

Comprehensive Analysis of Codon Usage Bias in Human Papillomavirus Type 51

XIAOCHUN TAN[#], SIWEN BAO[#], XIAOLEI LU[#], BINBIN LU, WEIFENG SHEN
and CHAOYUE JIANG^{*}

Department of Laboratory Medicine, The First Hospital of Jiaxing,
Affiliated Hospital of Jiaxing University, Jiaxing, China

Submitted 14 May 2024, accepted 3 September 2024, published online 28 October 2024

Abstract

Human papillomavirus type 51 (HPV-51) is associated with various cancers, including cervical cancer. Examining the codon usage bias of the organism can offer valuable insights into its evolutionary patterns and its relationship with the host. This study comprehensively analyzed codon usage bias in HPV-51 by examining 64 complete genome sequences sourced from the NCBI GenBank database. Our analysis revealed no noteworthy preference for codon usage in HPV-51 overall. However, there was a noticeable bias towards A/T-ending codons, accompanied by GC3s below 32%. Dinucleotide frequency analysis revealed reduced frequencies for ApA, CpG, and TpC dinucleotides, while CpA and TpG dinucleotides were more frequent than others. Relative Synonymous Codon Usage analysis revealed 30 favored codons, primarily concluding with A/T nucleotides. Further analysis using Parity Rule 2, Effective Number of Codons plot, and neutrality plot indicated a balance between mutational pressure and natural selection, with natural selection being the primary force shaping codon usage bias. The Isoacceptor tRNA Pool analysis indicates that HPV-51 has a higher translation efficiency within the human cellular translational system. Moreover, the Codon Adaptation Index and Relative Codon Deoptimization Index analyses suggested a moderate adaptation of HPV-51 to human codon preferences. Our discoveries offer valuable perspectives on how HPV-51 evolves and uses genetic codes, contributing to a deeper comprehension of its endurance and disease-causing potential.

Key words: codon usage bias, human papillomavirus type 51, natural selection, Relative Synonymous Codon Usage, host adaptation

Introduction

Human papillomavirus (HPV) belongs to the *Papillomaviridae* family, characterized by its double-stranded DNA structure, recognized for causing various diseases in humans, such as cervical cancer, genital warts, and other anogenital cancers (Muñoz et al. 2003; de Sanjose et al. 2010; Tan et al. 2022). HPV is categorized into high-risk and low-risk types based on the severity of its pathogenicity (Liu et al. 2021). The continued presence of high-risk human papillomavirus (HR-HPV) is strongly associated with the development and progression of cervical cancer (Zhao et al. 2009). HPV-51 falls under the high-risk category, indicating its potential to cause abnormal cell changes leading to cancer if untreated (de Sanjose et al. 2010). HPV-51

DNA has been identified in both low-grade and high-grade cervical lesions, indicating its involvement in the progression of cervical cancer (Smith et al. 2007).

Furthermore, HPV-51 has been implicated in various anogenital cancers and lesions, such as vulvar, vaginal, penile, and anal cancers (Muñoz et al. 2003). While HPV-16 and HPV-18 are commonly acknowledged as high-risk types, HPV-51 has been identified in a significant proportion of cervical cancer cases globally, emphasizing its relevance in HPV-related diseases. A study highlights that while HPV-51 is less common among high-risk types, it still accounts for a notable percentage of HPV-related cancers (Smith et al. 2007). The prevalence of HPV-51 varies by region, but it has been detected in cervical, anal, and oropharyngeal cancers. While HPV-51 may not be as widespread as some

[#] Xiaochun Tan, Siwen Bao, Xiaolei Lu contribute equally to this work.

^{*} Corresponding author: C. Jiang, Department of Laboratory Medicine, The First Hospital of Jiaxing, Affiliated Hospital of Jiaxing University, Jiaxing, China; e-mail: tg30328@163.com

© 2024 Xiaochun Tan et al.

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License (<https://creativecommons.org/licenses/by-nc-nd/4.0/>).

HPV types, its potential to cause cancer underscores its importance in HPV-related disease prevention and management (Smith et al. 2007).

Codon bias is the phenomenon wherein synonymous codons coding for the same amino acid are not uniformly utilized within an organism’s genome (Sharp and Li 1986). This phenomenon is prevalent across diverse organisms, including viruses (Sharp et al. 1988). Understanding codon bias offers insights into evolutionary dynamics, translational efficiency, and host adaptation strategies (Das et al. 2006). Several factors influence an organism’s codon usage pattern, such as the abundance of specific transfer RNAs (tRNAs) for each codon, mutation bias, and natural selection (Plotkin and Kudla 2011). High-frequency codons, commonly found in highly expressed genes, often align with abundant cellular tRNAs, promoting efficient translation. Conversely, low-frequency codons may lead to less efficient translation due to the limited availability of corresponding tRNAs, potentially affecting protein synthesis rates (Koonin and Novozhilov 2009). Viruses often adjust their codon usage to match that of their hosts, which improves viral replication and helps them avoid host immune responses (Hershberg and Petrov 2008; Chaney and Clark 2015). This adaptation underscores the importance of codon usage bias in understanding the evolutionary processes shaping genomes and optimizing heterologous gene expression for biotechnological applications (Quax et al. 2015). Codon usage can affect protein folding, stability, function, and the yield of recombinant proteins in expression systems. Consequently, codon optimization strategies are often employed to boost the expression of foreign genes in host organisms, especially in microbial systems used for biopharmaceutical production (Sharp and Devine 1989).

To date, there have been no reports on the codon preference of HPV-51. Given its clinical relevance and its association with various cancers, understanding its codon usage patterns could offer valuable insights into its evolutionary dynamics and interaction with host cellular machinery. This study aims to fill this gap by providing a comprehensive analysis of the codon usage bias of HPV-51.

Experimental

Materials and Methods

Data collection. We acquired HPV-51 genome sequences from the NCBI GenBank database. Strict criteria were applied to ensure quality: sequences were 7,000–8,500 bp long with <5 ambiguous sites; those with ≥70 consecutive “N” characters or duplicates

were removed. Each Coding Sequence (CDS) had to consist of a number of nucleotides divisible by three, starting with ATG and ending with TAG, TGA, or TAA. 64 HPV-51 genomes were analyzed, with accession numbers in the Table SI.

Nucleotide composition analysis. We first looked at the nucleotide content of HPV-51’s coding sequences throughout its genome. Then, we examined the distribution of nucleotides in the third position of synonymous codons, denoted as %A3s, %C3s, %T3s, and %G3s. The examination was carried out utilizing CodonW (<http://codonw.sourceforge.net>). Additionally, we calculated the G + C content in the first, second, and third positions of codons (GC1s, GC2s, and GC3s). GC values were determined using the EMBOSS explorer tool, which can be accessed at (<https://www.bioinformatics.nl/emboss-explorer>).

