VIROLOGICA SINICA, October 2010, 25 (5):329-340
DOI 10.1007/s12250-010-3123-3

Analysis of Synonymous Codon Usage Bias in 09H1N1

Zhen-peng LI[#], De-quan YING[#], Peng LI, Fei LI, Xiao-chen BO[**] and Sheng-qi WANG[**]

(Beijing Institute of Radiation Medicine, Beijing 100850, China)

Abstract: A novel subtype of influenza A virus 09H1N1 has rapidly spread across the world. Evolutionary analyses of this virus have revealed that 09H1N1 is a triple reassortant of segments from swine, avian and human influenza viruses. In this study, we investigated factors shaping the codon usage bias of 09H1N1 and carried out cluster analysis of 60 strains of influenza A virus from different subtypes based on their codon usage bias. We discovered that more preferentially used codons of 09H1N1 are A-ended or U-ended, and the intra-genomic codon usage bias of 09H1N1 is quite low. Base composition constraint, dinucleotide biases and translational selection are the main factors influencing the codon usage bias of 09H1N1. At the genome level, we find that the codon usage bias of 09H1N1 is similar to H1N1 (A/swine/Kansas/77778/2007H1N1), H9N2 from Asia, H1N2 from Asia and North America and H3N2 from North America. Our results provide insight for understanding the processes governing evolution, regulation of gene expression, and revealing the evolution of 09H1N1.

Key words: 09H1N1; Correspondence analysis; Codon usage bias

09H1N1, which is a combination of gene segments from both North American and Eurasian swine lineages [5,20,21], has crossed the species barrier to humans and has rapidly spread across the world. Several reports have illustrated the origin of this virus, which showed that all segments of 09H1N1 are directly related to swine influenza virus, including not only H1N1, but also the other subtypes of influenza A virus, mainly from America and Eurasian. The results also revealed that 09H1N1 is a triple reassortant of the segments from swine, avian and human influenza viruses.

It has been well established that synonymous codon usage varies both among genomes and within genomes. Several factors which can influence codon usage have been reported, such as mutational bias[9], translational selection[7], replicational and transcriptional selection [14], secondary structure of proteins [6], gene function[23], gene length[13] and environmental factors [2]. Codon usage biases of some organisms, such as bacteria, yeast, drosophila and mammals, have been examined in earlier research [15]. More recently, reports about the codon usage of RNA virus have also been reported [9, 19, 23], which show that intra-genomic synonymous codon usage bias (referred to as "codon usage bias"

Received: 2010-01-05, Accepted:2010-04-30

[#] These authors contributed equally to this work.

** Corresponding authors:
Xiao-chen BO: Phone: +86-10-66931422,
E-mail: boxc@bmi.ac.cn;
Sheng-qi WANG: Phone: +86-10-66932211,
E-mail:sqwang@bmi.ac.cn

for brevity hereafter) of most RNA viruses is quite low.

It is well known that a detailed knowledge of codon usage biases in RNA viruses can lead to a better understanding of the processes governing their evolution, particularly the role played by mutation pressure[9]. Such information can also provide clues to the mechanisms involved in the regulation of viral gene expression and the evolution of viruses.

MATERIALS AND METHODS

Materials

The sequences analyzed in this report were downloaded from the Influenza Virus Sequence Database [1]. Given that evolutionary analysis of 09H1N1 [5,20,21], showed that all segments of 09H1N1 are directly related to swine influenza A virus and 09H1N1 is a triple reassortant of the segments from swine, avian and human influenza viruses, three criteria were used to select the genomes. 1). All three hosts are considered suitable for H1N1: swine, avian and human, while swine is the only host for the other strains. 2). Only full length sequences were selected, i.e., they should contain 8 complete segments. 3). Except for H1N1, the sources of the strains were limited to three regions: Asia, Europe and North America. Based on this criteria, we then applied an additional 14 sub-criteria to screen the strains with the exception of 09H1N1 (Table 1). If there were more than 5 strains satisfying the sub-criteria, the 5 most recently sequenced samples were chosen. We also selected the earliest sequenced complete 09H1N1 strain (A/New York/1669/2009(H1N1)) as a reference genome. Based on these criteria, a total of 60 strains were selected. The sources of strains and the corresponding sub-criteria are listed in Table 1.

RSCU, RF [18]

Relative synonymous codon usage (RSCU) is defined as the ratio of observed codon counts to the counts expected where codon usage is uniform. The value of RSCU close to 1.0 indicates a lack of codon bias, while greater than 1.0 it means the corresponding codon is more frequently used than expected. The values are largely independent of amino acid composition, and widely used in comparing codon usage among gene segments that are different in size and amino acid composition.

Relative codon frequency (RF) is calculated as the ratio of the number of occurrences of codon to the sum of all synonymous codons of the same one. Compared with RSCU, RF does not consider the amino acid composition, but only reflects the codon bias in a specific amino acid.

$$\mathrm{RF}=\frac{n_{ac}}{\sum_{c=1}^{d_a} n_{ac}}$$

$$\mathrm{RSCU}=\frac{n_{ac}}{(1/d_a)\sum_{c=1}^{d_a} n_{ac}}$$

Where n_{ac} denotes the number of the cth codon of amino a in the gene segment or genome and d_a denotes the degree of codon degeneracy for amino acid a.

