCU values [36]. A web-based server CAIcal (<http://genomes.urv.cat/CAIcal/>, accessed on 11 November 2022) [28] was used to calculate CAI using the reference set of *Homo sapiens* and *Pan troglodytes* to determine the expression level of HCV in these two host organisms. The CAI value ranges between 0 and 1. For BovHepV, EqHV and RHV the reference set of *Bos taurus*, *Equus caballus* and *Peromyscus maniculatus* was used.

2.10. Parity Rule 2 (PR2) Bias Plot

A PR2 bias plot is suggestive of the comparative magnitude of mutation and selection forces acting together on the composition of a gene or genome. If the nucleobases are distributed proportionately across the plot, it indicates the influence of mutational force, while disproportionately distributed data points indicate the role of selection and mutation [32]. Here GC and AT bias at the third codon position is evaluated by plotting the GC bias [$G3 / (G3 + C3)$] on the X axis and the AT bias [$A3 / (A3 + T3)$] on the Y axis. If A and T and G and C bases are equal, it will result in a value of 0.5, and there is no bias between mutation and selection [37].

2.11. Codon Pair Context and Preferred Codons

Codon context is a preference of codon pairs in a given sequence. It is often linked to the accuracy of the translation [38] and speed [39]. The 5' codon context, rare codon and genome comparison analysis was performed using the ANACONDA[®] 2 software (<https://docs.anaconda.com/anaconda/install/hashe/wins-2-64/>, accessed on 10 November 2022) [38].

2.12. Protein Properties

General Average Hydropathicity (GRAVY) and Aromaticity (AROMA) are the frequency of hydrophobic and aromatic amino acids in a protein, and influence shape codon usage bias [40]. A positive GRAVY value indicates the dominance of hydrophobic amino acids, while a negative value suggests hydrophilic amino acids [41]. The correlation of these protein properties with GC%, GC3%, Nc, and CAI indicates the effect of selection force on codon usage bias [42].

2.13. Similarity Index Analysis

Similarity index analysis identifies the effect of host codon usage patterns on shaping codon usage of the pathogen. The similarity index was calculated by the formula given by Zhou and colleagues [43]. The values near Zero suggest a high similarity in codon usage between the host and pathogen, and values near 1 indicate a significant divergence [44].

2.14. The Relative Codon Deoptimization Index (RCDI) Analysis

The relative codon deoptimization index (RCDI) was calculated for the polyproteins of 64 HCV strains for chimpanzees and human hosts [45]. RCDI compares the similarities between the given gene set and the reference set and provides an idea about the rate of viral gene translation in the host. Values closer to 1 show higher translation rates and a more host-adapted codon usage pattern [30], while values higher than 1 show deoptimization in the codon usage pattern of the pathogen with that of a host. RCDI was calculated with *Pan troglodytes* and *Homo sapiens* as reference sequences using the RCDI/eRCDI server (<http://genomes.urv.cat/CAIcal/RCDI/>, accessed on 14 November 2022) [46].

2.15. Principal Component Analysis

The principal component analysis (PCA) is a multivariate analysis to analyze the major trends present between the variables. The PCA analysis of RSCU values of 64 coding sequences of HCV was done using Origin18 statistical software (<https://www.originlab.com/index.aspx?go=Support&pid=3301>, accessed on 10 November 2022). RSCU values of

each codon were distributed into 59 vectors, each corresponding to one codon (excluding Met, Trp, and stop codons). Here, RSCU values were converted into uncorrelated variables.

2.16. Phylogenetic Tree Construction

Based on the RSCU values, a phylogenetic tree was constructed to evaluate the relatedness among the HCV sequences using Ward's cluster analysis method. Past4 program (<https://www.nhm.uio.no/english/research/resources/past/>, Paleontological Statistics software version 4.03, accessed on 10 November 2022) was used for clustering, and the figure was generated in Mega10 software [47].

3. Results

3.1. Compositional Analysis

Compositional analysis revealed that the HCV genome is GC rich. The %A and %T composition was $20.18 \pm 0.37\%$ and $21.39 \pm 0.27\%$, respectively, while %C and %G were $30.12 \pm 0.84\%$ and $28.29 \pm 0.27\%$, respectively. Overall %GC composition was $58.42 \pm 1.05\%$ and %AT was $41.57 \pm 1.05\%$. A similar trend of the richness of C and G nucleotides was also observed at the third codon position, where %A3 and %T3 had $14.02 \pm 1.06\%$ and $18.21 \pm 1.9\%$ compositions, respectively. In contrast, for %C3 and %G3, it is $38.23 \pm 2.0\%$ and $29.52 \pm 0.85\%$ respectively. Comparative analysis of overall %GC composition and %GC composition at all the three codon positions reveals that among the three codon positions, %GC3 content was highest with an average value of $67.76 \pm 2.7\%$ while %GC2 was least with an average value of $50.57 \pm 0.31\%$.

Similar to HCV, in HCV-like viruses also, the sequences are GC-rich. Overall GC content is 51.92%, 50.31%, and 53.30% for BovHepV, EqHV, and RHV, respectively. A comparison of overall GC content at different codon positions was done for the HCV and envisaged HCV-like viruses. It revealed that overall, %GC is higher in HCV, RHVHV and EqHV compared to their respective hosts, except BovHepV. Furthermore, the percent GC2 is higher than the host's %GC2 in all viruses except for EqHV, which is the same for the virus and host (Figure 1).

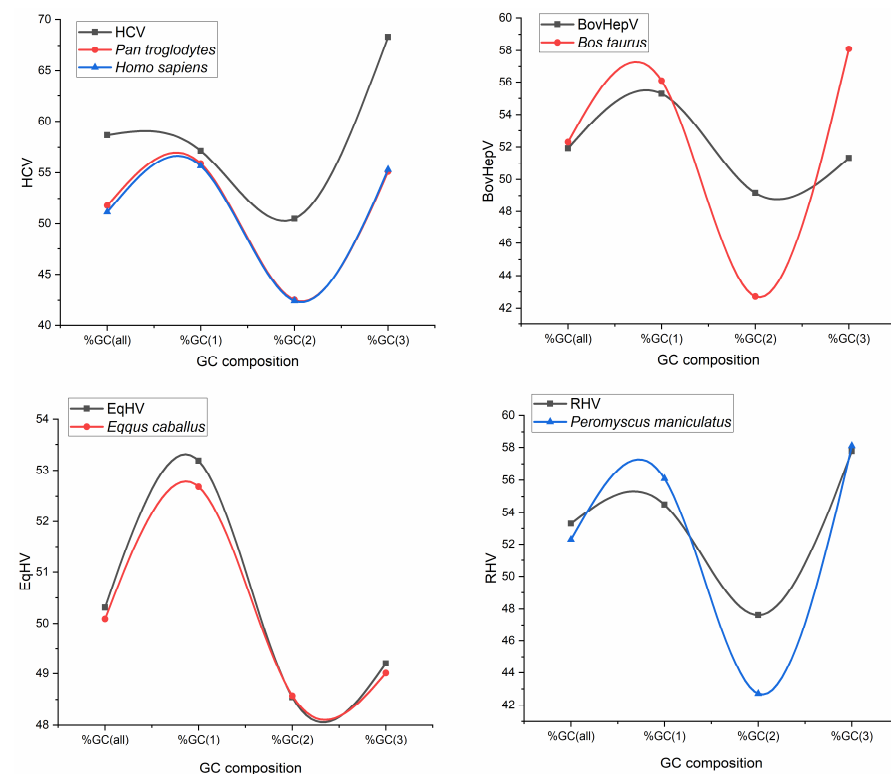


Figure 1. GC composition analysis for HCV and HCV like viruses with their respective hosts.

3.2. Odds Ratio Analysis Revealed Overrepresentation of GpG and CpC while Underrepresentation of TpA, ApA, TpT, and ApT

Dinucleotide odds ratio analysis indicated that four dinucleotides, TpA, ApA, TpT, and ApT were underrepresented (Odds ratio < 0.78) while two dinucleotides, GpG and CpC, were overrepresented (odds ratio > 1.23). CpG dinucleotide was randomly used (odds ratio 0.98) in HCV genomes. The odds ratio was compared with other HCV-like viruses' odds ratio, and the underrepresentation of TpA, ApA, TpT, and ApT was found to be a peculiar feature of HCV (Figure 2).

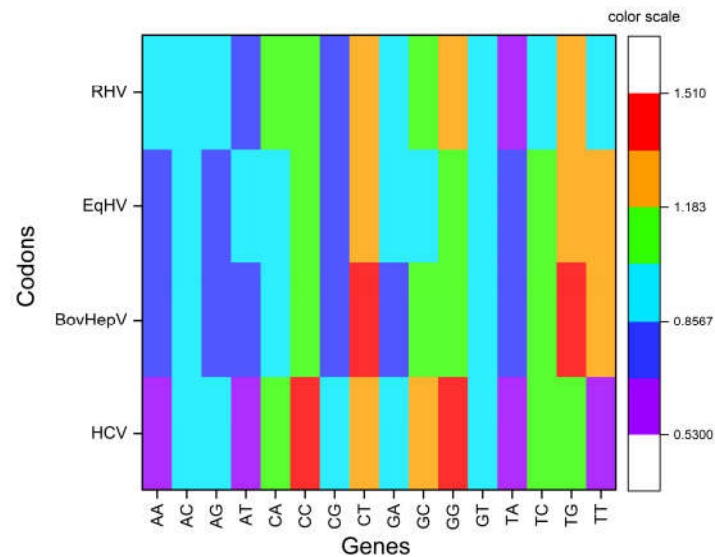


Figure 2. Dinucleotide odds ratio for HCV and HCV like viruses.

