Virus Genes (2007) 35:767–776
DOI 10.1007/s11262-007-0159-z

Mutation pressure shapes codon usage in the GC-Rich genome of foot-and-mouth disease virus

Jincheng Zhong · Yanmin Li · Sheng Zhao · Shenggang Liu · Zhidong Zhang

Received: 12 February 2007 / Accepted: 9 August 2007 / Published online: 1 September 2007

Abstract Foot-and-mouth disease (FMD) is economically the most important viral-induced livestock disease worldwide. In this study, we report the results of a survey of codon usage bias of FMD virus (FMDV) representing all seven serotypes (A, O, C, Asia 1, SAT 1, SAT 2, and SAT 3). Correspondence analysis, a commonly used multivariate statistical approach, was carried out to analyze synonymous codon usage bias. The analysis showed that the overall extent of codon usage bias in FMDV is low. Furthermore, the good correlation between the frequency of G + C at the synonymous third position of sense codons (GC_{3S}) content at silent sites of each sequence and codon usage bias suggested that mutation pressure rather than natural (translational) selection is the most important determinant of the codon bias observed. In addition, other factors, such as the lengths of open reading frame (ORF) and the hydrophobicity of genes also influence the codon usage variation among the genomes of FMDV in a minor way. The result of phylogenetic analyses based on the relative synonymous codon usage (RSCU) values indicated a few obvious phylogenetic incongruities, which suggest that more FMDV genome diversity may exist in nature than is currently indicated. Our work might give some clues to the features of FMDV genome and some evolutionary information of this virus.

Keywords FMDV · Synonymous codon usage · Mutational bias · GC content

Abbreviations

Bp	base pair
FMD	Foot-and-mouth disease
FMDV	Foot-and-mouth disease virus
RSCU	Relative synonymous codon usage
ENC	Effective number of codons
COA	Correspondence analysis
GC_{3S}	The frequency of G + C at the synonymous third position of sense codons
A_{3S}, T_{3S}, G_{3S} and C_{3S}	The adenine, thymine, guanine and cytosine content at synonymous third positions
ORF	Open reading frame
S.D.	Standard deviation

Introduction

The phenomenon of synonymous codon usage bias exists in a wide range of paradigms from prokaryote to eukaryote. Due to different genomes having their own characteristic patterns of synonymous codon usage [1], it has not been easy to provide a satisfactory explanation for the particular pattern found in a given genome. Compositional constraints and translational selection are thought to be the main factors accounting for codon usage variation among genes in different organisms such as *Escherichia coli*,

J. Zhong · S. Liu
University of Electronic Science and Technology of China, Chengdu, Sichuan 610054, P.R. China

Y. Li · Z. Zhang (✉)
Institute for Animal Health, Pirbright, Woking, Surrey GU24 0NF, UK
e-mail: Zhidong.Zhang@bbsrc.ac.uk

S. Zhao
Jingmen Technical College, Jingmen, Hubei 448000, P.R. China

Bacillus subtilis, *Saccharomyces cerevisiae*, *Dictyostelium discoideum* [2], *Drosophila melanogaster* [3] and *Caenorhabditis elegans* [4]. However, in some prokaryotes with extremely high A + T or G + C contents [2] and human [5], mutation bias is the major factor accounting for the variation in codon usage. In contrast, it was reported that translational selection at silent sites played the most important role in shaping codon usage in *Zea mays* [6] and *Arabidopsis thaliana* [7]. Recently, codon usage was suggested to be related to gene function [8] and protein secondary structure [9].