GC3s, GC, T3s, C3s, A3s and G3s

GC3s is defined as the ratio of G+C content of the synonymous third codon position. GC is the proportion of G and C in all of positions. T3s, C3s, A3s and G3s denote the ratios of the specified nucleotide at the third synonymous codon position to the maximum number possible for that nucleotide without altering the amino acid composition.

Table 1. Details of candidate data

Sub-criteria	Strain symbols	Sources
Reference strain	09H1N1	A/New York/1669/2009(H1N1)
H1N1 host as avian	H1N1_avian	A/duck/Italy/69238/2007(H1N1)
	H1N1_avian02	A/muscovy duck/New York/21211-5/2005(H1N1)
	H1N1_avian03	A/shorebird/Delaware/558/2006(H1N1))
	H1N1_avian04	A/northern pintail/Interior Alaska/1/2007(H1N1)
	H1N1_avian05	A/pintail/Alberta/21/2006(H1N1)
H1N1 host as swine	H1N1_swine	A/swine/Beijing/21/2008(H1N1)
	H1N1_swine02	A/swine/Beijing/26/2008(H1N1)
	H1N1_swine03	A/swine/Chachoengsao/NIAH587/2005(H1N1)
	H1N1_swine04	A/swine/IL/00685/2005(H1N1)
	H1N1_swine05	A/swine/Kansas/77778/2007(H1N1)
H1N1 host as human	H1N1_human	A/District of Columbia/WRAMC-1154047/2008(H1N1)
	H1N1_human02	A/Alabama/UR06-0455/2007(H1N1)
	H1N1_human03	A/Alabama/UR06-0536/2007(H1N1)
	H1N1_human04	A/Boston/35/2008(H1N1)
	H1N1_human05	A/England/26/2008(H1N1)
H1N2 from Asia	H1N2_Asia	A/swine/Miyazaki/1/2006(H1N2)
	H1N2_Asia02	A/swine/Shanghai/1/2007(H1N2)
	H1N2_Asia03	A/swine/Korea/CY08/2007(H1N2)
	H1N2_Asia04	A/swine/Hong Kong/1110/2006(H1N2)
	H1N2_Asia05	A/swine/Guangxi/13/2006(H1N2)
H1N1 from Europe	H1N2_Europe	A/swine/Doetlingen/IDT4735/2005(H1N2)
	H1N2_Europe02	A/swine/Cloppenburg/IDT4777/2005(H1N2)
	H1N2_Europe03	A/swine/Cotes d'Armor/790/97(H1N2)
H1N2 from North America	H1N2_NorthAmerica	A/swine/IL/07003243/2007(H1N2)
	H1N2_NorthAmerica02	A/swine/Oklahoma/032726/2008(H1N2)
	H1N2_NorthAmerica03	A/swine/Texas/050593/2008(H1N2)
	H1N2_NorthAmerica04	A/swine/Oklahoma/020734-2/2008(H1N2)
	H1N2_NorthAmerica05	A/swine/Oklahoma/020734-3/2008(H1N2)
H2N3 from North America	H2N3_NorthAmerica	A/swine/Missouri/4296424/2006(H2N3)
	H2N3_NorthAmerica02	A/swine/Missouri/2124514/2006(H2N3)
H3N2 from Asia	H3N2_Asia	A/Swine/Shandong/3/2005/H3N2
	H3N2_Asia02	A/swine/Jilin/37/2008(H3N2)
	H3N2_Asia03	A/swine/Jilin/19/2007(H3N2)
	H3N2_Asia04	A/swine/Jilin/5/2007(H3N2)
	H3N2_Asia05	A/swine/Nakhon pathom/NIAH586-1/2005(H3N2)
H3N2 from Europe	H3N2_Europe	A/swine/Spain/39139/2002(H3N2)
	H3N2_Europe02	A/swine/Hungary/13509/2007(H3N2)
	H3N2_Europe03	A/swine/Spain/54008/2004(H3N2)
	H3N2_Europe04	A/swine/Nordkirchen/IDT1993/2003(H3N2)
	H3N2_Europe05	A/swine/Spain/42386/2002(H3N2)
H3N2 from North America	H3N2_NorthAmerica	A/swine/Alberta/14722/2005(H3N2)
	H3N2_NorthAmerica02	A/swine/North Carolina/R08-001877-D08-013371/2008 (H3N2)
	H3N2_NorthAmerica03	A/swine/Minnesota/1145/2007(H3N2)
	H3N2_NorthAmerica04	A/swine/Oklahoma/008722/2007(H3N2)
	H3N2_NorthAmerica05	A/swine/British Columbia/28103/2005(H3N2)
H3N3 from North America	H3N3_NorthAmerica	A/swine/Ontario/42729A/01(H3N3)
	H3N3_NorthAmerica02	A/swine/Ontario/K01477/01(H3N3)
H5N1 from Asia	H5N1_Asia	A/swine/Guangxi/wz/2004(H5N1)
	H5N1_Asia02	A/swine/Anhui/cb/2004(H5N1)
	H5N1_Asia03	A/swine/Fujian/1/2003(H5N1)
	H5N1_Asia04	A/swine/Fujian/2001(H5N1)
	H5N1_Asia05	A/swine/Anhui/ca/2004(H5N1)
H5N2 from Asia	H5N2_Asia	A/swine/Korea/C12/2008(H5N2)
	H5N2_Asia02	A/swine/Korea/C13/2008(H5N2)
H9N2 from Asia	H9N2_Asia	A/swine/Shandong/w4/2003(H9N2)
	H9N2_Asia02	A/swine/Shandong/nc/2005(H9N2)
	H9N2_Asia03	A/swine/Guangxi/58/2005(H9N2)
	H9N2_Asia04	A/swine/Guangxi/FS2/2005(H9N2)
	H9N2_Asia05	A/swine/Henan/7/2004(H9N2)

GRAVY [11] and Aromo [12]

GRAVY denotes the general average hydrophobicity score for the conceptually translated gene product and is defined as the mean of the sum of the hydrophobic indices of each amino acid.

