

**Codon usage bias and evolutionary dynamics of porcine *Sapelovirus*: insights
into host adaptation**

Running title: Codon bias, evolution, and host adaptation in porcine *Sapelovirus*

Section: Original Research

doi: 10.22201/fmvz.24486760e.2026.1480

Mouna Venkataramappa

0009-0003-9353-3702

KP. Suresh

0000-0002-3005-0522

Neha Naik

0009-0001-7772-466X

J. Manjunatha

0009-0003-8272-1943

Akshatha Velankar

0009-0004-7322-9305

M. Vijay

0009-0001-3493-4800

Varsha Ramesh

0009-0005-8242-2833

M. Shijili

0009-0000-9717-6923

Jagadish Hire math

0000-0002-6606-3771

Siju S. Jacob

0000-0002-0254-0102

N. Shivasharanappa

0000-0003-3311-5886

BR. Gulati

0000-0002-4338-7941

Sharanagouda Patil*

0000-0002-3291-7513

ICAR-National Institute of Veterinary Epidemiology and Disease Informatics (NIVEDI).

Yelahanka, Bengaluru, India.

* Corresponding author: sharanspin13@gmail.com

Dates:

Submitted: 2025-02-05

Accepted: 2025-09-05

Published: 2026-01-22

Cite this as:

Venkataramappa M, Suresh K, Naik N, Manjunatha J, Velankar A, Vijay M, Ramesh V, Shijili M, Hire math J, Jacob SS, Shivasharanappa N, Gulati B, Patil S. Codon usage bias and evolutionary dynamics of porcine *Sapeloivirus*: insights into host adaptation. *Veterinaria México OA*. 2026;13. doi: 10.22201/fmvz.24486760e.2026.1480.

Codon usage bias and evolutionary dynamics of porcine *Sapelovirus*: insights into host adaptation

Running title: Codon bias, evolution, and host adaptation in porcine *Sapelovirus*

Section: Original Research

doi: 10.22201/fmvz.24486760e.2026.1480

Abstract

Porcine sapelovirus (PSV), a member of the *Sapelovirus* genus within the Picornaviridae family, is a swine pathogen causing respiratory diseases, polioencephalomyelitis, and gastroenteritis. The infection results in economic loss in the swine industry and often co-occurs with bacterial, viral, or fungal infections. Despite its impact, the evolutionary dynamics and adaptation mechanisms of PSV remain poorly understood. This study investigates the evolutionary forces shaping adaptation of the PSV polyprotein gene through codon usage bias and nucleotide composition analysis. A total of 34 polyprotein coding sequences of PSV were retrieved from the NCBI database and analyzed using bioinformatics tools. The nucleotide composition analysis revealed adenine as the most abundant nucleotide, with thymine predominating at the third codon positions. The Guanine-Cytocine (GC) content was balanced overall, with variations in GC content at the third codon position values suggesting mutational pressure. Relative synonymous codon usage analysis identified overrepresented and underrepresented codons, highlighting host-specific selection pressures. The Effective number of codons and neutrality plots indicated that natural selection predominantly influences codon usage bias in PSV, while mutational pressure contributes less. Chargaff's second parity rule analysis confirmed deviations influenced by these forces, while dinucleotide abundance analysis provided insights into codon usage trends. The codon adaptation index (CAI = 0.584) suggested

moderate adaptation of PSV to its natural host, *Sus scrofa domesticus*, reflecting evolutionary constraints on translational efficiency. Correspondence analysis highlighted factors driving viral evolution. These findings contribute to our understanding of PSV molecular evolution, supporting the development of antiviral strategies, vaccines, and diagnostic for disease control.

Keywords: Codon usage bias; Host adaptation; Molecular evolution; Porcine sapelovirus; Viral evolution.

Study contribution

This study makes significant contributions to our understanding of the evolutionary dynamics and host adaptation of Porcine *Sapelovirus* (PSV) by analyzing codon usage bias and nucleotide composition in its polyprotein gene. Through metrics such as relative synonymous codon usage, effective number of codons, codon adaptation index, and Chargaff's second parity rule, the research uncovers the roles of mutational pressure and natural selection in shaping codon usage patterns. The findings highlight regions with adenine and thymine nucleotides in the PSV genome, purine preference, and moderate codon usage bias, reflecting viral adaptation to its porcine host. Dinucleotide abundance analysis reveals genomic stability and evolutionary strategies, while neutrality and Chargaff's parity plots underscore the dominance of natural selection in codon usage evolution. These insights provide a foundation for understanding PSV molecular evolution and host-virus co-evolution, offering essential guidance for vaccine design, antiviral strategies, and disease control in the global swine industry.

Introduction

Sapelovirus, a recently recognized genus within the Picornaviridae family, comprises viruses that infect a wide range of animal species. This genus includes several species, most notably PSV, as well as simian, avian, and unclassified strains found in animals such as bats, marmots, and sea lions. PSV or *Sapelovirus anglia*, a positive-sense RNA virus, is a significant pathogen in swine, affecting both domestic pigs (*Sus scrofa domesticus*) and wild boar populations. This virus possesses a genome of approximately 7.5 kb, which encodes a single large open reading frame that is translated into a polyprotein. This polyprotein is subsequently cleaved by viral proteases into functional proteins, including both structural (VP1-VP4) and nonstructural (2A-3D) proteins, which are essential for viral replication and pathogenesis.^(1, 2)

The impact of PSV on swine health is significant, causing a range of clinical manifestations, from asymptomatic infections to severe diseases, including neurological disorders, respiratory distress, diarrhea, and reproductive failures.⁽¹⁻⁴⁾ The virus is primarily transmitted via the fecal-oral route and is often found in coinfections with other enteric pathogens, complicating diagnosis and control efforts. Despite its global distribution, the prevalence of PSV varies regionally, and the high rate of asymptomatic cases makes early detection and accurate diagnosis challenging.⁽⁴⁻⁷⁾

Codon usage bias (CUB), a non-random pattern in the use of synonymous codons encoding the same amino acid, plays a crucial role in the molecular evolution of viruses, including PSV. CUB is shaped by mutational pressure, natural selection, and genetic drift, and is influenced by factors such as GC content, gene length, RNA stability, and protein structure.⁽⁸⁻¹⁰⁾ In viral genomes, CUB is largely driven by the host's translational selection, which can influence key processes like mRNA synthesis, translation, protein folding, and

gene expression regulation.^(11–13) Despite its importance, the codon usage patterns in PSV have not been extensively studied, and this research seeks to fill that gap.

The polyprotein gene of PSV, which is critical for the viral life cycle, is an ideal target for studying codon usage bias. The translation of this polyprotein is crucial for the virus's ability to replicate and cause disease, making it a vital focus of molecular studies. This study aimed to investigate the codon usage patterns and nucleotide composition of PSV's polyprotein gene, providing a deeper understanding of the virus's evolutionary dynamics and its adaptation to its host, *Sus scrofa domesticus*. By analyzing parameters such as nucleotide composition, relative synonymous codon usage (RSCU), effective number of codons (ENC), and codon adaptation index (CAI), this research will contribute to a broader understanding of viral evolution and may aid in the development of improved diagnostic, therapeutic, and preventive strategies.

Materials and methods

Ethical statement

Ethical approval was not required for this study, as the data used in this investigation were publicly available.