Codon usage information has also been analyzed for viruses, such as human immunodeficiency virus (HIV) [10], nucleopolyhedroviruses [11], cauliflower mosaic virus (CMV) [12], human RNA viruses [13], H5N1 virus [14], and hepatitis A virus [15]. For example, HIV has a marked codon usage bias, due to its strong preference for the A nucleotide [10]. Rubella virus has a genomic G + C content of 0.70 [16]. Codon usage in Epstein-Barr virus (a DNA virus) may have an influence on regulation of latent versus productive infection [17]. In contrast, in nucleopolyhedroviruses codon usage appears to be simply a consequence of uneven base composition [11]. These published studies are mostly restricted to particular groups of viruses and have usually addressed phylogenetic questions [11, 18–20]. Moreover, mutational pressure rather than translational selection is the most important determinant of the codon bias in some RNA viruses [11, 13, 14, 21, 22]. However, a recent study showed that the G + C compositional constraint is the main factor that determines the codon usage bias in iridovirus genomes [23]. Clearly, studies of the synonymous codon usage in viruses can reveal information about the molecular evolution of individual genes and such information would be relevant in understanding the regulation of viral gene expression and also to vaccine design where the efficient expression of viral proteins may be required to generate immunity [24].

Foot-and-mouth disease (FMD) is a highly contagious disease of cattle and other cloven-hoofed animals. Apart from the influence on animal health and welfare, the economic impact of an outbreak of FMD can be of great importance for a country's export trade. The 2001 European outbreak of FMD which mainly affected the UK is estimated to have cost of 6000 million Euros [25]. Effective vaccines and stringent control measures have enabled FMD eradication in most developed countries, which maintain unvaccinated, seronegative herds in compliance with strict international trade policies. However, Outbreaks with devastating economic consequences still occur in many developing regions of Asia, Africa and South America, posing a serious problem for commercial trade with FMD-free countries. FMD is caused by FMD virus (FMDV), a small, non-enveloped virus that contains a single stranded positive-sense RNA genome of about 8500 nucleotides from the Aphthovirus genus of the *Picornaviridae* family. The capsid is made of four proteins (VP1, VP2, VP3 and VP4) and there are seven serotypes (A, O, C, Asia I, SAT 1, SAT 2, and SAT 3) that can be further divided into many genotypes according to nucleotide difference in the capsid proteins [26]. Although genome sequences of FMDV have been published and some studies have been performed on them in recent years [27–30], little codon usage analysis is available. Such information might give some clues to the features of virus biology and some evolutionary information of this virus, and also it is of interest to understand the factors that shape codon usage in this species. In the present study, the codon usage bias was analyzed in these seven serotypes of FMDV genomes. The key evolutionary determinants of codon usage bias in these viruses were also investigated.

Materials and methods

FMDV genome sequences

A total of 40 FMDV genomes were used in this study (Table 1), including 23 Type O genomes, four type A, six type C, two type Asia I , one Type SAT1, three Type SAT2, and one Type SAT3. The complete sequences these FMDV genomes were obtained from EMBI (http://www.ebi.ac.uk/cgi/) and NCBI (http://www.ncbi.nlm.nih.gov/). Serial number (SN), length of each genome, EMBI or GenBank accession numbers are listed in Table 1. FMDV genomes SN 6, 9, 13 and 19 have been reported to be from cattle. FMDV genomes SN 14, 23 were from pigs and FMDV genomes SN 15,16 and 17 have been reported to pig-adapted isolates carrying a deletion in 3A (codons93–102). Alterations or deletion in this region are associated with the reduced ability of the virus to cause FMD in cattle. A program based on Perl has been developed to extract the annotated ORF sequences from each genome.

Synonymous codon usage measures

In order to examine synonymous codon usage without the confounding influence of amino acid composition of different ORF samples, relative synonymous codon usage values (RSCU) of different codons in each ORF sample was calculated as described previously [31]. Additionally, the 'Effective Number of Codons' (ENC) was used to quantify the codon usage bias of a ORF [32], which is the best overall estimator of absolute synonymous codon usage bias (Comeron and Aguade, 1998) [51]. The reported value

Table 1 Genomes examined, strain, length and accession numbers[a]

