

Article

Comprehensive Analysis of Codon Usage on Porcine Astrovirus

Huiguang Wu^{1,2,3,4}, Zhengyu Bao¹, Chunxiao Mou^{1,2,3}, Zhenhai Chen^{1,2,3,4,*} and Jingwen Zhao^{5,*}

¹ College of Veterinary Medicine, Yangzhou University, Yangzhou 225009, China;

huiguang-wu@yzu.edu.cn (H.W.); bzy199703@163.com (Z.B.); ytxmxcx@163.com (C.M.)

² Institute of Comparative Medicine, Yangzhou University, Yangzhou 225009, China

³ Jiangsu Co-Innovation Center for Prevention and Control of Important Animal Infectious Diseases and Zoonoses, Yangzhou University, Yangzhou 225009, China

⁴ Joint International Research Laboratory of Agriculture and Agri-Product Safety, The Ministry of Education of China, Yangzhou University, Yangzhou 225009, China

⁵ College of Animal Science and Technology, Yangzhou University, Yangzhou 225009, China

* Correspondence: zhenhai@yzu.edu.cn (Z.C.); zhaojingwen79@163.com (J.Z.)

Received: 25 July 2020; Accepted: 3 September 2020; Published: 6 September 2020



Abstract: Porcine astrovirus (PAstV), associated with mild diarrhea and neurological disease, is transmitted in pig farms worldwide. The purpose of this study is to elucidate the main factors affecting codon usage to PAstVs. Phylogenetic analysis showed that the subtype PAstV-5 sat at the bottom of phylogenetic tree, followed by PAstV-3, PAstV-1, PAstV-2, and PAstV-4, indicating that the five existing subtypes (PAstV1-PAstV5) may be formed by multiple differentiations of PAstV ancestors. A codon usage bias was found in the PAstVs-2,3,4,5 from the analyses of effective number of codons (ENC) and relative synonymous codon usage (RSCU). Nucleotides A/U are more frequently used than nucleotides C/G in the genome CDSs of the PAstVs-3,4,5. Codon usage patterns of PAstV-5 are dominated by mutation pressure and natural selection, while natural selection is the main evolutionary force that affects the codon usage pattern of PAstVs-2,3,4. The analyses of codon adaptation index (CAI), relative codon deoptimization index (RCDI), and similarity index (SiD) showed the codon usage similarities between the PAstV and animals might contribute to the broad host range and the cross-species transmission of astrovirus. Our results provide insight into understanding the PAstV evolution and codon usage patterns.

Keywords: porcine astrovirus (PAstV); phylogenetic analysis; codon usage pattern; natural selection; mutation pressure; host adaptability

1. Introduction

Porcine astroviruses (PAstVs), comprising five distinct lineages (PAstV1-PAstV5), are highly prevalent in both diarrheic and clinically healthy pigs [1–3]. PAstV-3 is found in tissues from the central nervous system of piglets and sows with encephalomyelitis and neural necrosis [4,5]. Infection of piglets with PAstV-1 could cause mild diarrhea, growth retardation, and damage to the villi of the small intestinal mucosa [6]. PAstV-4 was detected in the nasal swabs [7] and the feces of pigs [8,9]. Co-infection of individual pigs with several lineages of PAstVs has also been observed [9]. Both PAstV-2 and PAstV-5 have been identified in the brains of newborn piglets with congenital tremors [10]. PAstV-2 and PAstV4 were simultaneously detected in the blood samples of apparently healthy domestic pigs, while the coexistence of PAstV-2, PAstV-4, and PAstV-5 has been observed in porcine fecal samples collected from the same farms [9,11].

PAstV is a non-enveloped, single-stranded, positive-sense RNA virus belonging to the genus *Mamastrovirus* within the family *Astroviridae*. The PAstV genome contains three overlapping open reading frames (ORFs) encoding nonstructural proteins (ORF1a), RNA-dependent RNA polymerase (ORF1b), and viral capsid proteins (ORF2) [9]. As a typical genome organization of the genus *Mamastrovirus*, the highly conserved ribosomal frameshifting signal (5'-AAAAAAC-3') is present precisely in the overlapping region of ORF1a and ORF1b of PAstV, which is essential for the translation of RNA polymerase [12]. The expression of ORF1b is mediated through programmed ribosomal frameshifting into ORF1b [13]. The PAstV capsid protein controls the initial phases of virus infection, including virus attachment, endocytosis, and genome release into the host cells [14].

The codon usage pattern reveals the basic features of molecular evolution [15]. Except for the codons of methionine (Met) and tryptophan (Trp), others encoding the same amino acid are termed synonymous codons. The codon usage of a species is not random, because some synonymous codons are more frequently used than others. This phenomenon of non-randomness in synonymous codon usage is called codon bias [16]. The viral codon usage may be influenced by the host, because the viral replication needs its host machinery. Although previous analysis revealed that host-affected nucleic acid composition and codon usage were the drivers of *Astroviridae* evolution [17], knowledge of codon usage patterns of PAstV is still very limited. Consequently, we employed a number of methods to investigate the evolutionary processes of PAstV by codon usage pattern analysis in this study.

2. Materials and Methods

2.1. Sequence Data

In this study, 69 complete coding sequences of PAstVs genomes were retrieved from the National Center for Biotechnology Information (NCBI) nucleotide database (<http://www.ncbi.nlm.nih.gov>) recorded up to January 2020. Considering the limited number of sequences, PAstV-1 was excluded from the study of codon usage. To investigate overall codon usage bias in the PAstVs genomes, stop codons of ORF1a and ORF1b were removed, and ORFs were concatenated in the following order: ORF1a-ORF1b-ORF2. Detailed information concerning PAstV genomes, including the accession number, strain name, geographical distribution of isolated strains, and the isolation year, is listed in Table S1.

2.2. Phylogenetic Analysis

The coding DNA sequences (CDSs) of PAstV genomes were aligned using MACSE [18] (version 2.03). The best-fit model of nucleotide substitution was identified using jModelTest2 [19] (version 2.1.10) according to corrected Akaike Information Criterion (AICc). The general time-reversible (GTR) with a gamma-distributed evolutionary rates (G) and invariant sites (I) (GTR + G + I) were selected as the best-fit model of nucleotide substitution based on the criteria mentioned above. The maximum likelihood (ML) phylogenetic tree was inferred using RAxML [20] (version 8.2.12) with model GTR + G + I. The node support was assessed by performing 10,000 bootstrap iterations. The Bayesian inference (BI) phylogeny was reconstructed using MrBayes [21] (version 3.2.7a) with the model GTR + G + I. The Markov Chain Monte Carlo (MCMC) search was conducted for 10,000,000 generations, and the posterior probabilities were estimated for each node. The first 25% of sampled trees were discarded as burn-in. The phylogenetic trees were viewed using Figtree [22] (version 1.4.4).

2.3. Nucleotide Composition

Considering the limited number of sequences, PAstV-1 was excluded, while the remaining four subtypes of PAstVs (PAstV-2 to PAstV-5) were investigated. The frequencies of mononucleotides (A, C, U, and G), GC, GC1s (GC content at the first codon positions), GC2s (GC content at the second codon positions), GC12s (the mean value of GC1s and GC2s), and GC3s (GC content at the third codon positions) were calculated using the seqinr package (version 3.6–1) [23] of R (version 3.6.2) [24]. The frequencies

of mononucleotides at the third synonymous codon position (A3s, C3s, U3s, and G3s) were calculated using CodonW software (version 1.4.2) developed by J. Peden (<http://codonw.sourceforge.net/>).

To evaluate the possible effects of relative dinucleotide abundances on codon usage of PAsTV, the odds ratio of observed to expected dinucleotide (O/E) frequency was computed according to the equation below [25]:

$$P_{xy} = \frac{f_{xy}}{f_x f_y} \quad (1)$$

where f_x , f_y , $f_x f_y$, and f_{xy} represent the frequency of nucleotide X, nucleotide Y, the expected frequency of dinucleotide XY, and the frequency of dinucleotide XY, respectively.

2.4. Effective Number of Codons (ENC)

The effective number of codons is designed to quantify how far the codon usage of a gene departs from equal usage of synonymous codons, regardless of the gene lengths and the number of amino acids [26]. The value of ENC ranges from 20 (if only one synonymous codon is exclusively used for the corresponding amino acid) to 61 (if all of the synonymous codons are used with no preference) [26,27]. The smaller the ENC value of a gene is, the stronger the extent of codon preference of this gene. ENC values were calculated using the following Equation [26]:

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

where F_i ($i = 2, 3, 4, 6$) represents the mean of F_i values for the i -fold degenerate amino acids. F_i can be calculated using the equation below [26]:

$$F_i = \frac{n \sum_{j=1}^i \left(\frac{n_j}{n}\right)^2 - 1}{n - 1} \quad (3)$$

where n represents the total number of observed codons for that amino acid; and n_j represents the total observed number of the j th codon for that amino acid. The ENC values for PAsTVs CDSs were calculated using the cordon package (version 1.4.0) [28] of R (version 3.6.2) [24].

2.5. Relative Synonymous Codon Usage (RSCU)

To measure the non-uniform usage of synonymous codons in a coding sequence, RSCU is defined as the ratio of observed to expected codon frequency under equal codon usage without being affected by the amino acid compositions or the CDS sizes of different gene samples [16]. Synonymous codons with RSCU values <1.0 , $=1.0$, and >1.0 represent negative codon usage bias, no bias, and positive codon usage bias, respectively. Furthermore, synonymous codons with RSCU values >1.6 and <0.6 were regarded as “overrepresented” and “underrepresented” codons, respectively [29]. The RSCU was calculated as:

$$RSCU = \frac{g_{ij}}{\sum_j^{n_i} g_{ij}} n_i \quad (4)$$

where g_{ij} represents the observed number of the i th codon for the j th amino acid, which is encoded by n_i synonymous codons [30]. The RSCU index was calculated for each sequence using the seqinr package (version 3.6–1) [23] of R (version 3.6.2) [24].

2.6. Principal Component Analysis (PCA)

Principal component analysis (PCA) is a multivariate statistical method that reduces data dimensionality by performing a covariance analysis between factors [31]. To investigate the dominant patterns and variations in the codon usage of PAsTVs CDSs, we performed a PCA with the RSCU values of the PAsTVs genome. For transforming the RSCU values into uncorrelated variables, the RSCU

value of each PAsTV sequence was distributed into a 59-dimensional vector corresponding to the 59 synonymous codons by excluding AUG, UGG, and three terminal codons. A matrix comprising 59 RSCU values of each sequence was built for the PCA and transformed into several major axes. PCA was performed on the obtained RSCU dataset by using the factextra package (version 1.0.6) [32] of R (version 3.6.2) [24].

