JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 11, 2009
© Mary Ann Liebert, Inc.
Pp. 1539–1547
DOI: 10.1089/cmb.2009.0046

Categorizing Host-Dependent RNA Viruses by Principal Component Analysis of Their Codon Usage Preferences

*MING-WEI SU,[1] *HSIU-MAN LIN,[1] HANNA S. YUAN,[2] and WOEI-CHYN CHU[1]

ABSTRACT

Viruses have to exploit host transcription and translation mechanisms to replicate in a hostile host cellular environment, and therefore, it is likely that the infected host may impose pressure on viral evolution. In this study, we investigated differences in codon usage preferences among the highly mutable single strain RNA viruses which infect vertebrate or invertebrate hosts, respectively. We incorporate principal component analysis (PCA) and k-mean methods to clustering viruses infected with different type of hosts. The relative synonymous codon usage (RSCU) indices of all genes in 32 RNA viruses were calculated, and the correlation of the RSCU indices among different viruses was analyzed by the PCA. Our results show a positive correlation in codon usage preferences among viruses that target the same host category. Results of k-means clustering analysis further confirmed the statistical significance of this study, demonstrating that viruses infecting vertebrate hosts have different codon usage preferences to those of invertebrate viruses. Based on the analysis of the effective number of codons (ENC) in relation to the GC-content at the synonymous third codon position (GC3s), we further identified that mutational pressure was the dominant evolution driving force in making the different codon usage preferences. This study suggests a new and effective way to characterize host-dependent RNA viruses based on the codon usage pattern.

Key words: codon usage bias, k-means clustering, principal component analysis, RNA viruses, RSCU.

1. INTRODUCTION

Codon usage preference refers to the bias shown by different organisms and by different genes in the codon choices among a synonymous group of codons that all code the same amino acid (Andersson and Kurland, 1990; Kurland, 1993). A consistency of codon choices and the fact that highly expressed genes have stronger selective preferences was first observed in bacteria (Gouy and Gautier, 1982; Grantham et al., 1980). Subsequently, species-specific codon usage preferences were identified in many other organisms, including

[1] Institutes of Biomedical Engineering, National Yang-Ming University, Taipei, Taiwan, Republic of China.
[2] Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan, Republic of China.

*The first two authors made equal contribution to this work.

yeast (*Saccharomyces cerevisiae*) (Bennetzen and Hall, 1982; Sharp et al., 1986), worm (*Caenorhabditis elegans*) (Stenico et al., 1994), plant (*Arabidopsis thaliana*) (Chiapello et al., 1998), fruit fly (*Drosophila melanogaster*), and human (Sharp et al., 1988). Moreover, closely related organisms have more similar patterns of codon usage; for example, the codon usage preferences in *Salmonella typhimurium* closely resemble those in *Escherichia coli*, while those of all mammalian species and human are similar (Sharp et al., 1988).

The diverse codon usage preferences may arise from translation selection as the populations of iso-accepting tRNA contents vary in different organisms and tissues (Ikemura, 1985; Dittmar, 2006). Alternatively, mutational pressures have been shown to play dominant roles in codon selection in mammalian genomes (Francino and Ochman, 1999; Karlin and Mrazek, 1996). According to the latter scenario, the GC content at a chromosomal location shapes the codon usage preferences at that location. A gene located at GC rich regions preferentially utilizes GC ending codons, so the codon usage bias is mainly determined by mutation pressure. Codon usage and genome GC content are highly correlated when synonymous codons is closely correlated with the *GC* compositions on the three codon position (Mooers and holmes, 2000). In vertebrate DNA viruses mutation pressure rather than translation selection explains virus codon usage (Shackelton et al., 2006; Tao et al., 2009). The relationship between codon usage and tRNA availability was identified in bovine papillomavirus type 1 late gene (Zhou et al., 1999). The classical swine fever virus has shown the correlation between base composition and codon usage bias suggested that mutation pressure is a main factor in shaping codon usage (Tao et al., 2009). In general, choice of synonymous codons in unicellular organisms appears to be mainly determined by tRNA availability and other factors related to translational efficiency. In the multicellular organisms, different cells produce different proteins, so the simple relationship between tRNA abundance and codon usage preference is unexpected.

