Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 406342, 10 pages
http://dx.doi.org/10.1155/2013/406342

Research Article

A Comparison of Synonymous Codon Usage Bias Patterns in DNA and RNA Virus Genomes: Quantifying the Relative Importance of Mutational Pressure and Natural Selection

Youhua Chen

Department of Zoology, University of British Columbia, Vancouver, BC, Canada V6T 1Z4

Correspondence should be addressed to Youhua Chen; haydi@126.com

Received 15 April 2013; Revised 30 June 2013; Accepted 4 August 2013

Academic Editor: Sankar Subramanian

Codon usage bias patterns have been broadly explored for many viruses. However, the relative importance of mutation pressure and natural selection is still under debate. In the present study, I tried to resolve controversial issues on determining the principal factors of codon usage patterns for DNA and RNA viruses, respectively, by examining over 38000 ORFs. By utilizing variation partitioning technique, the results showed that 27% and 21% of total variation could be attributed to mutational pressure, while 5% and 6% of total variation could be explained by natural selection for DNA and RNA viruses, respectively, in codon usage patterns. Furthermore, the combined effect of mutational pressure and natural selection on influencing codon usage patterns of viruses is substantial (explaining 10% and 8% of total variation of codon usage patterns). With respect to GC variation, GC content is always negatively and significantly correlated with aromaticity. Interestingly, the signs for the significant correlations between GC, gene lengths, and hydrophobicity are completely opposite between DNA and RNA viruses, being positive for DNA viruses while being negative for RNA viruses. At last, GC12 versus G3s plot suggests that natural selection is more important than mutational pressure on influencing the GC content in the first and second codon positions.

1. Introduction

Codon usage is not a random event [1]. Codon usage bias has been broadly observed, and different mechanisms have been proposed to explain the bias patterns, for example, mutation pressure, translational efficiency, gene length [2], dinulcoetide bias [3], tRNA abundance [4], organ specificity [5], and so on. Codon usage bias patterns have been broadly studied in recent years, especially for virus genomes [3, 6, 7]. However, most of these previous studies only consider a specific virus or a specific virus clade [8–10], a global comparison of virus codon usage bias patterns is still largely lacking, even though some literature had worked on many RNA and DNA viruses as whole [11–13]. A holistic observation and comparison of codon usage patterns over different clades of viruses would throw new insights into virus genome explicitly. To cope with such a knowledge gap, in the present study, I analyzed codon usage patterns for the available 2317 virus genomes for the purpose of providing a more robust and integrated understanding of synonymous codon usage patterns.

Given the accumulation of genome sequences from different viruses in GenBank database, another purpose of the present study is to quantify the relative contribution of mutation pressure and natural selection on influencing codon usage patterns of virus genomes. I could achieve such an objective by introducing a new statistical method called variance partitioning to quantitatively examine the separated role of different mechanisms on synonymous codon usage patterns of viruses.

2. Materials and Methods

2.1. Sequence Data. The complete genome sequences for 2317 different virus species were originally obtained from GenBank database (http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html). Because some viruses have been sequenced for multiple times using different strains, for avoiding sampling bias, only one from these multiple genomic sequences for the same virus is used. Furthermore, because RNA and DNA viruses are very different on their codon usage biase patterns [12], RNA and DNA viruses are analyzed separately. Genomes belonging to other types of viruses, such as Retro-transcribing viruses, are not considered.

Consequently, 786 DNA viruses and 725 RNA viruses are retained for all subsequent analyses, representing around 65% of the total virus species in the NCBI Genome database. By extracting all the valid open reading frames (ORFs) from each genome sequence and removing problematic ones (including short-length (less than 350 bp) ORFs, overlapping ORFs for different genes/transcripts, ORFs with nontranslatable codons, and ORFs without synonymous codons), 35818 ORFs for DNA viruses and 2743 ORFs for RNA viruses are kept for calculating codon usage indices and performing multivariate analyses.

2.2. Measures of Relative Synonymous Codon Usage (RSCU). Relative synonymous codon usage values of each codon in a gene are calculated to investigate the characteristics of synonymous codon usage. The RSCU index is calculated as follows [14]:

$$\text{RSCU} = \frac{g_{ij} \times n_j}{\sum_i^{n_j} g_{ij}}, \tag{1}$$

where g_{ij} is the observed number of the ith codon for the jth amino acid which has n_i kinds of synonymous codons. Codons with higher (or lower) selected frequencies have higher (or lower) RSCU values. When the corresponding RSCU values of a codon are close to 1, it is used randomly and evenly.

2.3. Effective Number of Codons. The effective number of codons (ENC) is a measure of bias from equal codon usage in a gene [15]. The calculation formula is

$$\text{ENC} = 2 + \frac{9}{\overline{F}_2} + \frac{1}{\overline{F}_3} + \frac{5}{\overline{F}_4} + \frac{3}{\overline{F}_6}, \tag{2}$$

where $\overline{F}_k$ ($k = 2, 3, 4, 6$) is the mean of F_k values for the k-fold degenerate amino acids, which is estimated using the formula as follows:

$$F_k = \frac{nS - 1}{n - 1}, \tag{3}$$

where n is the total number of occurrences of the codons for that amino acid and

$$S = \sum_{i=1}^{k} \left(\frac{n_i}{n}\right)^2, \tag{4}$$

where n_i is the total number of occurrences of the ith codon for that amino acid.

N_c ranges from 20 for the strongest bias (where only one codon is used for each amino acid) to 61 for no bias (where all synonymous codons are used equally).

For elucidating the relationship between GC3s and ENC values, the expected ENC values for different GC3s are calculated as follows:

$$\text{ENC}^{\text{expected}} = 2 + s + \frac{29}{s^2 + (1 - s)^2}, \tag{5}$$

where s denotes the value of GC3s [6]. The observed and expected ENC values are compared to determine the influence of nucleotide compositional constraint on structuring synonymous codon usage bias.

2.4. Codon Adaptation Index. The codon adaptation index (CAI) estimates the extent of bias toward codons that are known to be favored in highly expressed genes [16]. In the present study, for simplicity, the *Escherichia coli* optimal codons are used as the reference.

2.5. Indices for Measuring Chemical Properties of Amino Acids. Hydrophobicity (GRAVY) and aromaticity (AROMO) of conceptually translated gene product may be factors influencing codon usage bias patterns [17]. As such, I quantify both indices to reveal the evidence of natural selection on codon usage bias.

For hydrophobicity index [17], it is calculated as

$$\text{GRAVY} = \frac{1}{N} \sum_{i=1}^{N} k_i, \tag{6}$$

where N is the number of amino acids and k_i is the hydrophobic index of the ith amino acid.