Relative dinucleotide abundance analysis. Relative dinucleotide abundance analysis involves examining the frequency distribution of dinucleotide pairs within a given sequence. This analysis provides insights into the compositional biases and potential functional significance of dinucleotide patterns within nucleic acid sequences. The formula for calculating dinucleotide frequency is:

$$P_{xy} = \frac{f_{xy}}{f_x f_y}$$

In this equation, f_x and f_y denote the occurrence rates of nucleotides x and y , correspondingly, with f_{xy} indicating the occurrence rate of the dinucleotide xy . A dinucleotide is considered lowly represented when P_{xy} is <0.78 or highly represented when >1.23 (Butt et al. 2014). This analysis was conducted using R 4.1.1. (R Core Team 2021).

Relative Synonymous Codon Usage (RSCU). RSCU is a metric used to quantify the frequency bias of synonymous codons in a given gene or genome. It measures the relative abundance of each codon compared to the expected frequency under the assumption of equal usage of synonymous codons for the same amino acid.

The formula for calculating RSCU is:

$$RSCU = \frac{X_{ij}}{\sum_j^{ni} X_{ij}} ni$$

In this context, X_{ij} represents the frequency of the j_{th} codon that encodes the i_{th} amino acid, while ni stands for the count of synonymous codons that encode the same i_{th} amino acid. A RSCU value of 1 suggests no bias in codon usage. Values less than 0.6, between 0.6 and 1.6, and greater than 1.6 indicate low, normal, and excessive codon usage, respectively (Puigbò et al. 2008). The RSCU index computation was carried out using the seqinr package (version 3.6–1) in R 4.1.1.

Hydropathicity (GRAVY) and Aromaticity (AROMO). The GRAVY score computes the mean hydrophobicity across all amino acids within a sequence. The score spans from -2.0 to $+2.0$, where positive values denote hydrophobic proteins and negative values signify hydrophilic ones (Kyte and Doolittle 1982). GRAVY values were calculated using <http://www.gravy-calculator.de>.

The AROMO score evaluates the frequency of aromatic amino acids, specifically Phe, Tyr, and Trp, within a provided amino acid sequence. AROMO values were calculated using CodonW (<http://codonw.sourceforge.net>).

Effective Number of Codons Analysis (ENC). ENC is a metric used to assess how efficiently codons are utilized within a gene, indicating how far it deviates from random selection. The gene's length does not influence this measure. ENC values range from 20, where each amino acid corresponds to a single effective codon, to 61, indicating equal utilization of codons. Generally, an ENC value below 35 implies a bias towards particular codon preferences (Wright 1990).

The ENC is calculated using a distinct mathematical formula:

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

Here, F denotes the mean of F_i (where $i = 2, 3, 4, 6$) for amino acids with i -fold degeneracy.

The calculation of F_i values is given by:

$$F_i = \frac{n \sum_{j=1}^i \left(\frac{n_j}{n} \right)^2 - 1}{n - 1}$$

Parity Rule 2 Analysis (PR2). For this investigation, PR2 was employed to assess the influence of natural selection and mutational pressure on codon usage. The horizontal axis represents AT skew $[A3/(A3+T3)]$, while the vertical axis signifies GC skew $[G3/(G3+C3)]$. When data points converge around the midpoint $(0.5, 0.5)$, indicating $A=T$ and $G=C$, it implies an equilibrium between natural selection and mutational pressure. The dispersion of data points from this midpoint indicates the extent of deviation from PR2 (Sueoka 1995).

Neutrality plot analysis. The neutrality plot is a valuable tool in evaluating the impact of mutational pressure and natural selection on codon usage bias. It illustrates the average GC content of the first and second positions (GC12s) along the vertical axis, with the GC content of the third position (GC3s) represented on the horizontal axis. A linear regression approach is employed to analyze the data. A slope approaching 1 signifies that mutational pressure predominantly shapes codon usage patterns. Conversely, a slope near-

ing 0 indicates that natural selection significantly affects codon usage bias (Chen et al. 2014).

ENC-plot analysis. The ENC plot illustrates a standard curve representing the correlation between ENC and GC3s with no selective pressures.

The standard curve is determined by the following calculation:

$$ENC_{expected} = 2 + GC3s + \frac{29}{GC3s^2 + (1 - GC3s)^2}$$

If the ENC value observed falls beneath this curve, it suggests that influences beyond mutation pressure, such as natural selection, play a role in shaping codon usage bias (Wright 1990). This analysis was conducted using R 4.1.1.

Codon Adaptation Index (CAI) analysis. CAI elucidates the prevalence of preferred codons within highly expressed genes. Its scale spans from 0.0 to 1.0, where higher scores signify heightened potential for gene expression. Scores nearing 1 denote the utilization of codons with elevated RSCU within the gene. Notably, CAI remains unaffected by sequence length, relying solely on amino acid frequency (Xia 2007). In this study, we calculated the CAI values for the entire coding sequence of HPV-51 using the RSCU reference set tailored for Homo sapiens, obtained from the codon usage database at <http://www.kazusa.or.jp/codon>. CAI calculations were performed via <http://genomes.urv.es/CAIcal>.

Isoacceptor tRNA Pool analysis. Different isoacceptor tRNAs, derived from various tRNA types, crucially bind specific codons corresponding to amino acids, ensuring accurate incorporation into polypeptide chains during translation. The frequency of human tRNA genes is sourced from <http://gtrnadb.ucsc.edu>, offering insights into gene expression (Chan and Lowe 2016). Multiple isoacceptor tRNAs can bind the same codon, varying among species and impacting gene expression and protein synthesis efficiency. Our analysis focuses on human cell tRNA gene frequencies for HPV-51's preferentially used codons, highlighting viral evolutionary adaptations. HPV-51's abundant isoacceptor tRNAs efficiently recognize these codons, influencing translation selection and protein synthesis accuracy within host cells, which is crucial for understanding viral-host translational interactions.

Relative Codon Deoptimization Index (RCDI). The RCDI evaluates the similarity in codon usage between HPV and its human host, indicating the viral genome's translation efficiency within the host's genetic framework. A value nearing 1 signifies a significant congruence in codon usage, suggesting a robust adaptation of the virus to its host (Puigbò et al. 2010). RCDI calculations were performed via <http://genomes.urv.es/CAIcal>.