Previous studies of codon usage have focused on understanding the general cause of codon choices, and on using codon usage preference as an indicator of genome evolution (Karlin and Mrazek, 1996; Mooers and Holmes, 2000). The analysis or comparison of codon usage preferences in viral genomes has been investigated less extensively. Viruses are intracellular pathogens, so they have to exploit and co-evolve with host molecular mechanisms to prosper in a hostile cellular environment. It has been shown that papillomavirus capsid protein expression level depends on the match between the codon usage and tRNA availability in the host cells (Lukashov and Goudsmit, 2001). A study of flavivirus genomes also showed that tick-borne and mosquito-borne viruses have different base compositions and codon usage preferences (Jenkins et al., 2001). Another study of the codon usage pattern of human immunodeficiency virus type 1 (HIV-1) reported that the HIV-1 within a host changes codon usage patterns to more closely resemble human codon usage patterns (Meintjes and Rodrigo, 2005). Moreover, a survey of codon usage preference in human RNA viruses demonstrated that little variation exists among different genes and different viruses which targeting all to human (Jenkins and Holmes, 2003). These earlier results indicate that the codon usage preference of viruses do co-evolve with the host and viruses likely share similar codon usage bias to those of their hosts. Thus, these studies raise the possibility that the codon usage preference of a virus may be used as an indicator of its host categories.

Nevertheless, analysis of codon usage preferences among a number of species is complicated by the fact that there are 64 codons for 20 amino acids, and a vast number of genes in a single species. Earlier reports usually simplified the analysis by calculating only the preferences for specific nucleotides. For example, it has been shown that the HIV has a marked codon usage preference for the A nucleotide (van Hemert and Berkhout, 1995); pneumoviruses overall have less GC content (Pringle and Easton, 1997); and all RNA viruses are deficient in the dinucleotide CpG (Karlin et al., 1994). In this study, we calculated the relative synonymous codon usage (RSCU) index for each viral genome (Sharp et al., 1986). The RSCU indices of different viruses were then tabulated and analyzed by principal component analysis (PCA) (Hotelling, 1933). RSCU values reflect the preference for the use of a specific codon among other synonymous codons and have been used to analyze the codon usage preferences in influenza viruses (Zhou et al., 2005), HIV (Meintjes and Rodrigo, 2005), flaviviruses (Jenkins and Holmes, 2003), coronaviruses (Gu et al., 2004), Mimivirus (Sau et al., 2006), and bocavirus (Zhao et al., 2008). PCA, on the other hand, is a multivariate analysis method frequently used to highlight the similarities and differences of multivariate data (Hotelling, 1933).

Since RNA viruses are highly mutable, allowing for great adaptability and rapid evolution of RNA genomes (Steinhauer and Holland, 1987), we selected 17 vertebrate and 15 invertebrate RNA viruses for this study (Table 1). RSCU indices of these 32 viruses were calculated and analyzed by PCA and k-means

Table 1. The 32 Vertebrate and Invertebrate Host Viruses and Their GenBank Accession Numbers