Aromo is calculated as the ratio of aromatic amino-acids (Phe, Tyr, Trp) in the hypothetical translated gene product.

Effective number of codons (Nc) [22]

For a given gene sequence, Nc is defined by:

$$Nc = 2 + 9/\overline{F_2} + 1/\overline{F_3} + 5/\overline{F_4} + 3/\overline{F_6}$$

Amino acids are divided into several classes based on the degree of codon degeneracy, i.e., the amino acids with the same degree of codon degeneracy belong to the same class. $\overline{F_i}$ denotes the average homozygosity for the amino acid class whose degree of codon degeneracy is i. The coefficients 9, 1, 5 and 3 denote the number of amino acids belonging to different classes. Nc is generally used as an index to measure the bias of a gene, with a value from 20 to 61. The smaller Nc is, the more bias of the gene it denotes. As highly biased genes are also highly expressed, Nc is also used to evaluate the expression level of genes [17].

Distance measure and cluster analysis

We defined the distance of two genome considering the 59 codon usage biases by:

$$d_{ij} = \sum_{k=1}^{59} (R_{ik} - R_{jk})^2$$

Where R_{ik} and R_{jk} denote the RSCU values of codon k of genome i and j respectively. As 60 strains were selected, A 60×60 matrix was constructed based on the distance between each genome pair. We then used the *hclust* function in the R statistical software package to lay out a cluster map based on the 60×60 matrix. The complete linkage method was used.

Correspondence analysis (CA) and within correspondence analysis (WCA)

CA is a statistical visualization method for identifying associations between the levels of a two-way contingency table. It can transform high dimensional data into a series of axis, which contain the contribution of different factors responsible for the difference between variables. Each axis includes all of the contribution values of variables of the two-way contingency. Generally, the first two axes are used as the coordinate to layout a plot, so that the global view of the data can be easily interpreted by the distances of different variables in the plot. CA has been widely used in codon usage analysis to investigate the major trend in codon usage variation among genes. It has been proved that CA based on RSCU was biased [16] for introducing unjustified statistical weights on data, yielding biased results especially for codon usage in rare amino-acids such as Cysteine. As a variant of CA, WCA considers codons coding the same amino acid as a group, and centres the data based on the amino acid, leading to a removal of much of the bias caused by amino acid composition and codon degeneracy. The effectiveness of WCA has been demonstrated by H.Suzuki *et al* [18] when applying four different CA methods to 241 bacterial genomes including WCA.

Software and statistic method

Codon usage indices including RSCU, GC3s, T3s, C3s, A3s, G3s, GC, GC12s and dinucleotide frequencies were computed by CodonW 1.4.4 (http://codonw.sourceforge.net/). WCA, cluster analysis and statistical mapping were performed by R [8], where the *ade4* [4] and *seqinr* packages [3] in R were used for WCA. Correlation analysis was carried out using

Spearman's rank correlation method.

RESULTS

Synonymous codon usage of 09H1N1

The overall RSCU values of 59 codons (two codons with degree of codon degeneracy 1 and three termination codons were excluded from calculation) of 09H1N1 and the indices for genes of 09H1N1 such as Nc and GC3s, are listed in Table 2 and Table 3. From Table 2, we found that most preferentially used codons of 09H1N1 are A-ended and U-ended, including 10 A-ended codons and 6 U-ended codons of the 20 most preferentially used codons. It was also observed that Nc values varied from 46.24(HA) to 57.29(M1). Based on these results, it can be concluded that codon usage bias of 09H1N1 is quite low.

The causes of codon usage bias of 09H1N1

WCA on codon usage: To seek for the sources of

Table 2. Synonymous codon usage in 09H1N1

Amino	Codon	N	RSCU	Amino	Codon	N	RSCU	Amino	Codon	N	RSCU
Phe	UUU	76	0.87	Pro	CCU	45	1.05	Lys	AAA	147	**1.12**
	UUC	98	**1.13**		CCC	34	0.8		AAG	116	0.88
Leu	UUA	35	0.6		CCA	71	**1.66**	Asp	GAU	107	**1.05**
	UUG	62	1.06		CCG	21	0.49		GAC	97	0.95
	CUU	69	**1.18**	Thr	ACU	70	1.01	Glu	GAA	195	**1.2**
	CUC	54	0.92		ACC	55	0.79		GAG	131	0.8
	CUA	68	1.16		ACA	28	**1.85**	Cys	UGU	35	0.88
	CUG	63	1.08		ACG	24	0.35		UGC	45	**1.13**
Ile	AUU	107	1.08	Ala	GCU	63	0.99	Arg	CGU	7	0.14
	AUC	76	0.77		GCC	55	0.86		CGC	16	0.33
	AUA	115	**1.16**		GCA	118	1.85		CGA	30	0.62
Val	GUU	53	0.82		GCG	19	0.3		CGG	22	0.45
	GUC	50	0.77	Tyr	UAU	65	**1.03**		AGA	139	**2.87**
	GUA	73	**1.13**		UAC	61	0.97		AGG	77	1.59
	GUG	83	1.28	His	CAU	44	**1.19**	Gly	GGU	43	0.58
Ser	UCU	61	1.05		CAC	30	0.81		GGC	45	0.6
	UCC	43	0.74	Gln	CAA	86	**1.02**		GGA	130	**1.74**
	UCA	90	**1.56**		CAG	83	0.98		GGG	81	1.08
	UCG	20	0.35	Asn	AAU	145	**1.16**				
	AGU	66	1.14		AAC	106	0.84				
	AGC	67	**1.16**								