3.3. Selection Force Is Dominant Force in the Shaping Codon Usage

A neutrality plot between %GC3 and %GC12 is used to determine the level at which mutation and selection forces influence codon usage in any sequence. If there is a correlation between %GC12 and %GC3, it is suggestive of the role of mutation force in influencing codon bias at all three codon positions [48]. In the present study, we found a positive correlation between %GC12 and %GC3 ($r = 0.791$, $p < 0.0001$) in HCV. Relative neutrality was 12.72%, while selective constraint was 87.28%, suggested the prominent role of selection force in modeling codon bias in HCV polyprotein sequences. At the same time correlation between %GC12 and %GC3 indicates the role of mutational force. Therefore, it can be inferred that both the selection and mutational forces act on HCV polyprotein sequences. The R2 value of 0.5957 indicated that 59.57% variation in %GC12 is attributed to the %GC3 composition.

In BovHepV, we found a native correlation, however insignificant, between %GC12 and %GC3 ($r = -0.296$, $p = 0.13$). Relative neutrality was 7.21%, while selective constraint was 92.79%. R2 value 0.0881% suggested that 8.81% variation in %GC12 is due to %GC3.

For EqHV, no correlation between %GC12 and %GC3 was found. Relative neutrality was 6.69%, while selective constraint was 93.31%, suggestive of selection as the dominant force determining the sequence of EqHV. The R2 value 0.0603 suggested 6.03% variations in %GC12 is attributed to the %GC3 composition.

For RHV, a positive correlation ($r = 0.771$, $p < 0.0001$) between %GC3 and %GC12 has been observed. The relative neutrality of 16.93% and selective constraints of 83.07% suggested a dominant role in selection. The R2 value 0.5952 suggested 59.52% variation in %GC12 is attributed to the %GC3 composition. Overall analysis suggested a dominant role of selective forces acting on HCV and HCV-like viruses influencing codon usage [49] (Figure 3).

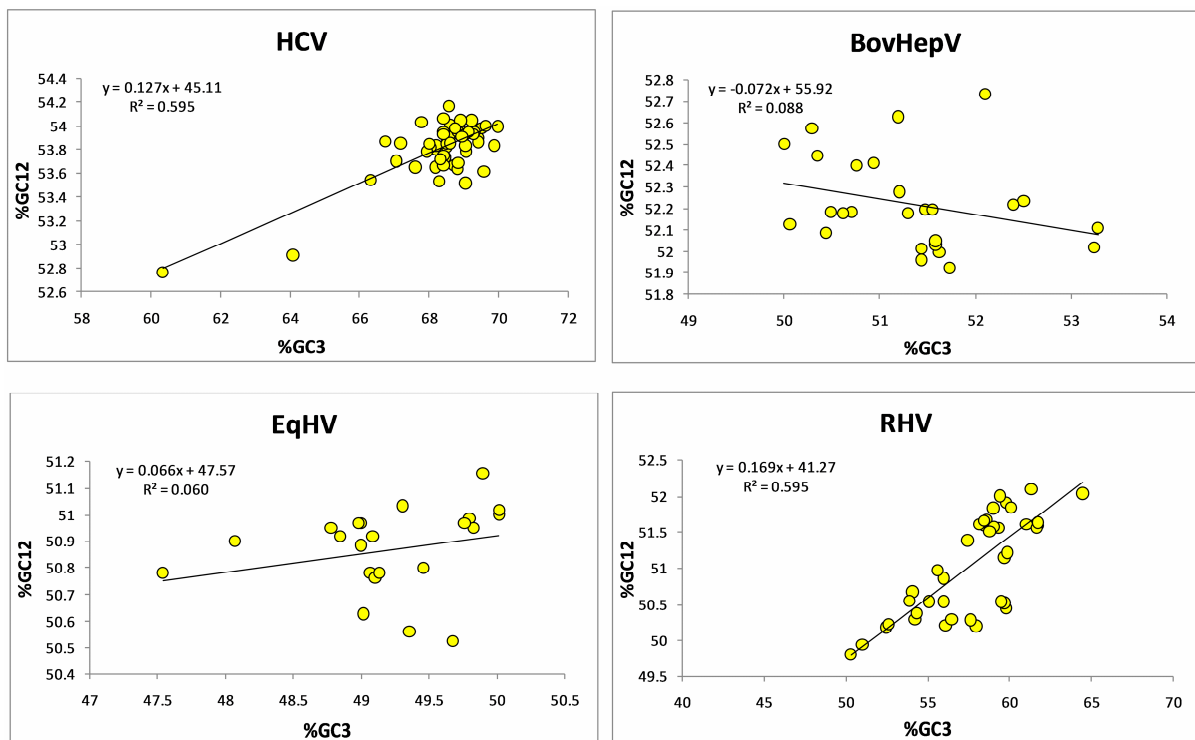


Figure 3. Neutrality plot analysis for HCV and HCV-like viruses suggests the dominant role of selection force in shaping codon usage.

3.4. Parity Plot Analysis Indicated Dominance of Pyrimidine over Purines

The parity plot is constructed by plotting GC bias ($G3 / G3 + C3$) at X-axis and AT bias ($A3 / A3 + T3$) at Y axis [50]. For HCV, the value of GC bias was 0.434 ± 0.007 , and AT bias was 0.433 ± 0.013 . GC bias was 0.443 ± 0.007 , 0.464 ± 0.007 , 0.475 ± 0.021 while AT bias was 0.297 ± 0.009 , 0.318 ± 0.008 , 0.421 ± 0.039 for BovHepV, EqHV and RHV, respectively. It concludes that nucleotides C and T are preferred over G and A at the third codon position and suggest selection force [51] in HCV and HCV-like viruses (Figure 4).

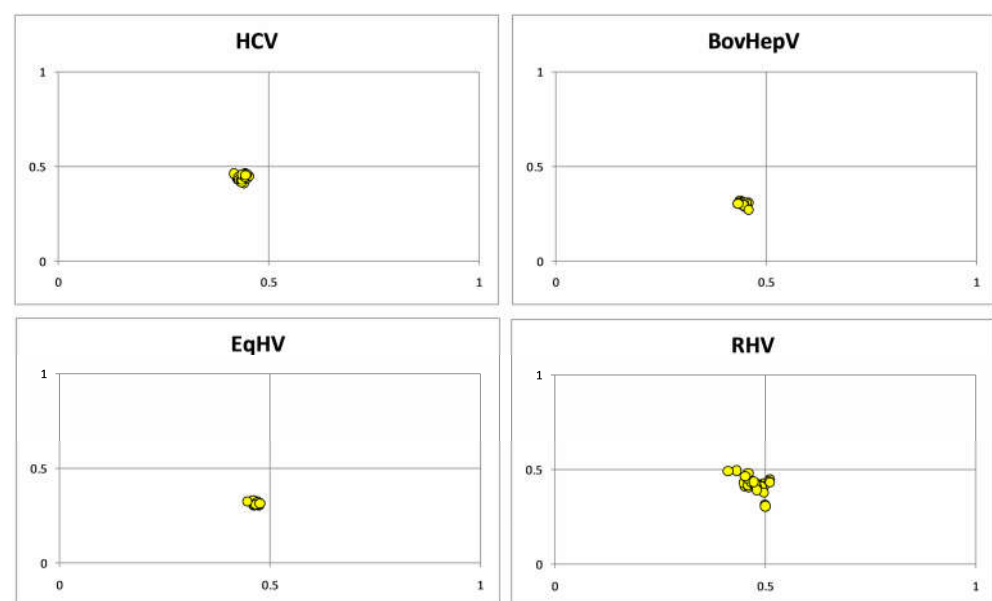


Figure 4. Parity plot for HCV and HCV like-viruses revealed preference of C and T over G and A.

3.5. Result of Skew

Skew is a disproportionate usage of the nucleotide. We performed the correlation analysis for skews with the gene expression and codon bias. Skew values were calculated for the 64 polypeptide sequences of HCV and were correlated with CAI and ENc. AT skew and GC skews are generally used to determine the compositional distributions of nucleotides [52]. In the present study, the mean AT and GC skew values were -0.028 and -0.033 , respectively and it suggested that T and C are preferred over A and G. The average values of other skews were also negative (-0.170 , -0.175 , -0.202 and -0.143 for purine, pyrimidine, amino and keto skews, respectively). An index related to gene expression CAI is shown to be influenced by nucleotide skew [53]. Except for AT skew, all the skews were significantly negatively correlated with gene expression ($p < 0.001$). ENc, a non-directional codon usage bias measure, was positively correlated with all the skews except for AT skew ($p < 0.001$). A significant correlation between codon bias and skew is found in Nipah virus also [54].

3.6. RSCU Analysis Revealed the Overrepresentation of G/C Ending Codons

RSCU analysis revealed that eight codons ending with G/C were overrepresented in the genes. TCC, ATC, CTC, AGG, CTG, ACC, CCC, and GGC were eight overexpressed codons (RSCU > 1.6) [55] in 100%, 96.29%, 96.29%, 92.59%, 85.18%, 70.37%, 57.4%, and 61.11% of HCV polypeptide sequences. On the other hand, thirteen A/T ending codons, TTA, CTA, ATT, GTT, GTA, AGT, AAT, GAT, GAA, TGT, CGT, CGA, and GGA, showed underrepresentation in 100%, 87.03%, 81.48%, 64.81%, 100%, 98.14%, 64.81%, 81.48%, 87.03%, 57.41%, 51.85%, 92.59%, and 72.22% of HCV genomes.

In HCV, out of 18 synonymous codons, all the 18 preferred codons were GC ending. In BovHepV, 07 GC ending and 11 AT-ending, in EqHV, 05 GC ending codons while 13 AT-ending, and in RHVHepV, 17 GC ending and 01 AT-ending codon were present. The results suggest different codon usage patterns in HCV and HCV-like viruses.

3.7. Protein Properties Are Dependent on the Composition and Codon Bias

We did correlation analysis for the nucleotide composition at all three codon positions. GRAVY and AROMA are the protein features that showed a positive, negative or no correlation. GRAVY showed a significant positive correlation with GC composition at all codon positions, while AROMA had no correlation (Table 1). Overall, the analysis revealed that composition significantly influences the hydropathicity and aromaticity of polypeptides encoded by the HCV genome [56].