For aromaticity index [17], it is calculated as

$$\text{AROMO} = \frac{1}{N} \sum_{i=1}^{N} v_i, \tag{7}$$

where v_i is either 1 (for an aromatic amino acid) or 0 (for a nonaromatic amino acid) and N is the number of amino acids.

2.6. Correspondence Analysis and Canonical Correspondence Analysis. In addition to utilizing conventional correspondence analysis (CA) [17], in the present study, I introduce a new method, namely, canonical correspondence analysis (CCA) [18], which could help reveal the principal trends of codon usage bias patterns and identify the most correlated variables simultaneously. CCA method has been broadly applied in ecological studies [18, 19]. However, it might be the first time to be applied to study synonymous codon usage patterns for viruses in the present study.

The mathematical formulation for CCA method [18, 19] is a bit complicated in comparison to its linear analogue redundancy analysis (RDA) [19, 20] because it requires data transformation. As such, the calculation steps for RDA are present here for demonstrating the calculation core steps of CCA.

Assuming that one has the codon usage matrix Y and the matrix of explanatory variables (codon usage indices) X (both have the same rows), then the RDA procedure is to predict the elements (codon usage values) in the matrix Y as

$$\widehat{Y} = X\left[X^T \ X\right]^{-1} X^T Y, \tag{8}$$

where the subscript T denotes the transpose of the matrix and −1 denotes the inverse of the matrix.

Thus the covariance matrix for the predicted codon usage matrix $\widehat{Y}$ is (n denotes the row number)

$$M = \frac{1}{n-1}\widehat{Y}^T\widehat{Y}. \tag{9}$$

The RDA or CCA method is to decompose the above matrix M into normalized eigenvalues E and normalized eigenvector matrix U. Elements from E ranked from high to low represent the explained proportion of total variation in the codon usage patterns, while the corresponding eigenvectors U can be used to obtain sample scores and biplots when generating the 2-dimensional plots.

2.7. Quantifying the Influence of Mutation Pressure and Selection Pressure Using Variation Partitioning. Variation partitioning is a relatively new method for helping elucidate the influence of each group of explanatory variables in multivariate statistics [21]. Variation partitioning has been broadly applied in ecological and evolutionary studies [19]. For quantifying the influence of mutation selection, I consider the metrics related to codon contents, like GC, GC3s, A3s, T3s, C3s, and G3s contents, as the factors reflecting mutational pressure. In contrast, the indices CAI, all kinds of protein properties, including hydrophobicity and aromaticity, are regarded as the representative of natural selection [3, 17, 22]. For simplicity, the mathematical formulation for variation partitioning technique is as follows [21, 23, 24].

Supposing that there are two groups of explanatory variables in two matrices X_1 and X_2, the total variation S in the codon usage table matrix Y with n rows is written as

$$S = \frac{1}{(n-1)}\operatorname{Trace}\left(\left(Y-\overline{Y}\right)^T\left(Y-\overline{Y}\right)\right), \tag{10}$$

where a hyphen above the variable(s) denotes the mean(s).

Then the proportion of variation R_1 only explained by the group of explanatory variables X_1 is obtained as

$$\begin{gathered}
Y_1 = X_2\left[X_2^T \quad X_2\right]^{-1}{X_2}^TY, \\
Y_1^{\text{res}} = Y - Y_1, \\
X_1^{\text{res}} = X_1 - X_2\left[X_2^T \quad X_2\right]^{-1}X_2^TX_1, \\
\widehat{Y}_1 = X_1^{\text{res}}\left[X_1^{\text{res}\,T}X_1^{\text{res}}\right]^{-1}X_1^{\text{res}\,T}{Y_1}^{\text{res}}, \\
R_1 = \frac{\operatorname{Trace}\left((1/(n-1))\left(\widehat{Y}_1-\overline{Y}_1\right)^T\left(\widehat{Y}_1-\overline{Y}_1\right)\right)}{S}.
\end{gathered} \tag{11}$$

The percentage of total variation R_2 attributed to the explanatory variable group X_2 is calculated following the same procedure as above.

Finally, the percentage of total variation explained by the interaction of the two variable groups X_1 and X_2 requires the determination of the percentage of variation (R_{12}) explained by all the variables X, the matrix of which combines matrices X_1 and X_2 together:

$$\begin{gathered}
Y_{12} = X\left[X^T \quad X\right]^{-1}X^TY, \\
R_{12} = \frac{\operatorname{Trace}\left((1/(n-1))\left(Y_{12}-\overline{Y}_{12}\right)^T\left(Y_{12}-\overline{Y}_{12}\right)\right)}{S}.
\end{gathered} \tag{12}$$

Thus, the proportion of variation that cannot be explained by any current explanatory variables is determined by

$$R_0 = 1 - R_{12}. \tag{13}$$

Then, the percentage of total variation explained by the interaction of the two variable groups is given by,

$$R_{1\cap 2} = R_{12} - R_1 - R_2. \tag{14}$$

In a summary, R_1, R_2, $R_{1\cap 2}$ and R_0 are the focused explained variation for the present study.

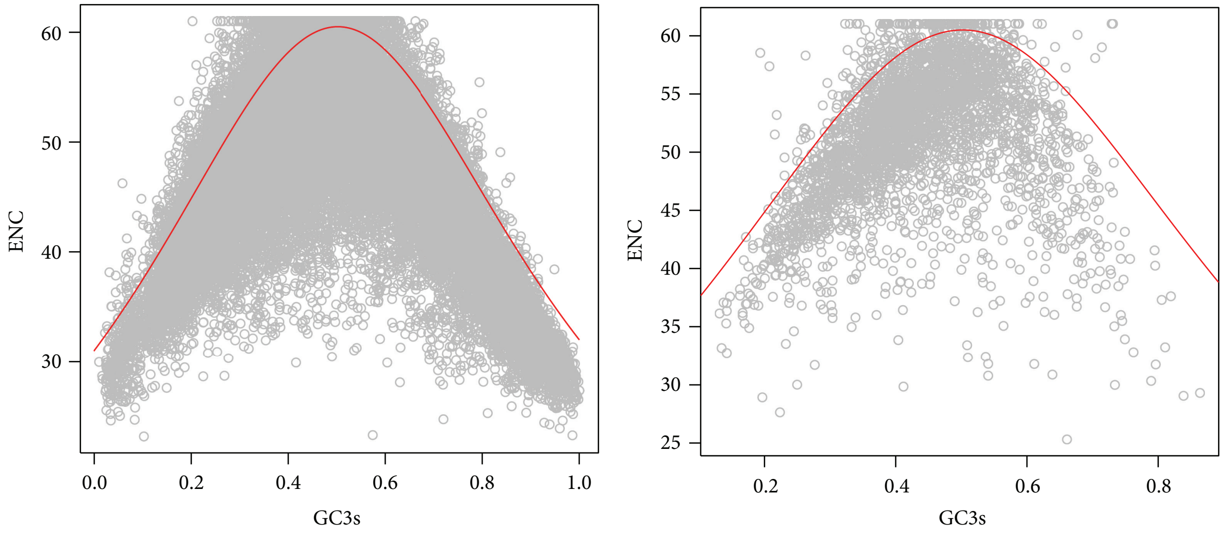
2.8. Statistical Programs. Multivariate analyses, including CA, CCA, and variation partitioning methods, were implemented in *R* [25] package "*vegan*" [26]. All other codon usage indices mentioned above were calculated using CodonW program [27].