Table 3. Indices for genes of 09H1N1

Gene Name	T3s	C3s	A3s	G3s	Nc	GC3s	GC	GC12s	Length	Gravy	Aromo
PB2	0.2543	0.2560	0.4420	0.2826	50.66	0.425	0.448	0.459	2280	-0.32082	0.065876
PB1	0.3017	0.2621	0.4618	0.2590	51.73	0.397	0.421	0.432	2274	-0.49617	0.088507
PA	0.3051	0.3013	0.4008	0.3095	54.99	0.452	0.443	0.438	2151	-0.46662	0.093575
HA	0.3151	0.2495	0.4926	0.2459	46.24	0.373	0.409	0.426	1701	-0.35159	0.09894
NP	0.2890	0.2685	0.4509	0.2334	50.68	0.397	0.462	0.492	1497	-0.5492	0.072289
NA	0.3597	0.2806	0.4190	0.2057	53.53	0.377	0.421	0.441	1410	-0.26546	0.102345
M1	0.3420	0.2591	0.3065	0.3175	57.29	0.464	0.488	0.499	759	-0.27302	0.051587
M2	0.4533	0.2533	0.2877	0.2923	53.69	0.409	0.443	0.46	294	-0.31753	0.092784
NS1	0.3081	0.3081	0.3976	0.2252	56.71	0.422	0.458	0.475	660	-0.25753	0.073059
NS2	0.2375	0.2375	0.4831	0.3415	46.43	0.431	0.413	0.406	366	-0.50248	0.090909

variation of codon usage bias in 09H1N1, WCA was implemented on genome of 09H1N1 based on the AF value of each gene and ten axes containing the positions of ten genes for each axis generated. As mentioned above, the axis of WCA denotes the source of the variation among a set of multivariate data points. The top four largest trends, which can explain 65.45% of the variation accumulatively, were observed: the first axis accounts for 24.61% of the variation, whereas the next three axes accounted for 18.08%, 11.58% and 11.18% of the variation respectively. Meanwhile, a scatter map of values of first two axes produced by WCA was plotted (Fig. 1). The distance between two points on a map can be used as an indicator to measure the similarity between them. Therefore, we can clearly investigate codons preferred by specific gene and the relationship between genes in the perspective of codon bias. We can observe from the distribution of codons in Fig 1 that genes in the positive axis have preference in using A-ended codons and have lower GC composition, while genes in the negative axis incline to use T-ended codons and have higher GC composition.

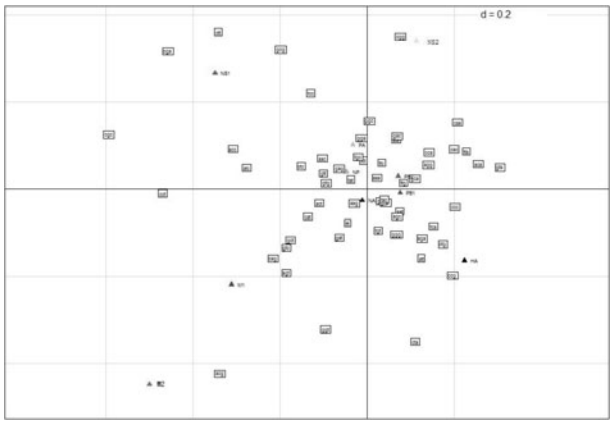

Fig.1. A scatter map of values of first two axes produced by WCA. The distance between genes can be seemed as a indicator for their codon usage similarity. A close distance means a close codon usage between genes. While the codons close to gene suggest the codons that the gene preferred.

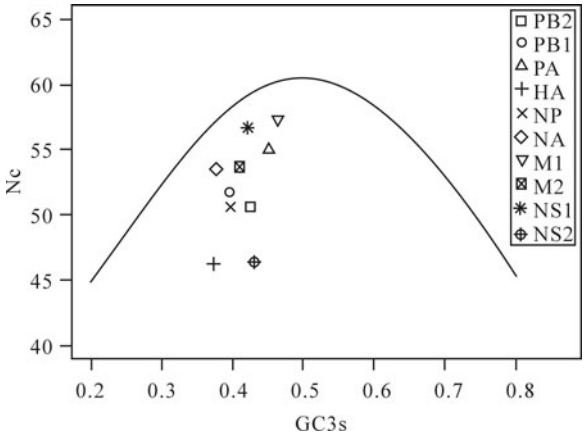

Fig.2. Nc-plot of genes of A(H1N1). A reference line indicates the expected codon usage if codon usage is only constrained by variation of GC3s.