Table 1. Pearson correlation analysis of GRAVY and AROMA with compositional components of HCV virus.

	%A	%A1	%A2	%A3	%T	%T1	%T2	%T3	%C	%C1
GRAVY (r value)	−0.680	−0.118	−0.487	−0.701	−0.384	−0.553	0.201	−0.303	0.617	0.635
p value	***	NS	***	***	**	***	NS	*	***	***
AROMA	−0.134	0.310	−0.270	−0.226	−0.175	−0.132	−0.073	−0.153	0.282	−0.001
p value	NS	*	*	NS	NS	NS	NS	NS	*	NS
	%C2	%C3	%G	%G1	%G2	%G3	%GC(all)	%GC(1)	%GC(2)	%GC(3)
GRAVY	−0.655	0.650	0.469	−0.110	0.641	0.176	0.602	0.510	0.277	0.592
p value	***	***	***	NS	***	NS	***	***	***	***
AROMA	−0.441	0.418	−0.041	−0.345	0.464	−0.282	0.169	−0.187	0.245	0.221
p value	***	**	NS	*	***	*	NS	NS	NS	NS

Level of significance *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; NS Non significant.

Hydropathicity and aromaticity are the protein properties that are thought to influence mRNA transcripts, thereby affecting codon bias [57]. Pearson correlation analysis revealed that codon bias index ENc was negatively correlated with the GRAVY and AROMA ($r = -0.581, p < 0.001$; $r = -0.281, p < 0.05$, respectively). The results suggest that the codon bias decreases with a decrease in the frequency of hydrophobic and aromatic amino acids. PCs are representative of dinucleotide and codon composition. Thus, we investigated the correlation of protein properties with PCs and found that GRAVY positively correlated with both PC1 and PC2 ($r = 0.633, p < 0.001$ and $r = 0.322, p < 0.05$). In contrast, AROMA showed a positive correlation with PC1 ($r = 0.553, p < 0.001$) and a negative correlation with PC2 ($r = -0.275, p < 0.05$). The correlation of GRAVY and AROMA with ENc and PCs showed that protein properties influence codon usage.

3.8. ENc Indicated Low Bias

ENc is the index to point to codon usage bias. Fifty-four HCV polyproteins ranged from 56.46 to 50.68, averaging 51.92 ± 0.88 . ENc values above 50 in the present study show weaker bias [58]. ENc values ranged between 56.07 to 54.43 (average 55.37 ± 0.36), 55.83 to 54.49 (average 55.25 ± 0.41), and 58.41 to 53.36 (average 56.48 ± 1.08) for BovHepV, EqHV and RHVs respectively. Overall, the analysis indicated a low bias in HCV and HCV-like sequences. Furthermore, the bias showed a negative correlation with the gene lengths ($r = -0.280, p < 0.05$) and is suggestive that the bias increases with the increase in the length of the HCV polyprotein sequence [59]. Contrary to HCV, RHV showed a positive correlation between ENc and length; thus, bias decreases with the increase in length in the case of RHV. In the case of both the BobHepV and EqHV, no correlation was observed between codon bias and length ($r = 0.351, p = 0.11$ for EqHV and $r = 0.041, p = 0.83$ for BovHepV).

3.9. Codon Context Analysis Revealed an Abundance of CTC-CTG Codon Pair and Rarity of CGA and TTA

Variations in the codon contexts in the top 20 codon pairs were determined for HCV. The trend of the favored codon pair is depicted in Figure 2. The HCV genome (polyprotein segment) had both the preferred (depicted with green color) and rejected (depicted with red color) codon pairs. CTC-CTG codon pair was most abundant, and six codon pairs were initiated with Val amino acid (GTC/GTG) in native HCVs. At the same time, the GCC-CTC pair was most abundant in recombinant HCVs, with no such preference in the codon pair initiation. GCT-GCT, GGC-GCT, and again the GCT-GCT codon pairs were most abundant in BovHepV, EqHV, and RHVHepV, with a maximum of six, five, and four alanine-initiated codons in BovHepV, EqHV, and RHV, respectively. The top 20 codon pairs are given in Table 2. Here it is interesting to see that HCV top 20 most occurring codon pairs encompassed six valine-initiated codons; contrarily, all other envisaged three HCV-like viruses have a preference for alanine-initiated codons.

A codon with a frequency of less than 0.5% was considered rare [60]. CGA (Arg) and TTA (Leu) codons were rarely used in the native HCV genome. In recombinant HCVs, GCA (Arg) and CGT (Arg) were rare. On the other hand, in HCV-like viruses like BovHepV, eqHV and RHV, none of the codons have a frequency of less than 0.5% (except for stop codons).

Codon pair bias was investigated for the HCV virus, and all three kinds of contexts (positive, negative or no bias) were present (Figure 5).

Table 2. Top 20 preferred codon pairs in HCV polyprotein sequences in naive HCV, recombinant HCV and HCV-like viruses.

S. No.	HCVs (Recombinants)		HCVs (Non-Recombinants)		Bovine Hepacivirus BovHepV		Equine Hepacivirus (EqHV)		Rodent Hepacivirus (RHV)	
	Codon Pair	Frequency	Codon Pair	Frequency	Codon Pair	Frequency	Codon Pair	Frequency	Codon Pair	Frequency
1	GCC-CTC	88	CTC-CTG	369	GCT-GCT	258	GGC-GCT	160	GCT-GCT	229
2	CCC-CCC	81	GTG-GCC	320	CTT-GAG	153	TGG-GCT	137	GAG-GAG	216
3	GTC-ATC	75	AAC-ACC	295	GTC-ACC	149	GCT-TGG	132	AAG-AAG	190
4	GGC-GCC	71	GCC-ATC	293	CTT-GCT	142	CTT-GCT	130	GCT-GGC	174
5	TAT-GAC	69	GTC-ACC	270	GTC-ACT	135	GCT-TCT	129	GAG-GAC	164
6	GAG-GTC	63	CTC-ACT	270	GTT-GCT	134	GAC-ACT	121	GCT-GCC	163
7	GCG-GCC	62	GTG-TGC	267	GGT-GCT	134	ACT-GGC	112	GCT-GAG	163
8	GTG-GAC	61	CTG-GAC	267	GGC-ACT	133	GAT-GTT	107	GAG-GCT	156
9	GAC-GCC	61	GCT-GCC	263	GCT-GTG	133	TTT-GAC	101	TTT-GAC	154
10	ACC-ATC	60	AAC-TGG	258	GCT-GTT	131	GCT-TTT	101	GTG-GTG	148
11	ACC-ACC	59	GTG-CGC	257	CCT-TAC	127	GCT-GTT	101	TTG-GCT	146
12	TGC-TCC	58	ATC-ACC	255	ACT-GCT	127	TCT-GTT	100	ACT-GGC	144
13	TGC-GGC	58	GTG-GGG	253	TGG-GCT	126	TGT-GGC	98	ACC-AAG	143
14	GAG-GAG	56	ATC-ATG	249	GAT-GTT	123	GCT-GTC	98	TAC-ACC	141
15	TAC-TCC	53	TGG-GCG	248	GGT-GCC	121	ACT-GTC	97	GAC-ACC	134
16	GAC-ATC	53	TAC-GTG	246	GCT-ACT	119	CCT-TAT	95	TGT-GAC	131
17	GGG-TAC	51	GCC-ACC	243	CCT-GCT	116	GGG-GAT	94	GTG-GCC	130
18	TCC-TGG	50	ATC-AAC	236	GCT-GGC	114	ATG-GGC	92	AAG-GAG	130
19	TAC-ATC	50	GTC-ATC	234	GTT-TGG	111	GAG-GAA	91	AAG-AAA	130
20	GGT-GTG	50	CTG-CTG	234	GCT-GTC	111	TAT-GAC	90	GGG-AAG	129

We then investigated the difference in codon context in different virus groups. To display the difference, we constructed a differential display map (DDM) between the sequences from two groups. No context or low context difference is depicted by residual values less than 20, while high context difference is depicted by residual values more than 100 [61] (Figure 6A). Comparison of HCV sequence with BovHepB (Figure 6B), EqHv (Figure 6C) and RHV (Figure 6D) are depicted below. From the results, it is clear that codon context or codon pair differences are prominent.

3.10. PCA Analysis

PCA is often used for reduction in dimensions. RSCU values of 59 synonymous codons were taken as 59 vectors. Most of the data points were clustered at three sites only, except three data points, and it is suggestive that most of the HCV genomes' codon usage mainly follows three kinds of trends. Only one data point was scattered (HCV 2k), indicating that its codon usage pattern differed from the rest of the HCV genomes. For native HCVs, primary and secondary axis contributed for 28.60%, and 19.84% variation, respectively. Some of the data points were scattered far from the axes, indicating low to moderate bias in codon usage in these genomes [13]. The graph also indicated more bias in a few HCV genomes than others (Figure 7).