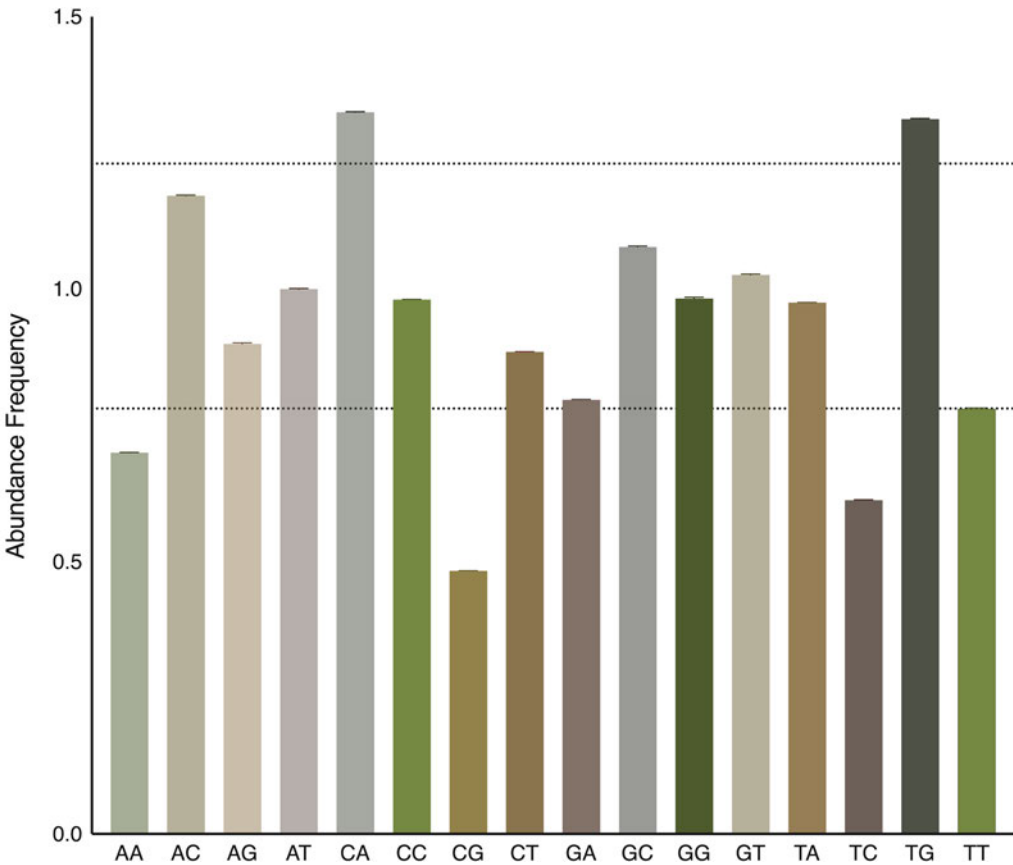


Fig. 1. Analysis of dinucleotide frequencies in the complete coding sequences of HPV-51. Dashed lines represent overrepresented values ($P_{xy} > 1.23$) or underrepresented values ($P_{xy} < 0.78$). The error bars are minimal due to the proximity of the data points.

Statistical analysis. The results are shown as means with standard deviations (SDs). Statistical analysis was conducted with SPSS® 22.0 software (IBM® SPSS® Statistics for Windows, version 22.0, IBM Corp., USA).

Results

Nucleotide composition analysis of HPV-51. In the entirety of HPV-51 genomes, the GC3s is below 32%, suggesting a bias towards A/T-ending codons (Table I). Additionally, GC1 is the most prevalent, whereas GC3s is the least. A3s and T3s appear more frequently than G3s and C3s positioned third among synonymous codons.

Examining dinucleotide influence on codon preference in HPV-51. The dinucleotide frequencies observed in HPV-51 differ from the anticipated value

of 1.0, indicating unique patterns influenced by translational selection. HPV-51 shows decreased occurrences of ApA, CpG, and TpC ($P_{xy} < 0.78$), alongside increased occurrences of CpA and TpG ($P_{xy} > 1.23$) (Fig. 1).

RSCU analysis. HPV-51 exhibited a preference for 30 specific codons ($RSCU > 1$; TTA[Leu], ACA[Thr], TTT[Phe], GCA[Ala], TAT[Tyr], AGA[Arg], AGT[Ser], GTA[Val], ATT[Ile], CCT[Pro], CCA[Arg], TCT[Ser], TGT[Cys], GAA[Glu], AAA[Lys], ATA[Ile], GAT[Asp], GCT[Ala], AAT[Asn], TTG[Leu], CAA[Gln], GGT[Gly], CGT[Arg], CAT[His], GTG[Val], AGG[Arg], ACT[Thr], GTT[Val], TCA[Ser], GGC[Gly]), with 26 of them ending in A/U. This includes eight over-represented codons ($RSCU > 1.6$): TTA, ACA, TTT, GCA, TAT, AGA, AGT, and GTA (Fig. 2).

Among the 59 synonymous codons, there are 15 under-represented codons ($RSCU < 0.6$), which include TTC [Phe], CTC [Leu], TCG [Ser], GTC [Val], ATC [Ile],

Table I
Nucleotide composition of complete genomes in HPV-51.

Genotype	GC1s*	GC2s	GC3s	T3s	C3s	A3s	G3s
HPV-51	46.23 ± 0.27	43.26 ± 0.22	31.03 ± 0.24	43.36 ± 5.56	17.08 ± 2.19	44.32 ± 5.68	19.59 ± 2.51

