Analysis of synonymous codon usage patterns in duck hepatitis A virus: a comparison on the roles of mutational pressure and natural selection

Youhua Chen · You-Fang Chen

Received: 4 October 2013 / Accepted: 6 January 2014 / Published online: 25 January 2014
© Indian Virological Society 2014

Abstract Codon usage patterns of duck hepatitis A virus (HAV) were studied in the present study. The major trends of codon usage patterns were analyzed using principal component analysis on the basis of the relative synonymous codon usage values. Correlation analysis was utilized to reveal the associations of the first two major axes of PCA and nucleotide- or amino acid-relevant indices. Our results showed that compositional constraint and/or mutational pressure are major factors influencing codon usage bias patterns of HAV. However, the influence of natural selection is also prevalent, as indicated by strongly significant correlations between the hydrophobicity, aromaticity, aliphaticity and ionization and the first axis of PCA. Also, ionization could characterize the second axis of PCA. At last, maximum likelihood phylogenetic analysis shows that there are no remarkable geographic clustering patterns of HAV strains in the phylogenetic tree. However, through MANOVA test, there are significant differences on the codon usage patterns among HAV strains from different countries. In conclusion, both mutational pressure and natural selection are of equally great importance to codon usage patterns of duck HAV genomes.

Keywords Synonymous substitution · Multivariate statistics · Optimal codons · Compositional constraint · Mutational pressure · Natural selection

Introduction

Synonymous codon usage is a phenomenon describing that different codons encode a same amino acid [2]. Synonymous codons are not used randomly [4, 16]. Many factors may contribute to synonymous codon usage bias, for example, mutational pressure [12], natural selection [21], and translational efficiency [16] and so on.

Duck hepatitis A virus (HAV), is the only species in the genus *Avihepatovirus* of the family *Picornaviridae*, containing a large open-reading frame (ORF) [23]. The genome size of hepatitis virus has around 7500 nucleotides. Synonymous codon usage bias patterns of hepatitis viruses have been extensively quantified in recent studies for all the five subtypes (A, B, C, D, and E) [1, 3, 7, 10, 23]. However, the analysis on HAV was incomplete [23], because in that previous study, only 21 genomes were analyzed. Up to date, there are much more genome sequences released in GenBank database for HAV. In a recent query of GenBank database, we found that over 100 genome sequences are available for HAV. As such, a comprehensive analysis on the codon usage bias patterns of duck HAV may be necessary to offer a holistic understanding of codon usage patterns in the virus evolution.

Although there are many literature contributing to understand codon usage bias patterns in virus evolution [9, 21], there is still a great lack of developing a quantitative procedure to analyze the relative importance of each influencing factor.

Therefore, there should have at least two novelties for the present study: (1) introduce two new chemical properties of

Electronic supplementary material The online version of this article (doi:10.1007/s13337-014-0191-2) contains supplementary material, which is available to authorized users.

Y. Chen (✉)
Department of Renewable Resources, University of Alberta,
Edmonton, AB T6G 2H1, Canada
e-mail: youhuach@gmail.com

Y.-F. Chen
College of Software, Harbin Normal University, Harbin,
Heilongjiang, China

Fig. 1 Maximum likelihood phylogenetic tree reconstructed for 90 HAV strains. Each clade is labeled with support percentage over the 1000 bootstrapping replicates. The labels on the tips indicate the GenBank accession numbers. At the end of these labels, the letters C, T, S and V indicate the strain origins: Mainland China, Taiwan, South Korea, and Vietnam

amino acids (ionization and aliphaticity) and investigate their potential roles in shaping codon usage bias patterns of virus evolution quantitatively; (2) provide an updated analysis of synonymous codon usage bias patterns for duck HAV and elucidate the underlying driving mechanisms of codon usage bias for the virus.

Materials and methods

Sequences

The 101 available complete genomes of duck HAV were obtained from GenBank (<http://www.ncbi.nlm.nih.gov/>). The coding regions of the genomes were retained for being analyzed on codon usage patterns. Among the 101 genomes, there are only 90 ones being unique from each other. As such, our subsequent analyses were performed on the 90 genomes, which were sampled from mainland China, South Korea, Taiwan of China and Vietnam respectively. The accession numbers for the genomes and other information were available in Supplementary material. By using the alignment of the whole genomic sequences, a phylogenetic tree reconstructed using maximum likelihood method [18] with 1000 bootstrapping replicates was presented in Fig. 1.