3. Results

3.1. ENC-GC3s Plot. As seen in Figure 1, the ENC-GC3s plot showed that most DNA or RNA virus genes lay on or slightly under the expected curve, indicating the extreme importance of mutational pressure for both groups of viruses. However, a great amount of points was laid under the curve as well for DNA (Figure 1(a)) and RNA (Figure 1(b)) viruses, suggesting that other factors, especially the influence of natural selection, were nontrivial.

3.2. Influence of Gene Lengths and Protein Properties on GC Variation. Based on the present observation, there was a significant and positive correlation between GC content and gene lengths for DNA viruses (Figure 2(a)). The log-transformation further enhanced the positive trend (Figure 2(b)). However, for RNA viruses, the patterns became reverse: GC was significantly and negatively correlated with gene lengths for either original (Figure 3(a)) and log-transformed data points (Figure 3(b)). These results thus were incongruent with many previous studies working on a single of virus or a clade of viruses, which suggested that there were no clear trends between GC and gene lengths, for example, influenza viruses [28], polioviruses [10], parvoviridae [29], and so on.

Similar to the relationship between GC and gene lengths, there was an opposite relationship between GC and hydrophobicity for different types of viruses as well (Figures 2(c) and 3(c)). For DNA viruses (Figure 2(c)), the correlation between the two quantities was positive and significant, while for RNA viruses (Figure 3(c)), the correlation became positive and significant (Figure 3(c)).



(a)

(b)

Figure 1: The relationship between ENC and GC3s for DNA (a) and RNA (b) virus genomes, respectively.

Table 1: Correlation analysis of the first two axes of CA and explanatory variables for codon usage bias patterns of virus genomes. For each axis, the correlation coefficients for the top three important variables are marked in boldface.

Variables	DNA viruses				RNA viruses			
	CA1	CA2	CCA1	CCA2	CA1	CA2	CCA1	CCA2
T3s	−0.919	−0.158	−0.94	**−0.238**	−0.774	**−0.501**	−0.807	**−0.57**
C3s	0.929	−0.031	0.946	0.052	**0.912**	0.006	**0.945**	0.077
A3s	**−0.941**	0.165	**−0.956**	0.148	−0.687	**0.585**	−0.691	**0.622**
G3s	0.858	0.099	0.899	0.155	0.24	0.203	0.317	0.176
GC3s	**0.991**	0.022	**0.993**	0.088	**0.935**	0.048	**0.957**	0.088
GC	**0.98**	−0.13	**0.982**	−0.104	**0.959**	−0.094	**0.971**	−0.041
CAI	0.422	**−0.607**	0.463	**−0.652**	0.38	**−0.4**	0.424	**−0.488**
ENC	−0.261	−0.056	−0.335	−0.08	0.217	0.051	0.34	0.162
GRAVY	0.028	**0.172**	0.086	0.169	−0.058	−0.358	−0.158	−0.481
AROMO	−0.328	**0.23**	−0.368	**0.287**	−0.362	−0.25	−0.404	−0.306

Finally, the relationship between GC and aromaticity was always negatively and significantly correlated for either DNA (Figure 2(d)) or RNA (Figure 3(d)) viruses. These significant correlations should suggest the signature of the influence of natural selection on codon usage patterns of viruses.

3.3. Quantifying the Relative Ratio between Mutation and Selection Using Neutrality Plot on the Three Positions of Codons of Viruses. As shown in Figure 4(a), for DNA viruses, the correlation of GC3s and GC12 was best fitted by a linear function as GC12 = 0.202 × GC3s + 0.203(R^2 = 0.461, $P < 0.0001$) for DNA viruses. For RNA viruses, the linear regression model was as similar as GC12 = 0.225 × GC3s + 0.206(R^2 = 0.461, $P < 0.0001$) (Figure 4(b)). The slope of the GC12-GC3s regression line indicated the relative mutaion pressure functioned on the first and second codon positions in relation to that on the third codon position [30–32]. As seen, GC12 was influenced by mutation pressure and natural selection with a ratio being 0.202/0.798 = 0.253 for DNA viruses and 0.225/0.775 = 0.29 for RNA viruses correspondingly. These results indicated that the natural selection was more important on structuring the first and second codon positions and had similar influences for both groups of viruses.

3.4. CA and CCA Analyses for Characterizing the Major Trends in Codon Usage Patterns of Viruses. For the ORFs of DNA viruses, the first (CA1) and second (CA2) axes of CA explained 34.5% and 5.6% of total variation in the codon usage patterns (Figure 5(a)). For RNA viruses, the first and second axes of CA explained 27.3% and 9.2% of the total variation, respectively, in synonymous codon usage patterns (Figure 5(b)). Thus, CA1 reflected the major trends for both DNA and RNA virus ORFs.

For DNA viruses, CA1 was strongly correlated with GC3s (r = 0.99), followed by GC (r = 0.98) and A3s (r = −0.94), while CA2 was strongly related to CAI (r = −0.61), followed by AROMO (r = 0.23) and GRAVY (r = 0.17) (Table 1). The patterns for RNA viruses were similar for the first axis, which