* – the values in the cells were represented as mean% ± SD%

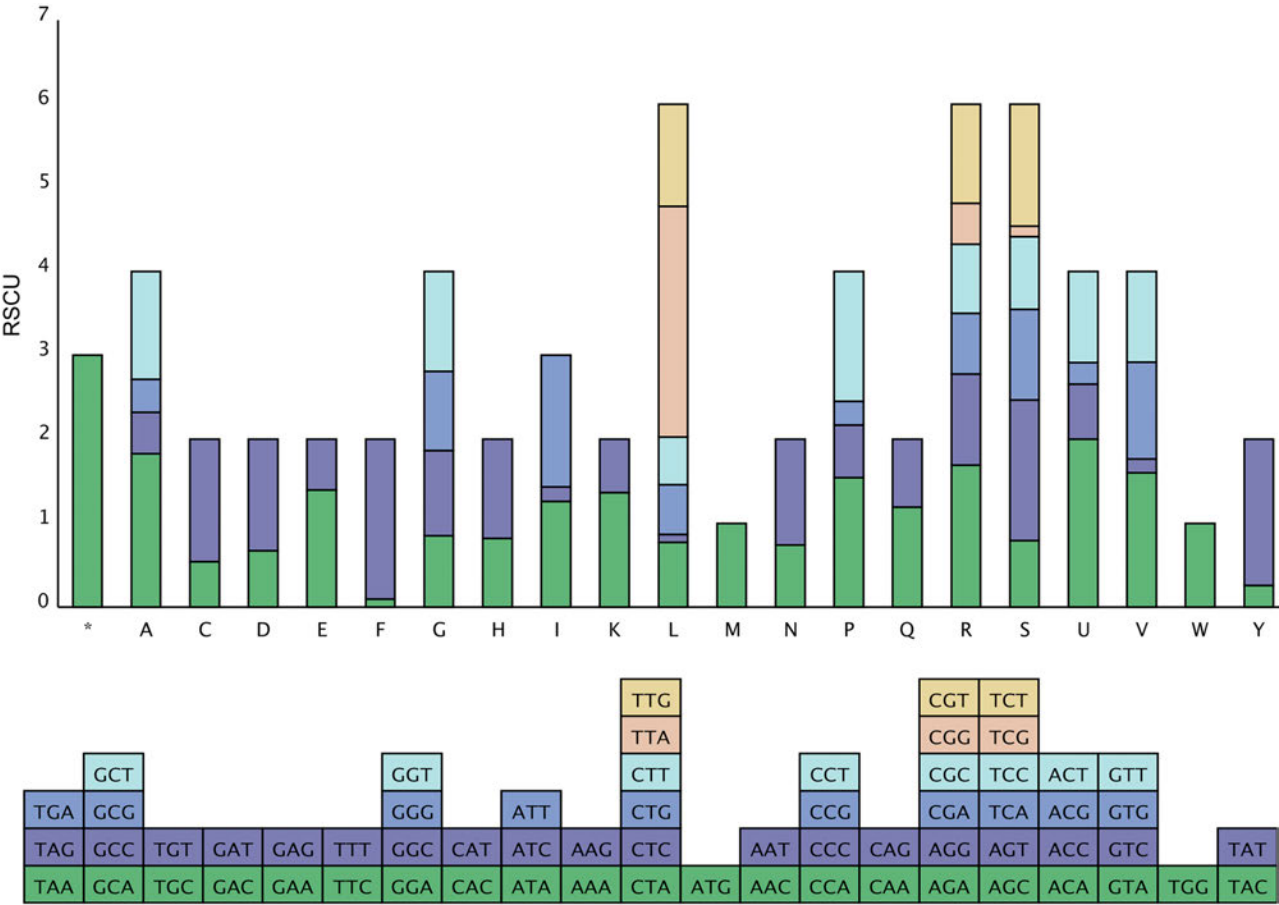


Fig. 2. Analysis of Relative Synonymous Codon Usage (RSCU) in the complete coding sequences of HPV-51. Codon values less than 0.6, between 0.6 and 1.6, and greater than 1.6 indicate low, normal, and excessive codon usage, respectively.

ACG [Thr], TAC [Tyr], CCG [Pro], GCG [Ala], CGG [Arg], GCC [Ala], TGC [Cys], TGC [Cys], CTT [Leu], and CTG [Leu]. All of these codons end with G or C.

Both RSCU and the composition of nucleotides suggest that HPV-51 tends to use codons ending with A/T. This implies that the nucleotide at the third position of codons influences the selection of synonymous codons in HPV-51 (Jenkins and Holmes 2003).

Impact of dinucleotide composition on codon usage preferences in HPV-51. To examine how dinucleotide usage affects codon preference, we compared the occurrence of overrepresented and underrepresented dinucleotides within both favored and less favored codons. All six codons containing ApA showed RSCU values below 1.6, indicating a suppression of this dinucleotide. Five of the eight CpG-containing codons were less represented (CGG, GCG, CCG, ACG, and TCG). Similarly, five of the eight TpC-containing codons were less represented (ATC, GTC, TCG, CTC, TTC), indicating reduced usage of CpG and TpC. For the eight codons with CpA, six (ACA, GCA, CCA, CAA, CAT, TCA) exhibit a preference for synonymous codons. Out of the five codons with TpG, three (TGT, TTG, GTG) predominantly utilized synonymous codons, showing a preference for

CpA and TpG. The dinucleotide composition of HPV-51 exhibits a low occurrence of ApA, CpG, and TpC, with CpA and TpG being more prevalent than others. Tri-nucleotide usage patterns align with these dinucleotide trends, as evidenced by RSCU data. These observations suggest that dinucleotide frequency heavily influences the codon usage preferences in HPV-51.

Exploration of influential factors on codon usage pattern of HPV-51. The PR2 analysis revealed that the codons of the entire genome of HPV-51 are closely clustered around the central point on the scatter plot, where A is approximately equal to T and C is approximately equal to G, this suggests a minor distinction between mutational pressure and natural selection in HPV-51 (Fig. 3).

We evaluated the ENC values for the entire genomes of HPV-51 to measure the extent of codon usage bias in its coding sequences. The average ENC was found to be 48.25 ± 0.20 , suggesting a predominantly low codon preference (Table II). When analyzing the complete HPV-51 genomes, they cluster slightly below the curve. This indicates that the codon usage pattern throughout the HPV-51 coding genome is notably shaped by natural selection (Fig. 4).

Table II
The Effective Number of Codons (ENC), Codon Adaptation Index (CAI), and Relative Codon Deoptimization Index (RCDI) values of complete genomes in HPV-51.

Genotype	ENC*	CAI	RCDI
HPV-51	48.25 ± 0.20	0.72 ± 0.00	± 0.01

* – the values in the cells were represented as mean% ± SD%

Neutral analysis offers further insights into the factors shaping the preference of HPV-51 for codon usage. Within the complete coding genomes of HPV-51, natu-

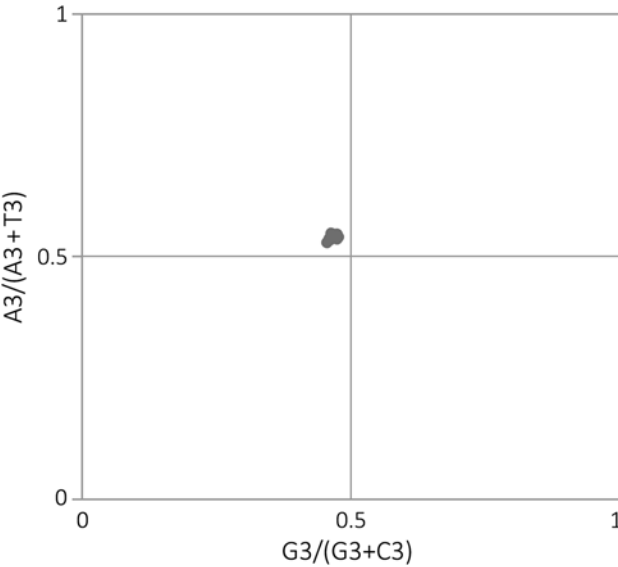


Fig. 3. Analysis of Parity Rule 2 (PR2) constructed from the complete coding sequences of HPV-51. The central point of the plot, represented by coordinates (0.5, 0.5), indicates a balance between mutation and selection rates.

ral selection predominantly shapes the pattern, contributing to about 73.37% of the observed bias (Fig. 5). In summary, the codon usage pattern observed in the HPV-51 genome shows a subtle balance between mutational pressure and natural selection, with the latter significantly molding the pattern.

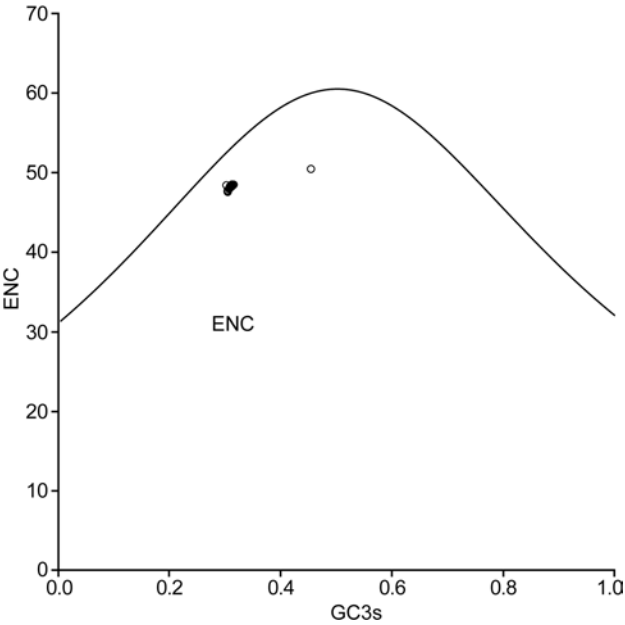


Fig. 4. Analysis of the Effective Number of Codons (ENC) plot for the complete coding sequences of HPV-51. ENC values are plotted against the GC3s content. The black line represents the standard curve based on codon usage bias determined solely by GC3s composition. This standard curve illustrates the expected relationship between ENC and GC3s without selective pressures. The deviation of actual ENC values below this curve indicates a more significant influence of natural selection on codon bias.