Measures of relative synonymous codon usage (RSCU)

Relative synonymous codon usage (RSCU) values of each codon in a gene were calculated to investigate the characteristics of synonymous codon usage. The RSCU index was calculated as follows [4, 14],

$$RSCU = \frac{g_{ij} \times n_j}{\sum_i g_{ij}}$$

where g_{ij} is the observed number of the i -th codon for the j -th amino acid which has n_j kinds of synonymous codons. Higher (or lower) RSCU values, higher (or lower) frequencies of the codons being chosen. When the RSCU values are close to 1, indicating that the codon is chosen randomly and evenly.

Effective number of codons

The effective number of codons (ENC) is a measure of bias from equal codon usage in a gene [22]. The calculation formula is,

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

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

$$F_k = \frac{nS - 1}{n - 1}$$

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

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

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

In our study, we calculated an improved ENC index [17]. ENC 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 revealing the relationship between GC3s and ENC values, the expected ENC values for different GC3s were calculated as follows,

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

where s denotes the value of GC3s [9]. The observed and expected ENC values were compared to determine the influence of nucleotide compositional constraint on structuring synonymous codon usage bias.

Codon adaptation index (CAI) and the correlation between CAI and ENC

The codon adaptation index (CAI) estimates the extent of bias toward codons that are known to be favored in highly expressed genes [15]. Thus, CAI index is one of the most powerful indices for quantifying the influence of natural selection on codon usage patterns. We simply USE the codon usage table for *Escherichia coli* as the reference to compute the CAI index for HAV.

Because ENC can be driven by both mutation and selection [20], it might not be a robust index to show the relative contribution of mutation and selection on structuring codon usage patterns. Moreover, the codon usage bias of base composition of the ORFs of a species with A/T biased genomes will behave differentially than those species with G/C biased genomes, as such, ENC-GC3s plot might be potentially misleading. In contrast, CAI is suggested as the most powerful index for showing the influence of natural selection on codon usage patterns [4, 15, 20]. If CAI value is high, then codon usage bias is extremely high and the influence of natural selection is

prevailing. Thus, some previous studies [20] suggested to utilize the correlation between CAI and ENC values to show the relative influence of mutation and selection. If selection is more important, the correlation (r) between the two quantities should be very high ($r \rightarrow -1$). In contrast, if mutation force is more important, r should approach 0 (no correlation).

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 of viruses [5, 11]. As such, we quantified both indices to reveal the evidence of natural selection on codon usage bias.

For hydrophobicity index [11], it is calculated as,

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

where N is the number of amino acids, and k_i is the hydrophobic index of the i -th amino acid.

For Aromo index [11], it is calculated as,

$$\text{Aromo} = \frac{1}{N} \sum_{i=1}^N v_i$$

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

In the present study, we also quantified other chemical properties of amino acids and reveal their possible influences on affecting codon usage biases, including the measures of aliphaticity and ionization (those amino acids with ionizable side chains).

For aliphaticity index, it is calculated as,

$$\text{Alipha} = \frac{1}{N} \sum_{i=1}^N l_i$$

where l_i is either 1 (for an aliphatic amino acid) or 0 (for a non-aliphatic amino acid), N is the number of amino acids.

For ionization index,

$$\text{Ion} = \frac{1}{N} \sum_{i=1}^N m_i$$

where m_i is either 1 (for an ionizable amino acid) or 0 (for a non-ionizable amino acid), N is the number of amino acids.