Vertebrate host viruses	*Hosts*	*Invertebrate host viruses*	*Hosts*
1. Norwalk (NV) AF093797	Human	18. Nodamura virus (NoV) AF174533	Butterflies, Moths, Aedes
2. Human astrovirus type 1 (HAstV) L23513	Human	19. Black beetle virus (BBV) X02396	Black beetle
3. Measles virus (MeV) K01711	Human	20. Pariacoto virus (PaV) AF171942	Spodoptera eridania
4. Mumps virus (MuV) AF201473	Human	21. Flock House virus (FHV) X77156	Hordeum vulgare, Saccharomyces cerevisiae, Galleria mellonella, Costelytra zealandica
5. Human parainfluenza virus 3 (HPIV-3) D84095	Human	22. Boolarra virus (BoV) AF329080	Hepialidae (ghost moths)
6. Human parainfluenza virus 1 (HPIV-1) AF457102	Human	23. Euprosterna elaeasa virus (EeV) AF461742	Euprosterna elaeasa
7. Human parainfluenza virus 2 (HPIV-2) X57559	Human	24. Nudaurelia capensis beta virus (NCβV) AF102884	Pine Emperor moth
8. Human respiratory syncytial virus (HRSV) U39661	Human	25. Dendrolimus punctatus tetravirus (DpTV) AY594352	Dendrolimus punctatus
9. Sudan ebolavirus (SEBOV) AY729654	Human, Bat	26. Aphid lethal paralysis virus (ALPV) NC_004365	Aphis
10. Zaire ebola virus (ZEBOV) AF086833	Human, Bat	27. Cricket paralysis virus (CrPV) NC_003924	black field cricket
11. Marburg virus (MARV) 450908	Human, Monkey	28. Himetobi P virus (HiPV) NC_003782	Laodelphax striatellus
12. Influenza A virus (FLUAV H1N1) NC_002023, NC_002021, NC_002022, NC_002017, NC_002019, NC_002018, NC_002016, NC_002020	Human, Avian, Swine	29. Plautia stali intestine virus (PSIV) NC_003779	Plautia stali
13. Influenza A virus (FLUAV H3N2) NC_007366, NC_007367, NC_007368, NC_007369, NC_007370, NC_007371, NC_007372, NC_007373	Human, Avian, Swine	30. Rhopalosiphum padi virus (RhPV) NC_001874	Rhopalosiphum padi
14. Influenza B virus (FLUBV) AF102006, AF101989, AF102023, X13553, AF100396, L49385, X67013, AF100378	Human, Avian, Swine	31. Triatoma virus (TrV) NC_003783	Triatoma infestans
15. Influenza C virus (FLUCV) U20228, M28060, M28062, K01689, M17700, M22038	Human, Avian, Swine	32. Black queen cell virus (BQCV) NC_003784	Apis mellifera
16. Lassa virus (LASV) U12396, D10370, K00610	Human, Rodent		
17. Lymphocytic choriomeningitis virus (LCMV) AF004519, M20869	Human, Rodent, Monkey		

clustering methods. We found that a positive correlation in codon usage preference does exist between viruses targeting different category of hosts. Our results provide a line of evidence for the related preference in codon usages of hosts and viruses and suggest the efficacy of codon usage preference as a method to identify the host category of a virus.

2. METHODS

2.1. Genome sequences and the relative synonymous codon usage

The genomic sequences of the 32 RNA viruses were selected from GenBank. Genes with overlapping reading frame being removed in our analysis. Virus sequence numbers for all of the viruses used in this study are listed in Table 1. These viruses were from different virus families with various sequence compositions, and thus they are not closely related with each other.

The use of RSCU index instead of simply counting the codon numbers has the advantage to avoid amino acid composition bias (Perriere and Thioulouse, 2002). The RSCU for a particular codon (i) is given by

$$RSCU_i = X_i \Big/ \left(\sum XI/n\right)$$

where X_i is the usage number of the i^{th} codon for a given amino acid; $\sum XI$ is the sum of the usage number for all the synonymous codons of the same amino acid; and n is the number of synonymous codons for that amino acid (Sharp et al., 1986). A 32×59 matrix was constructed with 59 columns of RSCU index for each codon and 32 rows of virus species (Supplementary Material 1) (see online Supplementary Material at www.liebertonline.com).

2.2. Principal component analysis method

PCA is an orthogonal linear transformation that transforms the original data set into a new coordinate system. The greatest variance represented by any projection of the data comes to lie on the first coordinate, so-called the first principal component (PC), the second greatest variance on the second PC, and so on. One can use a few PCs to represent the data instead of the large number of original variables (in this case, 59 variables). PCA implementation was divided into several steps. First, a zero-mean 32×59 RSCU data matrix was constructed. Of all the 32 rows, each row denoted the codon usage pattern of a specific virus, manifested by its RSCU index, for a virus listed in Table 1. Second, the covariance matrix was calculated with the ij^{th} entries representing the covariance between the i^{th} and j^{th} codon. Third, the 59 eigenvalues and corresponding eigenvectors of the covariance matrix were computed. The second and third steps were executed using MATLAB 7.0 (The MathWorks, Inc., Framingham, MA). Finally, the 59 uncorrelated PCs were determined and listed in descending order, with the PC containing the highest amount of data matrix variation listed first (Supplementary Material 2) (see online Supplementary Material at www.liebertonline.com).