Data assembly and sequencing editing

The National Center Biotechnology Information

(<https://www.ncbi.nlm.nih.gov/genbank/>)⁽¹⁴⁾ GenBank database yielded 34 full coding sequences (6996 bp) of the Porcine sapelovirus polyprotein gene, which were retrieved using the BLAST algorithm. Based on completeness, proper annotation, and geographic variety, these sequences, which were gathered from China, South Korea, Japan, and the USA with isolation years between 2005 and 2023, were selected and subjected to further bioinformatics analysis.⁽¹⁵⁾

Nucleotide composition analysis

A descriptive analysis of codon usage patterns was conducted using the CAI calculator program (<http://genomes.urv.es/CAIcal/>).⁽¹⁶⁾ The investigation focused on nucleotide composition by calculating the average frequencies of adenine (A), cytosine (C), guanine (G), and thymine (T), as well as the third codon regions (A3, C3, G3, T3). Likewise, the total GC content and the specific GC1, GC2, GC3, and GC12 at the first and second codon sites were examined, in addition to the computation of AT and GC biases.⁽¹⁷⁾

Dinucleotide abundance frequency analysis

Dinucleotide abundance frequencies were analyzed to reveal biases in dinucleotide selection affecting codon usage and genomic patterns, with overrepresentation defined as a frequency exceeding 1.23 and underrepresentation as a frequency below 0.78, determined by applying the formula:

$$P_{xy} = \frac{f_{xy}}{(f_x \times f_y)}$$

Where $f_x \times f_y$ indicates the expected frequency of the dinucleotide value. The observed frequency of the dinucleotide XY is represented by f_{xy} .^(18, 19)

Interpretation of relative synonymous codon usage (RSCU)

As this implies, RSCU analysis can assess how different genes use codons and offer proof of the factors influencing the intricate genetic code structure. Assuming either equal use of synonymous codons or the lack of bias in the coronagraph, RSCU for a specific codon will compare its observed frequency to the expected frequency over a specified set of synonymous codons. The number of times the amino acid it represents appears in the protein sequence and the total number of synonymous codons for that amino acid determine the anticipated frequency for a particular codon.⁽¹⁵⁾ The RSCU values have

been conducted using the formula $g_{ij}/(o_{ij}/n_i)$ where g_{ij} is the measured number of the i^{th} codon for the j^{th} amino acid, and n_i is the count of synonymous codons. Codons with RSCU values greater than 1.0 are considered positively biased (preferred codons), and those with values less than 1.0 are regarded as negatively biased (less preferred codons). Codons with RSCU values over 1.6 are seen as overrepresented, while those with values under 0.6 are seen as underrepresented; RSCU values between 0.6 and 1.6 indicate that the codons are chosen randomly or without bias.

Examination and visualization of the effective number of codons

An important metric for evaluating codon usage bias is the effective number of codons (ENC), which computes the difference between observed and expected standard usage. A score nearer 20 denotes a severe codon bias, with a single codon primarily encoding each amino acid, whereas a value nearer 61 suggests no bias, with synonymous codons employed evenly. An ENC score of 45 or fewer usually indicates significant codon usage bias, which is often the result of mutational pressure or natural selection. The value is assigned by an equation:

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

In this instance, the mean F_i for amino acids possessing i -fold degenerate codons is outlined to be F_i ($i = 2, 3, 4, \text{ and } 6$). Here, n_i denotes the number of synonymous codons for amino acids with i -fold degeneracy, and g_{ij} represents the observed frequency of the j -th codon for the corresponding amino acid. The formula for determining the F_i value is expressed as:

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

The *Sapelovirus* polyprotein gene's ENC values were ascertained using CodonW software. An ENC plot emerged with the aim of examining the relation between ENC values and GC3, which refers to the percentage of G and C at the third codon position. Furthermore, this graphic provides a means of detecting and quantifying absolute synonymous codon usage bias. The anticipated ENC values are calculated using the formula below:

$$ENC = 2 + s + \left(\frac{29}{s^2 + (1 - s)^2} \right)$$

Where s is the periodicity of G and C at synonymous codons' third position.

The factors influencing the codon usage bias are depicted in the ENC figure, which also demonstrates the link between GC3 (GC nucleotide at the third position) and ENC value. If the ENC values are on or close to the standard curve, the primary cause of codon usage bias is mutational pressure. Natural selection, on the other hand, is more significant in determining how the genes use their codons if the ENC values are far below the curve.^(18, 19)

Examination of Chargaff's second parity rule

The impact of natural selection and mutation pressure was assessed through the second parity rule (PR2) analysis of the third codon positions in organisms. The AT- and GC-bias values were calculated using the following formulas:

$$AT - bias = \frac{A_3}{A_3 + T_3}$$

And

$$GC - bias = \frac{G_3}{G_3 + C_3}$$

Where A_3 , T_3 , G_3 , and C_3 correspond to the third-position occurrences (frequencies or percentages) of adenine, thymine, guanine, and cytosine, respectively, in the analyzed gene sequences. The AT-bias is represented on the y-axis, while the GC-bias is represented on the x-axis. These values are illustrated in a PR2 bias plot rendered over R's ggplot2 resource. The neutral state of $A = T$ and $G = C$ was marked at the origin (0.5, 0.5). Greater deviation from this point highlighted coding preferences for purines or pyrimidines.^(18, 19)

Analysis of neutral evolution

The neutrality plot was constructed by plotting GC3 values on the x-axis against GC12 values on the y-axis, where GC12 represents the average of GC1 and GC2. This plot is used to assess the relative contributions of mutation pressure and natural selection to codon usage bias.^(18, 20) A regression line is fitted to the data, with a slope close to 1 indicating that mutation pressure predominantly shapes codon usage, while a slope near 0 suggests that natural selection is the primary factor. The analysis, performed using R software, highlights the linear relationship between GC3 and GC12 and quantifies the influence of each factor.

The codon adaptation index

To determine the degree to which highly expressed genes made use of favored codons, codon adaptation index (CAI) was computed.⁽²¹⁾ The CAI, which has a 0–1 range, indicates possible bias in gene expression and codon usage; larger values suggest a stronger preference for optimum codons. The *Sapelovirus* codon use database (<https://www.kazusa.or.jp/codon/>)⁽²²⁾ provided a reference set of RRSCU values that were used to calculate the CAI values for *Sapelovirus* coding sequences using the CAIcal

program.⁽¹⁶⁾ A comparison between the codon usage of *Sapelovirus* and that of its host, *Sus scrofa domesticus* (domestic pig), was also made easier by this database.

Correspondence analysis

Trends in codon use bias among genes were visualized using correspondence analysis (CA). A contingency table was used to express the data, with rows standing for genes and columns for codon counts.^(23, 24) Raw counts, counts corrected for amino acids, or relative synonymous codon usage values were used to extract trends. Therefore, in order to determine the primary causes of the patterns observed in codon usage, correspondence analysis was used, utilizing multivariate statistical methods.

Software and statistical packages

All statistical analyses and visualizations were performed using R software (version 4.3.2), with the following packages: CodonW for ENC calculation; FactoMineR and factoextra for correspondence analysis, and ggplot2 for generating PR2 bias plots, neutrality plots, ENC plots, and correspondence analysis plots.

Results

Nucleotide content and composition analysis

With an emphasis on the frequency of nucleotides (A, T, G, and C) and their third-position occurrences (A3, T3, G3, and C3), the nucleotide composition of the chosen polyprotein gene was reviewed using the CAIcal service, in furtherance of the GC content (GC, GC1, GC2, and GC3). The values presented reflect the mean \pm standard deviation across 34 full-length polyprotein gene sequences of *Sapelovirus*. The analysis demonstrated that adenine was the most frequent nucleotide, occurring at 31.09 ± 0.17 , followed by thymine at 28.72 ± 0.39 , guanine at 21.82 ± 0.16 , and cytosine at 18.35 ± 0.38 . In the third position of codons, thymine was the most commonly used nucleotide, with a frequency of

35.79 \pm 0.95, followed by adenine at 29.86 \pm 0.53, guanine at 18.71 \pm 0.51, and cytosine at 15.62 \pm 0.91. The mean GC content values were calculated as follows: GC1 (52.57 \pm 0.0038), GC2 (44.54 \pm 0.0019), GC3 (54.51 \pm 0.0090), and GC12 (48.56 \pm 0.0024)

[Figure 1].