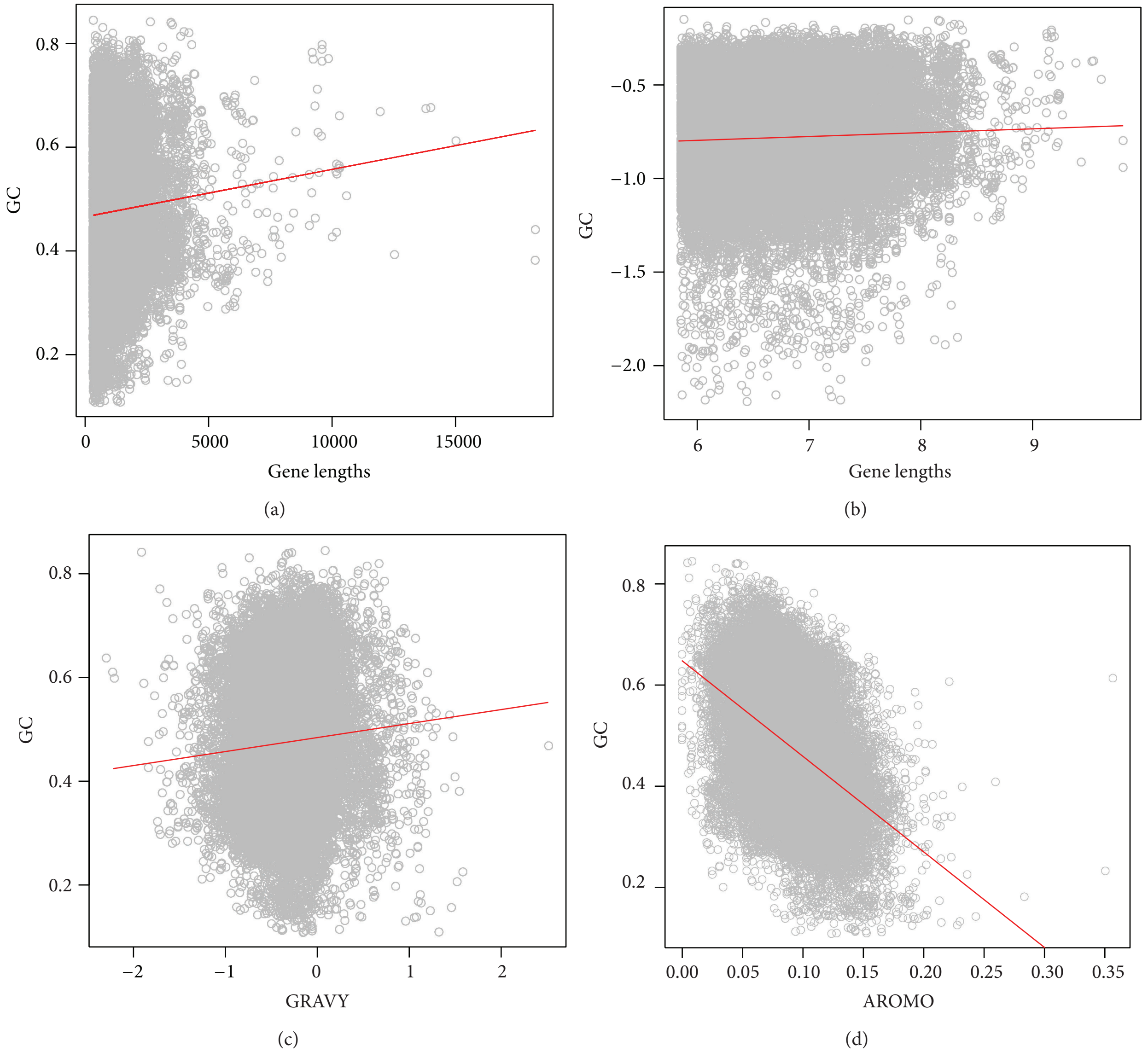


Figure 2: The relationships between GC content, gene length, and amino acid properties for DNA viruses. (a) GC-gene length relationship without transformation: GC = $9.14E-06 \times$ gene lengths $+ 0.466$ ($R^2 = 0.003, P < 0.0001$); (b) GC-gene length relationship with log-transformation: $\log_e(\mathrm{GC}) = 0.021 \times \log_e(\text{gene lengths}) - 0.921$ ($R^2 = 0.002, P < 0.0001$); (c) GC-hydrophobicity relationship: GC $= -0.027 \times$ GRAVY $+ 0.484$ ($R^2 = 0.005, P < 0.0001$); (d) GC-aromaticity relationship: GC $= -1.89 \times$ AROMO $+ 0.648$ ($R^2 = 0.19, P < 0.001$).

was most correlated with GC ($r = 0.96$), followed by GC3s ($r = 0.94$) and C3s ($r = 0.91$). However, the second axis CA2 was correlated with A3s ($r = 0.59$), T3s ($r = -0.50$), and CAI ($r = -0.4$) (Table 1).

For DNA viruses, the first (CCA1) and second (CCA2) axes of CCA explained 68.3% and 8.8% of the total variation in synonymous codon usage patterns (Figure 5(c)). Being identical to the correlation results for CA as described above, CCA1 was strongly correlated with GC3s, GC, and A3s, while CCA2 was strongly related to CAI, AROMO, and T3s (Table 1).

For RNA viruses, the first two axes explained 50.8% and 16.2% of total variation (Figure 5(d)). The most important variables correlated with CCA1 were identical as those for DNA viruses, while A3s, T3s, and CAI were the most important variables for CCA2 (Table 1).

When comparing both CA and CCA results, it was consistently found that the following factors are repeatedly identified as most correlated ones for the principal axes for both DNA and RNA viruses: GC3s, GC, and A3s (Table 1). Thus, these variables should be of great importance to influence codon usage bias patterns for viruses.

3.5. Quantifying the Relative Importance of Mutation Pressure and Natural Selection in Overall Codon Usage Patterns of Viruses. Based on the results of variation partitioning (Figure 6(a)), for DNA viruses, it was found that 27% of total variation could be attributed to mutational pressure, while