2.3. Cluster analysis

The PCA results were subsequently analyzed by k-mean clustering method. K-mean clustering algorithm was used to classify or to group data based on attributes or features into K groups. Clustering is done by minimizing the distances between data and the corresponding cluster center, making each cluster as close to each other as possible and far from data points in other clusters as possible. This algorithm uses iterative process over all K clusters. In this study, K = 2, i.e., vertebrate versus invertebrate host RNA viruses. The initial K centroids are randomly selected. In the first step, continuing reassigning data points to the nearest cluster centroid and recalculation the cluster centroids. Then again, the new data cluster will be reassigned based on the sum of their distances to the centroid. This iteration procedure stops when no new clustering events occurred.

2.4. Base compositional analysis

In order to determine the dominant driving force in shaping codon usage bias, the effective number of codons (ENC) and the GC-content at the synonymous third codon position (GC3s) were calculated and plotted (Wright, 1990). ENC is a general measurement of codon usage bias. Values range between 20 (only

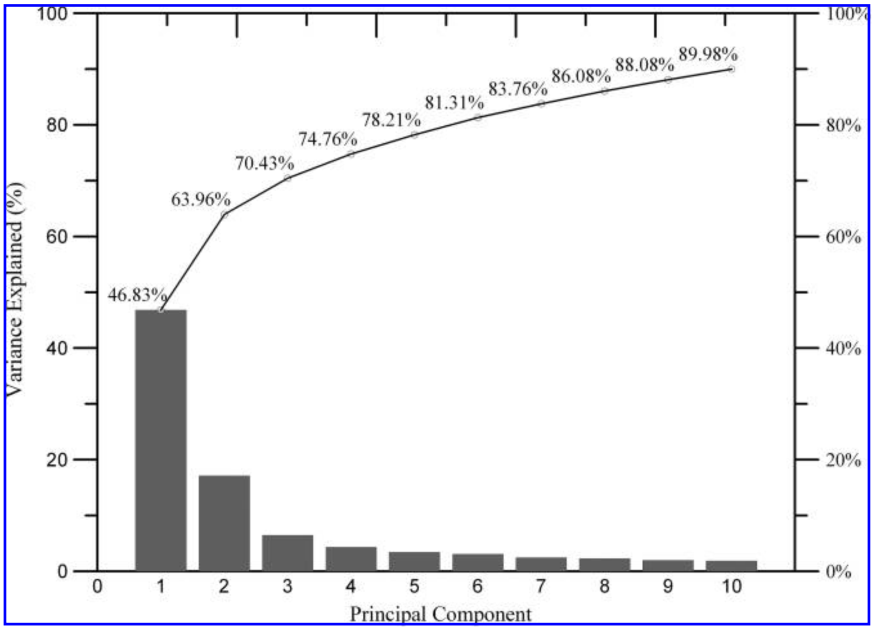

FIG. 1. Principal components (PCs) and variances explained in the analysis of the 59 relative synonymous codon usage (RSCU) indices. The first 10 PC vectors are listed on the right with the eigenvalue, variance explained (%), and accumulated variance (%). The plot on the left shows that the first three PCs have explained more than 70% of the variance of the original data.