Relative dinucleotide abundance

Because dinucleotide biases can affect codon usage bias, the relative abundance of dinucleotides (RAD) in the coding regions of the genomes was assessed using a

Table 1 Synonymous codon usage bias based on RSCU values in the genomes of duck hepatitis virus

Codon	Amino acid	ObsFreq	RSCU	Codon	Amino acid	ObsFreq	RSCU
UGA	*	63	2.1	AUG	M	4952	1
UAG	*	2	0.067	AAC	N	2715	0.593
UAA	*	25	0.833	AAU	N	6449	1.407
GCU	A	4196	1.432	CCA	P	5074	2.068
GCG	A	281	0.096	CCC	P	1722	0.702
GCC	A	3247	1.108	CCU	P	2866	1.168
GCA	A	3995	1.364	CCG	P	150	0.061
UGU	C	3461	1.412	CAA	Q	4546	1.045
UGC	C	1440	0.588	CAG	Q	4157	0.955
GAU	D	7169	1.296	AGA	R	2738	1.116
GAC	D	3898	0.704	AGG	R	2171	0.884
GAG	E	4746	0.865	CGA	R	1013	0.848
GAA	E	6231	1.135	CGC	R	1675	1.403
UUU	F	6656	1.413	CGG	R	878	0.735
UUC	F	2763	0.587	CGU	R	1211	1.014
GGU	G	3872	1.152	AGC	S	1018	0.599
GGG	G	2729	0.812	AGU	S	2381	1.401
GGC	G	3227	0.96	UCA	S	4078	1.565
GGA	G	3622	1.077	UCC	S	1629	0.625
CAC	H	1853	0.827	UCG	S	225	0.086
CAU	H	2626	1.173	UCU	S	4492	1.724
AUU	I	7056	1.781	ACC	T	2774	0.754
AUA	I	2409	0.608	ACA	T	5917	1.608
AUC	I	2420	0.611	ACG	T	357	0.097
AAA	K	6727	1.137	ACU	T	5675	1.542
AAG	K	5101	0.863	GUU	V	5227	1.479
CUA	L	1506	0.578	GUG	V	4407	1.247
CUC	L	2638	1.013	GUC	V	2673	0.756
CUG	L	2189	0.84	GUA	V	1830	0.518
CUU	L	4087	1.569	UGG	W	3361	1
UUA	L	2189	0.596	UAC	Y	1522	0.461
UUG	L	5154	1.404	UAU	Y	5085	1.539

Boldface indicated the most preferentially used codon for each amino acid (only synonymous codons are considered). ObsFreq indicated observed number of the codon in the genomes

previous method [8] with an odds ratio $\rho_{xy} = f_{xy}/f_x f_y$, where f_x and f_y denote the frequency of the nucleotide X and Y respectively. f_{xy} is the frequency of the dinucleotide XY (there are 16 dinucleotides in a total). When $\rho_{xy} > 1.23$ (or $\rho_{xy} < 0.78$), the dinucleotide pair XY is regarded as over-represented (or under-represented) on its relative abundance in the genomes [24].

MANOVA test on the codon usage difference among HAV strains from different regions

We performed a MANOVA test to check whether there were significant codon usage difference among the 90 HAV strains for each sampling country/region (mainland

China, Taiwan of China, South Korea and Vietnam). All the abovementioned codon usage indices were calculated for each sequence and used for MANOVA test. The Pillai statistic was calculated to test for significance.

Multivariate statistical analysis and correlation analysis

Principal component analysis (PCA) was conducted to investigate the major trend involved in the codon usage patterns of the genomes from different virus strains, which were measured by corresponding RSCU values for the 59 codons (after excluding Met, Trp, and three termination codons) [23].

Table 2 Correlation analyses between the A, U, C, G, GC content and A3s, U3s, C3s, G3s, GC3s content in all ORFs of HAV

	A3s	C3s	G3s	U3s	GC3s
A	0.993	-0.51	0.225	-0.786	-0.427
C	-0.523	0.98	-0.154	-0.058 ^{NS}	0.929
G	0.444	-0.301	0.904	-0.507	0.043 ^{NS}
U	-0.811	-0.007 ^{NS}	-0.426	0.992	-0.171 ^{NS}
GC	-0.294	0.825	0.313	-0.32	0.951

All the correlations are significant except the values marked with *NS* which indicated non-significant correlations ($p > 0.05$)

The first two axes of PCA (PC1 and PC2) were then subjected to correlation analysis. A Spearman's rank correlation analysis was used to infer the relationships between different codon usage indices and the first two principal components (PC1 and PC2) of duck HAV RSCU values.