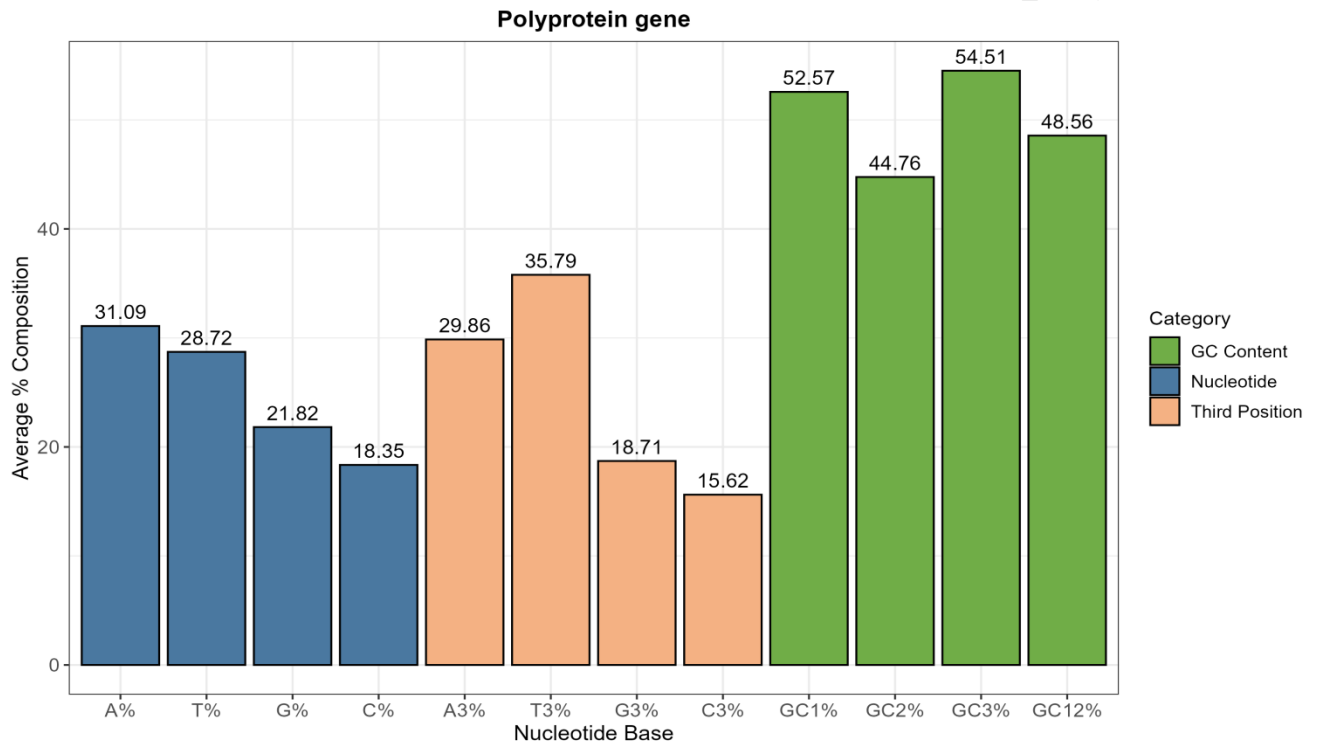


Figure 1. Bar graph representing the nucleotide composition of the polyprotein gene of *Sapelovirus*. The y-axis shows the mean percentage composition (\pm standard deviation) of each nucleotide and GC content metric, calculated across 34 full-length polyprotein gene sequences. Blue bars represent overall nucleotide frequencies (A %, T %, G %, C %), orange bars indicate third-position nucleotide usage (A3 %, T3 %, G3 %, C3 %), and green bars show GC content (GC1 %, GC2 %, GC3 %, GC12 %)

Dinucleotide abundance frequency analysis

Using RStudio software, the relative abundance frequencies of 16 dinucleotide pairs included in the *Sapelovirus* polyprotein gene were calculated. Dinucleotides having their abundance frequency greater than 1.23 were deemed overrepresented; however, frequencies below 0.78 were shown as underrepresented. Based on the results, CA (1.36), TG (1.35), and GC (1.46) were identified as being overrepresented, while CG (0.123) and TC (0.69) were shown as underrepresented (**Figure 2**).

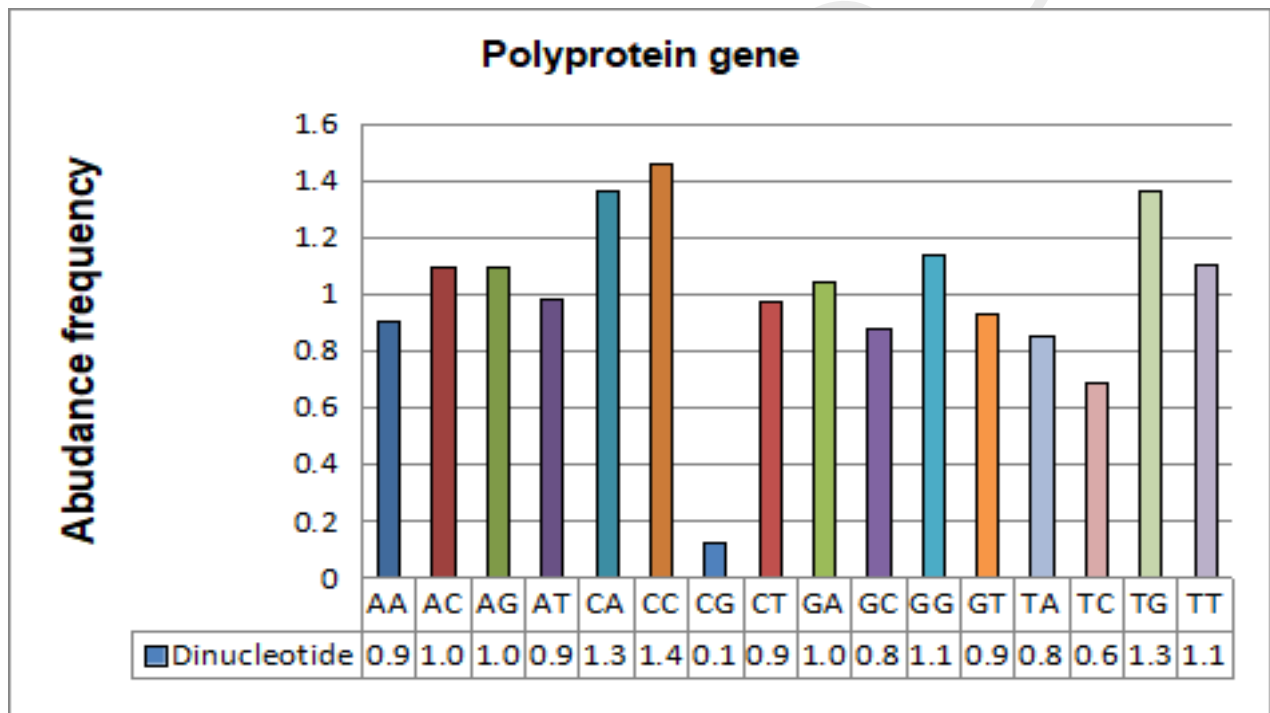


Figure 2. Relative dinucleotide abundance frequency of poly protein gene in *Sapelovirus*.

Sapelovirus's codon usage bias is influenced by the interaction of natural selection and mutational pressure.