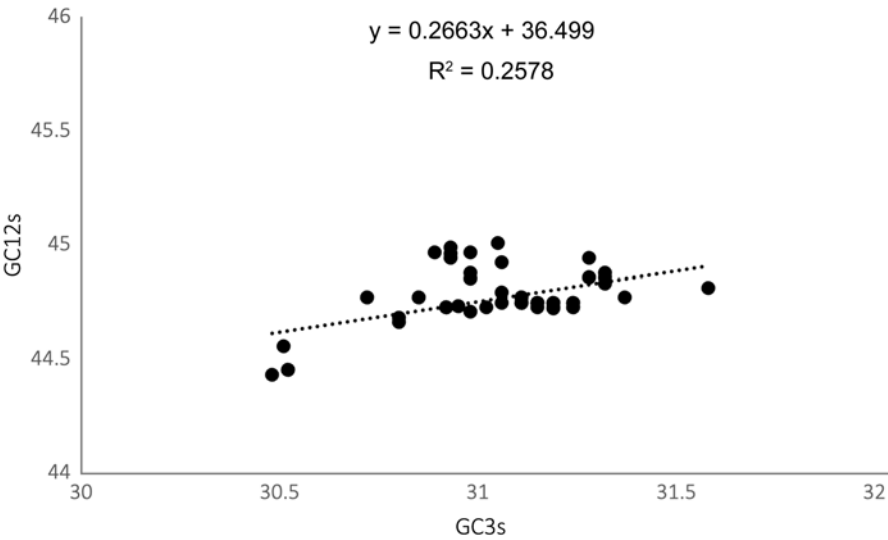


Fig. 5. Neutrality analysis of the complete coding sequences of HPV-51. The relationship between GC content at the first and second positions of the codon (GC12s) and the third position of the codon (GC3s) was analyzed. The dashed line depicts the correlation between GC12s and GC3s in HPV-51. A slope approaching 1 indicates that mutation pressure predominantly shapes the codon usage pattern, while a slope closer to 0 suggests a more substantial influence of natural selection.

Table III
Frequency of tRNA genes in human cells corresponding to the most preferentially used codons in HPV-51.

Amino acid	Most preferred codons in HPV-51	tRNA isotypes in human cells	Total count
Ala (A)	GCG	AGC (22), GGC (0), CGC (4), TGC (8)	34
Gly (G)	GGC	ACC (0), GCC (14), CCC (5), TCC (9)	28
Pro (P)	CCG	AGG (9), GGG (0), CGG (4), TGG (7)	20
Thr (T)	ACG	AGT (9), GGT (0), CGT (5), TGT (6)	20
Val (V)	GTG	AAC (9), GAC (0), CAC (11), TAC (5)	25
Ser (S)	TCG	AGA (9), GGA (0), CGA (4), TGA (4), ACT (0), GCT (8)	25
Arg (R)	AGG	ACG (7), GCG (0), CCG (4), TCG (6), CCT (5), TCT (6)	28
Leu (L)	CTG	AAG (9), GAG (0), CAG (9), TAG (3), CAA (6), TAA (4)	31
Phe (F)	TTC	AAA (0), GAA (10)	10
Asn (N)	AAT	ATT (0), GTT (20)	20
Lys (K)	AAG	CTT (15), TTT (12)	27
Asp (D)	GAC	ATC (0), GTC (13)	13
Glu (E)	GAG	CTC (8), TTC (7)	15
His (H)	CAC	ATG (0), GTG (10)	10
Gln (Q)	CAG	CTG (13), TTG (6)	19
Ile (I)	ATC	AAT (14), GAT (3), TAT (5)	22
Tyr (Y)	TAT	ATA (0), GTA (13)	13
Cys (C)	TGC	ACA (0), GCA (29)	29
Trp (W)	TGG	CCA (7)	7
Met (M)	ATG	CAT (9/10)	19

Isoacceptor tRNA Pool in HPV-51. Various isoacceptor tRNAs, originating from different tRNA types, bind to distinct codons representing specific amino acid residues. Table III presents the tRNA gene frequencies in human cells. It's worth noting that several isoacceptor tRNAs can recognize a single codon, and this phenomenon varies among different species. Recognizing the most frequently used codons by HPV-51 by the most abundant isoacceptor tRNAs influences translation selection (Kumar et al. 2016). For the 18 amino acids encoded by multiple codons, optimal codon-anticodon base pairs are employed for Gly, Val, Leu, Phe, Lys, Asp, Glu, His, Gln, and Cys, indicating higher translation efficiency of HPV-51 within the human cellular translational system.

Measures of virus adaptation of HPV-51. CAI values assess the expression level of viral proteins within the host and the virus's adaptability with its host. Sequences with elevated CAI values are deemed better suited for adapting to specific hosts compared to those with lower values. The CAI value for HPV-51 in relation to humans is 0.72 ± 0.00 , suggesting a degree of adaptation to human codon preferences (Table II).

Table IV presents a correlation examination among GC3s, CAI, GRAVY, and AROMO. GRAVY and AROMO are markers of natural selection influencing codon bias, while GC3s indicate compositional charac-

teristics (Barbhuiya et al. 2019). A positive correlation between GC3s and AROMO suggests that increased compositional bias is associated with increased aromaticity. The aim of assessing the impact of GRAVY and AROMO, markers of natural selection, on gene expression (represented by CAI), resulted in an investigation into their correlation. However, no significant correlation between CAI, GRAVY, and AROMO was detected.

Additionally, the RCDI values were utilized to evaluate the similarity in codon usage patterns between HPV-51 and its host. HPV-51 shows RCDI values approaching 2, suggesting a moderate similarity in codon preferences with humans (Table II).

These results indicate that HPV-51 adjusts to the host's translation process and minimizes competition with cellular proteins. These adaptive strategies may contribute to the persistent presence of these viruses in human hosts.

Table IV
Correlation analysis between GC3s, CAI, GRAVY, and AROMO.

	GC3s	CAI	GRAVY
CAI	-0.036		
GRAVY	0.143	0.168	
Aromo	0.818***	-0.135	-0.122

*** - $p < 0.001$

Discussion

HPV-51 belongs to the high-risk category of HPV, which is linked to multiple cancer types, such as cervical cancer. Although of clinical importance, scant attention has been paid to studying its codon usage bias. The research presents a comprehensive analysis of codon usage bias in HPV-51, shedding light on its evolutionary dynamics, host adaptation strategies, and potential translational efficiency within human cells. In our study, we thoroughly investigated codon usage patterns by analyzing coding sequences from 64 genomes of HPV-51.