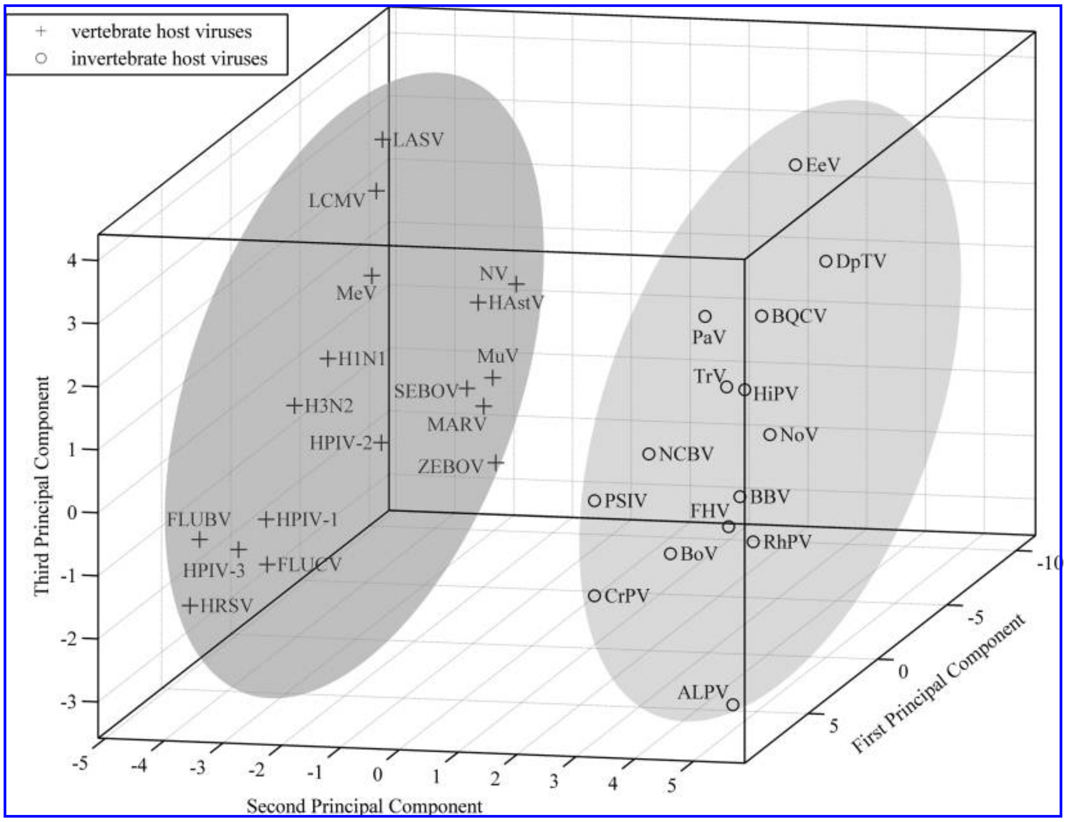

FIG. 2. Principal component analysis (PCA) plot for analysis of the relative synonymous codon usage (RSCU) indices of 32 RNA viruses. The PCA scores of the 32 viruses were plotted in a three-dimensional coordinate system using the first three principal component vectors as axes. The human epidemic vertebrate host RNA viruses (cross) and invertebrate RNA viruses (circle) are distinctively clustered into two different regions. The K-means clustering results were mapped onto the figure, displaying in dark- and light-shaded area.

one codon used by each amino acid) and 61 (except three stop codons, the remaining 61 out of 64 codons are used equally). If the sequence compositional constrain is the dominant driving force in shaping codon usage preference, all data points would lie on or below the expected curve.

3. RESULTS

3.1. Codon usage preferences of 32 RNA viruses

The genomic sequences of 32 RNA viruses (Table 1) were downloaded from NCBI GenBank. Seventeen of the 32 RNA viruses were genetically and ecologically diverse human epidemic RNA viruses (Jenkins and Holmes, 2003). The remaining 15 were invertebrate-host RNA viruses of *Nodaviridae, Dicistroviridae,* and *Tetraviridae*. The genomic sequences of these RNA viruses were downloaded from the taxonomic database provided by the International Committee on Taxonomy of Viruses (Büchen-Osmond, 2003). The human epidemic vertebrate host viruses were allotted a number of 1–17, whereas the invertebrate host viruses were allotted a number 18–32.

The coding regions of each viral genome were extracted and the RSCU index was calculated (Sharp et al., 1986). The RSCU index reflects the relative usage preference for a specific codon. The RSCU values larger than 1.0 indicate that a codon is favored over other synonymous codons; RSCU values of less than 1.0 indicate an unfavored codon; and RSCU values of exactly 1.0 indicate no preference. Because methionine and tryptophan are only associated with one single codon together with the three stop codons, were excluded from the analysis. A final 32×59 matrix was then constructed, in which the 59 columns listed the RSCU index for each codon and the 32 rows tabulated the different virus species (Supplementary Material 1) (see online Supplementary Material at www.liebertonline.com).

3.2. Identification of similarities and differences in codon usage preferences by PCA

To explore the codon usage pattern differences among these RNA viruses, the 32×59 RSCU matrix was processed by PCA to calculate the PCs in order to highlight the similarities and differences in codon usages. PCA is a classical data analysis method that identifies patterns and explores similarities and differences in a multivariate data set. Figure 1 shows the trend of the first 10 PCs. The first PC explained 46.83% of the variance among the 59 RSCU indices. The first two PCs accounted for 63.96% of the variance and the first three PCs accounted for 70.43% of the variance in codon usage. The variances of the total of 59 PCs generated from PCA are listed in Supplementary Material 2 (see online Supplementary Material at www.liebertonline.com).