2.7. ENC-Plot Analysis

In the ENC-plot analysis, the projection of ENC-values versus GC3s is commonly used to explore factors influencing the codon usage patterns, e.g., selection [26]. In an ENC plot, the observed and expected ENC values are compared to determine the influence of structuring synonymous codon usage bias. The expected ENC values for all of the GC3 compositions, ranging from 0 to 1, were calculated using the following equation:

$$ENC_{expected} = 2 + s + \frac{29}{s^2 + (1-s)^2} \quad (5)$$

where s is the frequency of G + C at the third codon position of synonymous codons. An expected curve was generated using the expected ENC values. In the ENC-GC3s plot, if observed ENC values fell on the curve of expected ENC values, it meant that mutation was the main force acting on third-position bases of codons, whereas if observed ENC values fell considerably below the expected curve, it meant that selection was the main force driving codon usage bias.

2.8. Neutrality Plot Analysis

A neutrality plot was used to identify the effects of natural selection and mutation pressure on the codon usage patterns [33]. The obtained GC3 and GC12 values (means of GC1 and GC2) of the synonymous codons were plotted on the horizontal and vertical axes, respectively, to produce a scatter diagram for the neutrality plot. The regression line was plotted between the GC3-variable and the GC12-variable. The slope (regression coefficient) of the regression line is regarded as the mutation-selection equilibrium coefficient [33]. If all of the points are distributed along the diagonal (slope = 1) and the correlation between GC3-variable and GC12-variable is statistically significant, this indicates that mutation is the main force shaping the codon usage. Alternatively, if the regression curve is parallel or tilted toward the horizontal axis (close to zero slope), selection is considered as the dominant factor. The regression analysis, which estimates the linear relationship between GC3-variable and GC12-variable, was performed by using R (version 3.6.2) [24].

2.9. Parity Rule 2 (PR2) Analysis

Parity rule 2 (PR2) plot analysis is another method used to investigate the influence of mutation and selection on codon usage. In the PR2 plot, AT bias [$A_3/(A_3 + T_3)$] and GC bias [$G_3/(G_3 + C_3)$] were chosen as the ordinate and abscissa, respectively. The center of the plot, i.e., $A = U$ and $G = C$ (PR2), defined as coordinates of the origin (0.5, 0.5), indicates no bias between the influences of mutation pressure and natural selection [34,35].

2.10. Codon Adaptation Index (CAI) Analysis

Codon adaptation index (CAI) is a quantitative measure for assessing the codon usage similarities between viral genes and their hosts [36]. The values of CAI range from 0 to 1. The virus sequences with higher CAI values are considered to be preferred over those with lower CAI values. The CAI analysis of the PAsTV coding sequences was performed with CAIcal [36]. The reference datasets of synonymous codon usage patterns of chicken (*Gallus gallus*), duck (*Anas platyrhynchos platyrhynchos*), human (*Homo sapiens*), dog (*Canis lupus familiaris*), horse (*Equus caballus*), mouse (*Mus musculus*), pig (*Sus scrofa*),

cat (*Felis catus*), cattle (*Bos taurus*), goat (*Capra hircus*) and sheep (*Ovis aries*) were downloaded from the Codon and Codon Pair Usage Tables (CoCoPUTs) database updated in January 2020 [37].

2.11. Relative Codon Deoptimization Index (RCDI)

RCDI is used to estimate the codon usage deoptimization trend of a virus to its host [38]. In a virus, the RCDI value of 1 indicates the virus has complete host-adapted codon usage pattern, while a value of RCDI higher than 1 indicates low adaptability to a host [39]. RCDI values of PAsTV sequences were calculated using CAIcal [36].

2.12. Similarity Index (SiD) Analysis

To measure the influence of the codon usage patterns of the host on codon usage bias of PAsTVs CDSs, a SiD analysis was performed. SiD is calculated using the following equation:

$$R(A, B) = \frac{\sum_{i=1}^{59} a_i \times b_i}{\sqrt{\sum_{i=1}^{59} a_i^2 \times \sum_{i=1}^{59} b_i^2}} \quad (6)$$

$$D(A, B) = \frac{1 - R(A, B)}{2} \quad (7)$$

where $R(A, B)$ is defined as the cosine of the angle included between the A and B spatial vectors; a_i means the RSCU value of 59 synonymous codons of the PAsTV coding sequence; b_i is the RSCU value of the same codon in the host; and $D(A, B)$ represents the potential effect of the overall codon usage of the host on that of PAsTV [40]. A high value of SiD indicates that the host has dominant effects on the codon usages of the virus.

2.13. Statistical Analysis

A non-parametric Kruskal–Wallis test was separately used to determine any significant differences between the values of CAI, RCDI and SiD of the four PAsTV subtypes. The p -values for Dunn's multiple comparisons were adjusted with the Benjamini–Hochberg method. The level of significance was set at $p < 0.05$. The statistical analysis was performed with the package `dunn.test` (version 1.3.5) [41] of R (version 3.6.2) [24].

3. Results

3.1. PAsTV-5 Subtype Has the Basal Position of Phylogenetic Tree

To explore the phylogeny among the five subtypes of PAsTV strains, we constructed the phylogenetic trees using ML and BI methods. Results showed the topologies of the BI and ML trees were identical (Figure 1). The subtype PAsTV-5 was a basal clade in the BI and ML trees, followed by PAsTV-3, PAsTV-1, PAsTV-2, and PAsTV-4, suggesting the ancestors of PAsTV may have undergone multiple differentiations before forming the five existing subtypes.

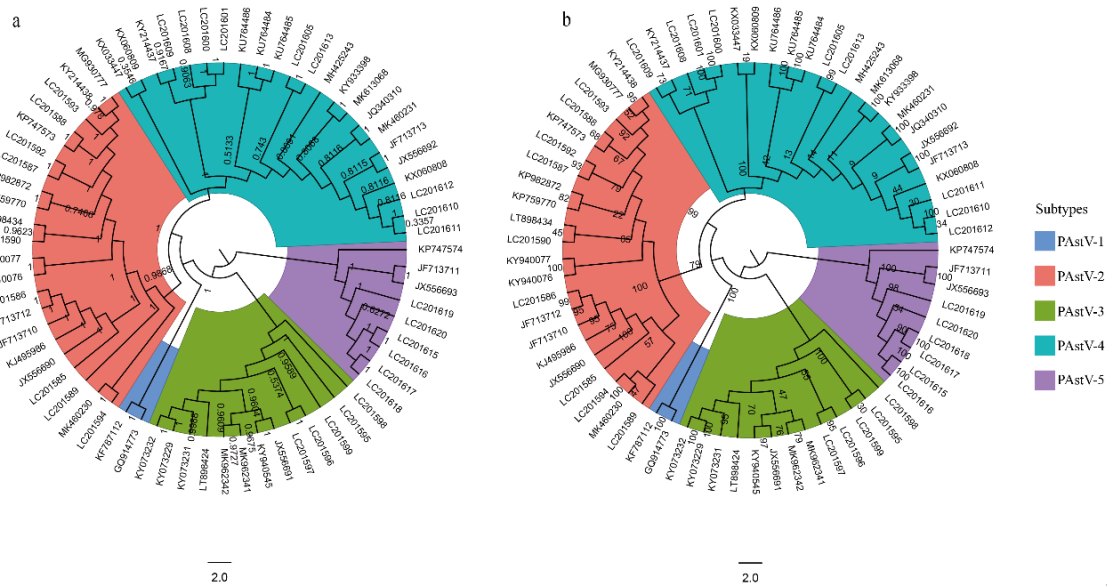


Figure 1. Phylogenetic trees of 69 complete genomes of PASTV. (a) BI tree of the PASTV genomes reconstructed by MrBayes. Posterior probability values are shown at each node. (b) ML tree of the PASTV genomes reconstructed by RAxML. Bootstrap support values are indicated on the tree as a percentage of 10,000 replicates. The colored circular sectors indicate the five subtypes of PASTV.

3.2. A and U Nucleotides Were More Frequent than C and G in the PASTV Coding Sequences

Nucleotide A was the most represented in the genome CDSs of PASTV-2 (0.268 ± 0.004) and PASTV-4 (0.301 ± 0.003) (Table 1 and Table S2). Content of nucleotide U was highest in the genome CDSs of PASTV-3 (0.279 ± 0.002) and PASTV-5 (0.269 ± 0.004). Further analysis of third-position wobble nucleotides revealed that U3 was the highest in the genome of PASTV-3 (0.433 ± 0.01), PASTV-4 (0.399 ± 0.011), and PASTV-5 (0.421 ± 0.009), whereas C3 was the most dominant in PASTV-2 (0.371 ± 0.011). The GC content in the coding sequence of PASTV was not uniformly distributed. In all of the four PASTV subtypes, the GC1 content was the highest compared to the contents of GC2, GC12, and GC3. The frequency values of GC3s of PASTV-2, PASTV-3, and PASTV-5 strains were higher than those of GC2s. The PASTV-2 has the highest ENC value (56.265 ± 0.602), followed by PASTV-5 (53.647 ± 0.316), PASTV-3 (53.059 ± 0.656), and PASTV-4 (52.007 ± 0.678). GC3 content was highest in PASTV-2 (0.534 ± 0.012), followed by PASTV-5 (0.467 ± 0.008), PASTV-3 (0.456 ± 0.01), and PASTV-4 (0.397 ± 0.014). High ENC values (>50) with a small amount of variation suggest a moderate bias but highly conserved codon usage pattern among the four subtypes of PASTVs. In summary, the overall and third-position nucleotide composition suggest that compositional constraints affect the codon usage pattern of PASTV genomes, and the PASTV CDSs were found to be rich in A/U nucleotides in comparison to G and C nucleotides.

Table 1. The nucleotide composition and properties of CDS of the PASTV genomes.

Catalogs	PASTV-2	PASTV-3	PASTV-4	PASTV-5	All
A	0.268 ± 0.004	0.247 ± 0.002	0.301 ± 0.003	0.238 ± 0.002	0.271 ± 0.024
C	0.263 ± 0.004	0.234 ± 0.003	0.222 ± 0.004	0.259 ± 0.005	0.243 ± 0.019
G	0.244 ± 0.003	0.24 ± 0.002	0.227 ± 0.003	0.234 ± 0.001	0.236 ± 0.008
U	0.225 ± 0.004	0.279 ± 0.002	0.25 ± 0.004	0.269 ± 0.004	0.25 ± 0.021
GC	0.507 ± 0.005	0.474 ± 0.004	0.449 ± 0.006	0.493 ± 0.005	0.479 ± 0.025
GC1s	0.549 ± 0.006	0.541 ± 0.003	0.528 ± 0.004	0.565 ± 0.007	0.543 ± 0.014
GC2s	0.438 ± 0.004	0.424 ± 0.003	0.422 ± 0.007	0.448 ± 0.003	0.431 ± 0.011
GC12s	0.494 ± 0.003	0.483 ± 0.003	0.475 ± 0.005	0.506 ± 0.004	0.487 ± 0.012
GC3s	0.534 ± 0.012	0.456 ± 0.01	0.397 ± 0.014	0.467 ± 0.008	0.463 ± 0.058

Table 1. Cont.