The analysis of codon usage bias in HPV-51 shows a slight preference for A/T-ending codons. We additionally computed RSCU values for 59 synonymous codons, indicating that the favored codons in HPV-51 ($RSCU > 1$) mainly terminate with A or T. Such bias is commonly associated with the interaction of mutational pressure and natural selection, reflecting the virus's evolutionary adaptation to optimize its replication and survival within the host (Sharp and Li 1987). The preference for A/T-ending codons may confer several advantages to HPV-51. First, these codons might be more abundant in the human genome, allowing the virus to camouflage its gene expression patterns and evade the host immune response (Karlin and Burge 1996). Second, in genomes with high AT content, a relationship frequently exists between codons ending in A or T and increased efficiencies in transcription and translation processes (Aktürk Dizman 2024). Additionally, the ENC values for HPV-51 suggest a relatively low overall codon usage preference. Identical findings of low codon usage bias were reported in other viruses like Zika virus (Tao and Yao 2020), mollusum contagiosum virus (Nair et al. 2021), and invertebrate iridescent virus (Aktürk Dizman 2023). The ENC value is an indicator of codon bias, with higher values signaling reduced bias and lower values suggesting a more pronounced bias towards particular codons. Previous studies have shown that viruses with higher ENC values tend to exhibit increased adaptability (Jenkins and Holmes 2003). Furthermore, a lower codon usage bias, as indicated by higher ENC values, may allow HPV-51 to conserve energy during viral biosynthesis and diminish interference with host protein synthesis, enhancing its overall fitness within the host environment (Wright 1990).

The frequency of dinucleotides significantly influences the preferences for codon usage in various organisms. Dinucleotide abundance reflects codon usage bias, mutational tendencies, and the impact of selective pressure and compositional constraints. Within viral genomes, dinucleotides' proportional distribution significantly influences codon usage patterns (Upadhyay and Vivekanandan 2015). In our research, we examined 16 different dinucleotide frequencies. Notably, dinucle-

otides ApA, CpG, and TpC were less common, whereas CpA and TpG were more frequently observed than others. CpG depletion is a common viral strategy to evade host immune detection (Karlin and Burge 1995). Natural selection drives this reduction in CpG-containing codons, as host defenses recognize unmethylated CpG sequences, triggering immune responses (Khandia et al. 2019). To evade these defenses, viruses often reduce their CpG content (Vetsigian and Goldenfeld 2009). The increase in CpA and TpG may be compensatory, balancing the reduced CpG levels (Rima and McFerran 1997). Comparing dinucleotide frequencies with RSCU values of HPV-51 reveals that trinucleotide usage often mirrors dinucleotide preferences. The abundance or scarcity of specific dinucleotides can significantly shape the overall codon usage bias (Karlin and Burge 1995). In summary, dinucleotide frequencies are foundational elements shaping codon usage bias, reflecting the complex interplay between viral evolutionary pressures and host interactions in HPV-51.

Then, to comprehensively understand the factors shaping codon usage in HPV-51, we conducted PR2 analysis, ENC plots, and neutrality analysis. These analyses revealed a slight bias in codon usage, indicative of the intertwined implications of mutational pressure and natural selection. Interestingly, natural selection has emerged as the primary factor influencing codon preferences within HPV-51. This phenomenon likely signifies the virus's adjustment to its host's cellular environment and ability to evade the host's immune responses. This preference might optimize viral protein synthesis and facilitate viral persistence. Our results align with previous research emphasizing the significance of natural selection in molding viral codon usage (Shackelton and Holmes 2008).

Additionally, we conducted an analysis of translational selection in HPV-51. In HPV-51, the most preferred codons utilize approximately half of the optimal codon-anticodon base pairings available in the human isoacceptor tRNA Pool. This preference for specific codons might indicate evolutionary pressures that optimize translation efficiency and accuracy (Kanaya et al. 2001). Utilizing the most compatible codons with the available tRNA molecules can enhance translation rates and reduce errors during protein synthesis (Ikemura 1985). The presence of approximately half of the optimal codon-anticodon pairs in the human isoacceptor tRNA pool facilitates efficient and accurate protein synthesis, thereby contributing to the organism's overall fitness. This preference for specific optimal codons is believed to result from co-evolution between codon usage and tRNA availability, ensuring an optimal balance between translational speed and accuracy (Akashi 2001).

The positive correlation between GC3s and AROMO values suggests an intriguing relationship between

compositional bias and amino acid aromaticity. This correlation implies that an increased compositional bias towards GC-rich codons is associated with higher aromaticity in proteins. One possible explanation for this correlation could be that GC-rich codons tend to encode aromatic amino acids more frequently, leading to proteins with higher aromaticity. Additionally, the selection pressures that favor GC-rich codons might also indirectly select proteins with increased aromaticity due to functional advantages conferred by aromatic amino acids (Duret 2002). The lower GC codon content in HPV-51, which is influenced more significantly by natural selection, supports this explanation. The interplay between codon bias, compositional, and amino acid properties is complex and multifaceted. Further research is needed to explore the mechanisms underlying these correlations and to elucidate the functional implications of increased aromaticity in proteins.

Finally, we evaluated the adaptation of HPV-51 to its human host using CAI and RCDI values. The CAI value for HPV-51 was determined to be 0.72 ± 0.00 , indicating a moderate alignment with human codon preferences. This suggests that the virus has undergone adaptive evolution to utilize the host's translational machinery efficiently. The RCDI value for HPV-51 was close to 2, further supporting its moderate adaptation to the human host. The concept of codon adaptation is crucial for understanding how viruses interact with their hosts at the molecular level. Viruses often optimize their codon usage to adapt to the translational machinery of the host cell, facilitating efficient production and duplication of viral proteins (Sharp et al. 2005). A high CAI value signifies a strong adaptation to host codon usage, whereas a low value may indicate a lack of adaptation or evolutionary divergence from the host. The RCDI value provides additional insights into the adaptation process by measuring the extent of codon deoptimization. A higher RCDI value suggests that the virus has minimized the use of non-preferred codons, which can lead to slower translation rates and increased competition with host mRNAs for ribosomal machinery (Plotkin and Kudla 2011). Therefore, an RCDI value close to 2 for HPV-51 indicates that the virus has struck a balance between efficiently utilizing the host's translational machinery and minimizing excessive competition with host mRNAs. Our findings suggest that HPV-51 has undergone adaptive evolution to align its codon usage with human preferences moderately. This adaptation likely contributes to the virus's ability to efficiently replicate and propagate within the human host, highlighting the importance of codon adaptation in viral-host interactions.