Relative synonymous codon analysis

The RSCU values for polyprotein genes were calculated using MEGA11's MODEL method, which demonstrated synonymous codon frequencies ranging from 0.6 to 1.6. The RSCU study only included 61 of the 64 codons because the stop codons UAA, UAG, and UGA do not code for amino acids. The frequency of low-frequency or negatively biased codons is less than 1.0, whereas that of high-frequency or positively biased codons is greater than 1.0 (**Figure 3**). The RSCU analysis of the polyprotein gene in pigs identified 33 positively biased and 28 negatively biased codons out of the 61 codons analyzed. Meanwhile, it was demonstrated that seven codons were over-represented and seventeen were under-represented (**Table 1**).

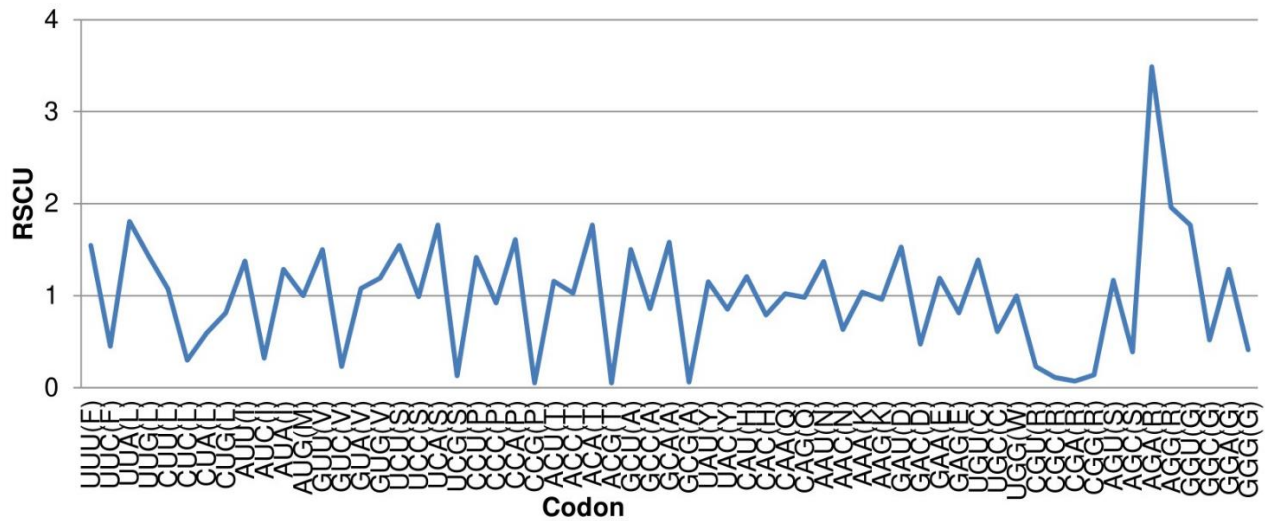


Figure 3. Overall frequencies of relative synonymous codon usage (RSCU) in *Sapelovirus* of polyprotein gene.

Table 1. Relative synonymous codon usage for each amino acid in the polyprotein gene of *Sapelovirus* in pig sequences

Aminoacid	Codon	RSCU	Aminoacid	Codon	RSCU	Aminoacid	Codon	RSCU
Phenylalanine	UUU	1.55	Proline	CCU	1.42	Lysine	AAA	1.04
	UUC	0.45		CCC	0.92		AAG	0.96
Leucine	UUA	1.81	Threonine	CCA	1.61	Aspartic acid	GAU	1.53
	UUG	1.43		CCG	0.05		GAC	0.47
	CUU	1.07		ACU	1.16	Glutamic acid	GAA	1.19
	CUC	0.3		ACC	1.03		GAG	0.81
	CUA	0.59		ACA	1.77	Cysteine	UGU	1.39
	CUG	0.81		ACG	0.05		UGC	0.61

Isoleucine	AUU	1.38	Alanine	GCU	1.5	Tryptophan	UGG	1	
	AUC	<i>0.32</i>		GCC	0.86		Arginine	CGU	<i>0.23</i>
	AUA	1.29		GCA	1.58			CGC	<i>0.11</i>
	AUG	1		GCG	<i>0.06</i>			CGA	<i>0.07</i>
Valine	GUU	1.5	Tyrosine	UAU	1.15	Glycine		CGG	<i>0.14</i>
	GUC	<i>0.23</i>		UAC	0.85		AGA	3.49	
	GUA	1.08		Histidine	CAU		1.21	AGG	1.96
	GUG	1.19			CAC		0.79	GGU	1.77
Serine	UCU	1.55	Glutamine	CAA	1.02	GGC	<i>0.52</i>		
	UCC	0.99		CAG	0.98	GGA	1.29		
	UCA	1.77	Asparagine	AAU	1.37	GGG	<i>0.41</i>		
	UCG	<i>0.13</i>		AAC	0.63				

Codons that are over-represented (> 1.6) are shown in **bold**, while those that are under-represented (< 0.6) are shown in *italic*.

Effective number of codons

The effective number of codons (ENC), a commonly used metric to assess codon usage bias, was calculated using codonW software to explore whether codon usage in specific genes is influenced by mutational or selective pressures. In this study, the ENC values of *Sapelovirus* coding sequences ranged from 44.73 to 47.69, with a mean value of 46.16, indicating a moderate level of codon usage bias. An ENC plot (**Figure 4**) was created using the R programming language to analyze the relative contributions of selection and mutational pressure. Each distinct colored point on the plot represents an analyzed gene. The consistent deviation of all data points below the standard ENC curve indicates that codon usage bias in *Sapelovirus* is not solely driven by mutational pressure. Instead, this pattern reflects a significant influence of natural selection acting on the polyprotein gene.

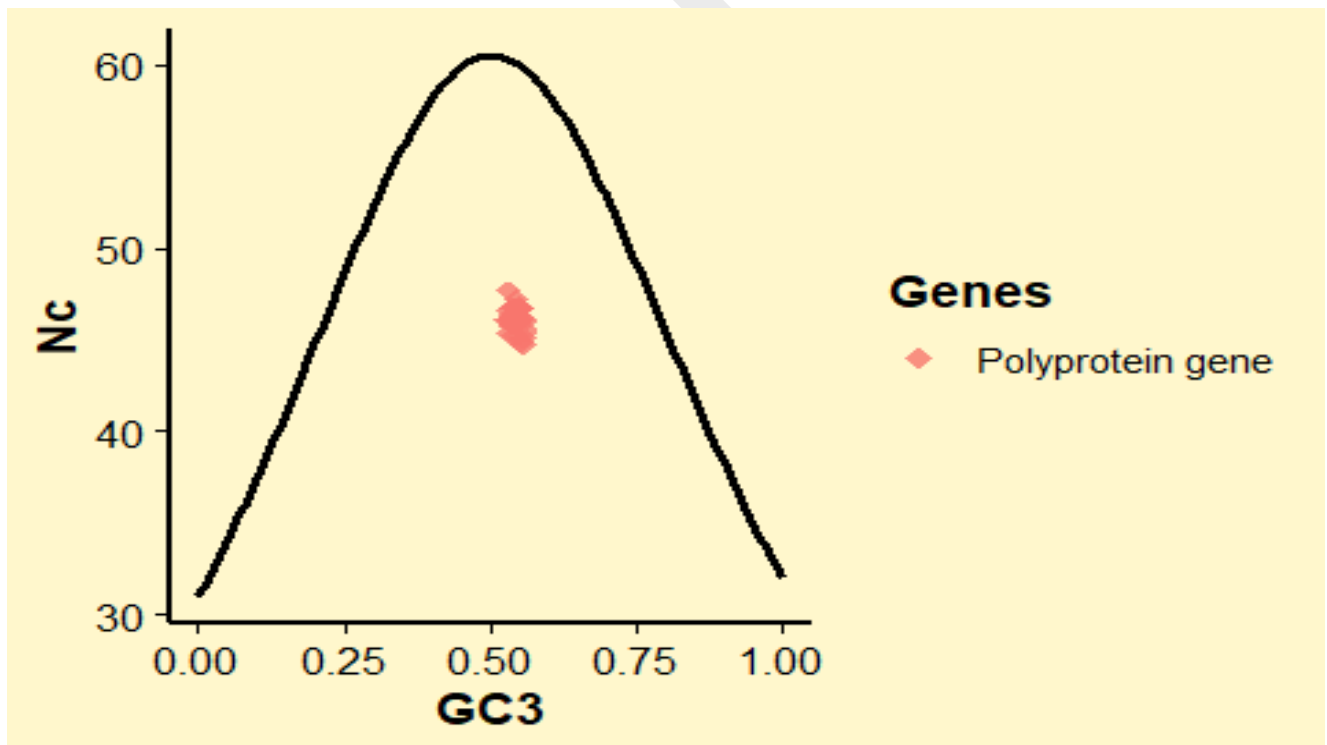


Figure 4. The association between codon frequency and GC3 values is illustrated through an ENC plot.