Results

Synonymous codon usage in duck hepatitis A virus

The overall RSCU values of all 64 codons in the 90 genomes of HAV were summarized in Table 1. All codons which are used preferentially are A or U-ended, except the two amino acids Arg and Val, which chose CGC and CUG as preferred codons. RSCU results are not surprising, because HAV genomes are A+U rich, the ranges of AU content for different virus strains are between 0.566 and 0.583, with an average 0.573 and SD 0.003.

Compositional properties of HAV ORFs, PCA analysis and correlation analysis

The values of A, U, C, G and GC were compared with the values of A3s, C3s, G3s, U3s and GC3s respectively (Table 2). Being inconsistent to the previous study [23], which presented complex correlations amongst these quantities, we found that most quantities were strongly significantly correlated, except for the insignificant correlations between U and C3s, U and GC3s, G and GC3s, C and U3s. This result implied that mutation pressure from the base composition was a principal driver structuring codon usage bias patterns of the virus because the effects of mutational pressure could happen over all the codon positions.

PCA showed that the first two axes could explain 75 % of the total variance involved in codon usage patterns of HAV (Fig. 2; Table 3). The first axis is paramount, explaining around 65 % of total variance. The correlation between these compositional quantities and the first two

Fig. 2 Scatter plot of different duck HAV ORFs on the first and second principal axes derived from PCA analysis. The first axis (PC1) explained 64.7 % of the total variance, while the second axis (PC2) accounted for 7.2 % of the total variance**Table 3** The correlation between the first two axes of PCA and different nucleotide-relevant indices

	PC1	PC2
Standard deviation	0.688	0.229
Explained variance (%)	64.80	7.20
Cumulative proportion of variance (%)	64.80	72
Correlation		
A	0.927	-0.03 ^{NS}
C	-0.491	-0.728
G	0.646	0.03 ^{NS}
T	-0.836	0.464
GC	-0.157 ^{NS}	-0.713
A3s	0.946	-0.023 ^{NS}
C3s	-0.507	-0.729
G3s	0.5	0.106 ^{NS}
T3s	-0.807	0.491
GC3s	-0.319	-0.693
ENC	0.398	-0.652
CAI	-0.976	-0.139 ^{NS}
GRAVY	-0.917	-0.048 ^{NS}
Aromo	-0.577	0.009 ^{NS}
Alipha	-0.755	0.008 ^{NS}
Ion	0.638	0.286

NS denotes non-significant correlations, others are significant correlations with $p < 0.05$. The three most important factors were marked in bold

axes of PCA showed that compositional constraint is a principal driver of codon usage bias patterns of HAV. As showed in Table 3, PC1 is significantly correlated with all

Fig. 3 ENC used in each ORF plotted against the GC3c content. The continuous curve indicates the null relationship between GC3s and ENC in the absence of selection but only mutation

Fig. 4 Correlation between CAI and ENC values of HAV ORFs

the quantities except GC content. In a comparison, PC2 is most correlated with C3s and GC3s, similar to the finding in the previous study [23]. It is found that the two new introduced indices, aliphaticity and ionization, are also significantly correlated with PCA major axes. As seen in Table 3, PC1 is negatively correlated with aliphaticity but positively correlated with ionization. Additionally, ionization could characterize the second axis of PCA as well with a correlation coefficient $r = 0.286$ ($p < 0.05$).

Analysis on ENC and CAI indices

The values of ENC (see Supplementary material) among the HAV ORFs are similar, varying from 49.6 to 52.4, with a

mean of 50.8 and SD of 0.62, suggesting that the codon usage bias is relatively low. The low codon usage bias is also proven by the low CAI indices over different genomes (see Supplementary material). The observed ENC values were all laid under the expected ENC curve considerably but not crossing or closing the null curve (Fig. 3), implying the importance of other factors influencing the codon usage pattern of the virus, in addition to mutational pressure. This result is basically analogous to the plot in a previous study [23].

Moreover, the correlation ($r = -0.29$, $p < 0.05$) between CAI and ENC (Fig. 4) implied that both selection and mutation play essential roles on structuring codon usage biases of HAV ORFs because as predicted above, when selection is the principal driver of codon usage biases, the correlation between CAI and ENC should be $r = -1$.