Catalogs	PAstV-2	PAstV-3	PAstV-4	PAstV-5	All
U3s	0.308 ± 0.014	0.433 ± 0.01	0.399 ± 0.011	0.421 ± 0.009	0.379 ± 0.052
C3s	0.371 ± 0.011	0.283 ± 0.008	0.241 ± 0.013	0.312 ± 0.011	0.302 ± 0.056
A3s	0.286 ± 0.011	0.256 ± 0.006	0.384 ± 0.01	0.245 ± 0.006	0.308 ± 0.057
G3s	0.276 ± 0.009	0.255 ± 0.008	0.235 ± 0.008	0.237 ± 0.005	0.253 ± 0.02
ENC	56.265 ± 0.602	53.059 ± 0.656	52.007 ± 0.678	53.647 ± 0.316	53.83 ± 1.902

3.3. RSCU Patterns of PAstV

The codons of UGU[Cys] and GAU[Asp] were preferably used by all four PAstV subtypes. The preferred codon usage profiles of PAstV-3 and PAstV-5 were very similar: 16 out of 18 preferred codons were commonly used, with exceptions for the preferred codons of isoleucine and glutamine (Table 2). Among the 18 most abundantly used codons in the PAstV-2 genomes, 13 (GCC, GAG, UUC, GGC, CAC, AUC, AAG, CUC, AAC, CAG, CGC, GUC, and UAC) codons were C/G-ended, and the remaining 5 (UGU, GAU, CCA, UCA, and ACA) codons were A/U-ended. The numbers of A/U-ended preferred codons in the PAstV-3, PAstV-4 and PAstV-5 were 16, 16, and 14, respectively. From the RSCU analysis, PAstV-2 exhibited comparatively higher codon usage bias towards C/G-end, and PAstV-3, PAstV-4, and PAstV-5 showed relatively higher codon usage bias towards A/U-ended codons. The results of the preferred codons were consistent with those of the nucleotide composition analysis. Among 59 codons, only two preferred codons, CUC[Leu] and CCA[Pro], were overrepresented (RSCU values >1.6) in the PAstV-2 genomes, while the PAstV-3, PAstV-4, and PAstV-5 genomes contained five (CUU[Leu], CCU[Pro], CGU[Arg], UCU[Ser], and GUU[Val]), six (GCA[Ala], CUU[Leu], CCA[Pro], AGG[Arg], UCA[Ser], and ACA[Thr]), and six (GGU[Gly], CUU[Leu], CGU[Arg], UCU[Ser], ACU[Thr], and GUU[Val]) overrepresented (RSCU values > 1.6) codons, respectively. PCA results showed that the first two principal axes accounted for 41.9% and 24.6% of the total variation of RSCUs (Figure S1). The points representing the four subgroups of PAstV genomes were mapped and clustered in clearly separate regions, although a small degree of overlap existed between PAstV-3 and PAstV-5 (Figure 2). Taken together, the RSCU analyses revealed the RSCU patterns of four PAstV genotypes, and compositional constraints of third position nucleotides in codons (G/C-ending codons of PAstV-2 versus A/T-ending codons of PAstV-3, PAstV-4 and PAstV-5) had the most influence on the selection of the preferred codons. The trend of the 59 synonymous codon usages indicated that the evolution of the four genotypes of PAstVs might be influenced.

Table 2. The relative synonymous codon usage (RSCU) of the CDS of PAstV strains. RSCU values of 59 synonymous codons excluding codons for Met and Trp are presented. The letter in brackets represents the abbreviation of the corresponding amino acid. The numbers of preferentially used synonymous codons are highlighted in green and orange colors. The RSCU values of the over-represented (RSCU > 1.6) codons are shaded using orange color.

Codon	PAstV-2	PAstV-3	PAstV-4	PAstV-5	All(G)
GCA(A)	1.129 ± 0.124	1.097 ± 0.084	1.813 ± 0.159	1.016 ± 0.05	1.343 ± 0.366
GCC(A)	1.331 ± 0.131	1.202 ± 0.074	0.874 ± 0.099	1.266 ± 0.103	1.14 ± 0.225
GCG(A)	0.568 ± 0.064	0.365 ± 0.061	0.34 ± 0.058	0.275 ± 0.049	0.411 ± 0.128
GCU(A)	0.972 ± 0.15	1.335 ± 0.09	0.972 ± 0.124	1.443 ± 0.082	1.106 ± 0.23
UGC(C)	0.98 ± 0.171	0.781 ± 0.187	0.71 ± 0.1	0.895 ± 0.18	0.837 ± 0.191
UGU(C)	1.02 ± 0.171	1.219 ± 0.187	1.29 ± 0.1	1.105 ± 0.18	1.163 ± 0.191
GAC(D)	0.961 ± 0.086	0.803 ± 0.06	0.692 ± 0.077	0.979 ± 0.102	0.84 ± 0.147
GAU(D)	1.039 ± 0.086	1.197 ± 0.06	1.308 ± 0.077	1.021 ± 0.102	1.16 ± 0.147
GAA(E)	0.834 ± 0.08	0.825 ± 0.08	1.163 ± 0.077	0.884 ± 0.075	0.952 ± 0.172
GAG(E)	1.166 ± 0.08	1.175 ± 0.08	0.837 ± 0.077	1.116 ± 0.075	1.048 ± 0.172
UUC(F)	1.072 ± 0.096	0.679 ± 0.055	0.702 ± 0.099	0.874 ± 0.121	0.842 ± 0.196
UUU(F)	0.928 ± 0.096	1.321 ± 0.055	1.298 ± 0.099	1.126 ± 0.121	1.158 ± 0.196

Table 2. Cont.

Codon	PAstV-2	PAstV-3	PAstV-4	PAstV-5	All(G)
GGA(G)	0.791 ± 0.118	0.531 ± 0.081	1.055 ± 0.092	0.486 ± 0.086	0.79 ± 0.244
GGC(G)	1.189 ± 0.117	1.332 ± 0.119	0.896 ± 0.119	1.092 ± 0.065	1.103 ± 0.2
GGG(G)	0.88 ± 0.138	0.629 ± 0.065	0.729 ± 0.079	0.624 ± 0.099	0.745 ± 0.144
GGU(G)	1.14 ± 0.146	1.508 ± 0.069	1.32 ± 0.124	1.798 ± 0.1	1.362 ± 0.247
CAC(H)	1.183 ± 0.127	0.898 ± 0.105	0.688 ± 0.087	0.775 ± 0.068	0.903 ± 0.234
CAU(H)	0.817 ± 0.127	1.102 ± 0.105	1.312 ± 0.087	1.225 ± 0.068	1.097 ± 0.234
AUA(I)	0.738 ± 0.109	0.652 ± 0.1	0.783 ± 0.11	0.889 ± 0.028	0.757 ± 0.121
AUC(I)	1.356 ± 0.107	0.93 ± 0.048	0.916 ± 0.132	1.07 ± 0.034	1.084 ± 0.222
AUU(I)	0.906 ± 0.116	1.418 ± 0.097	1.301 ± 0.105	1.041 ± 0.033	1.159 ± 0.231
AAA(K)	0.927 ± 0.093	0.86 ± 0.037	0.989 ± 0.08	0.954 ± 0.036	0.939 ± 0.087
AAG(K)	1.073 ± 0.093	1.14 ± 0.037	1.011 ± 0.08	1.046 ± 0.036	1.061 ± 0.087
CUA(L)	0.635 ± 0.163	0.501 ± 0.079	0.561 ± 0.128	0.425 ± 0.067	0.555 ± 0.144
CUC(L)	1.639 ± 0.161	0.849 ± 0.062	1.127 ± 0.141	1.361 ± 0.132	1.273 ± 0.326
CUG(L)	0.953 ± 0.17	0.668 ± 0.062	0.699 ± 0.083	0.83 ± 0.089	0.794 ± 0.168
CUU(L)	1.582 ± 0.155	1.915 ± 0.093	1.886 ± 0.156	1.814 ± 0.215	1.782 ± 0.209
UUA(L)	0.404 ± 0.093	0.889 ± 0.127	0.715 ± 0.09	0.458 ± 0.121	0.612 ± 0.216
UUG(L)	0.787 ± 0.141	1.178 ± 0.138	1.011 ± 0.137	1.112 ± 0.105	0.983 ± 0.2
AAC(N)	1.213 ± 0.107	0.819 ± 0.09	0.774 ± 0.102	0.868 ± 0.063	0.939 ± 0.217
AAU(N)	0.787 ± 0.107	1.181 ± 0.09	1.226 ± 0.102	1.132 ± 0.063	1.061 ± 0.217
CCA(P)	1.802 ± 0.115	1.084 ± 0.07	2.029 ± 0.111	1.271 ± 0.066	1.669 ± 0.387
CCC(P)	0.789 ± 0.102	0.876 ± 0.049	0.49 ± 0.112	0.903 ± 0.077	0.718 ± 0.195
CCG(P)	0.393 ± 0.099	0.308 ± 0.038	0.218 ± 0.053	0.296 ± 0.052	0.303 ± 0.1
CCU(P)	1.016 ± 0.114	1.733 ± 0.082	1.264 ± 0.139	1.531 ± 0.094	1.309 ± 0.29
CAA(Q)	0.836 ± 0.118	1.068 ± 0.087	1.048 ± 0.075	0.899 ± 0.098	0.962 ± 0.14
CAG(Q)	1.164 ± 0.118	0.932 ± 0.087	0.952 ± 0.075	1.101 ± 0.098	1.038 ± 0.14
AGA(R)	0.815 ± 0.183	0.555 ± 0.083	1.458 ± 0.166	0.726 ± 0.127	0.973 ± 0.395
AGG(R)	1.298 ± 0.119	1.43 ± 0.121	1.865 ± 0.209	1.183 ± 0.061	1.503 ± 0.311
CGA(R)	0.479 ± 0.119	0.375 ± 0.079	0.401 ± 0.118	0.354 ± 0.047	0.415 ± 0.113
CGC(R)	1.486 ± 0.225	1.292 ± 0.14	0.803 ± 0.136	1.478 ± 0.152	1.213 ± 0.35
CGG(R)	0.852 ± 0.156	0.609 ± 0.078	0.425 ± 0.091	0.585 ± 0.093	0.622 ± 0.21
CGU(R)	1.07 ± 0.174	1.739 ± 0.142	1.048 ± 0.188	1.674 ± 0.169	1.274 ± 0.353
AGC(S)	1.005 ± 0.136	0.222 ± 0.044	0.465 ± 0.125	0.386 ± 0.058	0.584 ± 0.327
AGU(S)	0.819 ± 0.16	0.313 ± 0.069	1.008 ± 0.11	0.651 ± 0.105	0.763 ± 0.279
UCA(S)	1.442 ± 0.227	1.549 ± 0.147	2.068 ± 0.238	1.5 ± 0.097	1.685 ± 0.345
UCC(S)	1.184 ± 0.2	1.404 ± 0.181	0.755 ± 0.158	1.157 ± 0.098	1.076 ± 0.3
UCG(S)	0.48 ± 0.145	0.511 ± 0.14	0.225 ± 0.065	0.434 ± 0.099	0.392 ± 0.168
UCU(S)	1.069 ± 0.194	2.001 ± 0.171	1.48 ± 0.231	1.873 ± 0.131	1.499 ± 0.409
ACA(U)	1.347 ± 0.194	0.995 ± 0.091	2.023 ± 0.133	0.854 ± 0.048	1.445 ± 0.476
ACC(U)	1.314 ± 0.157	1.099 ± 0.096	0.717 ± 0.094	1.266 ± 0.095	1.061 ± 0.287
ACG(U)	0.39 ± 0.099	0.348 ± 0.087	0.182 ± 0.053	0.279 ± 0.066	0.296 ± 0.118
ACU(U)	0.949 ± 0.143	1.559 ± 0.096	1.078 ± 0.106	1.601 ± 0.111	1.199 ± 0.295
GUA(V)	0.418 ± 0.113	0.406 ± 0.044	0.713 ± 0.141	0.241 ± 0.038	0.493 ± 0.2
GUC(V)	1.309 ± 0.162	0.894 ± 0.087	0.918 ± 0.139	1.079 ± 0.117	1.064 ± 0.226
GUG(V)	1.164 ± 0.117	0.936 ± 0.103	1.004 ± 0.133	0.834 ± 0.117	1.021 ± 0.164
GUU(V)	1.109 ± 0.16	1.764 ± 0.072	1.365 ± 0.144	1.847 ± 0.159	1.423 ± 0.317
UAC(Y)	1.238 ± 0.092	0.815 ± 0.127	0.775 ± 0.11	0.812 ± 0.038	0.94 ± 0.233
UAU(Y)	0.762 ± 0.092	1.185 ± 0.127	1.225 ± 0.11	1.188 ± 0.038	1.06 ± 0.233