Chargaff's rule of parity

A = T and G = C are examples of near parity in the nucleotide composition, according to the Chargaff's rule of parity (PR2) rule, also known as Chargaff's second parity rule. The PR2 bias plot is used to assess departures from this equilibrium at the third codon location in the context of codon usage bias. This map, which shows the relative contributions of selection and mutational pressure to codon usage, was made by graphing, on the x-axis:

$$\frac{G3}{G3 + C3}$$

And on the y-axis:

$$\frac{A3}{A3 + T3}$$

A bias favoring purine (A and G) over pyrimidines (T and C) was found via PR2 analysis. Adenine and thymine were more common than guanine and cytosine, according to the AT bias values, which varied from 0.633 to 0.686 (**Figure 5**).

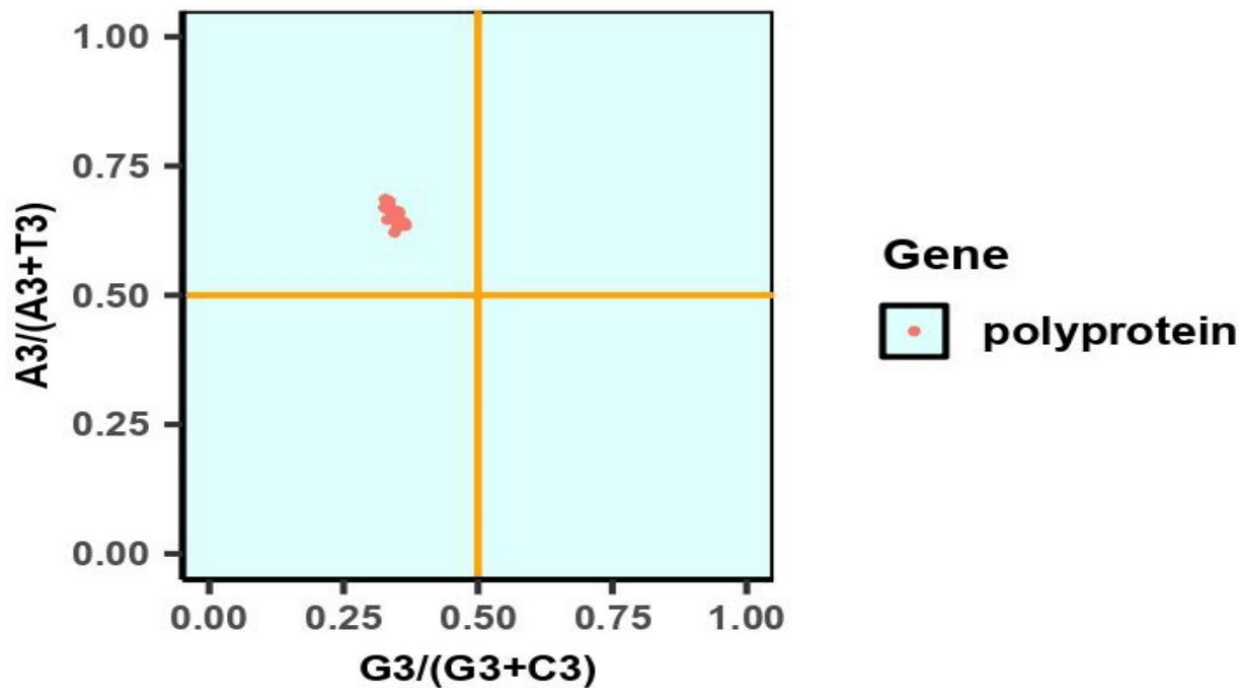


Figure 5. The parity rule 2 plot is used to visualize the relationship between AT-bias and GC-bias in the nucleotide sequences.

Analysis of neutral evolution

Connecting GC3 to GC12 (the mean of GC1 and GC2), the neutrality plot demonstrated a substantial positive regression:

$$y = 0.454 + 0.0578x$$

$$(R^2 = 0.05)$$

Where x represents the GC content at the third codon position (GC3), y denotes the mean GC content at the first and second codon positions (GC12), 0.454 is the intercept of the regression line (expected GC12 value when GC3 = 0), and 0.0578 is the slope indicating the rate of change in GC12 for each unit increase in GC3. The coefficient of determination ($R^2 = 0.05$) indicates that only 5 % of the variation in GC12 is explained by GC3. This

weak correlation suggests that mutational pressure alone does not entirely account for codon usage variation in the analyzed sequences, implying that natural selection and other evolutionary forces may also influence codon bias.

The research demonstrated that only 5.28 % of codon use bias is impacted by mutational pressure, whereas 94.72 % is driven by natural selection (**Figure 6**). This pattern depicts how natural selection outweighs mutational pressure.

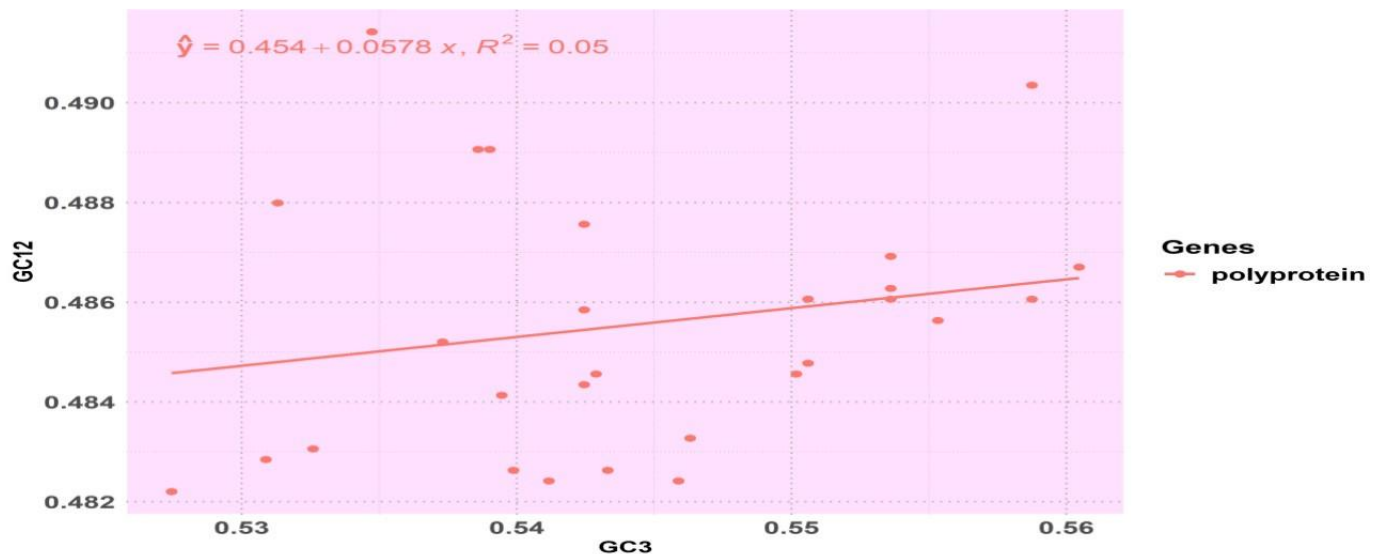


Figure 6. A neutrality plot depicts the connection between GC12 % and GC3 %, where the slope line signifies the role of natural selection in shaping the codon usage.