Relative abundance of dinucleotides

For the RAD, there were two highly chosen dinucleotides ($\rho_{xy} > 1.23$), UG and CA, and two under-represented ones ($\rho_{xy} < 0.78$), CG and UA (Table 4). The relative abundance of CpG showed that most remarkable low value (0.289 ± 0.038). The correlation between the RAD and the first two axes of PCA showed that most dinucleotides (14) are significantly correlated with PC1, while only 7 ones are significantly correlated with PC2 (Table 4).

Codon usage difference comparison among HAV sequences from different regions

Evolutionary divergence among HAV strains from different origins on the codon usage patterns is significant. Through MANOVA test (Table 5), it is found that the strains from different countries have significant differences ($F_{36,231} = 2.3$, $p < 0.05$). As such, HAV sequences from different origins presented significant codon usage divergence.

Discussion

Compositional constraint and/or mutational pressure greatly influences codon usage bias of HAV

For the first axis of PCA (Table 3), base A content and A3s content were the most important factors in representing the major trend of codon usage bias patterns of HAV. For the second axis of PCA (Table 3), it was found that both GC and GC3s were the characterizing factors the second gradient of codon usage bias patterns of HAV, as proven by the strong positive correlation between the two indices and PC2 values. As a consequence, the principal trends in

Table 4 Summary of correlation analysis between the first and second principal components in PCA analysis and relative abundances of 16 dinucleotides in HAV genomes

	AA	AC	AG	AU	CA	CC	CG	CU
ρ_{xy}	1.034 ± 0.049	1.072 ± 0.035	1.007 ± 0.089	1.015 ± 0.042	1.389 ± 0.04	1.143 ± 0.028	0.289 ± 0.038	1.035 ± 0.031
PC1	-0.977*	0.652*	0.988*	-0.846*	0.782 *	-0.019 ^{NS}	-0.596 *	-0.507*
PC2	0.026 ^{NS}	0.028 ^{NS}	-0.01 ^{NS}	0.104 ^{NS}	-0.245 *	0.199*	0.566 *	-0.014 ^{NS}
	GA	GC	GG	GU	UA	UC	UG	UU
ρ_{xy}	1.038 ± 0.018	0.935 ± 0.057	1.105 ± 0.033	0.879 ± 0.024	0.558 ± 0.03	0.913 ± 0.027	1.456 ± 0.031	1.103 ± 0.021
PC1	0.036 ^{NS}	-0.78*	-0.835*	0.393*	0.558 *	0.435*	-0.239 *	0.701*
PC2	-0.082 ^{NS}	-0.234*	0.095 ^{NS}	0.197*	0.326 *	-0.041 ^{NS}	-0.732 *	-0.035 ^{NS}

NS denotes non-significant correlations, while * denotes significant correlations with $p < 0.05$. ρ_{xy} for each dinucleotide is represented as mean ± SD. Values in boldface denote the over-represented ($\rho_{xy} > 1.33$) and under-represented ($\rho_{xy} < 0.78$) dinucleotides

codon usage patterns of HAV could be mainly explained by compositional constraint.

Further, it was found that the correlation between the overall base frequency and the base frequency at the third codon position are highly correlated (Table 2). As such, mutational pressure is very vital too for explaining codon usage patterns.

CpG suppression and relative abundance of dinucleotides could lead to codon usage bias in HAV genomes

CpG suppression could also contributed to codon usage bias [6, 19]. In our study, we further found such an evidence for HAV. PC1 are strongly correlated with the RAD AA and AG. Interestingly, it was found that there are strong correlations ($r > 0.9$, $p < 0.001$) between the relative abundances of AA and AG and the hydrophicity, the frequency of base A, A3s and CAI index, indicating the influence of mutational pressure and natural selection on codon usage patterns of HAV is synergistic.

To reveal the effects of CpG under-representation on codon usage bias, the RSCU values of the eight codons that contain CpG (CCG, GCG, UCG, ACG, CGC, CGG, CGU and CGA) were analyzed. Of these codons, we found that four ones, CCG (RSCU = 0.061), GCG (RSCU = 0.096), UCG (RSCU = 0.086), and ACG (RSCU = 0.097), were heavily suppressed, while two ones, CGG (RSCU = 0.735) and CGA (RSCU = 0.848) were slightly suppressed. Finally, only one codon CGU (RSCU = 1.014) is selected randomly and the last one CGC (RSCU = 1.403) is selected preferentially (Table 1).