Our previous research analyzed codon preference in HPV-33 and HPV-58 (Tan et al. 2024). Now, we compare the codon preference results of HPV-51 with

those of HPV-33 and HPV-58. All three types – HPV-33, HPV-58, and HPV-51 – show a preference for codons ending in A or T, with GC3s values below 35%. The analysis of RSCU of HPV-31 revealed 30 favored codons, 26 concluding with A/T nucleotides, while among the 26 favored codons in HPV-33 and HPV-58, 25 concluded with A/U. The mean ENC values were 44.19 for HPV-33, 46.90 for HPV-58, and 48.25 for HPV-51, all above 35, indicating a relatively low level of codon preference. HPV-33, HPV-58, and HPV-51 exhibit specific patterns in dinucleotide frequency, with lower frequencies of ApA, CpG, and TpC, and higher frequencies of CpA and TpG. Analyses, including the ENC plot, PR2, and neutrality analysis, suggest that natural selection is the primary factor influencing codon usage in HPV-33, HPV-58, and HPV-51. The RCDI values for HPV-33, HPV-58, and HPV-51 are around 2, indicating a similar degree of codon usage preference to humans. These comparisons suggest that HPV-33, HPV-58, and HPV-51 share similar codon preferences, indicating common evolutionary mechanisms and selective pressures influencing codon usage.

While our study provides valuable insights into HPV-51's codon usage bias, it is not without limitations. The analysis is based on genomic sequences from the NCBI GenBank database, which may not capture all HPV-51 types or regional variations (Van Doorslaer et al. 2013). Future studies could benefit from analyzing additional HPV-51 genomes, from diverse geographic regions and clinical samples, to further elucidate its codon usage patterns and evolutionary dynamics (Xi et al. 2006). Our thorough examination of codon utilization bias in HPV-51 reveals evolutionary patterns, host adaptation strategies, and translational efficiency within human cells. These findings enhance our understanding of HPV-51's persistence, pathogenicity, and potential implications for HPV-related diseases.

Availability of data and material

The supporting data for the findings presented in this study are accessible through the GenBank database at <http://www.ncbi.nlm.nih.gov/Genbank>. Reference numbers are provided in the Table S1 accompanying this article.

Funding

This study was supported by Jiaxing Key Discipline of Medicine-Clinical Diagnostics (2023-ZC-002), the "Venus" Talent Training of the First Hospital of Jiaxing of Zhejiang Province of China under Grant 2022-QXM-019 and 2021-QMX-20.

Conflict of interest

The authors do not report any financial or personal connections with other persons or organizations, which might negatively affect the contents of this publication and/or claim authorship rights to this publication.

Literature

- Akashi H.** Gene expression and molecular evolution. *Curr Opin Genet Dev.* 2001 Dec;11(6):660–666. [https://doi.org/10.1016/s0959-437x\(00\)00250-1](https://doi.org/10.1016/s0959-437x(00)00250-1)
- Aktürk Dizman Y.** Analysis of codon usage bias of exonuclease genes in invertebrate iridescent viruses. *Virology.* 2024 May; 593:110030. <https://doi.org/10.1016/j.virol.2024.110030>
- Aktürk Dizman Y.** Codon usage bias analysis of the gene encoding NAD⁺-dependent DNA ligase protein of Invertebrate iridescent virus 6. *Arch Microbiol.* 2023 Oct;205(11):352. <https://doi.org/10.1007/s00203-023-03688-5>
- Barbhuiya PA, Uddin A, Chakraborty S.** Compositional properties and codon usage of TP73 gene family. *Gene.* 2019 Jan;683:159–168. <https://doi.org/10.1016/j.gene.2018.10.030>
- Butt AM, Nasrullah I, Tong Y.** Genome-wide analysis of codon usage and influencing factors in chikungunya viruses. *PLoS One.* 2014 Mar;9(3):e90905. <https://doi.org/10.1371/journal.pone.0090905>
- Chan PP, Lowe TM.** Gttnadb 2.0: An expanded database of transfer rna genes identified in complete and draft genomes. *Nucleic Acids Res.* 2016 Jan;44(D1):D184–189. <https://doi.org/10.1093/nar/gkv1309>
- Chaney JL, Clark PL.** Roles for synonymous codon usage in protein biogenesis. *Annu Rev Biophys.* 2015;44:143–166. <https://doi.org/10.1146/annurev-biophys-060414-034333>
- Chen H, Sun S, Norenburg JL, Sundberg P.** Mutation and selection cause codon usage and bias in mitochondrial genomes of ribbon worms (Nemertea). *PLoS One.* 2014 Jan;9(1):e85631. <https://doi.org/10.1371/journal.pone.0085631>
- Das S, Paul S, Dutta C.** Synonymous codon usage in adenoviruses: influence of mutation, selection and protein hydrophathy. *Virus Res.* 2006 May;117(2):227–236. <https://doi.org/10.1016/j.virusres.2005.10.007>
- de Sanjose S, Quint WG, Alemany L, Geraets DT, Klaustermeier JE, Lloveras B, Tous S, Felix A, Bravo LE, Shin HR et al.; Retrospective International Survey and HPV Time Trends Study Group.** Human papillomavirus genotype attribution in invasive cervical cancer: A retrospective cross-sectional worldwide study. *Lancet Oncol.* 2010 Nov;11(11):1048–1056. [https://doi.org/10.1016/S1470-2045\(10\)70230-8](https://doi.org/10.1016/S1470-2045(10)70230-8)
- Duret L.** Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev.* 2002 Dec;12(6):640–649. [https://doi.org/10.1016/s0959-437x\(02\)00353-2](https://doi.org/10.1016/s0959-437x(02)00353-2)
- Hershberg R, Petrov DA.** Selection on codon bias. *Annu Rev Genet.* 2008;42:287–299. <https://doi.org/10.1146/annurev.genet.42.110807.091442>
- Ikemura T.** Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol.* 1985 Jan;2(1):13–34. <https://doi.org/10.1093/oxfordjournals.molbev.a040335>
- Jenkins GM, Holmes EC.** The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res.* 2003 Mar; 92(1):1–7. [https://doi.org/10.1016/s0168-1702\(02\)00309-x](https://doi.org/10.1016/s0168-1702(02)00309-x)
- Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T.** Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J Mol Evol.* 2001 Oct-Nov; 53(4–5):290–298. <https://doi.org/10.1007/s002390010219>
- Karlin S, Burge C.** Dinucleotide relative abundance extremes: A genomic signature. *Trends Genet.* 1995 Jul;11(7):283–290. [https://doi.org/10.1016/s0168-9525\(00\)89076-9](https://doi.org/10.1016/s0168-9525(00)89076-9)
- Karlin S, Burge C.** Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc Natl Acad Sci USA.* 1996 Feb;93(4):1560–1565. <https://doi.org/10.1073/pnas.93.4.1560>
- Khandia R, Singhal S, Kumar U, Ansari A, Tiwari R, Dhama K, Das J, Munjal A, Singh RK.** Analysis of Nipah virus codon usage and adaptation to hosts. *Front Microbiol.* 2019 May;10:886. <https://doi.org/10.3389/fmicb.2019.00886>
- Koonin EV, Novozhilov AS.** Origin and evolution of the genetic code: The universal enigma. *IUBMB Life.* 2009 Feb;61(2):99–111. <https://doi.org/10.1002/iub.146>
- Kumar N, Bera BC, Greenbaum BD, Bhatia S, Sood R, Selvaraj P, Anand T, Tripathi BN, Virmani N.** Revelation of influencing factors in overall codon usage bias of equine influenza viruses. *PLoS One.* 2016 Apr;11(4):e0154376. <https://doi.org/10.1371/journal.pone.0154376>
- Kyte J, Doolittle RF.** A simple method for displaying the hydrophobic character of a protein. *J Mol Biol.* 1982 May;157(1):105–132. [https://doi.org/10.1016/0022-2836\(82\)90515-0](https://doi.org/10.1016/0022-2836(82)90515-0)
- Liu W, Li J, Du H, Ou Z.** Mutation profiles, glycosylation site distribution and codon usage bias of human papillomavirus type 16. *Viruses.* 2021 Jun;13(7):1281. <https://doi.org/10.3390/v13071281>
- Muñoz N, Bosch FX, de Sanjosé S, Herrero R, Castellsagué X, Shah KV, Snijders PJ, Meijer CJ; International Agency for Research on Cancer Multicenter Cervical Cancer Study Group.** Epidemiologic classification of human papillomavirus types associated with cervical cancer. *N Engl J Med.* 2003 Feb;348(6):518–527. <https://doi.org/10.1056/NEJMoa021641>
- Nair RR, Mohan M, Rudramurthy GR, Vivekanandam R, Satheshkumar PS.** Strategies and patterns of codon bias in molluscum contagiosum virus. *Pathogens.* 2021 Dec;10(12):1649. <https://doi.org/10.3390/pathogens10121649>
- Plotkin JB, Kudla G.** Synonymous but not the same: The causes and consequences of codon bias. *Nat Rev Genet.* 2011 Jan;12(1):32–42. <https://doi.org/10.1038/nrg2899>
- Puigbò P, Aragonès L, Garcia-Vallvé S.** RCDI/eRCDI: A web-server to estimate codon usage deoptimization. *BMC Res Notes.* 2010 Mar;3:87. <https://doi.org/10.1186/1756-0500-3-87>
- Puigbò P, Bravo IG, Garcia-Vallvé S.** CAIcal: A combined set of tools to assess codon usage adaptation. *Biol Direct.* 2008 Sep;3:38. <https://doi.org/10.1186/1745-6150-3-38>
- Quax TE, Claassens NJ, Söll D, van der Oost J.** Codon bias as a means to fine-tune gene expression. *Mol Cell.* 2015 Jul;59(2):149–161. <https://doi.org/10.1016/j.molcel.2015.05.035>
- R Core Team.** A Language and Environment for Statistical Computing. Vienna (Austria): R Foundation for Statistical Computing; 2021 [cited 2024 Mar 21]. Available from <https://www.R-project.org>
- Rima BK, McFerran NV.** Dinucleotide and stop codon frequencies in single-stranded RNA viruses. *J Gen Virol.* 1997 Nov;78 (11): 2859–2870. <https://doi.org/10.1099/0022-1317-78-11-2859>
- Shackelton LA, Holmes EC.** The role of alternative genetic codes in viral evolution and emergence. *J Theor Biol.* 2008 Sep;254(1): 128–134. <https://doi.org/10.1016/j.jtbi.2008.05.024>
- Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE.** Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* 2005 Feb;33(4):1141–1153. <https://doi.org/10.1093/nar/gki242>
- Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F.** Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; A review of the considerable within-species diversity. *Nucleic Acids Res.* 1988 Sep;16(17):8207–8211. <https://doi.org/10.1093/nar/16.17.8207>
- Sharp PM, Devine KM.** Codon usage and gene expression level in *Dictyostelium discoideum*: Highly expressed genes do ‘prefer’ optimal codons. *Nucleic Acids Res.* 1989 Jul;17(13):5029–5039. <https://doi.org/10.1093/nar/17.13.5029>