SN	Virus	Strain	Length (bp)	Accession No.
1	FMDV-O_1	Campos/Brazil/58	8168	AJ320488
2	FMDV-O	Akesu/58	8147	AF511039
3	FMDV-O	OMIII (artificially attenuated from strain Akesu/58)	8083	AY359854
4	FMDV-O	CHA/1/99 (Tibet)	8173	AF506822
5	FMDV-O	Tibet/CHA/99	8183	AJ539138
6	FMDV-O	FRA/1/2001	8234	AJ633821
7	FMDV-O1	Kaufbeuren/FRG/66	7804	X00871
8	FMDV-O	HKN/2002	8104	AY317098
9	FMDV-O	JPN/2000 (Miyazaki, Japan)	7822	AB079061
10	FMDV-O	SAR/19/2000	8184	AJ539140
11	FMDV-O	SKR/2000	7813	AF377945
12	FMDV-O	SKR/2000	8182	AJ539139
13	FMDV-O	SKR/2000 (Chungju county; cattle	7799	AY312587
14	FMDV-O	SKR/2002 (pigs)	7803	AY312589
15	FMDV-O	Tau-YuanTW97 (Taiwan, 1997)	7739	AF154271
16	FMDV-O	Chu-Pei (Taiwan) (pig strain)	7733	AF026168
17	FMDV-O	Yunlin/Taiwan/97	8134	AF308157
18	FMDV-O	TAW/2/99 (TC)	8183	AJ539136
19	FMDV-O	TAW/2/99 (BOV)	8183	AJ539137
20	FMDV-O	NY00 (China?)	7731	AY333431
21	FMDV-O	HLJOC12/03(China)	7767	DQ119643
22	FMDV-O	lz	8104	DQ248888
23	FMDV-O	UKG/35/2001	8183	AJ539141
24	FMDV-A10	Argentina/61 (A61)	7107	X00429
25	FMDV-A12	119/Kent/UK/32 (ab variant)	7712	M10975
26	FMDV-A22	Azerbaijan/USSR/65	7820	X74812
27	FMDV-C3	Argentina/85	8161	AJ007572
28	FMDV-C_1	Santa Pau/Spain/70 (C-S8c1)	8115	AJ133357
29	FMDV-C_1	Santa Pau/Spain/70 (rp99)	8115	AJ133358
30	FMDV-C_1	Santa Pau/Spain/70 (rp146)	8115	AJ133359
31	FMDV-C_1	Santa Pau/Spain/70 (MARLS clone)	8115	AF274010
32	FMDV-Asia1	YNBS/China/58	8163	AY390432
33	FMDV-Asia1	IND/63/72	8167	AY304994
34	FMDV-SAT2	Kenya/3/57 (KEN/3/57)	7774	AJ251473
35	FMDV-SAT2	ZIM/7/83	8173	AF540910
36	FMDV-SAT1	SAT1–1bech	8173	NC_011451
37	FMDV-SAT2	SAT2–3kenya 11/60	8203	NC_003992
38	FMDV-SAT3	SAT3–2sa57/59	8170	NC_011452
39	FMDV-A	A10 Holland	8161	NC_011450
40	FMDV-C	From cell culture	8115	NC_002554

[a] Sequence numbers 1 to 35 are from EMBI; Sequence numbers 36 to 40 are from NCBI

of ENC is always between 20 (when only one codon is used for each amino acid) and 61 (when all codons are used equally) [32]. GC_{3S}, the frequency of the nucleotide G + C at the synonymous third codon position (excluding Met, Trp and the termination codons), was also used to calculate the extent of base composition bias. Similarly, GC_{1S} and GC_{2S} are the frequencies of the nucleotide G + C at the synonymous first and second position, respectively. In oder to examine the relationship between codon usage variation and compositional constraints, the GC_{1S}, GC_{2S} and GC_{3S} of each selected ORF have been calculated.

Correspondence analysis

Correspondence analysis (COA) was used to investigate the major trend in codon usage variation among ORFs. In order to minimize the effect of amino acid composition on

codon usage, each ORF is represented as a 59-dimensional vector. Each dimension corresponds to the RSCU value of one sense codon (excluding Met, Trp and the stop codons). The axis of a correspondence analysis identifies the source of the variation among a set of multivariate data points. This method has been successfully used to investigate the variation of RSCU values among ORFs [33–38].