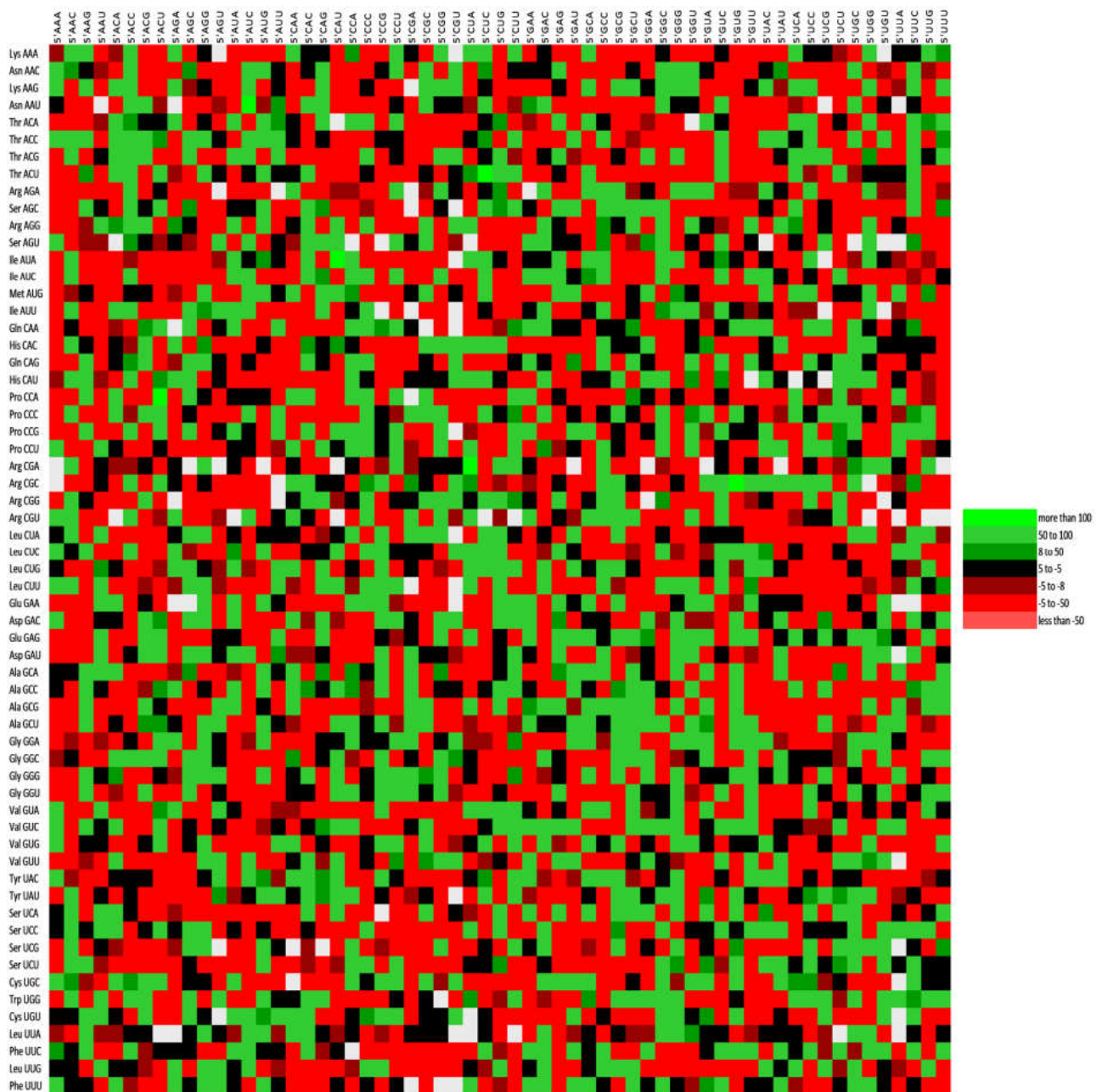


Figure 5. Codon context analysis for 54 polyprotein sequences of HCV. The green color shows a positive context, while red shows a negative context. Black color shows insignificant context, and grey indicates no context.

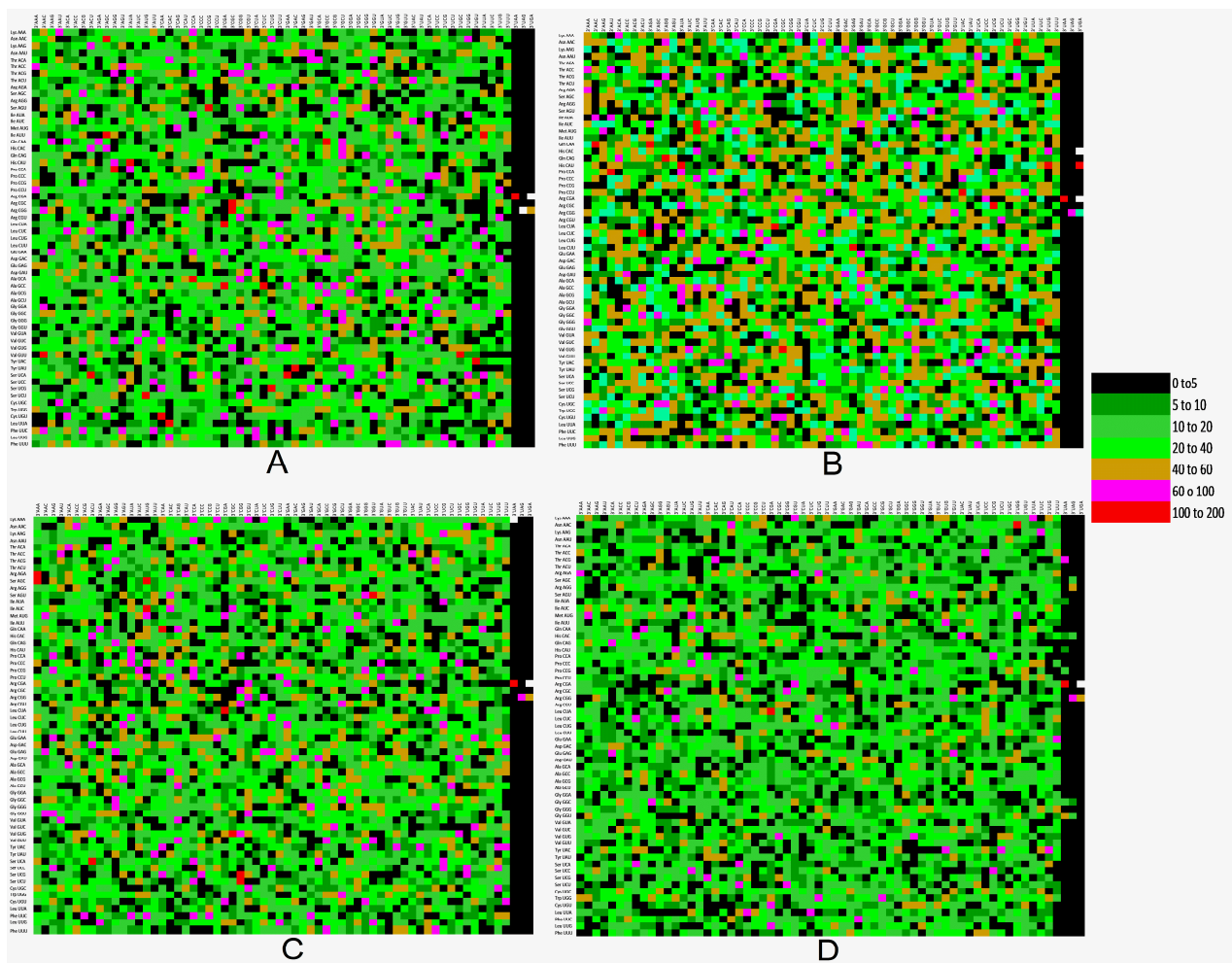


Figure 6. Differential display map for comparison of codon pair bias for genomes (A). HCV native and HCV recombinant (B). HCV and BovHepV (C). HCV and EqHV (D). HCV and RHV.

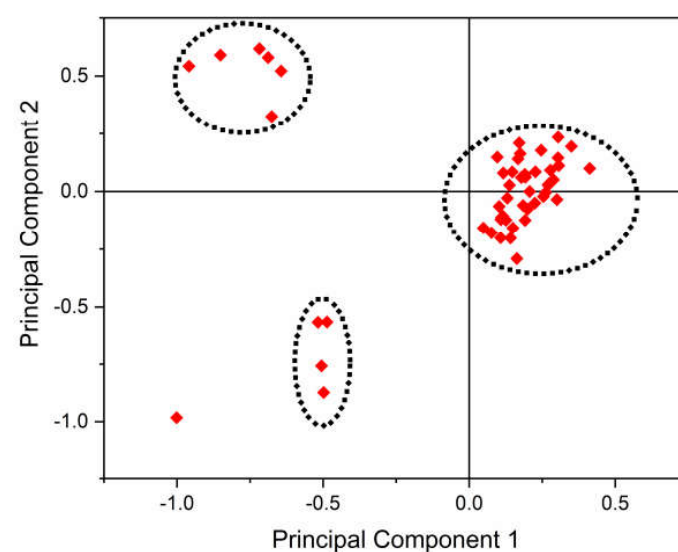


Figure 7. PCA analyses revealed low to moderate codon bias in HCV genomes. Red dots indicate each of the native HCVs. HCV 2k isolate took a different position on PCA.

3.11. Phylogeny Analysis

Phylogenetic analysis of 54 HCV genomes was carried out using Ward's hierarchical agglomerative clustering method with a 500-bootstrap value. The analysis revealed that genomes might be separated into two clusters. Fifteen polyprotein sequences formed one cluster, while the rest of the sequences made another cluster. The separation of clusters indicates that each cluster has different codon usage pattern (Figure 8).