Index of codon adaptation

The efficiency of codon usage and the degree of host adaptation by the virus were measured using the codon adaptation index. By contrasting the reference organism's codon usage pattern with the virus's, the pig (*Sus scrofa domestica*), the CAI values were calculated. A greater degree of host adaptation would be shown by CAI values near 1, while a poorer adaptation would be shown by CAI values around 0. CAI values for the

Sapelovirus polyprotein gene are 0.584 ± 0.05 , according to the results obtained. This number indicates a modest codon use bias in the virus and falls within the typical 0–1 range.

Correspondence analysis

The correspondence analysis of codon usage demonstrated significant patterns across four axes, with the following key observations: Axis 1 explained the most variation, with CGA (0.53665) having the highest positive score, reflecting a strong association with purine-rich codons. Axis 2 was influenced by codons with higher GC content, such as CGC (0.20038) and CCG (0.22064), showing strong positive loadings, while CGA (0.34506) presented a notable negative value. Axis 3 further separated codons under additional selective forces, with CGC (0.18865) and CGG (0.19533) displaying high positive scores. Axis 4 captured structural influences, with GCG (0.39534) showing the highest positive value, suggesting its unique role in the codon usage landscape. Overall, Axis 1 and Axis 2 explained the major trends in codon usage bias, while Axes 3 and 4 reflected more subtle evolutionary and functional constraints present in the dataset (**Figure 7** and **Table 2**).

Table 2. Correspondence analysis (COA) of codon usage in the polyprotein gene of *Sapelovirus*

Codon	Axis 1	Axis 2	Axis 3	Axis 4
UUU	0.00182	-0.02071	0.02035	0.00444
UCU	0.00704	-0.00789	0.01527	-0.04296
UAU	-0.02115	-0.0844	0.0251	-0.02971
UGU	0.01505	0.00019	0.01992	0.01228

UUC	-0.00626	0.07111	-0.06987	-0.01525
UCC	-0.06758	0.03352	-0.01371	0.05928
UAC	0.02834	0.1131	-0.03364	0.03982
UGC	-0.03411	-0.00043	-0.04517	-0.02784
UUA	0.02658	-0.05895	-0.04043	0.00719
UCA	-0.00678	-0.01834	0.01298	0.00677
UUG	0.00882	0.01066	0.04946	-0.00156
UCG	0.23682	-0.02045	0.02396	-0.20951
CUU	0.02208	-0.00239	0.00251	0.02983
CCU	0.05764	0.00624	-0.08341	0.03919
CAU	0.01196	-0.04003	0.01988	-0.00352
CGU	0.0363	-0.18395	-0.14782	-0.01358
CUC	0.06927	0.01947	0.03288	-0.07884
CCC	-0.06932	-0.06903	0.08309	-0.0139
CAC	-0.01846	0.06179	-0.03069	0.00543
CGC	-0.16162	0.20038	0.18865	-0.10777
CUA	-0.12766	0.02276	-0.07374	-0.00524
CCA	-0.00532	0.02627	0.02886	-0.01466
CAA	0.0354	-0.04252	-0.013	-0.00751
CGA	0.53665	-0.34506	-0.12618	-0.07738
CUG	-0.03629	0.0924	0.04164	-0.01989
CCG	-0.18011	0.22064	-0.07411	-0.35402
CAG	-0.03723	0.04472	0.01367	0.0079
CGG	-0.32673	-0.22933	0.19533	0.13985
AUU	0.02677	-0.02743	0.0099	0.01284
ACU	-0.01012	0.01052	-0.0295	0.00843
AAU	-0.01557	-0.04378	0.00785	0.0216
AGU	0.0546	0.02281	-0.01049	0.02318
AUC	0.01558	0.18437	-0.04802	0.08846
ACC	0.01084	-0.0103	0.00533	-0.00185

AAC	0.03364	0.09458	-0.01696	-0.04667
AGC	-0.06822	-0.03201	-0.06069	-0.01008
AUA	-0.03247	-0.01643	0.00134	-0.03566
ACA	-0.00083	0.00363	0.02142	-0.00196
AAA	-0.01703	-0.0228	0.02017	-0.00147
AGA	0.07255	0.01396	0.0284	0.01313
ACG	0.04638	-0.17821	-0.20069	-0.09772
AAG	0.01825	0.02443	-0.02162	0.00158
AGG	-0.12194	0.01453	-0.05218	-0.02276
GUU	0.02331	0.00946	0.00323	0.00805
GCU	-0.0332	-0.04647	-0.04726	0.00014
GAU	-0.00653	-0.01448	0.0139	0.01682
GGU	0.01537	0.01621	-0.01314	-0.01869
GUC	0.00725	0.06144	-0.10292	-0.0366
GCC	0.01672	0.03068	0.07116	-0.02345
GAC	0.0212	0.047	-0.0451	-0.05459
GGC	-0.0368	0.03716	0.01486	0.03891
GUA	-0.0001	-0.00152	0.01321	0.05095
GCA	0.02676	0.03212	0.00622	-0.00207
GAA	-0.0109	0.00571	-0.01409	0.01555
GGA	0.01177	0.00109	0.0341	0.01984
GUG	-0.03064	-0.02237	0.00384	-0.04902
GCG	-0.12103	-0.13285	-0.01125	0.39534
GAG	0.01585	-0.00831	0.02051	-0.02263
GGG	-0.05634	-0.12071	-0.06979	-0.03164

Axis 1–4 scores for individual codons.