Statistical analysis

Correlation analysis was carried out using the Spearman's rank correlation analysis method. Cluster analysis was done by Hierarchical cluster method and the distances between selected sequences were calculated by the Euclidean distance method.

Analysis tools

The RSCU, GC3s, ENC, G + C, GRAVY, Length value, and COA were calculated using the program CodonW version 1.4 (http://codonw.sourceforge.net). The correlation analysis and Cluster analysis were carried out by using the multianalysis software SPSS version 13.0 (http://spss.com).

Results

Synonymous codon usage in FMD

In order to better understand the synonymous codon usage variation among FMDV isolates, the average codon usage was analyzed for these 40 FMDV genomes. As shown in Table 2 (see http://222.210.17.171/ yak/FMDV/ Table 2. doc for details)., the codons ending in G or C are favored and the global pattern of codon usage is very similar among all FMDV genome examined, which indicates that there have not been significant compositional changes since these species diverged from their last common ancestor.

Compositional properties of coding sequences

In order to investigate if these 40 FMD viruses' coding sequences examined display similar compositional features, mean ENC and GC_{3S}, were calculated and summarized in Table 3. The values of ENC among these FMDV genomes examined are very similar, which vary from 50.02 to 52.79 with a mean value of 51.58 and S.D. of 0.67. All the ENC values of these strains are more than 50. The data suggest the homogeneity of synonymous codon usage among FMDV genomes examined. The concept is further supported by the GC_{3S} values for each FMDV strain, which range from 61.00 to 68.00% with a mean of 63.80% and S.D. of 0.01. Therefore, taken together with published data of codon usage bias among some RNA viruses [5, 13, 39–41], we could conclude that codon usage bias in FMDV genomes is less biased, and there is no significant variation of synonymous codon usage among FMDV seven serotypes.

Correspondence analysis on codon usage

In order to investigate the variation of RSCU values among ORFs, correspondence analysis (COA) was implemented on these 40 FMDV genomes examined as a single-dataset-based on the RSCU value of each strain' ORFs. As mentioned, the axis of a correspondence analysis identifies the source of the variation among a set of multivariate data points. The four largest trends in codon usage among these ORFs were observed: the first axis accounts for 31.20% of all variation among genomes, whereas the next three axes accounts for 16.50%, 12.40%, and 8.10%, respectively.

Effect of mutational bias on codon usage

In order to investigate if the evolution of codon usage bias is controlled by mutation pressure or by natural selection, firstly, G + C content at the first and second codon positions ($G + C_{12}$) was compared with that at synonymous third codon positions ($G + C_{3S}$) (Fig. 1). A highly significant correlation was observed ($r = 0.432$, $P < 0.05$), indicating that patterns of base composition are most likely the result of mutation pressure, and not natural selection, since the effects are present at all codon positions. Second, for each strain, actual codon bias was plotted against both $G + C_{3S}$ and the expected ENC value if codon usage bias is solely due to biased base composition (i.e., G + C content). Result showed that the actual codon usage indices are close to the values expected from their G + C composition, although all are slightly lower (Fig. 2). Thirdly, we plotted the first and second axis values in COA and GC_{3S} values of each strains (Fig. 3A,B). The patterns of codon usage in different strains also appear to be closely related to the GC content on the third codon position. Linear regression analysis has been implemented to each virus genome to find some correlation between synonymous codon usage and nucleotide compositions of the ORFs. We also found that axis 1 coordinates are correlated with GC_{3S} and GC ($r = 0.843$, $P < 0.01$; $r = 0.813$, $P < 0.01$), while there are significant correlations between axis 2 value and GC_{3S}($r = 0.355$, $P < 0.05$). Taken together, these analyses

Table 2 Synonymous codon usage in different Serotypes of FMDV[a]