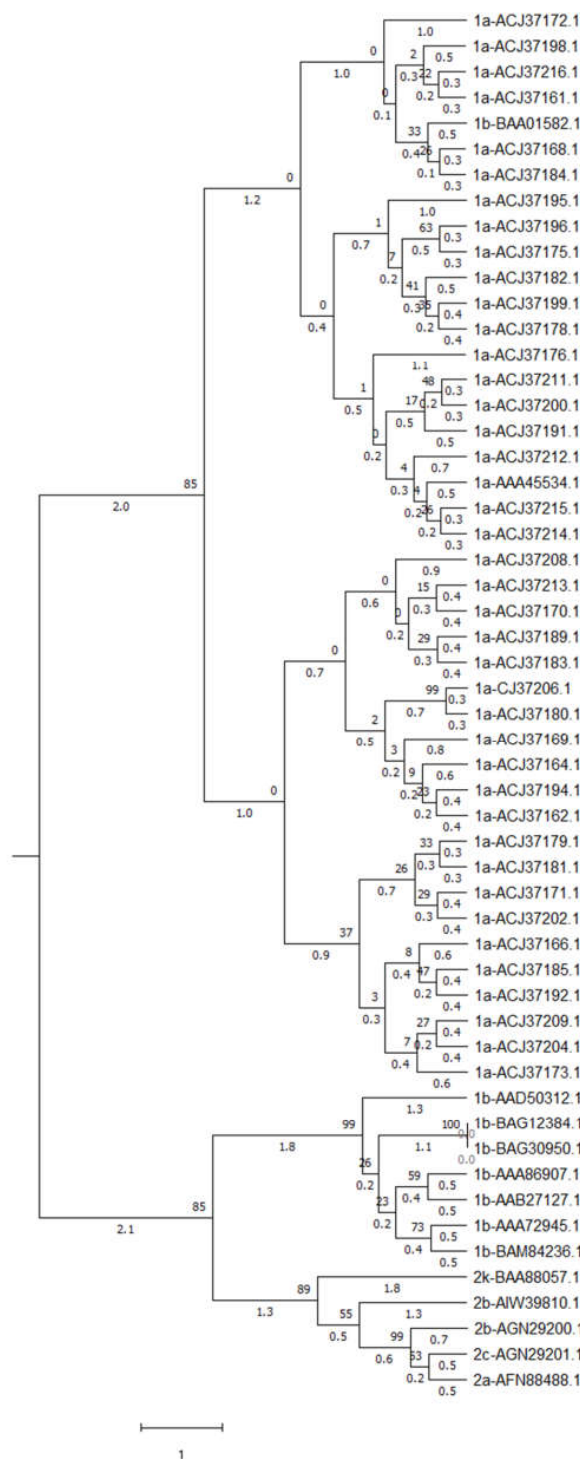


Figure 8. Phylogenetic analysis for 54 polyprotein sequences of HCV reveals that the sequences may be separated into two clusters based on Ward's hierarchical agglomerative clustering.

3.12. Adaptability of HCV Genome for its Hosts Human and Chimpanzee

3.12.1. The Codon Adaptation Index Reveals More Adaptability of HCV for Humans Compared to Chimpanzees

CAI values are used to determine the expression level of a pathogen in a host, or in other words, it indicates the adaptation of a virus into its hosts [36]. The genes or pathogens with higher CAI values in a host are considered more adapted than those with lower CAI values. The obtained CAI values in human and chimpanzee is given in supplementary file S3. For chimpanzees, it ranged between 0.665 to 0.721, while for humans, it ranged between 0.722 to 0.758. The average CAI values were 0.714 ± 0.006 and 0.751 ± 0.004 for chimpanzees and humans. The results indicate that the virus is more adapted to humans than chimpanzees. CAI values were 0.696 ± 0.003 , 0.639 ± 0.003 , and 0.402 ± 0.017 for BovHepV, EqHV, and RHV, respectively, for their respective hosts. Results suggested that the HCV virus is well adapted to humans and chimpanzees compared to other animal HCV-like viruses in their respective hosts. Adaptation is the least for RHV in its host, *Peromyscus maniculatus*. The CAI results are in concordance with the data of RCDI.

3.12.2. Codon Usage Pattern of HCV Is More Similar with That of Chimpanzee Codon Usage Pattern

We compared the codon usage per thousand for HCV, humans, and chimpanzees; the results are given in Figure 5. In addition, correlation analysis of codon usage between the hosts human and chimpanzee, and pathogen HCV was done to determine to which host the codon usage pattern of HCV matches more. The linear Pearson correlation analysis revealed that codon usage of HCV is near to the codon usage of chimpanzees ($r = 0.712$, $p < 0.001$) than humans ($r = 0.59$, $p < 0.001$) and is suggestive of chimpanzees as the primary host. Similarly, a statistically significant Pearson correlation was present between codon usage of *Bos taurus* and BovHepV ($r = 0.605$, $p < 0.001$), *Equus caballus* and EqHV ($r = 0.948$, $p < 0.001$), and *Peromyscus maniculatus* and RHV ($r = 0.648$, $p < 0.001$). The higher correlation of codon usage between the host and the pathogen is suggestive of adaptation.

3.12.3. HCV Displays the Highest Codon Usage Deoptimization for Human

The higher similarity of codon usage between host and pathogen is depicted by RCDI values closer to 1 [45]. It can be used to improve protein expression in a heterologous expression system. A lower RCDI is suggestive of more adaptation, and at the same time a higher RCDI also suggests that few of the genes are expressed in latency phases, or maybe the virus is present with a low replication rate. Here the RCDI value was 1.08 ± 0.01 and 1.11 ± 0.01 for chimpanzees and humans, respectively, suggesting that HCV codons are more deoptimized to humans. The RCDI value of 1.19 ± 0.01 , 1.19 ± 0.01 , and 1.81 ± 0.17 for BovHepV, EqHV and RHV showed that HCV-like viruses BovHepV, and EqHV are better adapted to their respective hosts, *Bos taurus* and *Equus caballus*, compared to RHV in one of its hosts *Peromyscus maniculatus*. Unfortunately, we could not compare the RCDI of RHV for its more common host *Lophuromys* (*L. dudu*, *L. machangui*, *L. stanleyi*, *L. laticeps*), since the genomic sequences were not available for them.

3.12.4. Similarity Index showed Pan Troglodytes Is Primary Host

The host with a lower similarity index will have more similar codon usage than the host with a higher similarity index [62]. The similarity index was 0.03 and 0.039 for *Pan troglodytes* and *Homo sapiens*, respectively. The results indicate that HCV codon usage similarity is more with *Pan troglodytes* than *Homo sapiens*. The similarity index was 0.066, 0.001, and 0.031 for BovHepV, EqHV and RHV in *Bos taurus*, *Equus caballus* and *Peromyscus maniculatus*, respectively. Results suggested that EqHV is most adapted to its host *Equus caballus*.

4. Discussion

HCV virus causes chronic and fatal hepatic liver problems leading to liver cirrhosis and hepatocellular carcinoma, with increased mortality and dismal survival worldwide [2,3]. In the present study, we performed codon usage analysis, codon pair, and other molecular patterns analysis that helped determine the host and the necessary information that might be useful in generating synthetic biology-based vaccine candidates. Since 2011, many HCV-like viruses have been identified in several hosts, including dogs [63], equine [64], bats [65], cattle [66], rodents [67], and monkeys [68] in a mammalian group. Full-length genome sequences were available: canine 01, equine 26, bat 06, bovine 27, monkey 01, and rodent 41. To add statistically correct controls, we compared HCV polyprotein sequences with qualified HCV-like animal viruses; those are BovHepV (27 sequences), EqHV (23 sequences) and RHV (41 sequences).

In an attempt to see the effects of skew variation, a short stretch of the HIV-I pol region was systematically manipulated by adding and removing A nucleotide. As a result, nucleotide A content was altered from 40.2% (wild type) to increase to 46.9% or reduced to 31.7% and 26.3%. Here AG skew was affected dramatically, and a reduced viral replication has been observed in the virus having maximum nucleotide A content [69]. Hence, it is depicted that skews decide viral fitness also. The skew may be manipulated by using optimized or deoptimized codons or codon pairs while making a vaccine candidate, depending on the type of vaccine candidate.

In each organism, the odds ratio is unique, and CpG dinucleotide bias has been observed in genomes of humans and mice [70] and may be used as a molecular signature. In HCV, the odds ratio analysis indicated the underrepresentation of TpA, ApA, TpT and ApT dinucleotide while GpG and CpC were overrepresented. CpG odds ratio is in normal ranges for large DNA viruses while small DNA viruses have low CpG [71,72]. All papillomaviruses and polyomaviruses are CpG depleted [73]. In case of RNA viruses, generally an underrepresentation of CpG is found, with the exception of rubella virus where CpG dinucleotide is present in normal ranges owing to the exceptionally high GC (up to 70%) content [74]. The dinucleotide odds ratio of a pathogen is also an indicator of host pathogen interactions. The members of Flaviviridae viruses infecting vertebrates exhibit depletion of both the CpG (possibly induced by the methylation-deamination process) and TpA, while those infecting non-vertebrates had only TpA depletion [75]. In the present study, HCV is a RNA virus and in the HCV genome the GC content is high. Thus, it is reflected on the dinucleotide content and the CpG dinucleotide, which is generally otherwise an underrepresented dinucleotide, and is present in normal ranges (between 0.78 to 1.3) here. The normal range CpG odds ratio is likely owing to the high GC content in the HCV genome. Based on nucleotide composition, the overrepresentation of GpG and CpC and the underrepresentation of TpA, ApA, TpT and ApT dinucleotides may be explained. Dinucleotide composition becomes symmetrical for complementary dinucleotide for all of the four underrepresented (TpA/ApT and ApA/TpT) and overrepresented (GpG/CpC) dinucleotides due to the double-stranded nature of DNA, which is here in case of HCV. Interestingly, HCV is an RNA virus, and thus we are unable to explain the reason behind this specific dinucleotide pattern. The high bias towards CpC (and GpG owing to complementary to the CpC dinucleotide) may also be explained as a fine-tuning process of protein expression [76]. Remarkably, underrepresentation of TpA/ApT and ApA/TpT and overrepresentation of GpG/CpC was a feature of HCV viruses which is absent in HCV like viruses.