Figure 2 is the three-dimensional PCA plot using the first three PCs of these 32 viruses as axes (the corresponding PCA coordinates are listed in Supplementary Material 3) (see Supplementary Material at www.liebertonline.com). This PCA score diagram showed that the vertebrate host viruses and invertebrate host viruses were clustered into two separate regions. All vertebrate-infected viruses displayed negative values on the second PC axes. The invertebrate-infected viruses were clustered in a region with positive values on the second PC axis.

K-mean clustering method was used to determine the statistical significance of the PCA results. K-mean results indicated that the 32 viruses can be distinctly clustered into two groups, vertebrate-host viruses, and invertebrate-host viruses, as shown in Figure 2. These results showed that the codon usage preference categorized by the first three PCs of PCA possessed sufficient information to differentiate RNA viruses that affect vertebrate hosts and invertebrate hosts, respectively.

3.3. Mutational pressures play a dominant role in shaping codon usage preferences of RNA viruses

To evaluate the dominant factor in shaping the codon usage preferences among the examined RNA viruses, we calculated the ENC in relation to the GC-content at the GC3s (Wright, 1990). ENC is an indicator for the extent of codon preference, ranging from 20 (maximum bias, only one codon used for each amino acid) to 61 (no bias, all synonymous codons are equally used for each amino acid). GC3s, on the other hand, is an indicator of sequence composition bias, ranging from 0% (no G or C at the third codon position) to 100% (only G or C is found at the third codon position). ENC versus GC3s plot reveals the influence of base composition constraints imposed on codon usage preferences.

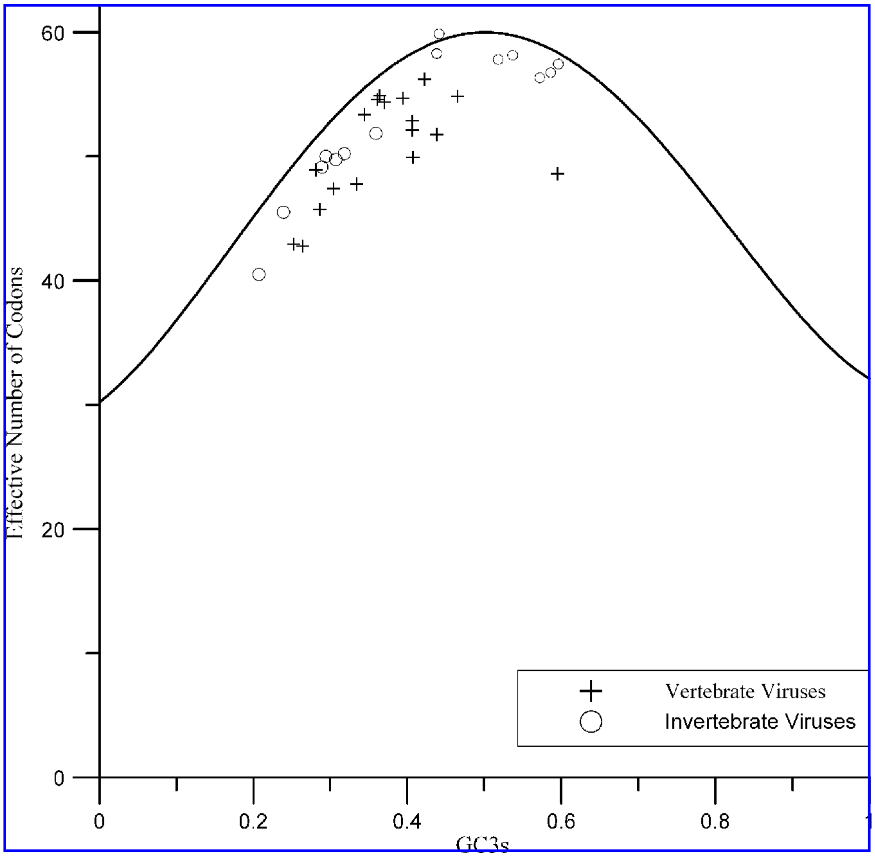

FIG. 3. Distribution of the effective number of codons (ENC) in relation to the GC-content at the synonymous third codon position (GC3s) of the 32 RNA viruses. The curve indicates the expected ENC with respect to GC3s when only the sequence compositional constraints account for the codon usage preferences.