		FMDV Serotypes													
		O		A		C		SAT1		SAT2		SAT1		Asia 1	
AA	Codon	No.	RSCU	No.	RSCU	No.	RSCU	No.	RSCU	No.	RSCU	No.	RSCU	No.	RSCU
Phe	UUU	45.22	0.82	51.25	0.95	43.17	0.84	41.00	0.75	39.33	0.71	29.00	0.54	47.50	0.89
	UUC	65.17	1.18	57.00	1.05	60.33	1.16	68.00	1.25	72.67	1.29	79.00	1.46	59.00	1.11
Leu	UUA	2.39	0.07	3.00	0.09	0.33	0.01	1.00	0.03	2.67	0.09	2.00	0.06	1.50	0.05
	UUG	37.87	1.14	42.75	1.32	32.83	1.03	34.00	1.00	32.33	1.02	31.00	0.91	33.00	1.02
	CUU	37.04	1.11	42.00	1.30	42.67	1.33	26.00	0.76	28.67	0.90	36.00	1.06	45.00	1.39
	CUC	62.70	1.88	58.75	1.82	62.83	1.97	74.00	2.17	64.00	2.00	64.00	1.88	60.00	1.86
	CUA	5.65	0.17	6.25	0.19	8.83	0.28	4.00	0.12	6.33	0.20	8.00	0.24	7.50	0.23
	CUG	54.13	1.63	41.25	1.28	44.33	1.38	66.00	1.93	58.00	1.81	63.00	1.85	47.00	1.45
Ile	AUU	36.96	1.08	43.75	1.23	39.83	1.16	37.00	1.10	29.00	0.87	29.00	0.85	42.00	1.24
	AUC	57.22	1.68	55.50	1.56	57.67	1.68	61.00	1.81	65.00	1.95	68.00	2.00	53.50	1.57
	AUA	8.17	0.24	7.75	0.22	5.67	0.16	3.00	0.09	6.00	0.18	5.00	0.15	6.50	0.19
Val	GUU	40.04	0.89	42.25	0.95	44.67	0.98	44.00	0.97	38.00	0.84	46.00	0.98	44.50	0.99
	GUC	49.65	1.10	47.75	1.07	49.33	1.08	61.00	1.35	58.67	1.29	57.00	1.22	46.00	1.02
	GUA	13.30	0.30	14.00	0.32	12.50	0.27	11.00	0.24	12.33	0.27	11.00	0.24	12.00	0.27
	GUG	76.70	1.71	74.50	1.67	76.83	1.68	65.00	1.44	73.33	1.61	73.00	1.56	78.50	1.74
Ser	UCU	13.48	0.66	15.75	0.73	19.67	0.95	18.00	0.83	16.67	0.76	15.00	0.69	18.00	0.86
	UCC	30.83	1.52	35.50	1.64	30.17	1.46	31.00	1.43	39.67	1.80	34.00	1.57	30.50	1.45
	UCA	21.26	1.05	23.50	1.09	21.83	1.06	19.00	0.88	17.67	0.80	17.00	0.78	22.50	1.07
	UCG	17.39	0.86	19.50	0.90	16.50	0.80	25.00	1.15	21.67	0.98	23.00	1.06	18.50	0.88
Pro	CCU	35.04	1.07	35.25	1.11	32.33	1.00	41.00	1.27	37.00	1.16	33.00	1.03	36.00	1.12
	CCC	40.22	1.22	37.75	1.19	38.67	1.20	40.00	1.24	37.33	1.17	44.00	1.38	38.50	1.20
	CCA	25.87	0.79	26.50	0.84	29.50	0.91	29.00	0.90	34.67	1.09	28.00	0.88	27.00	0.84
	CCG	30.22	0.92	27.50	0.87	28.67	0.89	19.00	0.59	18.33	0.58	23.00	0.72	27.50	0.85
Thr	ACU	39.00	0.94	39.00	0.93	41.17	0.98	42.00	1.01	39.00	0.94	34.00	0.84	41.00	0.97