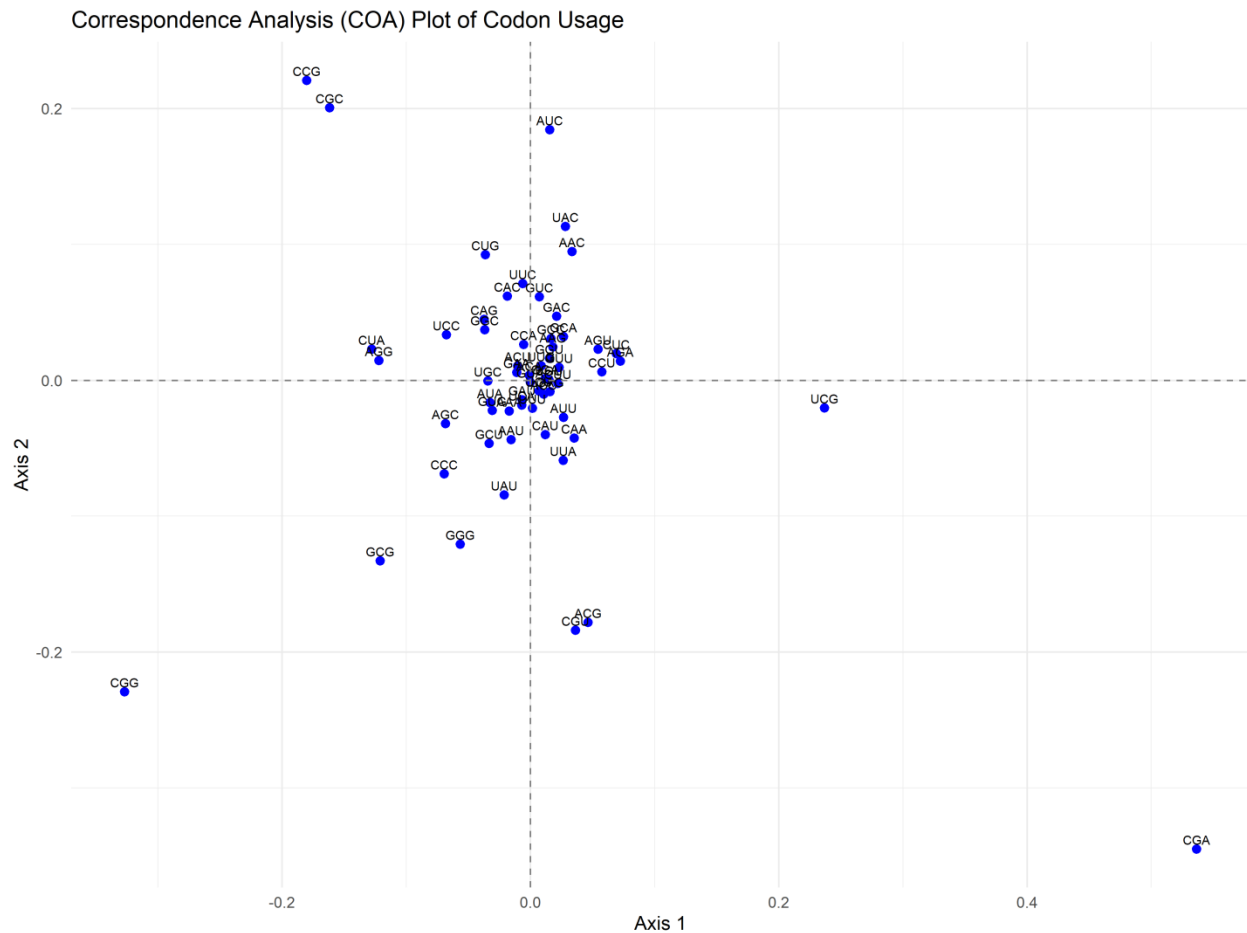


Figure 7. The correspondence analysis plot of the relative synonymous codon usage values for the polyprotein gene in *Sapelovirus* indicates that Axis 1 and Axis 2 explain the largest proportion of variation in codon usage patterns.

Discussion

The analysis of nucleotide composition, dinucleotide abundance, codon usage bias, and various associated metrics has provided significant insights into the molecular evolution and adaptation mechanisms of the polyprotein gene of *Sapelovirus*. This discussion synthesizes the key findings and interprets their biological significance. The nucleotide composition of the polyprotein gene reveals a predominance of adenine and thymine, with

adenine being the most frequent nucleotide. The third codon position emphasizes thymine as the most frequently used base, confirming the genome's AT-rich nature. This AT richness is consistent with observations in many RNA viruses, where mutational pressures and genome organization often contribute to such nucleotide patterns.^(10, 25)

The GC content values show positional variation, with GC1 and GC3 higher than GC2, suggesting that mutational forces may influence codon structure. Although ENC values indicate a moderate level of codon usage bias, the consistent deviation of all points below the expected ENC curve suggests that codon usage in *Sapelovirus* is not solely influenced by mutational pressure. This pattern, together with neutrality plot findings, supports that natural selection plays the predominant role, while mutational forces contribute to a lesser extent.

The relative synonymous codon usage (RSCU) analysis demonstrated distinct codon preferences, with some codons being overrepresented (e.g. AGA for arginine, GUU for valine) and others underrepresented (e.g. CCG for proline, CGC for arginine). These preferences may reflect adaptation to the host's translational machinery. The moderate codon adaptation index (CAI) value of 0.584 suggests that while the virus demonstrates some adaptation to its porcine host, it does not fully optimize codon usage to host preferences. This may indicate a balance between efficient translation and other evolutionary constraints.⁽¹⁶⁾

The analysis of dinucleotide abundance demonstrated biases in *Sapelovirus*, including overrepresentation of CA, TG, and GC dinucleotides, which may be important for maintaining genome stability or influencing RNA secondary structure.⁽²⁶⁾ The significant underrepresentation of CG dinucleotides aligns with patterns observed in many viruses, as methylation of CG sites can trigger host immune responses.⁽²⁵⁾ These biases likely

reflect evolutionary pressures aimed at evading host defenses while maintaining functional genome integrity.

The neutrality plot suggests that natural selection is the dominant force in shaping codon usage bias, with only 5.28 % of variation attributable to mutational pressure. This underscores the importance of adaptive evolution in *Sapeloivirus*, particularly in optimizing gene expression and protein function within the porcine host environment.⁽²⁷⁾ The parity rule 2 (PR2) plot further supports the notion that selection favors purines (A and G) over pyrimidines (T and C), a common feature in viral genomes under host-driven selection pressures.⁽²⁸⁾ Correspondence analysis of RSCU values provided insights into broader trends and finer details of codon usage bias. The separation of codons along Axis 1 and Axis 2, accounting for most of the variance, indicates that specific codons (e.g. CGA, GCG) are strongly influenced by selective forces and nucleotide composition. The clustering of codons based on GC content and purine preference highlights the interplay between mutational biases and selection-driven adaptation.

Implications for viral evolution and host adaptation

The findings suggest that both mutational pressures and natural selection shape the codon usage patterns of the polyprotein gene in *Sapeloivirus*. The moderate codon usage bias, coupled with evidence of adaptation to the porcine host, highlights the virus's evolutionary strategy to balance efficient replication with the evasion of host immune responses. The overrepresentation and underrepresentation of specific codons and dinucleotides likely contribute to optimizing viral fitness within the host environment.⁽¹⁰⁾ The moderate CAI values suggest that *Sapeloivirus* maintains flexibility in its codon usage, enabling it to adapt to different host strains or environmental conditions. This adaptability

may play a crucial role in the virus's persistence and transmission within swine populations.⁽¹⁶⁾

Future perspectives

The insights from this study have significant implications for understanding the evolution and host adaptation of *Sapelovirus*. Future research could explore the following:

- Functional implications: investigating the impact of codon usage bias on viral protein structure, function, and host immune recognition.
- Host-virus interactions: analyzing the co-evolutionary dynamics between *Sapelovirus* and its porcine host, particularly in regions of high codon bias.
- Therapeutic applications: codon usage insights for vaccine design and antiviral drug development, potentially targeting underrepresented or highly biased codons.
- Comparative analysis: examining codon usage patterns across different *Sapelovirus* strains to uncover conserved and divergent evolutionary trends.

Conclusions

In summary, the codon usage patterns of the *Sapelovirus* polyprotein gene are primarily driven by natural selection, with a moderate contribution from mutational pressure. The AT-rich genome, suppression of CG dinucleotides, and over-representation of specific codons reflect the influence of selective pressures to optimize translation and evade host immune responses. These findings enhance our understanding of viral evolution and adaptation mechanisms, providing insights for vaccine development and antiviral strategies.

Data availability

Additional resources are available from the corresponding author upon request.

Acknowledgments

The authors thank ICAR and the Director, ICAR-NIVEDI, for providing the necessary infrastructure and constant support throughout the study.

Funding statement

The compilation and evaluation of data in this study did not necessitate significant financial resources

Conflicts of interest

The authors declare that they have no conflicts of interest related to this study

Author contributions

Conceptualization: SS Patil, BR Gulati.

Data curation: V Mouna.

Investigation: SS Patil, KP Suresh

Methodology: V Mouna.

Project administration: SS Patil.

Resources: KP Suresh, Varsha Ramesh, M Shijili.

Supervision: SS Patil, KP Suresh.