As shown in Figure 3, the ENC values of the vertebrate-host viruses are more closely grouped together as compared to those of invertebrate viruses. Also, the GC3s of the invertebrate-host RNA viruses spanned a twice wider range, from 20.7% to 59.6%, as opposed to the GC3s of the vertebrate-host RNA viruses of 25.2% to 46.5%. The plotted reference curve in Figure 3 shows the expected ENC value with respect to GC3s when only the sequence compositional constraints account for the codon usage preferences.

4. DISCUSSION

We have used PCA method to analyze codon usage preference among vertebrate and invertebrate RNA viruses. PCA method is known for reducing vector space dimensions and to locate PCs that best represent the differences in a multivariate data set. With PCA, our results showed more than 70% of the variances of the 59 codon variables have been adequately represented by the first three PCs (Fig. 1). The PCA plot of the 32 RNA viruses (Fig. 2) showed the host-dependent RNA viruses being categorized into two distinct groups, indicating these viruses bear different codon usage preferences and the difference can be used to distinguish their host-dependency. Such clustering result has also been confirmed by the k-mean clustering method.

It is important to identify the determinants of codon choices in order to obtain a better understanding of viral evolution. Based on the ENC-GC3 plot (Fig. 3), the closely grouped and sparsely distributed ENC values and GC3s for the respective vertebrate-host and invertebrate-host RNA viruses suggest that the vertebrate-host viruses share similar codon usage preference than those invertebrate-host counterparts. Previous studies on human RNA viruses (Jenkins and Holmes, 2003) and *Flavivirus* (Jenkins et al., 2001) reported that mutation pressure is an important determinant of the codon bias observed. However, their results showed that weak translational selection may also have some influence in shaping codon usage bias. Because the ENCs of both the vertebrate-host or invertebrate-host RNA viruses are located under or on the expected curve, their codon usage preferences are presumably resulted from mutational pressures (Francino and Ochman, 1999).

Future studies would be necessary to further characterize the relationship of codon choices between hosts and viruses. For example, the analysis of codon usage preferences for the viruses from the same family, but targeting different hosts, may better reveal the influences of a host imposed onto a virus; the analysis of codon usage difference between viral structural and non-structural genes, or between highly expressed and lowly expressed genes may clarify the roles of translation and mutation pressures on codon choices. Moreover, microarray analysis on tRNA isoacceptors have identified not only a large diversity of tRNA genes, and also that the amounts of tRNA within the total cellular RNA vary widely among different human tissues (Goodenbour and Pan, 2006; Dittmar et al., 2006). Therefore, the studies of codon usage bias for the viruses targeting different human tissues may further enlighten the role of host tRNA levels imposed on viral genomic evolution. This study thus paves a way for the analysis of the genomic codon usage preferences among different viruses and hosts.

In conclusion, the PCA results show RNA viruses targeting vertebrate hosts share similar codon usage preferences, which are distinct from those of RNA viruses targeting invertebrate hosts. The ENC-GC3 plot demonstrates codon usage preferences in the studied RNA viruses may likely be influenced by their host species by mutational pressures.

ACKNOWLEDGMENTS

This work was supported by research grants from Ministry of Education aim for the Top University Plan, the National Science Council (NSC 97-2320-B-010-003-MY3), and the Summit Projects of Academia Sinica of the Republic of China. We would like to thank Prof. Carmay Lim for reading and discussion of this manuscript.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Andersson, S.G., and Kurland, C.G. 1990. Codon preferences in free-living microorganisms. *Microbiol. Rev*. 54, 198–210.

Büchen-Osmond, C. 2003. The Universal Virus Database ICTVdB. *Comput. Sci. Eng*. 5, 16–25.

Bennetzen, J.L., and Hall, B.D. 1982. Codon selection in yeast. *J. Biol. Chem*. 257, 3026–3031.

Chiapello, H., Lisacek, F., Caboche, M., and Hénaut, A. 1998. Codon usage and gene function are related in sequences of *Arabidopsis thaliana. Gene* 209, GC1–GC38.

Dittmar, K.A., Goodenbour, J.M., and Pan, T. 2006. Tissue-specific differences in human transfer RNA expression. *PLoS Genet*. 2, 2107–2114.