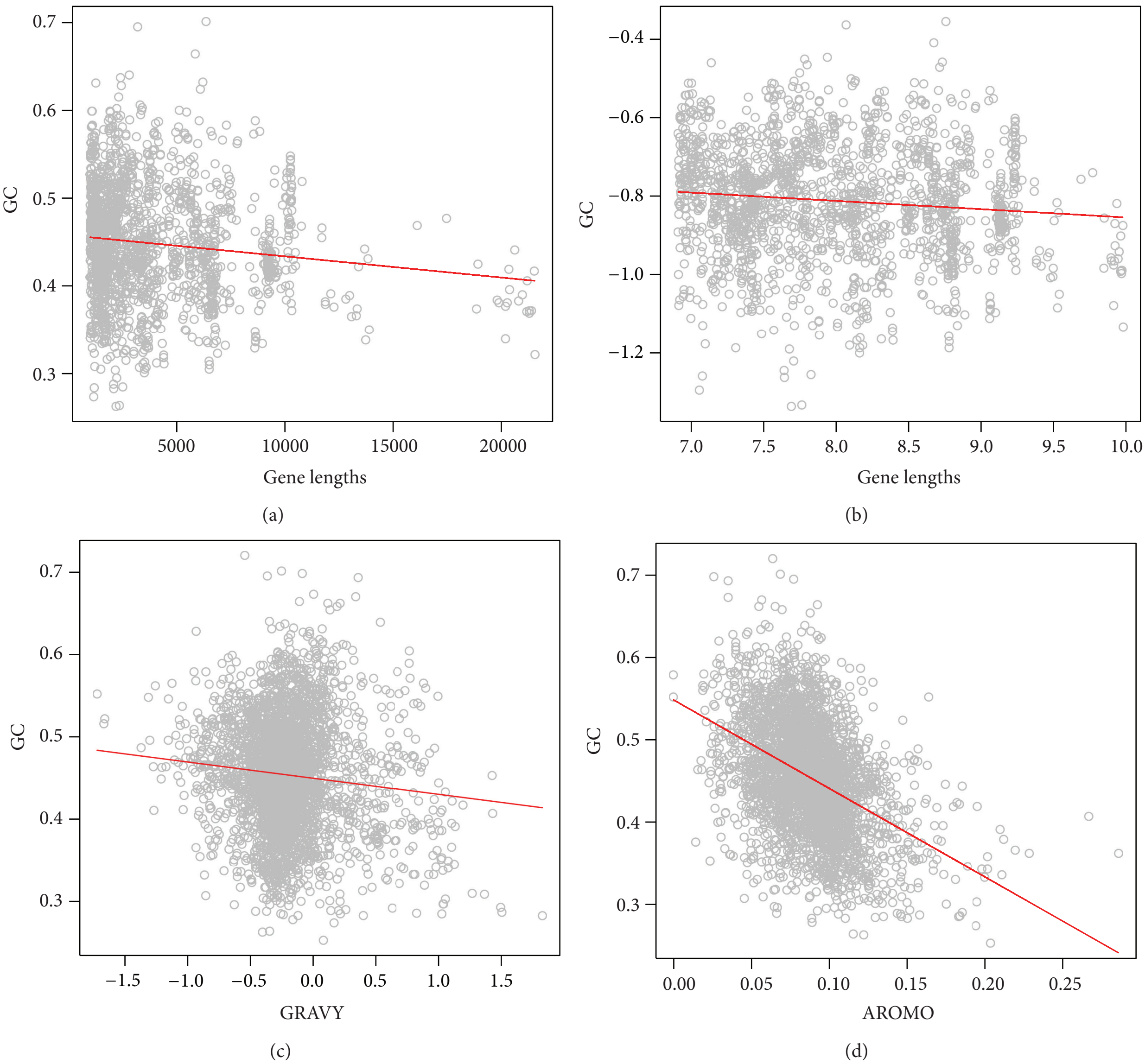


FIGURE 3: The relationships between GC content, gene length, and amino acid properties for RNA viruses. (a) GC-gene length relationship without transformation: GC = $-3.19E-06 \times$ gene lengths + 0.463 ($R^2 = 0.021, P < 0.0001$); (b) GC-gene length relationship with log-transformation: $\log_e$(GC) = $-0.0227 \times \log_e$(gene lengths) − 0.629 ($R^2 = 0.021, P < 0.001$); (c) GC-hydrophobicity relationship: GC = $-0.009 \times$ GRAVY + 0.452 ($R^2 = 0.01, P < 0.05$); (d) GC-aromaticity relationship: GC = $-1.154 \times$ AROMO + 0.555 ($R^2 = 0.159, P < 0.001$).

only 5% of total variation of codon usage patterns attributed to selection pressure. The interaction between mutation and selection further explained 10% of total variation. Very similarly, for RNA viruses (Figure 6(b)), mutational pressure explained 21% of the total variation in codon usage patterns, while natural selection explained 6% of the total variation. The interaction of both mechanisms further explained 8% of the total variation.

4. Discussion

4.1. The Relationship between GC Variation and Codon Usage Factors. Interestingly, it is found that the correlation between GC content and hydrophobicity and gene lengths is positive and significant for DNA viruses (Figures 2(a)–2(c)), while being negative and significant of RNA viruses (Figures 3(a)–3(c)). In contrast, the tendency between GC versus aromaticity is always negative (Figures 2(d) and 3(d)). The positive correlation between GC and hydrophobicity for RNA viruses is contradictory to some previous studies working on specific RNA virus species or clades, which argued that the correlation should be positive [33]. Moreover, it is still controversial whether there is a clear correlation for a specific virus or a clade of viruses. Some previous studies [3, 34] concluded that there was no clear relationship between these two quantities.

I do not observe a congruent relationship between GC content and gene lengths for DNA and RNA viruses

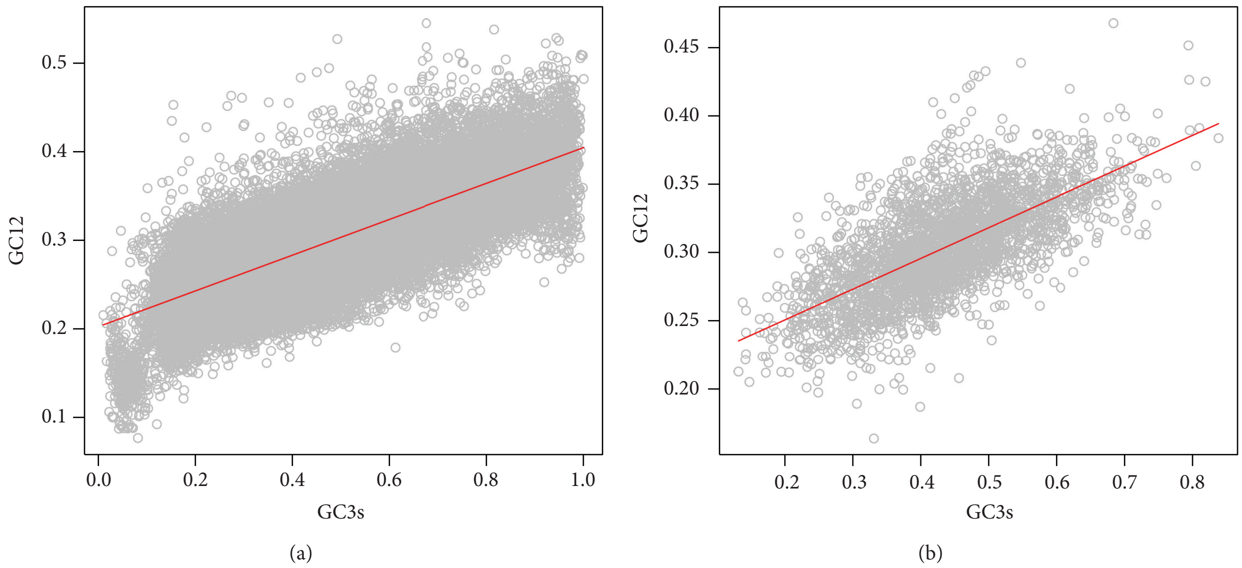


Figure 4: The relationship between GC12 and GC3s of DNA (a) and RNA (b) virus genomes. The fitted regression line has the formula as $\mathrm{GC12} = 0.202 \times \mathrm{GC3s} + 0.203 (R^2 = 0.461, P < 0.0001)$ for DNA viruses and $\mathrm{GC12} = 0.225 \times \mathrm{GC3s} + 0.206 (R^2 = 0.461, P < 0.0001)$ for RNA viruses, respectively.