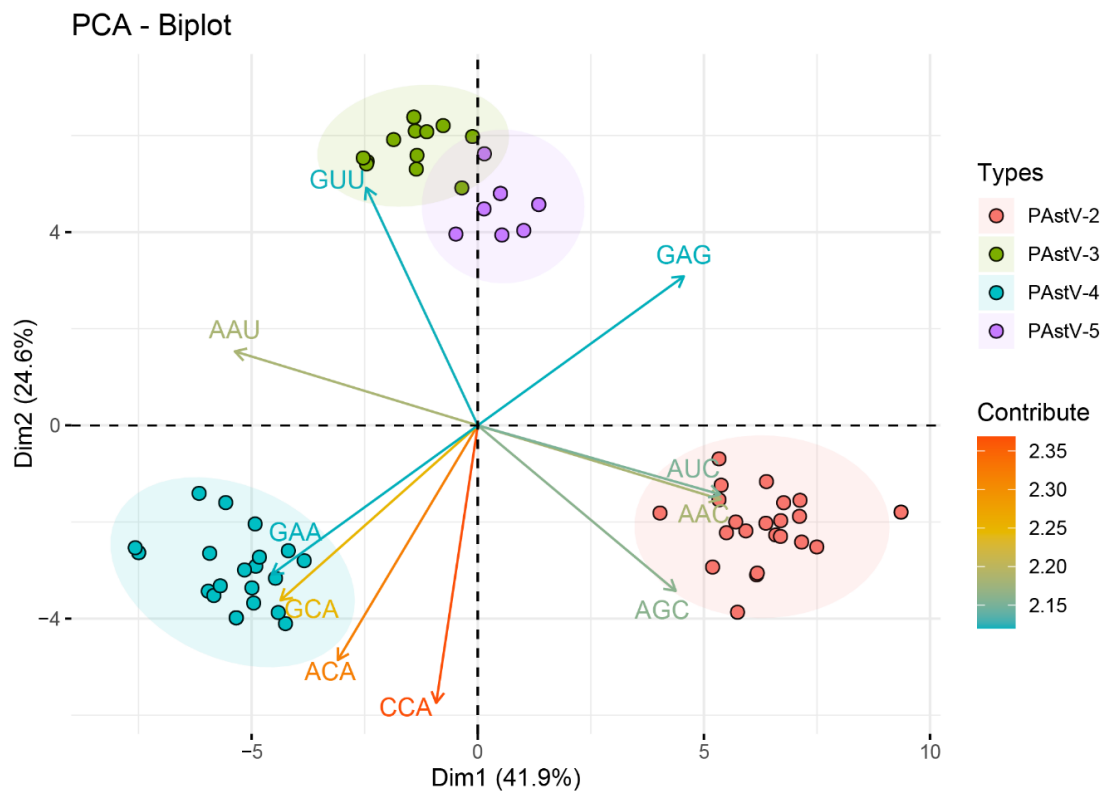


Figure 2. Principal component analysis (PCA) of the PASTV complete coding genomes. The first dimension was plotted against the second dimension. PASTV-2, PASTV-3, PASTV-4, and PASTV-5 are represented in orange, green, blue, and purple, respectively.

3.4. Dinucleotide Frequency Abundance Influences the Codon Usage Bias of PASTV

We performed a dinucleotide analysis on the four subtypes of PASTVs to understand the possible effect of dinucleotide frequencies on the codon usage. Dinucleotides UpG were overrepresented ($P_{xy} \geq 1.23$), whereas dinucleotides CpG and UpA were underrepresented ($P_{xy} \leq 0.78$) in the genome CDSs of the four subtypes of PASTVs (Table 3 and Figure 3). Additionally, dinucleotide CpA was overrepresented ($P_{xy} \geq 1.23$) in the genome CDSs of PASTV-2 and PASTV-4, and dinucleotide CpU was over-represented in the genome CDSs of PASTV-5. These results showed that significant biases of the dinucleotide content variation were observed in the four subtypes of PASTVs.

In order to determine the effect of dinucleotide usage on codon usage bias, we compared the over-representative and under-representative dinucleotides with preferred and under-representative codons. Among eight CpG-containing codons, five codons (GCG, CCG, CGA, UCG, and ACG) were under-represented (RSCU value < 0.6), indicating that dinucleotide CpG were inhibited. Furthermore, the RSCU values of all six UpA-containing codons (AUA, CUA, UUA, GUA, UAC, and UAU) were < 1.6 , suggesting that dinucleotide UpA were inhibited. Among all five UpG dinucleotides-containing codons, UGU codon was found to be a preferred codon, and RSCU values of all UpG dinucleotides-containing codons were > 0.6 . For all eight CpA dinucleotides-containing codons, five (CAC, CCA, CAG, UCA, and ACA) and six (GCA, CAU, CCA, CAA, UCA, and ACA) codons were preferentially used synonymous codons in PASTV-2 and PASTV-4, respectively. Of the eight CpU dinucleotides-containing codons, five codons (GCU, CUU, CCU, UCU, and ACU) were preferentially used in PASTV-2. These results indicated that dinucleotide abundance influences the codon usage bias of the four subtypes of PASTVs.

Table 3. Relative dinucleotides frequencies of CDS of PAsV strains. The odds ratios of over-represented (≥ 1.23) and the under-represented (≤ 0.78) dinucleotides are highlighted using orange and blue, respectively.

	PAstV-2	PAstV-3	PAstV-4	PAstV-5	All(G)
AA	1.001 \pm 0.033	1.121 \pm 0.015	1.039 \pm 0.021	1.076 \pm 0.021	1.047 \pm 0.05
AC	1.049 \pm 0.022	1.025 \pm 0.021	0.991 \pm 0.02	1.092 \pm 0.017	1.03 \pm 0.04
AG	0.977 \pm 0.025	1 \pm 0.018	1.011 \pm 0.02	1.014 \pm 0.014	0.998 \pm 0.026
AU	0.967 \pm 0.025	0.872 \pm 0.021	0.949 \pm 0.019	0.83 \pm 0.026	0.924 \pm 0.055
CA	1.239 \pm 0.023	1.137 \pm 0.027	1.279 \pm 0.022	1.135 \pm 0.022	1.219 \pm 0.065
CC	1.028 \pm 0.022	1.124 \pm 0.028	1.11 \pm 0.021	1.003 \pm 0.021	1.072 \pm 0.053
CG	0.642 \pm 0.027	0.528 \pm 0.023	0.384 \pm 0.015	0.595 \pm 0.01	0.525 \pm 0.112
CU	1.072 \pm 0.03	1.181 \pm 0.02	1.126 \pm 0.041	1.23 \pm 0.02	1.133 \pm 0.063
GA	1.055 \pm 0.024	1.032 \pm 0.023	1.062 \pm 0.021	1.1 \pm 0.011	1.059 \pm 0.029
GC	0.928 \pm 0.021	0.955 \pm 0.023	0.954 \pm 0.029	0.937 \pm 0.017	0.943 \pm 0.026
GG	1.09 \pm 0.026	1.142 \pm 0.009	1.129 \pm 0.025	1.118 \pm 0.026	1.117 \pm 0.031
GU	0.92 \pm 0.03	0.886 \pm 0.022	0.851 \pm 0.029	0.871 \pm 0.016	0.883 \pm 0.039
UA	0.659 \pm 0.021	0.748 \pm 0.023	0.647 \pm 0.022	0.713 \pm 0.02	0.679 \pm 0.045
UC	0.987 \pm 0.032	0.913 \pm 0.014	0.956 \pm 0.028	0.971 \pm 0.027	0.96 \pm 0.038
UG	1.349 \pm 0.031	1.274 \pm 0.02	1.418 \pm 0.025	1.277 \pm 0.018	1.349 \pm 0.064
UU	1.043 \pm 0.028	1.061 \pm 0.019	1.085 \pm 0.031	1.041 \pm 0.024	1.061 \pm 0.033

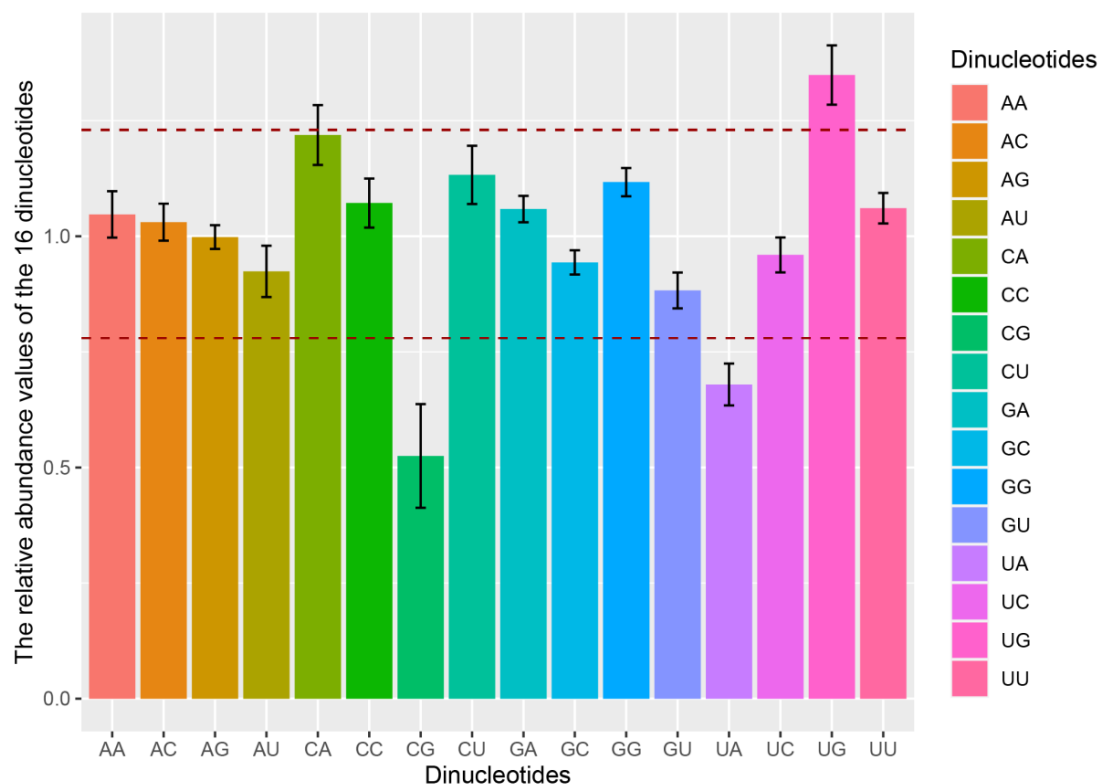


Figure 3. Dinucleotide abundance of the complete CDS of PAsV. The different colors represent the different dinucleotides. Dinucleotides are regarded as underrepresented or overrepresented if the relative abundance values are below 0.78 or over 1.23 (dashed lines), respectively.