Base composition constraint as a key factor can influence the codon usage bias of 09H1N1: Wright *et al* put forward the Nc-plot (Nc plotted against GC3s) as part of a general strategy to investigate patterns of synonymous codon usage[22]. A reference line was displayed to show the expected position of genes whose codon usage is only constrained by variation of GC3s. Genes, of which codon choice is constrained by GC3s originating from mutational pressure, will lie on or just below the reference line of the predicted values. It can be seen clearly from Fig. 2 that most of the points lay well below the expected curve, which indicates that the codon choice constrained by GC3s mutation bias is small and uneven. To further validate this result, the correlation coefficient between GC12s and GC3s was also calculated ($r = 0.25$, $p = 0.49$), the different base composition between intron and synonymous sites also suggested that codon usage bias of 09H1N1 is not likely the result of mutation pressure, since the effects are not the same at all codon positions[9]. To validate the other factors shaping the codon usage bias of 09H1N1, correlation

coefficients between the position of genes along the first four axis and A3s, T3s, G3s, C3s, GC3s, GC12s, GC were calculated. The results are listed in Table 4. As shown from Table 4, the positions of genes along the axis 1, which accounts for substantial amount of synonymous codon usage bias of intra-genome, are significantly correlated with A3s (r= 0.91, p<0.01), GC12s (r =-0.830, p<0.01) and also slightly correlated with GC composition (r=0.75, p<0.05), while the second axis of the WCA is slightly negatively correlated with T3s (r=-0.76, p<0.05). The other indices, such as G3s, C3s, GC3s, have little relationship with the first and second axis. Taken together, the close relationship between codon usage bias and composition constraint indicated that base composition constraint is a crucial factor contributing to codon usage bias.

Dinucleotide biases are also a key factor determining the codon usage bias of 09H1N1: Dinucleotide biases, which are independent of the overall base composition, are present in virtually all viruses[10] and may also affect codon bias. Thus, to investigate relationship between composition of dinucleotides and the codon usage bias of 09H1N1, 16 dinucleotide frequencies were calculated for each gene (Table 5), we found that majority of dinucleotide frequencies deviate from the mean value except TT and GG. The two highest dinucleotide frequencies are AA and GA, and the first two lowest dinucleotide frequencies are CG, CC. We also calculated the correlation coefficients between the 16 dinucleotide frequencies and the position of genes along the first four axes generated by WCA (Table 6).The results suggested that TC, CG, AC and AA are correlated with the position of genes along Axis 1, among which TC, AA,CG,GC are significantly correlated (p<0.05). There are two (GT, GA) and one (GG) dinucleotides correlated with the

Table 4. Correlation coefficients between values of the top four axes and indices of 09H1N1 genes

	T3s	C3s	A3s	G3s	GC3s	GC	GC12s	Nc	Length	Gravy	Aromo
Axis1	-0.491	-0.467	0.915**	0.067	-0.31	-0.75*	-0.830**	-0.806**	0.491	-0.479	0.309
Axis2	-0.758*	0.285	0.321	0.03	0.286	-0.018	-0.261	-0.151	0.067	-0.345	-0.103
Axis3	-0.236	-0.551	0.127	0.588	0.152	-0.439	-0.515	-0.381	-0.018	-0.491	0.273
Axis4	0.030	0.285	-0.188	-0.467	-0.395	0.323	0.394	0.030	0.176	-0.115	-0.176

*p<0.05; **p<0.01

Table 5. Frequencies of 16 dinucleotides of the 10 genes respectively

	TT	TC	TA	TG	CT	CC	CA	CG	AT	AC	AA	AG	GT	GC	GA	GG
PB2	0.053	0.047	0.044	0.074	0.042	0.041	0.086	0.019	0.075	0.056	0.109	0.096	0.049	0.044	0.096	0.069
PB1	0.056	0.049	0.05	0.07	0.049	0.04	0.087	0.016	0.085	0.065	0.123	0.081	0.035	0.038	0.094	0.061
PA	0.064	0.05	0.04	0.08	0.054	0.044	0.072	0.022	0.081	0.054	0.115	0.081	0.036	0.043	0.103	0.061
HA	0.057	0.048	0.063	0.068	0.046	0.037	0.087	0.015	0.086	0.062	0.129	0.079	0.048	0.038	0.077	0.059
NP	0.048	0.054	0.031	0.077	0.044	0.046	0.083	0.024	0.078	0.047	0.111	0.093	0.039	0.05	0.104	0.07
NA	0.064	0.055	0.058	0.082	0.048	0.036	0.08	0.02	0.088	0.055	0.106	0.066	0.059	0.038	0.071	0.074
M1	0.043	0.053	0.042	0.078	0.066	0.043	0.085	0.026	0.06	0.061	0.087	0.091	0.047	0.063	0.084	0.074
M2	0.082	0.068	0.048	0.082	0.051	0.027	0.061	0.041	0.085	0.038	0.075	0.082	0.061	0.048	0.096	0.055
NS1	0.068	0.062	0.041	0.078	0.072	0.049	0.062	0.032	0.074	0.058	0.088	0.07	0.035	0.046	0.1	0.064
NS2	0.077	0.046	0.051	0.067	0.06	0.022	0.067	0.022	0.067	0.063	0.12	0.099	0.036	0.041	0.111	0.051
mean	0.061	0.053	0.047	0.076	0.053	0.038	0.077	0.024	0.078	0.056	0.106	0.084	0.044	0.045	0.094	0.064
sd	0.012	0.007	0.009	0.006	0.01	0.008	0.01	0.008	0.009	0.008	0.018	0.011	0.01	0.008	0.013	0.008

Table 6. Correlation coefficients between values of the top four axes and the frequency of sixteen dinucleotides