(Figures 2(a), 2(b), 3(a), and 3(b)). Based on some predictions, GC content should be correlated with gene length since selection should be stronger in longer genes, causing the directional change of GC content [35–38]. Indeed, GC has been thought to relate to gene lengths in prokaryotes, plants, nematodes, or insects [2, 22, 35, 36, 39], although the relationship among these taxa is still debatable [40]. On the basis of the results for DNA and RNA virus genomes at the present study, I argue that there is no consistent relationship between GC profiling and gene lengths for viruses. As shown in Figures 2(a) and 2(b), for DNA viruses, the relationship between GC and gene lengths is positive, implying the imprint of natural selection. However, for RNA viruses (Figures 3(a) and 3(b)), the relationship becomes negative, being opposite to the prediction of natural selection. This is not surprising, because RNA viruses are believed to have much higher mutation rates than DNA viruses [12, 28, 41]. At this perspective, viruses are different to other life forms from other kingdoms, in which natural selection plays differential roles to influence the stability of longer genes of DNA and RNA viruses.

It is found that the prescreening and removal of short-length ORFs are very crucial to obtain accurate trends between GC and codon usage indices. For example, the negative relationship between GC and hydrophobicity may become obscured when more short-length ORFs (less than 350 bp) are included in the study for DNA viruses. The correlation will become nonsignificant (results not showed here).

4.2. Virus Genomes Are Profoundly Influenced by Mutation Pressure. Based on the results of variation partitioning, the present study identifies that mutational pressure is the most prevailing mechanism driving the codon usage bias patterns for both DNA and RNA viruses (Figure 6), because it can explain 27% and 21% of total variation, respectively, in codon usage patterns of both groups of viruses. In contrast, the influence of natural selection is very minor, only explaining 5% and 6% of the total variation, respectively, for both groups of viruses. Previous studies on a single or a clade of viruses largely have confirmed the dominating influence of mutational pressure [6, 10, 42, 43], but many studies also mentioned the considerable importance of selection [29, 44]. Thus, through the present intergenomic analysis, I have a chance to quantify the relative importance of mutation versus selection on structuring codon usage patterns of viruses, and the similar conclusion is enforced: natural selection is not so important in comparison to mutational pressure in synonymous codon usage patterns of viruses.

Through correlation analysis between codon usage indices and major axes from CA and CCA analyses, the present study identifies that the three most important indices are GC, GC3s, and A3s. Thus, the present results are contradictory with a previous study [13] which showed that genomic nucleotide content was the most important factor predicting synonymous codon usage patterns in RNA viruses using randomization techniques. The difference raised may be partially due to the studied data size. In the previous study [13], only 29 RNA virus species were examined. This number is a very small number in comparison to the present study which works on 725 RNA viruses. Thus, the conclusion from the previous study [13] stating that GC content was a poor predictor of codon usage patterns of RNA viruses may be challenging if the authors can work on a large number of RNA viruses.

Finally, as implied by a large fraction of the unexplained variation (over 50%) for both groups of viruses, my quantification of mutational pressure and natural selection by utilizing the present ten codon usage variables might not be sufficient to quantify the relative importance of mutational pressure and natural selection. Other more important variables, especially those for characterizing natural selection

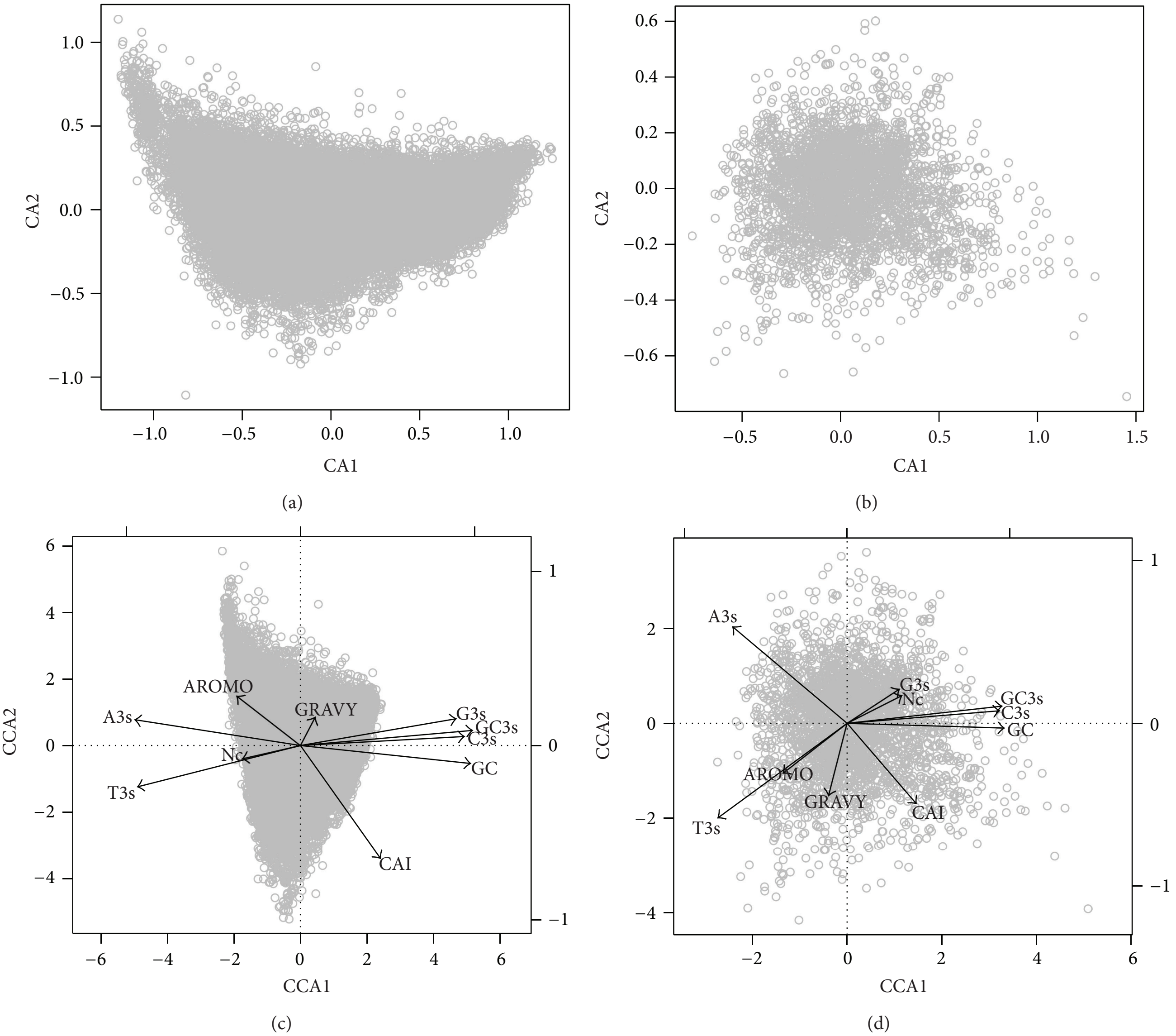


Figure 5: CA plots and CCA biplots for showing the major trends of codon usage patterns of the ORFs for DNA and RNA viruses. (a) CA plot for DNA viruses; (b) CA plot for RNA viruses; (c) CCA biplot for DNA viruses; (d) CCA biplot for RNA viruses.