Apart from codon bias, codon pair bias also exists, demonstrating the likelihood of two codons' presence together. Few codon pairs are favored [77,78] while few are avoided [79] than expected in a protein-coding region. In the context of viruses, codon optimization in gag and pol genes of HIV-1 did not improve the viral replication, but optimization results in virus attenuation [80]. However, contrary results were seen by Jordan-Paiz [81] in the HIV-1 envelope gene, where codon pair deoptimization doesn't necessarily generate attenuation while optimizing attenuated virus replication in MT-4 cells. Hence, the effect

of codon pair optimization or deoptimization is gene-specific and can't be generalized. Codon pair analysis revealed that for HCV, CTC-CTG, while for BovHepV, EqHV, and RHV, GCT-GCT, GGC-GCT and GCT-GCT codon pairs were most preferred. In the present study, it was found that out of the most preferred 20 codon pairs in HCV and HCV-like viruses, HCV had maximum valine amino acid initiated codon pairs. In contrast, other envisaged HCV-like viruses preferred alanine-initiated codon pairs. It might relate to the initiation of protein synthesis with alanine without the involvement of met-tRNA, as present in the *Drosophila C virus*, *Himetobi P virus* and *Rhopalosiphum padi virus* [82].

Since lower RCDI is suggestive of better adaptation to the host [36] and RCDI for *Pan troglodytes* is lower than *Homo sapiens*, it may be inferred that in chimpanzees, the HCV is better adapted than humans. Our results concord with the data obtained with other HCV-like viruses where low RCDI suggested a better adaptation of BovHepV and EqHV in their respective hosts, *Bos taurus* and *Equus caballus* owing to lower RCDI. On the other hand, RHV is less adapted in *Peromyscus maniculatus* compared to different species of *Lophuromys* (*L. dudui*, *L. machangui*, *L. stanleyi*, *L. laticeps*) [83].

CAI is a quantitative measure of the expression level of a gene in the host using a highly expressed gene set as a reference [28]. A comparison of CAI values, which are the indicators of gene expression in the host, revealed that HCV has higher expression in humans (0.75 ± 0.005) than in chimpanzees (0.71 ± 0.009). This observation contradicts the results we obtained based on RCDI, similarity index, and correlation analysis of codon usage between HCV and hosts. This deviation is possibly due to the reason that here for calculating CAI as a reference set, data from only highly expressed genes is taken as a reference which may not necessarily explain the whole codon usage. Thus, based on our observations, we conclude that *Pan troglodytes* may possibly be the primary host for HCV with high codon usage similarities (based on similarity index), less codon deoptimization (low RCDI), and a higher correlation of codon usage of HCV RSCU with *Pan troglodytes*.

The expression of viruses in the different hosts is different, as evidenced by different CAI and RCDI values of the Nipah virus in 10 different host species. One virus may be adapted differently in various host species. Based on the RCDI and CAI analysis, Nipah virus is best adapted in African green monkeys. Codon pattern analysis helps evaluate the clinical outcome of a pathogen infection where high virus adaptation will result in higher replication, increased infectivity, and clinically pathogenic outcomes of the virus [36]. Although viral fitness and virulence are often coupled, sometimes deviations are also observed owing to complex virus–host interactions. For example, an experiment conducted in the vesicular stomatitis virus (VSV) to evaluate the relationship between viral fitness and virulence revealed that overall, there was a positive correlation between the two was present. However, few outliers were also present, with the higher fitness and low virulence and low fitness with no effect in virulence [84].

Similarly, bats are shown to be highly adapted to the Nipah virus, yet avirulent to the host, with that a low calculated fitness, yet high virulence in a ferret model is exhibited. Although, such examples are rare and result from the complex host–pathogen interaction, and are presented as only a few outliers. Still, the codon usage pattern analysis helps determine the host of a virus [36]. Thus, based on our observations, it is clear that codon usage analysis can likely determine the hosts and clinical outcome of infection for a virus.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/pathogens12020325/s1>.

Author Contributions: Conceptualization, R.K., I.L., N.K. and P.G.; methodology, R.K., N.K., P.G.; software, A.A.K.; validation, I.V.R., A.A.K. and; formal analysis, R.K.; investigation, I.V.R., N.K.; resources, A.A.K.; data curation, R.K., N.K.; writing—original draft preparation, I.V.R.; writing—review and editing, R.K.; visualization, A.A.K., I.L., R.K.; supervision, R.K., I.L. and A.A.K.; project administration, R.K.; funding acquisition, A.A.K., I.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not involved.

Informed Consent Statement: Not required.

Data Availability Statement: Available upon request.

Acknowledgments: This work was funded by the Researchers Supporting Project Number (RSP2023R339) King Saud University, Riyadh, Saudi Arabia and project R-1007/2018, University of Murcia, Spain.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liang, Y.; Zhang, G.; Li, Q.; Han, L.; Hu, X.; Guo, Y.; Tao, W.; Zhao, X.; Guo, M.; Gan, T.; et al. TRIM26 Is a Critical Host Factor for HCV Replication and Contributes to Host Tropism. *Sci. Adv.* **2021**, *7*, eabd9732. [\[CrossRef\]](#) [\[PubMed\]](#)
2. El-Serag, H.B. Epidemiology of Viral Hepatitis and Hepatocellular Carcinoma. *Gastroenterology* **2012**, *142*, 1264–1273.e1. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Pinheiro, P.S.; Callahan, K.E.; Jones, P.D.; Morris, C.; Ransdell, J.M.; Kwon, D.; Brown, C.P.; Kobetz, E.N. Liver Cancer: A Leading Cause of Cancer Death in the United States and the Role of the 1945–1965 Birth Cohort by Ethnicity. *JHEP Rep.* **2019**, *1*, 162–169. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Soriano, V.; Tefferi, A. Prevention of Liver Cancer with New Curative Hepatitis C Antivirals: Real-World Challenges. *Cancer* **2018**, *124*, 1647–1649. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Forman, L.M.; Lewis, J.D.; Berlin, J.A.; Feldman, H.I.; Lucey, M.R. The Association between Hepatitis C Infection and Survival after Orthotopic Liver Transplantation. *Gastroenterology* **2002**, *122*, 889–896. [\[CrossRef\]](#)
6. Couto, L.B.; Kolykhalov, A.A. Animal Models for HCV Study. In *Hepatitis C Viruses: Genomes and Molecular Biology*; Tan, S.-L., Ed.; Horizon Bioscience: Norfolk, UK, 2006; ISBN 978-1-904933-20-5.
7. Kames, J.; Alexaki, A.; Holcomb, D.D.; Santana-Quintero, L.V.; Athey, J.C.; Hamasaki-Katagiri, N.; Katneni, U.; Golikov, A.; Ibla, J.C.; Bar, H.; et al. TissueCoCoPUTs: Novel Human Tissue-Specific Codon and Codon-Pair Usage Tables Based on Differential Tissue Gene Expression. *J. Mol. Biol.* **2020**, *432*, 3369–3378. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Camiolo, S.; Farina, L.; Porceddu, A. The Relation of Codon Bias to Tissue-Specific Gene Expression in Arabidopsis Thaliana. *Genetics* **2012**, *192*, 641–649. [\[CrossRef\]](#)
9. Parvathy, S.T.; Udayasuriyan, V.; Bhadana, V. Codon Usage Bias. *Mol. Biol. Rep.* **2022**, *49*, 539–565. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Arella, D.; Dilucca, M.; Giansanti, A. Codon Usage Bias and Environmental Adaptation in Microbial Organisms. *Mol. Genet. Genom.* **2021**, *296*, 751–762. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Zhou, J.-H.; Gao, Z.-L.; Zhang, J.; Ding, Y.-Z.; Stipkovits, L.; Szathmari, S.; Pejsak, Z.; Liu, Y.-S. The Analysis of Codon Bias of Foot-and-Mouth Disease Virus and the Adaptation of This Virus to the Hosts. *Infect. Genet. Evol.* **2013**, *14*, 105–110. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Quax, T.E.F.; Claassens, N.J.; Söll, D.; van der Oost, J. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol. Cell* **2015**, *59*, 149–161. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Khandia, R.; Saeed, M.; Alharbi, A.M.; Ashraf, G.M.; Greig, N.H.; Kamal, M.A. Codon Usage Bias Correlates With Gene Length in Neurodegeneration Associated Genes. *Front. Neurosci.* **2022**, *16*, 895607. [\[CrossRef\]](#)
14. Chan, C.T.Y.; Deng, W.; Li, F.; DeMott, M.S.; Babu, I.R.; Begley, T.J.; Dedon, P.C. Highly Predictive Reprogramming of tRNA Modifications Is Linked to Selective Expression of Codon-Biased Genes. *Chem. Res. Toxicol.* **2015**, *28*, 978–988. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Hia, F.; Takeuchi, O. The Effects of Codon Bias and Optimality on mRNA and Protein Regulation. *Cell. Mol. Life Sci.* **2021**, *78*, 1909–1928. [\[CrossRef\]](#)
16. Sheikh, A.; Al-Taher, A.; Al-Nazawi, M.; Al-Mubarak, A.I.; Kandeel, M. Analysis of Preferred Codon Usage in the Coronavirus N Genes and Their Implications for Genome Evolution and Vaccine Design. *J. Virol. Methods* **2020**, *277*, 113806. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Sexton, N.R.; Ebel, G.D. Effects of Arbovirus Multi-Host Life Cycles on Dinucleotide and Codon Usage Patterns. *Viruses* **2019**, *11*, E643. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Engel, A.J.; Kithil, M.; Langhans, M.; Rauh, O.; Cartolano, M.; Van Etten, J.L.; Moroni, A.; Thiel, G. Codon Bias Can Determine Sorting of a Potassium Channel Protein. *Cells* **2021**, *10*, 1128. [\[CrossRef\]](#)
19. Mittal, P.; Brindle, J.; Stephen, J.; Plotkin, J.B.; Kudla, G. Codon Usage Influences Fitness through RNA Toxicity. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 8639–8644. [\[CrossRef\]](#)
20. Yang, Q.; Yu, C.-H.; Zhao, F.; Dang, Y.; Wu, C.; Xie, P.; Sachs, M.S.; Liu, Y. ERF1 Mediates Codon Usage Effects on mRNA Translation Efficiency through Premature Termination at Rare Codons. *Nucleic Acids Res.* **2019**, *47*, 9243–9258. [\[CrossRef\]](#)
21. Liu, Y.; Yang, Q.; Zhao, F. Synonymous but Not Silent: The Codon Usage Code for Gene Expression and Protein Folding. *Annu. Rev. Biochem.* **2021**, *90*, 375–401. [\[CrossRef\]](#)
22. Fu, Y.; Huang, Y.; Rao, J.; Zeng, F.; Yang, R.; Tan, H.; Liu, Z.; Du, W.; Liu, L. Host Adaptation of Codon Usage in SARS-CoV-2 from Mammals Indicates Potential Natural Selection and Viral Fitness. *Arch. Virol.* **2022**, *167*, 2677–2688. [\[CrossRef\]](#)