3.5. Identification of the Forces Influencing Codon Usage Patterns

To evaluate the forces shaping the codon usage patterns in the four genotypes of PAsVs, PR2 bias, ENC plots, and neutrality analyses were carried out. In the PR2 bias analysis, significant deviations from the parity rules were observed ($A \neq U$, $C \neq G$) (Figure 4), indicating that the extent of the evolutionary forces shaping the codon usage patterns of the four subtypes of PAsVs were not unique.

In the ENC plot, all of the ENC values of PAsTV strains fell below but were close to the expected ENC curve (Figure 5). Additionally, sequences of PAsTV-2 and PAsTV-4 were clustered separately, whereas sequences of PAsTV-3 and PAsTV-5 were clustered together in the ENC plots. These results indicate that mutation pressure and natural selection led to the codon usage bias of the four genotypes of PAsTVs.

The neutrality analysis between the GC3s and GC12s values was employed to determine the extent of the two evolutionary forces on the codon usage pattern of PAsTV strains. A significant correlation between GC3s and GC12s was observed in the PAsTV-5 strains ($y = 0.4089x + 0.3156$; $R^2 = 0.657$; $p < 0.01$) (Figure 6). Thus, the percentage of constraints of natural selection was 59.11% for the PAsTV-5 strains. No significant correlation between GC3s and GC12s was observed in the genomes of PAsTV-2 ($p = 0.2639$; $R^2 = 0.062$), PAsTV-3 ($p = 0.679$; $R^2 = 0.0161$), or PAsTV-4 ($p = 0.158$; $R^2 = 0.0925$) strains. Therefore, natural selection plays a dominant role in driving codon usage bias for these three subtypes. Overall, the above results indicate that the effect of directional mutation pressure is present in the codon usage of PAsTV-5, but natural selection dominates the evolution of codon usage of the four subtypes of PAsTVs.

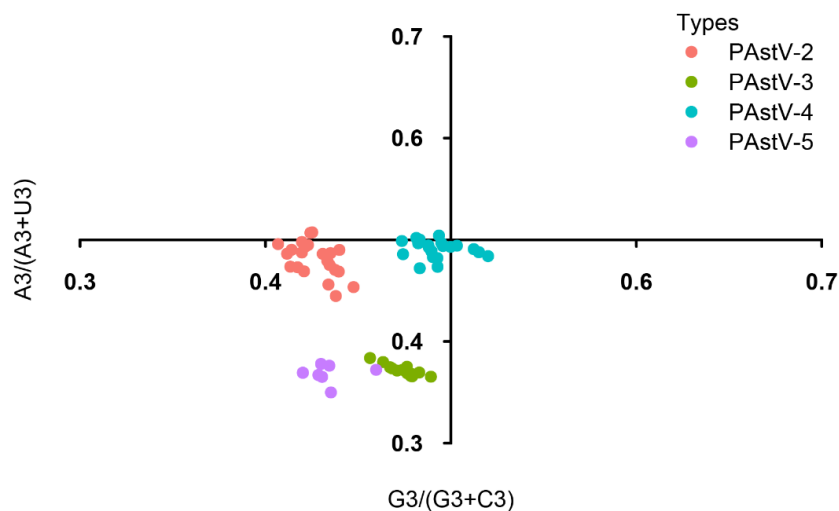


Figure 4. PR2 plot analysis of the PAsTV complete coding genomes. PAsTV-2, PAsTV-3, PAsTV-4, and PAsTV-5 strains are represented in orange, green, blue, and purple, respectively.

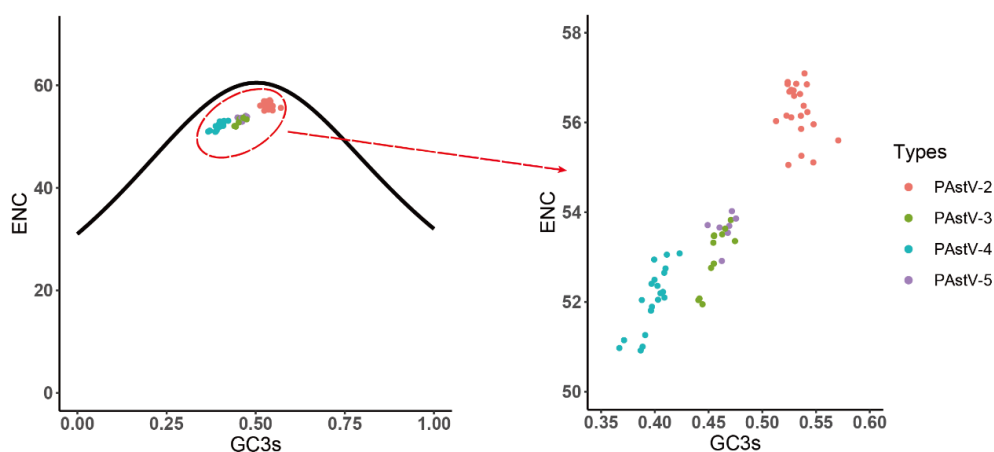


Figure 5. ENC plot analysis of the complete coding genomes of PAsTV. The ENC plot displayed the relationships between ENC and GC content at the third codon position (GC3s) of protein-coding sequences. The curve represents the expected ENC values for all GC3 compositions. PAsTV-2, PAsTV-3, PAsTV-4, and PAsTV-5 strains are represented in orange, green, blue, and purple, respectively.

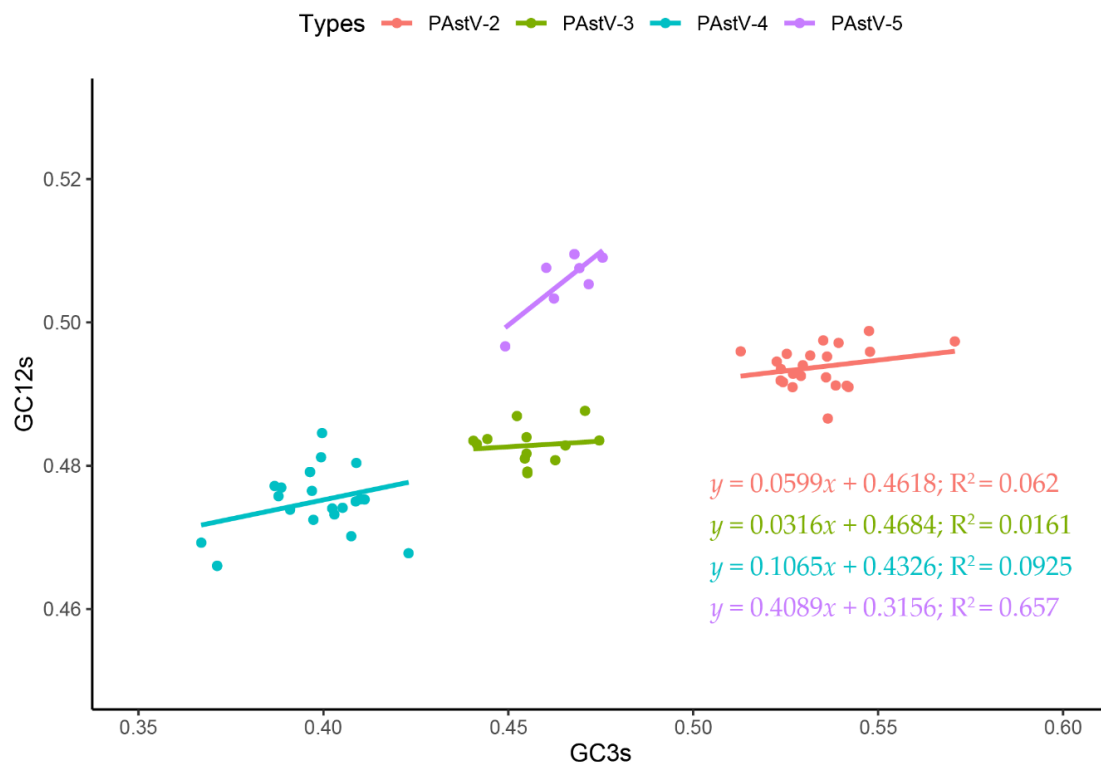


Figure 6. Neutrality analysis of the PAsTV complete coding genomes. The neutrality plot displays the correlation between GC content at synonymous positions (GC12s) and GC content at non-synonymous positions (GC3s). PAsTV-2, PAsTV-3, PAsTV-4, and PAsTV-5 strains are represented in orange, green, blue, and purple, respectively.

3.6. PAsTV Strains Adaptation to Host Species

The analyses of CAI, RCDI, and SiD values were employed to evaluate the codon usage similarities between the PAsTV strains and potential host species. The results based on CAI values show that PAsTV presented the highest CAI value to ducks, followed by chickens, humans, dogs, horses, mice, pigs, cats, cattle, and sheep, while it was comparatively unsuitable for growth in the goat (Figure S2). PAsTV-2 displayed the significant higher CAI values to pig compared with the other three subtypes of PAsTVs (Figure 7). Comparable RCDI analysis showed that the mean RCDI values of PAsTV-3, PAsTV-4, PAsTV-5 were significantly higher than PAsTV-2 (Figure 8), suggesting the codon deoptimization of PAsTV-2 is less than PAsTV-3, PAsTV-4, and PAsTV-5. PAsTV-3 and PAsTV-4 were significantly higher than the PAsTV-2 and PAsTV-5 in SiDs (Figure 9), indicating that the pigs had a significantly deeper effect on PAsTV-3 and PAsTV-4 than PAsTV-2 and PAsTV-5.