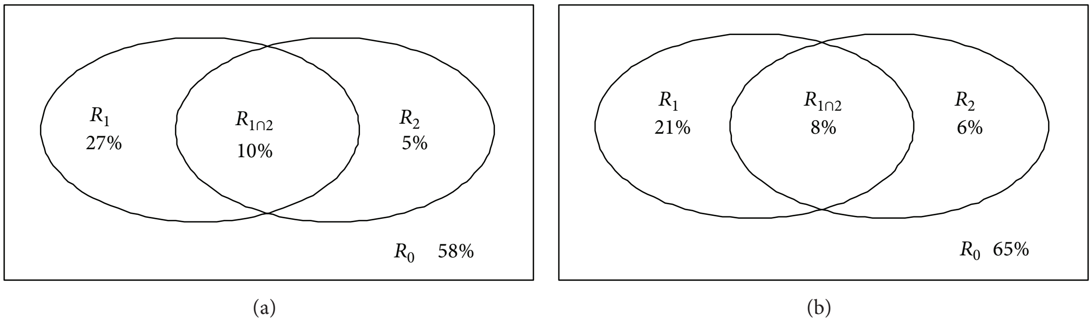


Figure 6: Variation partitioning of codon usage patterns attributed to mutation and selection for DNA (a) and RNA (b) viruses. The meaning of each part of the variation is interpreted as follows, R_1: proportion of variation explained by mutation; R_2: proportion of variation explained by selection; $R_{1\cap 2}$: proportion of variation explained by the interaction of selection and mutation; R_0: proportion of unexplained variation.

(e.g., the frequency of usage of optimal codons [45]), might increase the explanatory power of natural selection on codon usage patterns of virus genomes.

4.3. Limitations of the Present Study. I have to acknowledge that the recombination events happened at either gene or genome levels can influence the codon usage bias patterns to some extent, as evidenced by some previous studies [46–48]. Virus genomes have been broadly observed to process some degrees of homologous recombination [49–52]. As such, it would be a contribution when ones eliminate the influence of homologous recombination in the virus genomes before analyzing codon usage patterns to accurately disentangle the relative importance of mutation and selection.

Acknowledgments

The author thanks the Associate Editor Dr. Sankar Subramanian and two anonymous reviewers for their constructive and critical comments on helping improve the quality of the text. This work is supported by China Scholarship Council (CSC).

References

[1] M. Archetti, "Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code," *Journal of Molecular Evolution*, vol. 59, no. 2, pp. 258–266, 2004.

[2] L. Duret and D. Mouchiroud, "Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 8, pp. 4482–4487, 1999.

[3] P. Tao, L. Dai, M. Luo, F. Tang, P. Tien, and Z. Pan, "Analysis of synonymous codon usage in classical swine fever virus," *Virus Genes*, vol. 38, no. 1, pp. 104–112, 2009.

[4] E. N. Moriyama and J. R. Powell, "Codon usage bias and tRNA abundance in Drosophila," *Journal of Molecular Evolution*, vol. 45, no. 5, pp. 514–523, 1997.

[5] G. P. Holmquist and J. Filipski, "Organization of mutations along the genome: a prime determinant of genome evolution," *Trends in Ecology and Evolution*, vol. 9, no. 2, pp. 65–69, 1994.

[6] X. Liu, C. Wu, and A. Y.-H. Chen, "Codon usage bias and recombination events for neuraminidase and hemagglutinin genes in Chinese isolates of influenza A virus subtype H9N2," *Archives of Virology*, vol. 155, no. 5, pp. 685–693, 2010.

[7] Y. Zhang, Y. Liu, W. Liu et al., "Analysis of synonymous codon usage in Hepatitis A virus," *Virology Journal*, vol. 8, article 174, 2011.

[8] F. P. Lobo, B. E. F. Mota, S. D. J. Pena et al., "Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts," *PLoS ONE*, vol. 4, no. 7, Article ID e6282, 2009.

[9] A. Lukashev, J. Drexler, I. Belalov, M. Eschbach-Bludau, S. Baumgrate, and C. Drosten, "Genetic variation and recombination in Aichi virus," *Journal of General Virology*, vol. 93, pp. 1226–1235, 2012.

[10] J. Zhang, M. Wang, W.-Q. Liu et al., "Analysis of codon usage and nucleotide composition bias in polioviruses," *Virology Journal*, vol. 8, article 146, 2011.

[11] L. A. Shackelton and E. C. Holmes, "The evolution of large DNA viruses: combining genomic information of viruses and their hosts," *Trends in Microbiology*, vol. 12, no. 10, pp. 458–465, 2004.

[12] G. M. Jenkins and E. C. Holmes, "The extent of codon usage bias in human RNA viruses and its evolutionary origin," *Virus Research*, vol. 92, no. 1, pp. 1–7, 2003.

[13] I. Belalov and A. Lukashev, "Causes and implications of codon usage bias in RNA viruses," *PLoS ONE*, vol. 8, Article ID e56642, 2013.

[14] P. M. Sharp and W.-H. Li, "Codon usage in regulatory genes in Escherichia coli does not reflect selection for 'rare' codons," *Nucleic Acids Research*, vol. 14, no. 19, pp. 7737–7749, 1986.

[15] F. Wright, "The 'effective number of codons' used in a gene"," *Gene*, vol. 87, pp. 23–29, 1990.

[16] P. M. Sharp and W.-H. Li, "The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications," *Nucleic Acids Research*, vol. 15, no. 3, pp. 1281–1295, 1987.

[17] J. Peden, *Analysis of codon usage [Ph.D. thesis]*, Department of Genetics; Unviersity of Norttingham, 1999.

[18] C. J. F. Ter Braak, "Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis," *Ecology*, vol. 67, no. 5, pp. 1167–1179, 1986.