Besides, as evidenced by the correlation analysis between the first two axes of PCA and the relative abundances of dinucleotides, we found that other dinucleotides were also influential to affect codon usage bias patterns of HAV. For example, when we checked the under-representation of UpA, of the eight codons that contain UpA, we found 7 ones were markedly or slightly suppressed (RSCU < 1), except for the codon UAU, which became a preferred optimal codon (RSCU = 1.539) for amino acid Tyr. Interestingly, the over-representation of two

Table 5 MANOVA test on the codon usage difference among HAV sequences from different regions

	DF	Pillai	F	DF	Prob(>F)
Strain origin	3	0.79	2.34	36/231	<0.05
Residuals	86				

All the codon usage indices presented in Supplementary material were used in the test but excluding A, T, G and C contents to avoid multicollinearity issue

dinucleotides UpG and CpA had different stories. Initially, we would expect that most codons containing UpG or CpA should be largely preferred, however, among the seven codons containing UpG (excluding one stop codon), it was found that only three ones (UUG, UGU, GUG) had $RSCU > 1.2$. In contrast, the codons containing CpA was consistent with our prediction, for all the eight codons, 4 were selected as preferred optimal codons, and the other 4 ones were selected as sub-optimal codons (the ones being secondly preferred) (Table 1).

The signature of natural selection on structuring codon usage bias in HAV genomes

Aromo and hydrophobicity are two of the factors in influencing codon usage of HAV. The PC1 values have significant negative correlations with both Aromo and hydrophobicity values for HAV, suggesting that natural selection should be involved as well for the evolution of codon usage of hepatitis A virus. The importance of Aromo has been revealed in some other viruses, for example, bovine viral diarrhea virus [21] and classical swine fever virus [19]. However, the emerging influence of hydrophobicity is rarely reported previously. Our study clearly showed that hydrophobicity could also affect HAV codon usage patterns. Moreover, it is found that both aliphaticity and ionization are significantly correlated with PC1 and ionization is significantly correlated with PC2. More importantly, the most powerful index for quantifying natural selection, CAI, is found to be the top one important variable correlated with PC1 ($r = -0.976$; Table 3). As such, all these significant correlations solidly imply the importance of natural selection on structuring codon usage patterns of HAV.

For human RNA viruses, it was found that mutational pressure is much more influential than natural selection in influencing codon usage patterns because RAN viruses have higher mutation rates [23]. However, in our study, by using the updated genome information of HAV, we found a prevailing influence of natural selection on structuring codon usage bias of HAV.

Geographic variation and evolutionary divergence of codon usage bias patterns among HAV strains

The above analysis and discussion have holistically revealed the codon usage bias patterns for HAV genomes. However, it has not been determined yet whether the codon usage bias patterns can be linked to the HAV evolution and whether different HAV strains sampled from different geographic regions should differentiation on the codon usage. As such, this part of the discussion focuses on the

geographic and phylogenetic patterns of codon usage patterns of HAV strains.

It is commonly known that most HAV strains are from East Asian region, especially China [13]. A previous study [13] on the phylogenetic analysis of HAV strains in the region show that there are no remarkable geographic differentiation among HAV strains. The strains from different countries (e.g., South Korea, China and Vietnam) usually are very similar from each other. Our phylogenetic analysis (Fig. 1) using maximum likelihood method further confirms this finding: sequences sampled from different regions tend to group with each other. For example, DQ256133 from South Korea is very close to NC009750 from mainland China (Fig. 1). Thus, no distinct geographic clades are found. In contrast, evolutionary divergence among HAV strains on the codon usage patterns is significant (Table 5), even though there are no remarkable geographic clustering patterns on the phylogenetic tree (Fig. 1).