23. Franzo, G.; Tucciarone, C.M.; Legnardi, M.; Cecchinato, M. Effect of Genome Composition and Codon Bias on Infectious Bronchitis Virus Evolution and Adaptation to Target Tissues. *BMC Genom.* **2021**, *22*, 244. [[CrossRef](#)] [[PubMed](#)]
24. Karlin, S.; Mrázek, J. Compositional Differences within and between Eukaryotic Genomes. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 10227–10232. [[CrossRef](#)] [[PubMed](#)]
25. Hubert, B. *SkewDB: A Comprehensive Database of GC and 10 Other Skews for over 28,000 Chromosomes and Plasmids.* *bioRxiv* **2021**. [[CrossRef](#)]
26. Necşulea, A.; Lobry, J.R. A New Method for Assessing the Effect of Replication on DNA Base Composition Asymmetry. *Mol. Biol. Evol.* **2007**, *24*, 2169–2179. [[CrossRef](#)]
27. Uddin, A.; Chakraborty, S. Codon Usage Trend in Mitochondrial CYB Gene. *Gene* **2016**, *586*, 105–114. [[CrossRef](#)]
28. Puigbò, P.; Bravo, I.G.; Garcia-Vallve, S. CAIcal: A Combined Set of Tools to Assess Codon Usage Adaptation. *Biol. Direct* **2008**, *3*, 38. [[CrossRef](#)]
29. Yu, X.; Liu, J.; Li, H.; Liu, B.; Zhao, B.; Ning, Z. Comprehensive Analysis of Synonymous Codon Usage Patterns and Influencing Factors of Porcine Epidemic Diarrhea Virus. *Arch. Virol.* **2021**, *166*, 157–165. [[CrossRef](#)]
30. Butt, A.M.; Nasrullah, I.; Qamar, R.; Tong, Y. Evolution of Codon Usage in Zika Virus Genomes Is Host and Vector Specific. *Emerg. Microbes Infect.* **2016**, *5*, e107. [[CrossRef](#)]
31. Singh, R.K.; Pandey, S.P. Phylogenetic and Evolutionary Analysis of Plant ARGONAUTES. *Methods Mol. Biol.* **2017**, *1640*, 267–294. [[CrossRef](#)]
32. Deb, B.; Uddin, A.; Chakraborty, S. Codon Usage Pattern and Its Influencing Factors in Different Genomes of Hepadnaviruses. *Arch. Virol.* **2020**, *165*, 557–570. [[CrossRef](#)]
33. Wright, F. The “effective Number of Codons” Used in a Gene. *Gene* **1990**, *87*, 23–29. [[CrossRef](#)] [[PubMed](#)]
34. Bourret, J.; Alizon, S.; Bravo, I.G. COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences. *Genome Biol. Evol.* **2019**, *11*, 3523–3528. [[CrossRef](#)]
35. Henry, I.; Sharp, P.M. Predicting Gene Expression Level from Codon Usage Bias. *Mol. Biol. Evol.* **2007**, *24*, 10–12. [[CrossRef](#)] [[PubMed](#)]
36. Khandia, R.; Singhal, S.; Kumar, U.; Ansari, A.; Tiwari, R.; Dhama, K.; Das, J.; Munjal, A.; Singh, R.K. Analysis of Nipah Virus Codon Usage and Adaptation to Hosts. *Front. Microbiol.* **2019**, *10*, 886. [[CrossRef](#)]
37. Wang, Q.; Lyu, X.; Cheng, J.; Fu, Y.; Lin, Y.; Abdoulaye, A.H.; Jiang, D.; Xie, J. Codon Usage Provides Insights into the Adaptive Evolution of Mycoviruses in Their Associated Fungi Host. *Int. J. Mol. Sci.* **2022**, *23*, 7441. [[CrossRef](#)]
38. Moura, G.; Pinheiro, M.; Silva, R.; Miranda, I.; Afreixo, V.; Dias, G.; Freitas, A.; Oliveira, J.L.; Santos, M.A. Comparative Context Analysis of Codon Pairs on an ORFeome Scale. *Genome Biol.* **2005**, *6*, R28. [[CrossRef](#)] [[PubMed](#)]
39. Baeza, M.; Alcaíno, J.; Barahona, S.; Sepúlveda, D.; Cifuentes, V. Codon Usage and Codon Context Bias in Xanthophyllomyces Dendrorhous. *BMC Genom.* **2015**, *16*, 293. [[CrossRef](#)] [[PubMed](#)]
40. Chen, Y.; Li, X.; Chi, X.; Wang, S.; Ma, Y.; Chen, J. Comprehensive Analysis of the Codon Usage Patterns in the Envelope Glycoprotein E2 Gene of the Classical Swine Fever Virus. *PLoS One* **2017**, *12*, e0183646. [[CrossRef](#)] [[PubMed](#)]
41. Kyte, J.; Doolittle, R.F. A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.* **1982**, *157*, 105–132. [[CrossRef](#)]
42. Chandan, J.; Gupta, S.; Babu, V.; Singh, D.; Singh, R. Comprehensive Analysis of Codon Usage Pattern in Withania Somnifera and Its Associated Pathogens: Meloidogyne Incognita and Alternaria Alternata. *Genetica* **2022**, *150*, 129–144. [[CrossRef](#)] [[PubMed](#)]
43. Zhou, J.; Zhang, J.; Sun, D.; Ma, Q.; Chen, H.; Ma, L.; Ding, Y.; Liu, Y. The Distribution of Synonymous Codon Choice in the Translation Initiation Region of Dengue Virus. *PLoS One* **2013**, *8*, e77239. [[CrossRef](#)]
44. Zhou, J.-H.; Li, X.-R.; Lan, X.; Han, S.-Y.; Wang, Y.-N.; Hu, Y.; Pan, Q. The Genetic Divergences of Codon Usage Shed New Lights on Transmission of Hepatitis E Virus from Swine to Human. *Infect. Genet. Evol.* **2019**, *68*, 23–29. [[CrossRef](#)] [[PubMed](#)]
45. Mueller, S.; Papamichail, D.; Coleman, J.R.; Skiena, S.; Wimmer, E. Reduction of the Rate of Poliovirus Protein Synthesis through Large-Scale Codon Deoptimization Causes Attenuation of Viral Virulence by Lowering Specific Infectivity. *J. Virol.* **2006**, *80*, 9687–9696. [[CrossRef](#)] [[PubMed](#)]
46. Puigbò, P.; Aragonès, L.; Garcia-Vallvé, S. RCDI/ERCDI: A Web-Server to Estimate Codon Usage Deoptimization. *BMC Res. Notes* **2010**, *3*, 87. [[CrossRef](#)] [[PubMed](#)]
47. Murtagh, F.; Legendre, P. Ward’s Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward’s Criterion? *J. Classif.* **2014**, *31*, 274–295. [[CrossRef](#)]
48. Jenkins, G.M.; Holmes, E.C. The Extent of Codon Usage Bias in Human RNA Viruses and Its Evolutionary Origin. *Virus Res.* **2003**, *92*, 1–7. [[CrossRef](#)]
49. Munjal, A.; Khandia, R.; Shende, K.K.; Das, J. Mycobacterium Lepromatosis Genome Exhibits Unusually High CpG Dinucleotide Content and Selection Is Key Force in Shaping Codon Usage. *Infect. Genet. Evol.* **2020**, *84*, 104399. [[CrossRef](#)]
50. Khandia, R.; Sharma, A.; Alqahtani, T.; Alqahtani, A.M.; Asiri, Y.I.; Alqahtani, S.; Alharbi, A.M.; Kamal, M.A. Strong Selectional Forces Fine-Tune CpG Content in Genes Involved in Neurological Disorders as Revealed by Codon Usage Patterns. *Front. Neurosci.* **2022**, *16*, 596. [[CrossRef](#)]
51. Gao, Y.; Lu, Y.; Song, Y.; Jing, L. Analysis of Codon Usage Bias of WRKY Transcription Factors in Helianthus Annuus. *BMC Genom. Data* **2022**, *23*, 46. [[CrossRef](#)] [[PubMed](#)]