- Sharp PM, Li WH.** An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol.* 1986;24(1–2):28–38. <https://doi.org/10.1007/BF02099948>
- Sharp PM, Li WH.** The codon adaptation index – A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 1987 Feb;15(3):1281–1295. <https://doi.org/10.1093/nar/15.3.1281>
- Smith JS, Lindsay L, Hoots B, Keys J, Franceschi S, Winer R, Clifford GM.** Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical lesions: A meta-analysis update. *Int J Cancer.* 2007 Aug;121(3):621–632. <https://doi.org/10.1002/ijc.22527>
- Sueoka N.** Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evol.* 1995 Mar;40(3):318–325. <https://doi.org/10.1007/BF00163236>
- Tan X, Xie Y, Jiang C, Li H, Lu Y, Shen W, Chen J.** Codon usage bias of human papillomavirus type 33 and 58: A comprehensive analysis. *J Basic Microbiol.* 2024 May;64(5):e2300636. <https://doi.org/10.1002/jobm.202300636>
- Tan X, Yang J, Jiang J, Wang W, Ren J, Li Q, Xie Z, Chen X, Zhang L, Li W.** Significant growth inhibition by a bispecific affibody targeting oncoprotein E7 in both HPV16 and 18 positive cervical cancer *in vitro* and *in vivo*. *Eur J Pharm Sci.* 2022 May;172:106156. <https://doi.org/10.1016/j.ejps.2022.106156>
- Tao J, Yao H.** Comprehensive analysis of the codon usage patterns of polyprotein of Zika virus. *Prog Biophys Mol Biol.* 2020 Jan;150:43–49. <https://doi.org/10.1016/j.pbiomolbio.2019.05.001>
- Upadhyay M, Vivekanandan P.** Depletion of CpG dinucleotides in papillomaviruses and polyomaviruses: A role for divergent evolutionary pressures. *PLoS One.* 2015 Nov 6;10(11):e0142368. <https://doi.org/10.1371/journal.pone.0142368>
- Van Doorslaer K, Tan Q, Xirasagar S, Bandaru S, Gopalan V, Mohamoud Y, Huyen Y, McBride AA.** The Papillomavirus Episteme: A central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res.* 2013 Jan;41(D1):D571–D578. <https://doi.org/10.1093/nar/gks984>
- Vetsigian K, Goldenfeld N.** Genome rhetoric and the emergence of compositional bias. *Proc Natl Acad Sci USA.* 2009 Jan;106(1):215–220. <https://doi.org/10.1073/pnas.0810122106>
- Wright F.** The ‘effective number of codons’ used in a gene. *Gene.* 1990 Mar;87(1):23–29. [https://doi.org/10.1016/0378-1119\(90\)90491-9](https://doi.org/10.1016/0378-1119(90)90491-9)
- Xi LF, Kiviat NB, Hildesheim A, Galloway DA, Wheeler CM, Ho J, Koutsky LA.** Human papillomavirus type 16 and 18 variants: Race-related distribution and persistence. *J Natl Cancer Inst.* 2006 Aug;98(15):1045–1052. <https://doi.org/10.1093/jnci/djj297>
- Xia X.** An improved implementation of codon adaptation index. *Evol Bioinform Online.* 2007 May 17;3:53–58.
- Zhao R, Zhang WY, Wu MH, Zhang SW, Pan J, Zhu L, Zhang YP, Li H, Gu YS, Liu XZ.** Human papillomavirus infection in Beijing, People’s Republic of China: A population-based study. *Br J Cancer.* 2009 Nov;101(9):1635–1640. <https://doi.org/10.1038/sj.bjc.6605351>

Supplementary materials are available on the journal’s website.