Francino, M.P., and Ochman, H. 1999. Isochores result from mutation not selection. *Nature* 400, 30–31.

Goodenbour, J.M., and Pan, T. 2006. Diversity of tRNA genes in eukaryotes. *Nucleic Acids Res*. 34, 6137–6146.

Gouy, M., and Gautier, C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res*. 10, 7055–7074.

Grantham, R., Gautier, C., Gouy, M., et al. 1980. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res*. 8, r49–r62.

Gu, W., Zhou, T., Ma, J., et al. 2004. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. *Virus Res*. 101, 155–161.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol*. 24, 498–520.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol*. 2, 13–34.

Jenkins, G.M., and Holmes, E.C. 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res. 92, 1–7.

Jenkins, G.M., Pagel, M., Gould, E.A., et al. 2001. Evolution of base composition and codon usage bias in the genus Flavivirus. J. Mol. Evol. 52, 383–390.

Karlin, S., Doerfler, W., and Cardon, L.R. 1994. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J. Virol. 68, 2889–2897.

Karlin, S., and Mrazek, J. 1996. What drives codon choices in human genes? J. Mol. Biol. 262, 459–472.

Kurland, C.G. 1993. Major codon preference: theme and variations. Biochem. Soc. Trans. 21, 841–846.

Lukashov, V.V., and Goudsmit, J. 2001. Evolutionary relationships among parvoviruses: virus-host coevolution among autonomous primate parvoviruses and links between adeno-associated and avian parvoviruses. J. Virol. 75, 2729–2740.

Meintjes, P.L., and Rodrigo, A.G. 2005. Evolution of relative synonymous codon usage in human immunodeficiency virus type-1. J. Bioinform. Comput. Biol. 3, 157–168.

Mooers, A.O., and Holmes, E.C. 2000. The evolution of base composition and phylogenetic inference. Trends Ecol. Evol. 15, 365–369.

Perriere, G., and Thioulouse, J. 2002. Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Res. 30, 4548–4555.

Pringle, C.R., and Easton, A.J. 1997. Monopartite negative strand RNA genomes. Semin. Virol. 8, 49–57.

Sau, K., Gupta, S.K., Sau, S., et al. 2006. Factors influencing synonymous codon and amino acid usage biases in Mimivirus. Biosystems 85, 107–113.

Shackelton, L.A., Parrish, C.R., and Holmes, E.C. 2006. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. J. Mol. Evol. 62, 551–563.

Sharp, P.M., Cowe, E., Higgins, D.G., et al. 1988. Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. Nucleic Acids Res. 16, 8207–8211.

Sharp, P.M., Tuohy, T.M., and Mosurski, K.R. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14, 5125–5143.

Steinhauer, D.A., and Holland, J.J. 1987. Rapid evolution of RNA viruses. *Annu.* Rev. Microbiol. 41, 409–433.

Stenico, M., Lloyd, A.T., and Sharp, P.M. 1994. Codon usage in *Caenorhabditis elegans:* delineation of translational selection and mutational biases. Nucleic Acids Res. 22, 2437–2446.

Tao, P., Dai, L., Luo, M., et al. 2009. Analysis of synonymous codon usage in classical swine fever virus. Virus Genes 38, 104–112.

van Hemert, F.J., and Berkhout, B. 1995. The tendency of lentiviral open reading frames to become A-rich: constraints imposed by viral genome organization and cellular tRNA availability. J. Mol. Evol. 41, 132–140.

Wright, F. 1990. The "effective number of codons" used in a gene. Gene 87, 23–29.

Zhao, S., Zhang, Q., Liu, X., et al. 2008. Analysis of synonymous codon usage in 11 human bocavirus isolates. Biosystems 92, 207–214.

Zhou, J., Liu, W.J., Peng, S.W., et al. 1999. Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. J. Virol. 73, 4972–4982.

Zhou, T., Gu, W., Ma, J., et al. 2005. Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses. Biosystems 81, 77–86.

Address correspondence to:
Dr. Woei-Chyn Chu
Institutes of Biomedical Engineering
National Yang-Ming University
155 Sec. 2, Linong Street
Peitou 112, Taipei, Taiwan ROC

E-mail: wchu@ym.edu.tw

and

Dr. Hanna S. Yuan
Institute of Molecular Biology
Academia Sinica
Taipei, Taiwan ROC