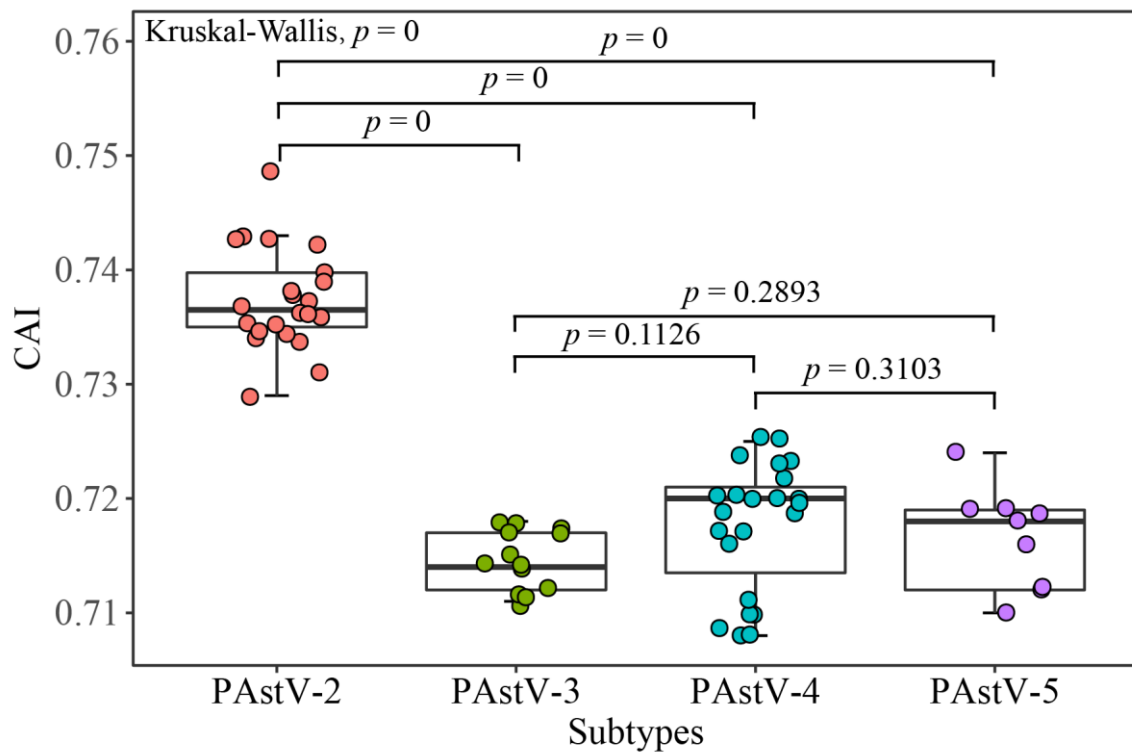


Figure 7. Codon adaptation index (CAI) of the PASTV complete coding genomes. PASTV-2, PASTV-3, PASTV-4, and PASTV-5 strains are represented in orange, green, blue, and purple, respectively.

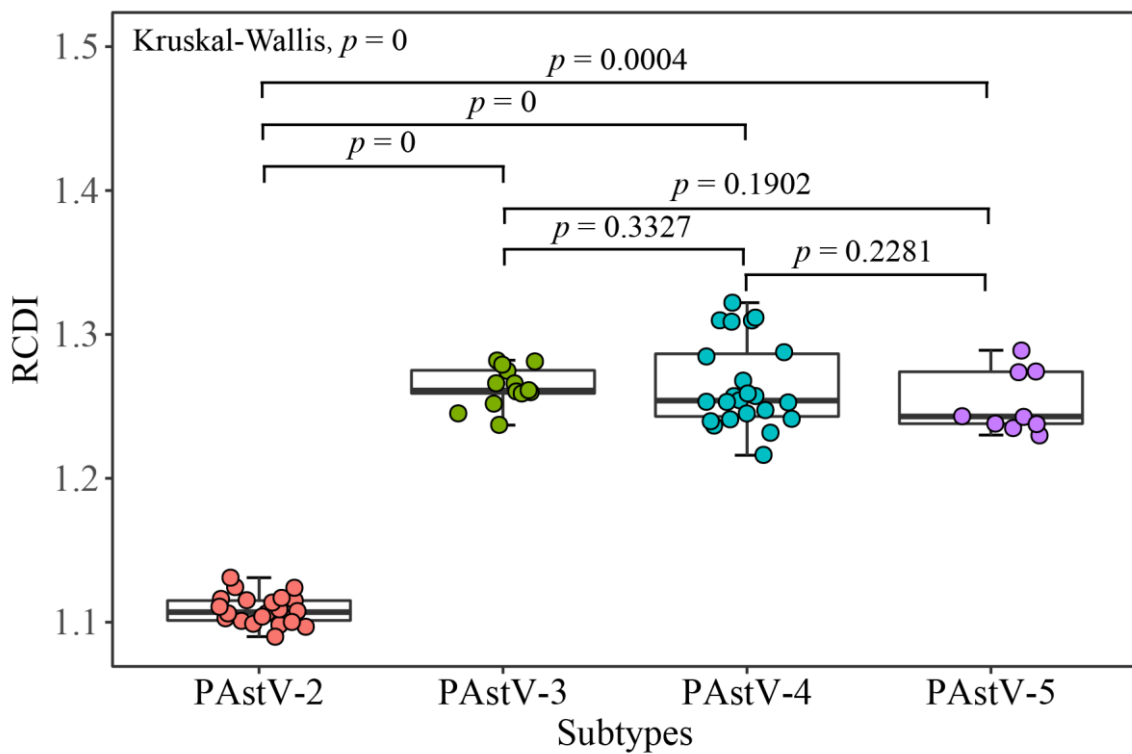


Figure 8. Relative codon deoptimization index (RCDI) of the PASTV complete coding genomes. PASTV-2, PASTV-3, PASTV-4, and PASTV-5 strains are represented in orange, green, blue, and purple, respectively.

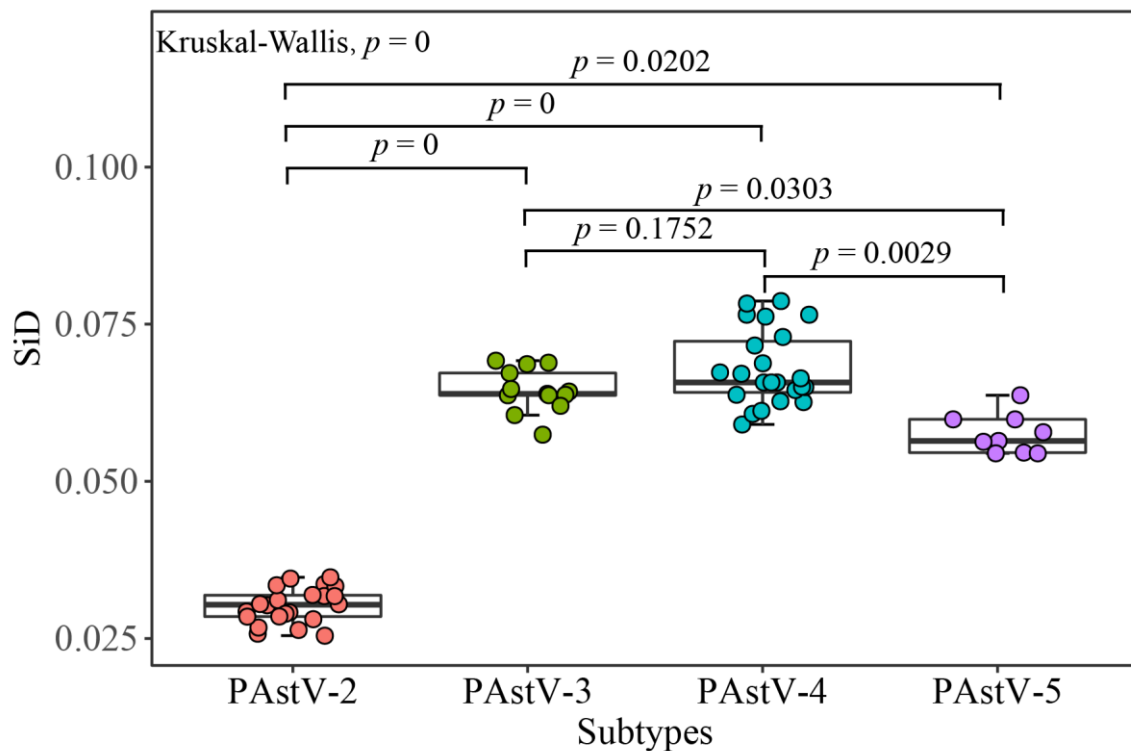


Figure 9. Similarity index (SiD) of the PASTV complete coding genomes. PASTV-2, PASTV-3, PASTV-4, and PASTV-5 strains are represented in orange, green, blue, and purple, respectively.

4. Discussion

In this study, we analyzed the phylogenetic relationship of PASTV. Phylogenetic analysis demonstrated that the PASTV-5 occupied the basal position in the phylogenetic tree, indicating the multiple differentiations of PASTV. Given that the phylogenetic differentiation in porcine astrovirus might imply its evolutionary history, the identification of the phylogeny of porcine astrovirus provides valuable insight into the origin and evolution of porcine astrovirus.

To adapt to changes in the host and the environment, RNA viruses evolve by altering the composition of their genomes [42]. As an important indicator of viral evolution, codon usage preference is affected by many factors, including natural selection, mutation pressure, composition of the genomes or genome regions, and gene length [43]. To dissect evolutionary forces of codon usage bias, a total of 67 complete coding sequences of PASTV genomes were used to perform a comprehensive analysis of codon usage among PASTV-2, PASTV-3, PASTV-4, and PASTV-5.

The genotype-specific preferences of the four subtypes of PASTVs were observed in the third nucleotide position of the codons. More specifically, PASTV-2 tends to use the G/C ending codons, whereas PASTV-3, PASTV-4, and PASTV-5 prefer A/U ending codons. Codon usage bias, largely determined by the nucleotide at the third position of the codon, allows a different perspective on the evolution of the virus [29]. Differences in the nucleotide usage of genome CDSs demonstrate that nucleotide composition indeed affects codon usage bias of the four subtypes of PASTVs.

The effective number of codons (ENC) was calculated to identify bias in the use of synonymous codons. High ENC values (>40) have been identified in many animal viruses, such as porcine circovirus 3 [44], porcine deltacoronavirus [45], and rabies virus [46]. In our study, the mean ENC values of PASTV-2, PASTV-3, PASTV-4, and PASTV-5 were 56.265 ± 0.602 , 53.059 ± 0.656 , 52.007 ± 0.678 , and 53.647 ± 0.316 , respectively, demonstrating that a low codon preference was present in the four subtypes of PASTVs. As suggested by previous reports [47,48], the four subtypes of PASTVs with low codon bias may have a selective advantage for their efficient replication in pigs.