[19] P. Legendre and L. Legendre, *Numerical Ecology*, Elsevier Science BV, Amsterdam, The Netherlands, 1998.

[20] P. Legendre, J. Oksanen, and C. Ter Braak, "Testing the significance of canonical axes in redundancy analysis," *Methods in Ecology and Evolution*, vol. 2, pp. 269–277, 2011.

[21] P. R. Peres-Neto, P. Legendre, S. Dray, and D. Borcard, "Variation partitioning of species data matrices: estimation and comparison of fractions," *Ecology*, vol. 87, no. 10, pp. 2614–2625, 2006.

[22] H. Liu, R. He, H. Zhang, Y. Huang, M. Tian, and J. Zhang, "Analysis of synonymous codon usage in Zea mays," *Molecular Biology Reports*, vol. 37, no. 2, pp. 677–684, 2010.

[23] P. Legendre, D. Borcard, and P. R. Peres-Neto, "Analyzing beta diversity: partitioning the spatial variation of community composition data," *Ecological Monographs*, vol. 75, no. 4, pp. 435–450, 2005.

[24] D. Borcard, P. Legendre, and P. Drapeau, "Partialling out the spatial component of ecological variation," *Ecology*, vol. 73, no. 3, pp. 1045–1055, 1992.

[25] R. Development Core Team, "R: A Language and Environment for Statistical Computing, Vienna, Austria," R Foundation for Statistical Computing, Vienna, Austria, 2011, http://www.R-project.org.

[26] J. Oksanen, G. Blanchet, R. Kindt et al., "Vegan: Community Ecology Package. R package version 2. 0-4," 2012.

[27] J. Peden, "CodonW," 2005, http://codonw.sourceforge.net/.

[28] T. Zhou, W. Gu, J. Ma, X. Sun, and Z. Lu, "Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses," *BioSystems*, vol. 81, no. 1, pp. 77–86, 2005.

[29] S. Shi, Y. Jiang, Y. Liu, R. Xia, and L. Qin, "Selective pressure dominates the synonymous codon usage in parvoviridae," *Virus Genes*, vol. 40, pp. 10–19, 2013.

[30] C. Fu, J. Xiong, and W. Miao, "Genome-wide identification and characterization of cytochrome P450 monooxygenase genes in the ciliate Tetrahymena thermophila," *BMC Genomics*, vol. 10, article 208, 2009.

[31] N. Sueoka, "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*, vol. 238, no. 1, pp. 53–58, 1999.

[32] N. Sueoka, "Intrastrand parity rules of DNA base composition and usage biases of synonymous codons," *Journal of Molecular Evolution*, vol. 40, no. 3, pp. 318–325, 1995.

[33] M. Wang, J. Zhang, J.-H. Zhou et al., "Analysis of codon usage in bovine viral diarrhea virus," *Archives of Virology*, vol. 156, no. 1, pp. 153–160, 2011.

[34] H. W. Cao, H. Zhang, Y. Liu, and D. S. Li, "Synonymous codon usage bias of spike genes of porcine epidemic diarrhea virus," *African Journal of Microbiology Research*, vol. 5, no. 22, pp. 3784–3789, 2011.

[35] N. Stoletzki, "The surprising negative correlation of gene length and optimal codon use—disentangling translational selection from GC-biased gene conversion in yeast," *BMC Evolutionary Biology*, vol. 11, no. 1, article 93, 2011.

[36] N. Stoletzki and A. Eyre-Walker, "Synonymous codon usage in Escherichia coli: selection for translational accuracy," *Molecular Biology and Evolution*, vol. 24, no. 2, pp. 374–381, 2007.

[37] M. D. Ermolaeva, "Synonymous codon usage in bacteria," *Current Issues in Molecular Biology*, vol. 3, no. 4, pp. 91–97, 2001.

[38] A. Eyre-Walker, "Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy?" *Molecular Biology and Evolution*, vol. 13, no. 6, pp. 864–872, 1996.

[39] G. Marais and L. Duret, "Synonymous codon usage, accuracy of translation, and gene length in Caenorhabditis elegans," *Journal of Molecular Evolution*, vol. 52, no. 3, pp. 275–280, 2001.

[40] T. Zhou, X. Sun, and Z. Lu, "Synonymous codon usage in environmental chlamydia UWE25 reflects an evolutional divergence from pathogenic chlamydiae," *Gene*, vol. 368, no. 1-2, pp. 117–125, 2006.

[41] J. W. Drake and J. J. Holland, "Mutation rates among RNA viruses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 24, pp. 13910–13913, 1999.

[42] V. V. Khrustalev and E. V. Barkovsky, "Mutational pressure is a cause of inter- and intragenomic differences in GC-content of simplex and varicello viruses," *Computational Biology and Chemistry*, vol. 33, no. 4, pp. 295–302, 2009.

[43] W.-Q. Liu, J. Zhang, Y.-Q. Zhang et al., "Compare the differences of synonymous codon usage between the two species within cardiovirus," *Virology Journal*, vol. 8, article 325, 2011.

[44] J. Zhou, Z. Gao, J. Sun et al., "A comparative analysis on the synonymous codon usage pattern in viral functional genes and their translational initiation region of ASFV," *Virus Genes*, vol. 46, no. 2, pp. 271–279, 2013.

[45] T. Ikemura, "Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system," *Journal of Molecular Biology*, vol. 151, no. 3, pp. 389–409, 1981.

[46] T. Zhou, Z. H. Lu, and X. Sun, "The correlation between recombination rate and codon bias in yeast mainly results from mutational bias associated with recombination rather than hill-robertson interference," in *Proceedings of the 27th Annual International Conference of the Engineering in Medicine and Biology Society (IEEE-EMBS '05)*, pp. 4787–4790, September 2005.

[47] G. Marais, D. Mouchiroud, and L. Duret, "Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 10, pp. 5688–5692, 2001.

[48] S. Behura and D. Severson, "Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes," *Biological Reviews*, vol. 88, pp. 49–61, 2013.

[49] M. Worobey, "Extensive homologous recombination among widely divergent TT viruses," *Journal of Virology*, vol. 74, no. 16, pp. 7666–7670, 2000.

[50] E. van der Walt, E. P. Rybicki, A. Varsani et al., "Rapid host adaptation by extensive recombination," *Journal of General Virology*, vol. 90, no. 3, pp. 734–746, 2009.

[51] G.-Z. Han and M. Worobey, "Homologous recombination in negative sense RNA viruses," *Viruses*, vol. 3, no. 8, pp. 1358–1373, 2011.

[52] C.-Q. He, Z.-X. Xie, G.-Z. Han et al., "Homologous recombination as an evolutionary force in the avian influenza A virus," *Molecular Biology and Evolution*, vol. 26, no. 1, pp. 177–187, 2009.