	TT	TC	TA	TG	CT	CC	CA	CG
Axis1	-0.122	-0.855**	0.636	-0.762*	-0.406	-0.418	0.626	-0.906**
Axis2	0.201	-0.333	-0.333	-0.360	0.212	0.309	-0.261	-0.006
Axis3	0.462	-0.333	0.176	-0.128	-0.176	-0.588	-0.219	-0.073
Axis4	-0.085	0.479	-0.37	0.287	-0.467	0.261	-0.036	0.182
	AT	AC	AA	AG	GT	GC	GA	GG
Axis1	0.243	0.661*	0.879**	0.097	-0.128	-0.779**	-0.140	-0.299
Axis2	-0.419	0.176	0.345	0.207	-0.689*	-0.104	0.754*	-0.171
Axis3	0.140	-0.152	0.212	0.48	0.213	-0.166	0.377	-0.744*
Axis4	0.316	-0.527	-0.224	-0.018	0.122	0.239	0.103	0.134

*p<0.05;**p<0.01

position of genes along the Axis 2 and Axis 3, respectively (p<0.1). However, there is no dinucleotide correlated with the position of genes along the Axis4. These observations suggested that the composition of dinucleotides also as a factor can determine the codon usage bias of 09H1N1.

Translational selection can also drive the codon usage bias of 09H1N1: To investigate whether translational selection, gene length, hydrophobicity of proteins and aromaticity of amino acids contribute to the codon usage bias of 09H1N1, we calculated correlation coefficients between Nc, Length, Gravy, Aromo and the position of genes along the first four axis respectively. The results indicate that only Nc (r=-0.81, p<0.01) has a significantly negative correlation with the position of genes along the Axis 1. Because Nc can be used to evaluate the expression level of genes[17], the result suggests a close relationship between translational selection and codon usage bias. As Wright proposed that [22] the calculations of Nc for short sequences (e.g. less than 200 codons long) where there are some amino acids that are not used, can lead to inaccurate performance. At the same time, NS1 and M2, with only 121 and 97 codons, are short compared with other genes. Moreover, all of the three genes have unusual amino acid compositions, for example, NS2 has the lowest proportion of 4-fold and highest proportion of 6-fold amino acids; M2 has the second lowest proportion of 1-fold and 4-fold amino acids, and PA has the highest proportion of 2-fold amino acids. Correlation coefficient would increase to 0.93 if excluding these three genes. It has been shown that gene function also affected the inter-genome codon usage biases of different subtype influenza A virus[23]. To reveal whether genes' functions contribute to codon usage bias of 09H1N1, Ten proteins of 09H1N1 were classified into two major types: Structural proteins (HA, NA, NP, M1 and M2) and nonstructural proteins(PBl, PB2, PA, NS1 and NS2). Structural proteins participate in the assembly of virus particles and play crucial roles in the maintenance of the virus morphology, virus particles assembly, adsorption, invasion and release. While nonstructural proteins play important regulatory roles in the virus replicate cycle, without participating in the assembly of virus particles. It can be seen clearly from Fig 1 that there doesn't exist a clear boundary between these two classes. This is likely because of the existence of other factors, such as mutational pressure and translational selection, which also contribute and it is difficult to ascertain the exact correlation between codon usage bias and gene function.