In order to understand the codon usage patterns, RSCU values of 59 synonymous codons were estimated. The results of the RSCU analysis revealed that A/U-ended codons were preferentially used over G/C-ended codons in the genomes of PAsTVs-3,4,5, while PAsTV-2 tended to use G/C-ending preferred codons. The PCA plot showed a clear separation among different PAsTV subtypes, indicating that synonymous codon usage is distinct for each subtype of PAsTV strains. These results showed that despite being a single-stranded RNA virus with a very high mutation rate, PAsTV has a relatively stable synonymous codon usage at a subtype level.

Although RSCU analysis is generally used to investigate synonymous codon usage patterns, it has limitations in revealing the forces that affect codon usage [49]. Therefore, the codon usage analysis was further carried out on the dinucleotides of the four subtypes of PAsTVs. The results indicated remarkable divergence of dinucleotide patterns among the four subtypes of PAsTVs. In coding sequences of PAsTV genomes, dinucleotides CpG and UpA were underrepresented, and dinucleotide UpG was overrepresented. Dinucleotide CpA was specifically overrepresented in the genome CDSs of PAsTV-2 and PAsTV-4. The frequency of dinucleotides is affected by codon usage, mutation pressure, and natural selection [50]. CpA and UpG increases are regarded as a compensatory mechanism for both CpG and UpA reduction [51,52]. Low CpG content in viruses is usually considered to function to evade host defense and to be affected by natural selection [50,53]. UpA is another dinucleotide that is commonly underrepresented in viral genomes due to natural selection [48]. Decreasing the content of UpA can reduce the sensitivity of ribonuclease, which is conducive to maintaining the stability of mRNA [50] and to avoiding energy instability [54]. The results demonstrated that CpG and UpA were underrepresented in the four subtypes of PAsTVs, suggesting that natural selection may have an important role in modeling the codon usage patterns of PAsTV strains.

To better understand the roles of mutation pressure and natural selection in shaping the codon usage, PR2 analysis, ENC-GC3 plots, and neutrality analysis were performed. We found a non-proportional distribution from the parity rules, suggesting that both mutation pressure and natural selection contributed to codon usage bias of the four subtypes of PAsTVs. The ENC-GC3 plot showed that the points representing PAsTV sequences fell below the expected ENC curve. For the ENC-GC3s correlation analysis, if the codons were only affected by the mutation pressure, the actual ENC observations would fall above the ENC expectation curve on the plot of ENC against GC3s [55]. Conversely, if the actual observations of ENC values fall far below the expected curve of ENC values, it means that natural selection has played a major role in codon usage patterns [50]. Therefore, the analysis using the ENC plot indicated mutation pressure and natural selection has driven the codon usage bias of the four subtypes of PAsTVs. Although the ENC-GC3s analysis provides a method for quantitative analysis of codon usage bias, this method does not accurately measure the contributions of natural selection and mutation pressure to the codon usage bias of a species [56,57]. To provide more information on this issue, neutral evolution analysis was performed. According to the neutrality plot, the codon usage patterns of PAsTV subgroups were determined under different evolutionary pressures. Specifically, mutation pressure and natural selection contributed 40.89% and 59.11% to shaping codon usage pattern of PAsTV-5. Natural selection accounted for 94.01%, 96.84%, and 89.35% driving the codon usage bias of PAsTV-2, PAsTV-3, and PAsTV-4, respectively. Taken together, these results suggest that different evolutionary pressures are acting on the four subtypes of PAsTVs. Both mutation pressure and natural selection influence codon usage patterns of PAsTV-5, while natural selection is the dominant evolutionary force driving codon usage bias of PAsTV-2, PAsTV-3, and PAsTV-4.

The emergence, dynamics, and evolution of viral diseases are determined by host-virus interactions. Intriguingly, all of the four subtypes of PAsTVs have the highest CAI value to ducks among the 10 tested hosts. Multiple interspecies transmission events have occurred among human astroviruses, non-human mammalian astroviruses, and avian astroviruses [58]. There have been reports of *Avastrovirus* infecting mammalian species in ecotones, such as small and medium sized farms that rear multiple species [58,59]. The prevalent interspecies transmission of astroviruses reflects their varying origins. Codon usage study can easily predict the carrier hosts that may act as a source of infection in other co-circulating

species [50]. The high CAI value of four subtypes of PAsTVs to animals indicated a similar codon usage pattern between animals and PAsTVs. We proposed the similar codon usage between PAsTV and hosts might advance the cross-species transmission of astrovirus. The values of CAI, RCDI, and SiD may reveal the different adaptabilities of four subtypes of PAsTVs to pigs. Of these, PAsTV-2 may be most adaptive to pigs in theory than the others, in view of its high value of CAI and low values of RCDI and SiD. This might explain to some extent why PAsTV-2 was found as the predominant genotype in many countries [60]. Future studies are warranted to pay more attention to the epidemiology and pathogenicity of PAsTV-2 strains.

Codon usage analysis could be used to design the protein-based vaccine against pathogenic viruses based on the control of viral protein expression. Attenuation by the deoptimizations of dinucleotides and/or codons has achieved as a rapid and efficient strategy for attenuation of various small RNA viruses which causes attenuation of viral virulence, and is used to the development of live, attenuated RNA virus vaccines with superior genetic stabilities [61–64]. Conversely, the optimizations of dinucleotides and/or codons in viral genes increase the protein expression level dramatically and are often performed for vaccine research to increase the immunogenicity of the target [64]. Besides, codon usage bias provides a theoretical basis for studying the transcript regulation, function, and pathological relevance of viral protein. A new transcription regulation was found in some persistent viruses which use poor codons in a distinctive way to temporarily regulate late expression of structural gene products [65]. Information regarding the codon usage pattern and host adaptability of the four subtypes of PAsTVs may be useful to identify the potential hosts and the suitable experimental animal models for pathogenesis and vaccine researches.

5. Conclusions

To our knowledge, this study is for the first time to reveal the codon usage pattern for PAsTVs. Phylogenetic analysis result showed the clade PAsTV-5 occupies the basal position of the phylogenetic tree. The results from nucleotide composition analysis show that the genome CDSs of PAsTVs-3,4,5 are rich in A/U nucleotides in comparison to G/C nucleotides. The C/G-ended codons are the preferentially used synonymous codons in the PAsTV-2, whereas AU-ended codons were the preferred synonymous codons in the PAsTV-3, PAsTV-4, and PAsTV-5. Natural selection and mutation pressure are the main factors influencing the codon usage bias in the PAsTV-5 genome. The codon usages of PAsTV-2, PAsTV-3, and PAsTV-4 are mainly constrained by selection pressure. The high similar codon usage between PAsTV and animals might account for the broad host range and the cross-species transmission of astrovirus. Overall, the information from this study provides new insights for understanding PAsTV evolution regarding codon usage pattern and host adaptability.

Supplementary Materials: The following are available online at <http://www.mdpi.com/1999-4915/12/9/991/s1>, Figure S1: Scree plot of percentage of explained variances for each principal component of the RSCU values of PAsTV complete CDS. Figure S2: CAI analysis of the PAsTV complete coding genomes in relation to potential host species. Table S1: The detailed information of PAsTV strains in NCBI nucleic acid database. Table S2: The nucleotide composition and properties of complete CDS of the PAsTVs.

Author Contributions: H.W. analyzed the data, and wrote and finalized the manuscript. Z.B. and C.M. collected the data. Z.C. and J.Z. proposed the work. Z.B., C.M., Z.C., and J.Z. revised the important technical content of the manuscript, and finalized the manuscript. All authors approve the version to be published, and agree to take responsibility for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China [2017YFD0500104] to Z.C.; the National Natural Science Foundation of China [Grant No. 31772748; 31660648] to Z.C. and H.W.; and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) to Z.C.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Reuter, G.; Knowles, N.J. Astroviruses. *Dis. Swine* **2019**, 457–460. [CrossRef]
2. Bridger, J. Detection by electron microscopy of caliciviruses, astroviruses and rotavirus-like particles in the faeces of piglets with diarrhoea. *Vet. Rec.* **1980**, *107*, 532–533. [PubMed]
3. Luo, Z.; Roi, S.; Dastor, M.; Gallice, E.; Laurin, M.-A.; L'Homme, Y. Multiple novel and prevalent astroviruses in pigs. *Vet. Microbiol.* **2011**, *149*, 316–323. [CrossRef] [PubMed]
4. Bailey Arruda, P.A.; Hensch, M.; Chen, Q.; Zheng, Y.; Yang, C.; Gatto, I.R.H.; Ferreyra, F.M.; Gauger, P.; Schwartz, K.; Bradner, L. Porcine astrovirus type 3 in central nervous system of swine with polyoencephalomyelitis. *Emerg. Infect. Dis.* **2017**, *23*, 2097. [CrossRef]
5. Boros, Á.; Albert, M.; Pankovics, P.; Bíró, H.; Pesavento, P.A.; Phan, T.G.; Delwart, E.; Reuter, G. Outbreaks of neuroinvasive astrovirus associated with encephalomyelitis, weakness, and paralysis among weaned pigs, Hungary. *Emerg. Infect. Dis.* **2017**, *23*, 1982. [CrossRef] [PubMed]
6. Fang, Q.; Wang, C.; Liu, H.; Wu, Q.; Liang, S.; Cen, M.; Dong, Q.; Wei, Y.; Chen, Y.; Ouyang, K. Pathogenic Characteristics of a Porcine Astrovirus Strain Isolated in China. *Viruses* **2019**, *11*, 1156. [CrossRef]
7. Padmanabhan, A.; Hause, B.M. Detection and characterization of a novel genotype of porcine astrovirus 4 from nasal swabs from pigs with acute respiratory disease. *Arch. Virol.* **2016**, *161*, 2575–2579. [CrossRef]
8. Kumthip, K.; Khamrin, P.; Saikruang, W.; Kongkaew, A.; Vachirachewin, R.; Ushijima, H.; Maneekarn, N. Detection and genetic characterization of porcine astroviruses in piglets with and without diarrhea in Thailand. *Arch. Virol.* **2018**, *163*, 1823–1829. [CrossRef]
9. Xiao, C.-T.; Gimenez-Lirola, L.G.; Gerber, P.F.; Jiang, Y.-H.; Halbur, P.G.; Opriessnig, T. Identification and characterization of novel porcine astroviruses (PAstVs) with high prevalence and frequent co-infection of individual pigs with multiple PAstV types. *J. Gen. Virol.* **2013**, *94*, 570–582. [CrossRef]
10. Blomström, A.-L.; Ley, C.; Jacobson, M. Astrovirus as a possible cause of congenital tremor type AII in piglets? *Acta Vet. Scand.* **2014**, *56*, 82. [CrossRef]
11. Brnić, D.; Prpić, J.; Keros, T.; Roić, B.; Starešina, V.; Jemeršić, L. Porcine astrovirus viremia and high genetic variability in pigs on large holdings in Croatia. *Infect. Genet. Evol.* **2013**, *14*, 258–264. [CrossRef] [PubMed]
12. Jiang, B.; Monroe, S.S.; Koonin, E.V.; Stine, S.E.; Glass, R.I. RNA sequence of astrovirus: Distinctive genomic organization and a putative retrovirus-like ribosomal frameshifting signal that directs the viral replicase synthesis. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 10539–10543. [CrossRef] [PubMed]
13. Marczinke, B.; Bloys, A.J.; Brown, T.; Willcocks, M.M.; Carter, M.J.; Brierley, I. The human astrovirus RNA-dependent RNA polymerase coding region is expressed by ribosomal frameshifting. *J. Virol.* **1994**, *68*, 5588–5595. [CrossRef] [PubMed]
14. Arias, C.F.; DuBois, R.M. The astrovirus capsid: A review. *Viruses* **2017**, *9*, 15. [CrossRef]
15. Sharp, P.M.; Matassi, G. Codon usage and genome evolution. *Curr. Opin. Genet. Dev.* **1994**, *4*, 851–860. [CrossRef]
16. Sharp, P.M.; Li, W.-H. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **1986**, *24*, 28–38. [CrossRef]
17. Van Hemert, F.J.; Berkhout, B.; Lukashov, V.V. Host-related nucleotide composition and codon usage as driving forces in the recent evolution of the Astroviridae. *Virology* **2007**, *361*, 447–454. [CrossRef]
18. Ranwez, V.; Harispe, S.; Delsuc, F.; Douzery, E.J. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS ONE* **2011**, *6*, e22594. [CrossRef]
19. Darriba, D.; Taboada, G.L.; Doallo, R.; Posada, D. jModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* **2012**, *9*, 772. [CrossRef]
20. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef]
21. Ronquist, F.; Teslenko, M.; Van Der Mark, P.; Ayres, D.L.; Darling, A.; Höhna, S.; Larget, B.; Liu, L.; Suchard, M.A.; Huelsenbeck, J.P. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **2012**, *61*, 539–542. [CrossRef] [PubMed]
22. Rambaut, A. 2012. FigTree v1. 4. Available online: <http://tree.bio.ed.ac.uk/software/figtree/> (accessed on 5 September 2020).