Validation: SS Patil, BR Gulati.

Visualization: KP Suresh, Siju S Jacob, Jagadish Hiremath.

Writing-original draft: V Mouna.

Writing-review and editing: SS Patil, BR Gulati.

References

1. Ibrahim YM, Zhang W, Werid GM, Zhang H, Feng Y, Pan Y, *et al.* Isolation, characterization, and molecular detection of Porcine Sapelovirus. *Viruses*. 2022;14(2):349. doi: 10.3390/v14020349.
2. Yang T, Yu X, Yan M, Luo B, Li R, Qu T, *et al.* Molecular characterization of Porcine sapelovirus in Hunan, China. *Journal of General Virology*. 2017;98:2738–2747. doi: 10.1099/jgv.0.000951.
3. Kim DS, Kang MI, Son KY, Bak GY, Park JG, Hosmillo M, *et al.* Pathogenesis of Korean Sapelovirus A in piglets and chicks. *Journal of General Virology*. 2016;97(10):2566–2574. doi: 10.1099/jgv.0.000571.
4. Alexandersen S, Knowles NJ, Dekker A, Belsham GJ, Zhang Z, Koenen F. Picornaviruses. In: JJ Zimmerman, LA Karriker, A Ramirez, KJ Schwartz, GW Stevenson, editors. *Diseases of Swine*. 10th edition. West Sussex, UK: Wiley-Blackwell; 2012. pp. 587–620.
5. Donin DG, Leme RD, Alfieri AF, Alberton GC, Alfieri AA. First report of Porcine Teschovirus (PTV), Porcine Sapelovirus (PSV), and Enterovirus G (EV-G) in pig herds of Brazil. *Tropical Animal Health and Production*. 2014;46(3):523–528. doi: 10.1007/s11250-013-0523-z.
6. Cano-Gómez C, García-Casado MA, Soriguer R, Palero F, Jiménez-Clavero MA. Teschoviruses and Sapeloviruses in fecal samples from wild boar in Spain. *Veterinary Microbiology*. 2013;165(1–2):115–122. doi: 10.1016/j.vetmic.2012.11.022.
7. Prodělalová J. The survey of porcine teschoviruses, sapeloviruses, and enteroviruses B infecting domestic pigs and wild boars in the Czech Republic

- between 2005 and 2011. *Infection, Genetics and Evolution*. 2012;12(7):1447–1451.
doi: 10.1016/j.meegid.2012.04.025.
8. Qin H, Wu WB, Comeron JM, Kreitman M, Li WH. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics*. 2004;168(4):2245–2260. doi: 10.1534/genetics.104.030866.
 9. Deb B, Uddin A, Chakraborty S. Codon usage pattern and its influencing factors in different genomes of hepadnaviruses. *Archives of Virology*. 2020;165(3):557–570. doi: 10.1007/s00705-020-04533-6.
 10. Sharp PM, Li WH. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for ‘rare’ codons. *Nucleic Acids Research*. 1986;14(19):7737–7749. doi: 10.1093/nar/14.19.7737.
 11. Liu Y, Yang Q, Zhao F. Synonymous but not silent: the codon usage code for gene expression and protein folding. *Annual Review of Biochemistry*. 2021;90:375–401. doi: 10.1146/annurev-biochem-071320-112701.
 12. Parvathy ST, Udayasuriyan V, Bhadana V. Codon usage bias. *Molecular Biology Reports*. 2022;49(1):539–565. doi: 10.1007/s11033-021-06749-4.
 13. Zhang Z, Dong Y, Zhang M, Liu L, Chen S. Codon usage is an important determinant of gene expression levels. *Proceedings of the National Academy of Sciences of the United States of America*. 2016;113(50):14252–14257. doi: 10.1073/pnas.1606724113.
 14. National Center for Biotechnology Information. GenBank [internet]. Bethesda, Maryland, US: National Library of Medicine. <https://www.ncbi.nlm.nih.gov/genbank/>
 15. Ramesh V, Suresh KP, Mambully S, Rani S, Ojha R, Kumar KV, *et al.* Dynamic evolution of peste des petits ruminants virus in sheep and goat hosts across India

- reveals the swift surge of the F gene. *Virus Disease*. 2024;35(3):505–519. doi: 10.1007/s13337-024-00890.
16. Puigbò P, Bravo IG, Garcia-Vallvé S. CAIcal: a combined set of tools to assess codon usage adaptation. *Biology Direct*. 2008; 3:38. doi: 10.1186/1745-6150-3-38 Available from: <http://genomes.urv.es/CAIcal/>
17. Mandya Naganayak M, Kuralayanapalya Puttahonnappa S, Indrabalan UB, Paramanandham K, Jacob SS, Subramaniam S, Patil SS, *et al.* An extensive analysis of codon usage pattern, evolutionary rate, and phylogeographic reconstruction in Foot and Mouth Disease (FMD) serotypes (A, Asia 1, and O) of six major climatic zones of India: a comparative study. *Acta Tropica*. 2022(Dec8);236:106674. doi: 10.1016/j.actatropica.2022.106674.
18. Patil SS, Indrabalan UB, Suresh KP, Shome BR. Analysis of codon usage bias of classical swine fever virus. *Veterinary World*. 2021;14(6):1415–1422. doi: 10.14202/vetworld.2021.1415-1422.
19. Rani S, Mamathashree MN, Bharthi UI, Patil SS, Krishnamoorthy P, Shueb M, *et al.* Comprehensive examination on codon usage bias pattern of the bovine ephemeral fever virus. *Journal of Biomolecular Structure and Dynamics*. 2023;15(1):123. doi: 10.1080/07391102.2023.2258220.
20. Indrabalan UB, Suresh KP, Shivamallu C, Patil SS. An extensive evaluation of codon usage pattern and bias of structural proteins p30, p54, and p72 of the African swine fever virus (ASFV). *Virus Disease*. 2021;7:1–13. doi: 10.1007/s13337-021-0071
21. Sharp PM, Li WH. The codon adaptation index: a measure of directional synonymous codon usage bias and its potential applications. *Nucleic Acids Research*. 1987;15(3):1281–1295. doi: 10.1093/nar/15.3.1281.

22. Kazusa DNA Research Institute. Codon Usage Database [internet]. NCBI-GenBank Flat File Release 160.0 [June 15, 2007]. Kisarazu, Japan. <https://www.kazusa.or.jp/codon/>
23. Shueb M, Prasad SK, Suresh KP, Indrabalan UB, Beelagi MS, Shivamallu C, *et al.* The first study on analysis of the codon usage bias and evolutionary analysis of the glycoprotein envelope E2 gene of seven *Pestiviruses*. *Veterinary World*. 2022;15(7):1857–1868. doi: 10.14202/vetworld.2022.1857-1868.
24. Greenacre MJ. *Theory and Applications of Correspondence Analysis*. London, UK: Academic Press; 1984.
25. Greenbaum BD, Levine AJ, Bhanot G, Rabadan R. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PLoS Pathogens*. 2008;4(6):e1000079. doi: 10.1371/journal.ppat.1000079.
26. Karlin S, Mrázek J. Compositional differences within and between eukaryotic genomes. *Proceedings of the National Academy of Sciences of the United States of America*. 1997;94(19):10227–10232. doi: 10.1073/pnas.94.19.10227.
27. Butt AM, Nasrullah I, Qamar R, Tong Y. Evolution of codon usage bias in Zika virus genomes is host and vector-specific. *Emerging Microbes & Infections*. 2016;5(10):e107. doi: 10.1038/emi.2016.117.
28. Belalov IS, Lukashev AN. Causes and implications of codon usage bias in RNA viruses. *PLoS ONE*. 2013;8(2):e56642. doi: 10.1371/journal.pone.0056642.