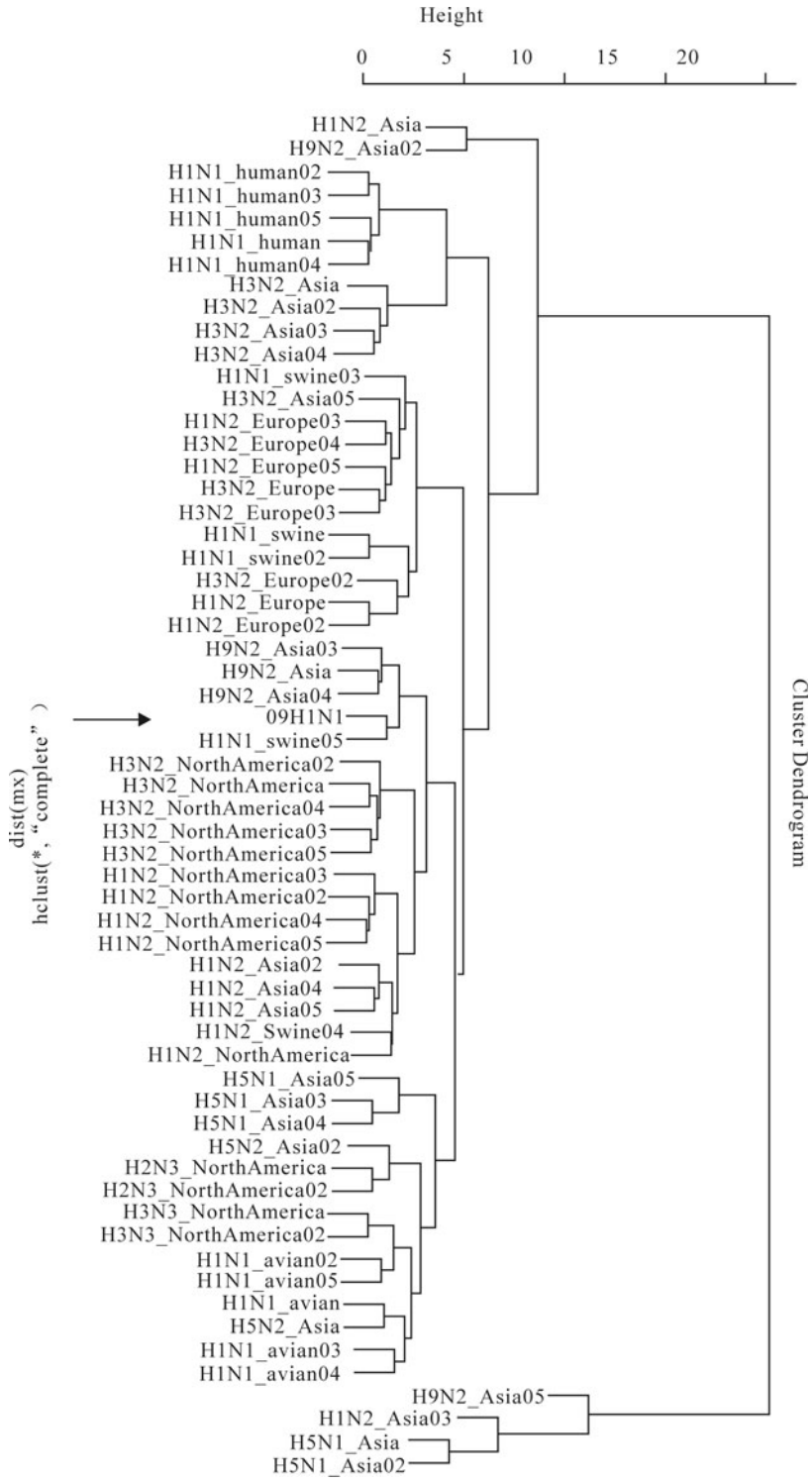

Fig. 3. Cluster map for 60 strains based on the overall RSCU of their genome. The cluster was implemented based on the codons' RSCU of each genome; meanwhile, the sources of virus were listed in Table 1.

Cluster analysis based on codon usage bias: We calculated the overall RSCU of 59 sense codons for each genome of 60 strains, and then plotted the cluster map for the strains, as shown in Fig. 3. It can be seen that 09H1N1 and H1N1_swine05 (A/swine/Kansas/77778/2007(H1N1)) are clustered together in Fig 3, indicating that they have a close relationship. Also H1N2 from Asia and North America, H3N2 from

North America and H9N2 from Asia are also clustered in a large branch containing 09H1N1, suggesting a close phylogenetical relationship among them.

DISCUSSION

In this study, we investigated the codon usage bias of 09H1N1. Through codon usage analysis, we found that the most preferentially used codons of 09H1N1 are A-ended and U-ended codons and the codon usage bias of 09H1N1 is quite low. After long term coexisting with a host, the codon usage patterns of the virus may adapt to its host. It is believed that the codon usage patterns of host may become an obstacle to block the virus to transmit to another species with codon usage patterns quite different from its natural host. The low codon usage bias suggests a more uniformed synonymous codon selection of 09H1N1, which may endow 09H1N1 the advantage to transmit across the species barriers.

Codon bias is likely a product of various kinds of mutational and selective forces. We try to investigate the various factors shaping the codon bias of 09H1N1, rather, we need to mention that there may exist other factors influencing the codon bias of 09H1N1 that are not detected. Through Nc-plot and the computation of correlation coefficient between the position of genes along the first two axis of WCA and indices related to base composition, we found that base composition constrains is a key factor driving the codon usage bias of 09H1N1, while the low correlation coefficient between GC3s and the position of genes along the first two axes suggested that GC3s mutational bias is small and uneven in shaping the codon usage bias of 09H1N1, which is consistent with the Nc-plot and correlation extent between GC12s and GC3s. As it has been proved that mutational bias is the main factor determines the codon usage bias of influenza A virus [23], the uneven and small effect of mutational bias on 09H1N1 may give indirect support for its complex genome origins. Meanwhile, the correlation between Nc and Axis 1 deriving from WCA suggests a close relationship between translational selection and codon usage bias. Other factors, such as gene length, hydrophobicity of proteins and aromaticity of amino acid have no significant correlation with the codon usage bias of 09H1N1. As there doesn't exist a clear boundary between structural proteins and nonstructural proteins in Fig 1, it is likely that gene function is entangled with other factors, it's hard to ascertain the exact correlation between bias and gene's function.

The correlation relationships between the 16 dinucleotide frequencies and the first three axes derived from WCA suggests that dinucleotide biases, which are independent of the overall base composition, can also affect the codon usage bias of 09H1N1. The relationship between dinucleotide frequencies and codon usage bias is evident in some cases. For example, the AA dinucleotide has the highest mean frequency in table 5, there are six codons including AA, related to coding four amines, i.e. Gln, Asn, Lys and Glu, the most preferentially used codons of all these four aminos are all AA-including codons. The situation is similar to GA, which has the second highest mean frequency. In contrast to AA and GA, CG has the lowest mean frequence, of the eight codons containing CG, which relate to encode 5 amino, only have a mean RSCU value of 0.38, meanwhile, the least preferentially used codons of these five amino all contain CG. The significant CG deficiency is a common phenomenon in small

eukaryotic viruses. Thus, it could be a strategy for viruses to resist host defense as CpGs may be recognized by the host's innate immune system as pathogen signature [19].