52. McLean, M.J.; Wolfe, K.H.; Devine, K.M. Base Composition Skews, Replication Orientation, and Gene Orientation in 12 Prokaryote Genomes. *J. Mol. Evol.* **1998**, *47*, 691–696. [\[CrossRef\]](#) [\[PubMed\]](#)
53. Alqahtani, T.; Khandia, R.; Puranik, N.; Alqahtani, A.M.; Chidambaram, K.; Kamal, M.A. Codon Usage Is Influenced by Compositional Constraints in Genes Associated with Dementia. *Front. Genet.* **2022**, *13*, 884348. [\[CrossRef\]](#) [\[PubMed\]](#)
54. Chakraborty, S.; Deb, B.; Barbhuiya, P.A.; Uddin, A. Analysis of Codon Usage Patterns and Influencing Factors in Nipah Virus. *Virus Res.* **2019**, *263*, 129–138. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Alqahtani, T.; Khandia, R.; Puranik, N.; Alqahtani, A.M.; Almikhlaifi, M.A.; Alqahtany, M.A. Leucine Encoding Codon TTG Shows an Inverse Relationship with GC Content in Genes Involved in Neurodegeneration with Iron Accumulation. *J. Integr. Neurosci.* **2021**, *20*, 905–918. [\[CrossRef\]](#)
56. Khandia, R.; Pandey, M.K.; Rzhepakovsky, I.V.; Khan, A.A.; Alexiou, A. Synonymous Codon Variant Analysis for Autophagic Genes Dysregulated in Neurodegeneration. *Mol. Neurobiol.* **2023**. [\[CrossRef\]](#)
57. Deb, B.; Uddin, A.; Chakraborty, S. Genome-Wide Analysis of Codon Usage Pattern in Herpesviruses and Its Relation to Evolution. *Virus Res.* **2021**, *292*, 198248. [\[CrossRef\]](#)
58. Liu, Y.; Huo, N.; Dong, L.; Wang, Y.; Zhang, S.; Young, H.A.; Feng, X.; Gu, Y.Q. Complete Chloroplast Genome Sequences of Mongolia Medicine Artemisia Frigida and Phylogenetic Relationships with Other Plants. *PLoS One* **2013**, *8*, e57533. [\[CrossRef\]](#)
59. Li, Y.; Khandia, R.; Papadakis, M.; Alexiou, A.; Simonov, A.N.; Khan, A.A. An Investigation of Codon Usage Pattern Analysis in Pancreatitis Associated Genes. *BMC Genom. Data* **2022**, *23*, 81. [\[CrossRef\]](#)
60. Khandia, R.; Pandey, M.K.; Khan, A.A.; Rzhepakovsky, I.V.; Gurjar, P.; Karobari, M.I. Codon Usage and Context Analysis of Genes Modulated during SARS-CoV-2 Infection and Dental Inflammation. *Vaccines* **2022**, *10*, 1874. [\[CrossRef\]](#)
61. Moura, G.; Pinheiro, M.; Arrais, J.; Gomes, A.C.; Carreto, L.; Freitas, A.; Oliveira, J.L.; Santos, M.A.S. Large Scale Comparative Codon-Pair Context Analysis Unveils General Rules That Fine-Tune Evolution of mRNA Primary Structure. *PLoS One* **2007**, *2*, e847. [\[CrossRef\]](#)
62. Silverj, A.; Rota-Stabelli, O. On the Correct Interpretation of Similarity Index in Codon Usage Studies: Comparison with Four Other Metrics and Implications for Zika and West Nile Virus. *Virus Res.* **2020**, *286*, 198097. [\[CrossRef\]](#)
63. Kapoor, A.; Simmonds, P.; Gerold, G.; Qaisar, N.; Jain, K.; Henriquez, J.A.; Firth, C.; Hirschberg, D.L.; Rice, C.M.; Shields, S.; et al. Characterization of a Canine Homolog of Hepatitis C Virus. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 11608–11613. [\[CrossRef\]](#)
64. Burbelo, P.D.; Dubovi, E.J.; Simmonds, P.; Medina, J.L.; Henriquez, J.A.; Mishra, N.; Wagner, J.; Tokarz, R.; Cullen, J.M.; Iadarola, M.J.; et al. Serology-Enabled Discovery of Genetically Diverse Hepaciviruses in a New Host. *J. Virol.* **2012**, *86*, 6171–6178. [\[CrossRef\]](#)
65. Quan, P.-L.; Firth, C.; Conte, J.M.; Williams, S.H.; Zambrana-Torrel, C.M.; Anthony, S.J.; Ellison, J.A.; Gilbert, A.T.; Kuzmin, I.V.; Niezgoda, M.; et al. Bats Are a Major Natural Reservoir for Hepaciviruses and Pegiviruses. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 8194–8199. [\[CrossRef\]](#)
66. Baechlein, C.; Fischer, N.; Grundhoff, A.; Alawi, M.; Indenbirken, D.; Postel, A.; Baron, A.L.; Offinger, J.; Becker, K.; Beineke, A.; et al. Identification of a Novel Hepacivirus in Domestic Cattle from Germany. *J. Virol.* **2015**, *89*, 7007–7015. [\[CrossRef\]](#) [\[PubMed\]](#)
67. Drexler, J.F.; Corman, V.M.; Müller, M.A.; Lukashev, A.N.; Gmyl, A.; Coutard, B.; Adam, A.; Ritz, D.; Leijten, L.M.; van Riel, D.; et al. Evidence for Novel Hepaciviruses in Rodents. *PLoS Pathog.* **2013**, *9*, e1003438. [\[CrossRef\]](#) [\[PubMed\]](#)
68. Lauck, M.; Sibley, S.D.; Lara, J.; Purdy, M.A.; Khudyakov, Y.; Hyeroba, D.; Tumukunde, A.; Weny, G.; Switzer, W.M.; Chapman, C.A.; et al. A Novel Hepacivirus with an Unusually Long and Intrinsically Disordered NS5A Protein in a Wild Old World Primate. *J. Virol.* **2013**, *87*, 8971–8981. [\[CrossRef\]](#) [\[PubMed\]](#)
69. Klaver, B.; van der Velden, Y.; van Hemert, F.; van der Kuyl, A.C.; Berkhout, B. HIV-1 Tolerates Changes in A-Count in a Small Segment of the Pol Gene. *Retrovirology* **2017**, *14*, 43. [\[CrossRef\]](#)
70. Ulveling, D.; Dinger, M.E.; Francastel, C.; Hubé, F. Identification of a Dinucleotide Signature That Discriminates Coding from Non-Coding Long RNAs. *Front. Genet.* **2014**, *5*, 316. [\[CrossRef\]](#)
71. Shackelton, L.A.; Parrish, C.R.; Holmes, E.C. Evolutionary Basis of Codon Usage and Nucleotide Composition Bias in Vertebrate DNA Viruses. *J. Mol. Evol.* **2006**, *62*, 551–563. [\[CrossRef\]](#)
72. Karlin, S.; Doerfler, W.; Cardon, L.R. Why Is CpG Suppressed in the Genomes of Virtually All Small Eukaryotic Viruses but Not in Those of Large Eukaryotic Viruses? *J. Virol.* **1994**, *68*, 2889–2897. [\[CrossRef\]](#) [\[PubMed\]](#)
73. Upadhyay, M.; Vivekanandan, P. Depletion of CpG Dinucleotides in Papillomaviruses and Polyomaviruses: A Role for Divergent Evolutionary Pressures. *PLoS One* **2015**, *10*, e0142368. [\[CrossRef\]](#) [\[PubMed\]](#)
74. Rima, B.K.; McFerran, N.V. Dinucleotide and Stop Codon Frequencies in Single-Stranded RNA Viruses. *J. Gen. Virol.* **1997**, *78 Pt 11*, 2859–2870. [\[CrossRef\]](#) [\[PubMed\]](#)
75. Lobo, F.P.; Mota, B.E.F.; Pena, S.D.J.; Azevedo, V.; Macedo, A.M.; Tauch, A.; Machado, C.R.; Franco, G.R. Virus-Host Coevolution: Common Patterns of Nucleotide Motif Usage in Flaviviridae and Their Hosts. *PLoS One* **2009**, *4*, e6282. [\[CrossRef\]](#)
76. Khandia, R.; Ali Khan, A.; Alexiou, A.; Povetkin, S.N.; Verevkin, M.N. Codon Usage Analysis of Pro-Apoptotic Bim Gene Isoforms. *J. Alzheimer's Dis.* **2022**, *86*, 1711–1725. [\[CrossRef\]](#)
77. Trinh, R.; Gurbaxani, B.; Morrison, S.L.; Seyfzadeh, M. Optimization of Codon Pair Use within the (GGGGS)₃ Linker Sequence Results in Enhanced Protein Expression. *Mol. Immunol.* **2004**, *40*, 717–722. [\[CrossRef\]](#)

78. Huang, Y.; Lin, T.; Lu, L.; Cai, F.; Lin, J.; Jiang, Y.E.; Lin, Y. Codon Pair Optimization (CPO): A Software Tool for Synthetic Gene Design Based on Codon Pair Bias to Improve the Expression of Recombinant Proteins in *Pichia Pastoris*. *Microb. Cell Fact.* **2021**, *20*, 209. [[CrossRef](#)]
79. Kunec, D.; Osterrieder, N. Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias. *Cell Rep.* **2016**, *14*, 55–67. [[CrossRef](#)]
80. Martrus, G.; Nevot, M.; Andres, C.; Clotet, B.; Martinez, M.A. Changes in Codon-Pair Bias of Human Immunodeficiency Virus Type 1 Have Profound Effects on Virus Replication in Cell Culture. *Retrovirology* **2013**, *10*, 78. [[CrossRef](#)]
81. Jordan-Paiz, A.; Franco, S.; Martinez, M.A. Synonymous Codon Pair Recoding of the HIV-1 Env Gene Affects Virus Replication Capacity. *Cells* **2021**, *10*, 1636. [[CrossRef](#)]
82. RajBhandary, U.L. More Surprises in Translation: Initiation without the Initiator tRNA. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 1325–1327. [[CrossRef](#)] [[PubMed](#)]
83. Bletsa, M.; Vrancken, B.; Gryseels, S.; Boonen, I.; Fikatas, A.; Li, Y.; Laudisoit, A.; Lequime, S.; Bryja, J.; Makundi, R.; et al. Molecular Detection and Genomic Characterization of Diverse Hepaciviruses in African Rodents. *Virus Evol.* **2021**, *7*, veab036. [[CrossRef](#)] [[PubMed](#)]
84. Furió, V.; Garijo, R.; Durán, M.; Moya, A.; Bell, J.C.; Sanjuán, R. Relationship between Within-Host Fitness and Virulence in the Vesicular Stomatitis Virus: Correlation with Partial Decoupling. *J. Virol.* **2012**, *86*, 12228–12236. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.