Our study might be a comprehensive study on codon usage patterns of HAV up to date. Codon usage bias is relatively low in HAV genomes. Both mutational pressure and natural selection are of great importance on structuring the codon usage patterns of the virus. The introduction of new chemical properties of amino acids (aliphaticity and ionization) would allow ones to further quantify the importance of natural selection on codon usage bias patterns. At last, there are no remarkable associations between phylogenetic positioning and geographic distribution of HAV strains, but there are significant differences on the codon usage patterns among HAV strains from different countries through MANOVA test.

Acknowledgments We acknowledge two anonymous reviewers for their insightful comments on helping improve the quality of the work. This work is supported by China Scholarship Council.

References

1. Bashal A, Mukherjee R, Chakraborty C. Synonymous codon usage pattern analysis of hepatitis D virus. *Virus Res.* 2013; 173(2):350–3. doi:10.1016/j.virusres.2013.01.007.
2. Botzman M, Margalit H. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biol.* 2011;12:R109.
3. Bouquet J, Cherel P, Pavio N. Genetic characterization and codon usage bias of full-length hepatitis E virus sequences shed new lights on genotypic distribution, host restriction and genome evolution. *Infect Genet Evol.* 2012;12:1842–53.
4. Chen Y. A comparison of synonymous codon usage bias patterns in DNA and RNA virus genomes: quantifying the relative importance of mutational pressure and natural selection. *Biomed Res Int.* 2013;2013:406342.
5. Das S, Paul S, Dutta C. Synonymous codon usage in adenoviruses: influence of mutation, selection and protein hydrophathy. *Virus Res.* 2006;117:227–36.

6. Gong X, Fan S, Cui Z, Li X. The CpG suppression of polymerase segments and its impact on codon usage bias in H1N1 influenza virus. *Acta Biophys Sin.* 2011;27:537–44.
7. Hu J, Wang Q, Zhang J, et al. The characteristic of codon usage pattern and its evolution of hepatitis C virus. *Infect Genet Evol.* 2011;11:2098–102.
8. Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* 1995;11:283–90.
9. Liu X, Wu C, Chen A. Codon usage bias and recombination events for neuraminidase and hemagglutinin genes in Chinese isolates of influenza A virus subtype H9N2. *Arch Virol.* 2010;155:685–93.
10. Ma M, Ha X, Ling H, et al. The characteristics of the synonymous codon usage in hepatitis B virus and the effects of host on the virus in codon usage pattern. *Virol J.* 2011;8:544.
11. Peden J. Analysis of codon usage. PhD thesis. Department of Genetics, University of Nottingham; 1999. p. 1–215.
12. Rao Y, Wu G, Wang Z, et al. Mutation bias is the driving force of codon usage in the *Gallus gallus* genome. *DNA Res.* 2011;18:499–512.
13. Ren L, Li J, Bi Y, et al. Overview on duck virus hepatitis A. *Chin J Biotechnol.* 2012;28:789–99.
14. Sharp P, Li W. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for “rare” codon. *Nucleic Acids Res.* 1986;14:7737–49.
15. Sharp P, Li W. The codon adaptation index a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 1987;15:1281–95.
16. Stoletski N, Eyre-Walker A. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol Biol Evol.* 2007;24:374–81.
17. Sun X, Yang Q, Xia X. An improved implementation of effective number of codons (ENC). *Mol Biol Evol.* 2013;30:191–6.
18. Tamura K, Dudley J, Nei M, Kumar S. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol.* 2007;24:1596–9.
19. Tao P, Dai L, Luo M, et al. Analysis of synonymous codon usage in classical swine fever virus. *Virus Genes.* 2009;38:104–12.
20. Vicario S, Moriyama E, Powell J. Codon usage in twelve species of *Drosophila*. *BMC Evol Biol.* 2007;7:226.
21. Wang M, Zhang J, Zhou J, et al. Analysis of codon usage in bovine viral diarrhea virus. *Arch Virol.* 2011;156:153–60.
22. Wright F. The effective number of codons used in a gene. *Gene.* 1990;87:23–9.
23. Zhang Y, Liu Y, Liu W, et al. Analysis of synonymous codon usage in hepatitis A virus. *Virol J.* 2011;8:174.
24. Zhang J, Wang M, Liu W, et al. Analysis of codon usage and nucleotide composition bias in polioviruses. *Virol J.* 2011;8:146.