Other than the analysis mentioned above, the cluster map among genomes of 60 different strains was also plotted. 09H1N1 and swine05 (A/swine/ Kansas/77778/2007(H1N1)) are closely clustered in the cluster map among genomes, which suggests that they have similar codon usage bias. Early evolutional original analysis of 09H1N1 had revealed that six of eight segments have high similarity with the swine H1N2 influenza A viruses isolated in North America and Asia [20,21]. Together with the fact that H1N2 is descendant of the triple-reassortant swine H3N2 isolated in North America[21], it is under- standable that H1N2 from North America and Asia, H3N2 from North America are clustered in the same big branch as 09H1N1. It is worth noting that H9N2 from Asia are also in the same big branch as 09H1N1, which was ignored in the original analysis [20,21]. As codon usage bias may be related to gene' function, expression level and protein structure, and cluster analysis based on codon usage bias may provide additional information when compared with sequence analysis. Further experiments or data are needed to verify whether there exist a biological relationship between 09H1N1 and H9N2 from Asia.

Our results will provide a complement to phy- logenetic studies of 09H1N1. Furthermore, a better knowledge of codon usage biases in RNA viruses will provide necessary information, which is useful to understand the processes governing their evolution, such as mutation pressure. At last, such information can provide relevant clues to grasp the regulation of viral gene expression and evolutionary origin of different genes of 09H1N1.

References

1. **Bao Y, Bolotov P, Dernovoy D, *et al.*** 2008. The Influenza Virus Resource at the National Center for Biotechnology Information. **J Virol,** 82 (2): 596-601.
2. **Basak S, Banerjee T, Gupta S K**. 2004. Investigation on the causes of codon and amino acid usages variation between thermophilic Aquifex aeolicus and mesophilic Bacillus subtilis. **J Biomol Struct Dyn**, 22 (2): 205-214.
3. **Charif D, Lobry J.** 2007. SeqinR 1.0-2. A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In: **Structural Approaches to Sequence Evolution** (Bastolla U, Porto M, Roman E, Vendruscolo M, eds.), Berlin Heidelberg: Springer, p207-232.
4. **Dray S, Dufour A B**. 2007. The ade4 package: implementing the duality diagram for ecologists. **J Stat Softw**, 22 (4): 1-20.
5. **Garten R J, Davis C T, Russell C A.** 2009. Antigenic and Genetic Characteristics of Swine-Origin 2009 A (H1N1) influenza Viruses Circulating in Humans. **Science**, 325 (5937): 197-201.
6. **Gu W, Zhou T, Ma J**. 2004. The relationship between synonymous codon usage and protein structure in Escherichia coli and Homo sapiens. **BioSystems**,73 (2): 89-97.
7. **Gupta S K, Ghosh T C.** 2001.Gene expressivity is the main factor in dictating the codon usage variation among the genes in Pseudomonas aeruginosa. **Gene**, 273 (1): 63-70.
8. **Ihaka R, Gentleman R**. 1996. R: A language for data analysis and graphics. **J Comp Graph Stat**, 5 (3): 299-314.
9. **Jenkins G M, Holmes E C**. 2003. The extent of codon usage biases in human RNA viruses and its evolutionary origin. **Virus Res**, 92 (1): 1-7.
10. **Karlin S, Doerfler W, Cardon L R**. 2007. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? **J Virol**, 68 (5): 2889-2897.
11. **Kyte J, Doolittle R F**. 1982. A simple method for displaying

the hydropathic character of a protein. **J Mol Biol**, 157 (1): 105-32.

12. **Lobry J R, Gautier C.** 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino acid usage in 999 Escherichia coli chromosome encoded genes. **Nucl Acids Res,** 22 (15): 3174-3180.
13. **Marais G.** 2001.Duret L. Synonymous codon usage, accuracy of translation, and gene length in Caenorhabditis elegans**. J Mol Evol,** 52 (3): 275-280.
14. **McInerney J O.** 1998. Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. **Proc Natl Acad Sci USA**, 95 (18): 10698-10703.
15. **Mooers A Ø, Holmes E C.** 2000. The evolution of base composition and phylogenetic inference. **Trends Ecol Evol (Amst.)**, 15 (9): 365-369.
16. **Perriere G, Thioulouse J**. 2002. Use and misuse of correspondence analysis in codon usage studies, **Nucl Acids Res**, 30 (20): 4548-4555.
17. **Sharp P M, Tuohy T M, Mosurski K R, *et al***. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. **Nucl Acids Res**, 14 (13): 5125-5143.
18. **Suzuki H, Brown C J, Forney L J**. 2008. Comparison of correspondence analysis methods for synonymous codon usage in bacteria. **DNA Res**, 15 (6): 357-365.
19. **Tao P, Dai L, Luo M.** 2009. Analysis of synonymous codon usage in classical swine fever virus. **Virus Genes**, 38 (1): 104-112.
20. **Trifonov V, Khiabanian H, Greenbaum B, *et al***. 2009. The origin of the recent swine influenza A (H1N1) virus infecting humans. **Euro Surveill**, 14 (17) : pii=19193. Available online. http//www.eurosurveillance. org/ View-Article.aspx? Article Id=19193.
21. **Trifonov V, Khiabanian H, Rabadan R**. 2009. Geographic Dependence, Surveillance, and Origins of the 2009 Influenza A (H1N1) Virus. **N Engl J Med**, 361 (2): 115-119.
22. **Wright F.** 1990. The 'effective number of codons' used in a gene. **Gene,** 87 (1): 23-29.
23. **Zhou T, Gu W, Ma J, *et al.*** 2005. Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses. **BioSystems**, 81 (1): 77-86.