23. Charif, D.; Lobry, J.R. SeqinR 1.0-2: A contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In *Structural Approaches to Sequence Evolution*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 207–232.
24. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019; Available online: <https://www.R-project.org/> (accessed on 5 September 2020).
25. Kariin, S.; Burge, C. Dinucleotide relative abundance extremes: A genomic signature. *Trends Genet.* **1995**, *11*, 283–290. [[CrossRef](#)]
26. Wright, F. The ‘effective number of codons’ used in a gene. *Gene* **1990**, *87*, 23–29. [[CrossRef](#)]
27. Fuglsang, A. The ‘effective number of codons’ revisited. *Biochem. Biophys. Res. Commun.* **2004**, *317*, 957–964. [[CrossRef](#)]
28. Elek, A.; Kuzman, M.; Vlahoviček, K. coRdon: Codon Usage Analysis and Prediction of Gene Expressivity. R Package Version 1.4.0. 2019. Available online: <https://github.com/BioinfoHR/coRdon> (accessed on 5 September 2020).
29. Wong, E.H.; Smith, D.K.; Rabadan, R.; Peiris, M.; Poon, L.L. Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus. *BMC Evol. Biol.* **2010**, *10*, 253. [[CrossRef](#)]
30. Sharp, P.M.; Tuohy, T.M.; Mosurski, K.R. Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* **1986**, *14*, 5125–5143. [[CrossRef](#)]
31. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [[CrossRef](#)]
32. Kassambara, A.; Mundt, F. Factoextra: Extract and visualize the results of multivariate data analyses. R Package Version **2017**, *1*, 2017.
33. Sueoka, N. Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 2653–2657. [[CrossRef](#)]
34. Sueoka, N. Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G+ C content of third codon position. *Gene* **1999**, *238*, 53–58. [[CrossRef](#)]
35. Sueoka, N. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J. Mol. Evol.* **1995**, *40*, 318–325. [[CrossRef](#)] [[PubMed](#)]
36. 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](#)] [[PubMed](#)]
37. Athey, J.; Alexaki, A.; Osipova, E.; Rostovtsev, A.; Santana-Quintero, L.V.; Katneni, U.; Simonyan, V.; Kimchi-Sarfaty, C. A new and updated resource for codon usage tables. *BMC Bioinform.* **2017**, *18*, 391. [[CrossRef](#)] [[PubMed](#)]
38. 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](#)]
39. Puigbò, P.; Aragonès, L.; Garcia-Vallvé, S. RCDI/eRCDI: A web-server to estimate codon usage deoptimization. *BMC Res. Notes* **2010**, *3*, 87. [[CrossRef](#)]
40. 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*. [[CrossRef](#)]
41. Dinno, A. Dunn.Test: Dunn’s Test of Multiple Comparisons Using Rank Sums. R Package Version 1.3.5. 2017. Available online: <https://CRAN.R-project.org/package=dunn.test> (accessed on 5 September 2020).
42. Streicker, D.G. From persistence to cross-species emergence of a viral zoonosis. *Science* **2013**, *342*, 1185–1186. [[CrossRef](#)]
43. Salim, H.M.; Cavalcanti, A.R. Factors influencing codon usage bias in genomes. *J. Braz. Chem. Soc.* **2008**, *19*, 257–262. [[CrossRef](#)]
44. Li, G.; Wang, H.; Wang, S.; Xing, G.; Zhang, C.; Zhang, W.; Liu, J.; Zhang, J.; Su, S.; Zhou, J. Insights into the genetic and host adaptability of emerging porcine circovirus 3. *Virulence* **2018**, *9*, 1301–1313. [[CrossRef](#)]
45. He, W.; Wang, N.; Tan, J.; Wang, R.; Yang, Y.; Li, G.; Guan, H.; Zheng, Y.; Shi, X.; Ye, R. Comprehensive codon usage analysis of porcine deltacoronavirus. *Mol. Phylogenet. Evol.* **2019**, *141*, 106618. [[CrossRef](#)]
46. Zhang, X.; Cai, Y.; Zhai, X.; Liu, J.; Zhao, W.; Ji, S.; Su, S.; Zhou, J. Comprehensive analysis of codon usage on rabies virus and other lyssaviruses. *Int. J. Mol. Sci.* **2018**, *19*, 2397. [[CrossRef](#)] [[PubMed](#)]
47. 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*, 1–14. [[CrossRef](#)] [[PubMed](#)]

48. Kumar, N.; Bera, B.C.; Greenbaum, B.D.; Bhatia, S.; Sood, R.; Selvaraj, P.; Anand, T.; Tripathi, B.N.; Virmani, N. Revelation of influencing factors in overall codon usage bias of equine influenza viruses. *PLoS ONE* **2016**, *11*. [[CrossRef](#)] [[PubMed](#)]
49. Butt, A.M.; Nasrullah, I.; Tong, Y. Genome-wide analysis of codon usage and influencing factors in chikungunya viruses. *PLoS ONE* **2014**, *9*. [[CrossRef](#)]
50. Khandia, R.; Singhal, S.; Kumar, U.; Ansari, A.I.; 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](#)]
51. Yomo, T.; Ohno, S. Concordant evolution of coding and noncoding regions of DNA made possible by the universal rule of TA/CG deficiency-TG/CT excess. *Proc. Natl. Acad. Sci. USA* **1989**, *86*, 8452–8456. [[CrossRef](#)]
52. Rima, B.K.; McFerran, N.V. Dinucleotide and stop codon frequencies in single-stranded RNA viruses. *J. Gen. Virol.* **1997**, *78*, 2859–2870. [[CrossRef](#)]
53. Vetsigian, K.; Goldenfeld, N. Genome rhetoric and the emergence of compositional bias. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 215–220. [[CrossRef](#)]
54. Breslauer, K.J.; Frank, R.; Blöcker, H.; Marky, L.A. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA* **1986**, *83*, 3746–3750. [[CrossRef](#)]
55. He, B.; Dong, H.; Jiang, C.; Cao, F.; Tao, S.; Xu, L. Analysis of codon usage patterns in Ginkgo biloba reveals codon usage tendency from A/U-ending to G/C-ending. *Sci. Rep.* **2016**, *6*, 35927. [[CrossRef](#)]
56. Jia, X.; Liu, S.; Zheng, H.; Li, B.; Qi, Q.; Wei, L.; Zhao, T.; He, J.; Sun, J. Non-uniqueness of factors constraint on the codon usage in Bombyx mori. *BMC Genom.* **2015**, *16*, 356. [[CrossRef](#)] [[PubMed](#)]
57. Kawabe, A.; Miyashita, N.T. Patterns of codon usage bias in three dicot and four monocot plant species. *Genes Genet. Syst.* **2003**, *78*, 343–352. [[CrossRef](#)] [[PubMed](#)]
58. Donato, C.; Vijaykrishna, D. The broad host range and genetic diversity of mammalian and avian astroviruses. *Viruses* **2017**, *9*, 102. [[CrossRef](#)] [[PubMed](#)]
59. Wohlgenuth, N.; Honce, R.; Schultz-Cherry, S. Astrovirus evolution and emergence. *Infect. Genet. Evol.* **2019**, *69*, 30–37. [[CrossRef](#)] [[PubMed](#)]
60. Ito, M.; Kuroda, M.; Masuda, T.; Akagami, M.; Haga, K.; Tsuchiaka, S.; Kishimoto, M.; Naoi, Y.; Sano, K.; Omatsu, T. Whole genome analysis of porcine astroviruses detected in Japanese pigs reveals genetic diversity and possible intra-genotypic recombination. *Infect. Genet. Evol.* **2017**, *50*, 38–48. [[CrossRef](#)] [[PubMed](#)]
61. Kunec, D.; Osterrieder, N. Codon pair bias is a direct consequence of dinucleotide bias. *Cell Rep.* **2016**, *14*, 55–67. [[CrossRef](#)] [[PubMed](#)]
62. Gustafsson, C.; Govindarajan, S.; Minshull, J. Codon bias and heterologous protein expression. *Trends Biotechnol.* **2004**, *22*, 346–353. [[CrossRef](#)]
63. Coleman, J.R.; Papamichail, D.; Skiena, S.; Fitcher, B.; Wimmer, E.; Mueller, S. Virus attenuation by genome-scale changes in codon pair bias. *Science* **2008**, *320*, 1784–1787. [[CrossRef](#)]
64. Burns, C.C.; Shaw, J.; Campagnoli, R.; Jorba, J.; Vincent, A.; Quay, J.; Kew, O. Modulation of poliovirus replicative fitness in HeLa cells by deoptimization of synonymous codon usage in the capsid region. *J. Virol.* **2006**, *80*, 3259–3272. [[CrossRef](#)]
65. Shin, Y.C.; Bischof, G.F.; Lauer, W.A.; Desrosiers, R.C. Importance of codon usage for the temporal regulation of viral gene expression. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 14030–14035. [[CrossRef](#)]