	ACC	71.61	1.74	65.50	1.56	62.50	1.49	64.00	1.54	67.00	1.62	55.00	1.37	64.50	1.53
	ACA	29.70	0.72	39.75	0.94	37.17	0.89	41.00	0.99	33.67	0.82	47.00	1.17	36.00	0.86
	ACG	24.91	0.60	24.25	0.58	27.33	0.65	19.00	0.46	25.67	0.62	25.00	0.62	27.50	0.65
Ala	GCU	45.30	0.91	43.25	0.91	49.00	0.96	44.00	0.96	35.67	0.77	41.00	0.89	51.50	1.03
	GCC	71.39	1.44	70.50	1.48	74.17	1.46	64.00	1.40	70.67	1.52	63.00	1.36	70.50	1.41
	GCA	49.17	0.99	52.25	1.10	50.33	0.99	54.00	1.18	48.33	1.04	54.00	1.17	47.00	0.94
	GCG	33.00	0.66	24.25	0.51	30.17	0.59	21.00	0.46	31.67	0.68	27.00	0.58	31.00	0.62
Tyr	UAU	9.39	0.23	10.25	0.25	13.67	0.34	14.00	0.32	12.33	0.29	14.00	0.34	11.50	0.29
	UAC	71.26	1.77	73.75	1.76	68.17	1.67	73.00	1.68	73.67	1.71	68.00	1.66	70.00	1.72
His	CAU	5.43	0.18	6.75	0.21	7.50	0.22	11.00	0.34	11.00	0.34	10.00	0.30	7.50	0.23
	CAC	56.26	1.82	57.00	1.79	60.83	1.78	54.00	1.66	54.00	1.66	56.00	1.70	58.00	1.77
Gln	CAA	28.74	0.74	33.00	0.83	35.50	0.91	40.00	1.05	32.33	0.86	38.00	0.97	31.50	0.82
	CAG	49.35	1.26	46.75	1.17	42.83	1.10	36.00	0.95	43.00	1.14	40.00	1.03	45.50	1.19
Asn	AAU	15.74	0.31	13.25	0.27	16.33	0.33	20.00	0.37	14.00	0.27	8.00	0.16	15.50	0.32
	AAC	86.17	1.69	86.75	1.73	84.17	1.67	87.00	1.63	88.33	1.73	92.00	1.84	83.50	1.69
Lys	AAA	61.17	0.86	62.25	0.86	59.83	0.85	68.00	0.92	62.67	0.85	57.00	0.79	60.00	0.86
	AAG	81.57	1.14	82.75	1.14	81.00	1.15	80.00	1.08	85.00	1.15	87.00	1.21	80.50	1.15
Asp	GAU	30.83	0.43	38.00	0.52	34.17	0.48	35.00	0.49	39.67	0.53	39.00	0.53	35.50	0.50
	GAC	112.43	1.57	107.75	1.48	108.33	1.52	107.00	1.51	110.67	1.47	109.00	1.47	107.50	1.51
Glu	GAA	43.48	0.66	40.75	0.63	45.17	0.70	51.00	0.76	42.67	0.66	54.00	0.81	43.50	0.67
	GAG	87.61	1.34	88.25	1.37	83.00	1.30	84.00	1.24	87.33	1.35	80.00	1.19	86.50	1.33

Table 2 continued

		FMDV Serotypes													
		O		A		C		SAT1		SAT2		SAT1		Asia 1	
AA	Codon	No.	RSCU	No.	RSCU	No.	RSCU	No.	RSCU	No.	RSCU	No.	RSCU	No.	RSCU
Cys	UGU	14.70	0.85	16.50	0.99	15.17	0.94	11.00	0.65	17.67	0.96	16.00	0.82	13.00	0.81
	UGC	20.09	1.15	17.00	1.02	17.33	1.07	23.00	1.35	19.00	1.04	23.00	1.18	19.50	1.20
Arg	CGU	15.83	0.91	16.25	0.91	14.33	0.81	22.00	1.32	22.00	1.28	17.00	0.94	15.50	0.86
	CGC	26.26	1.51	28.00	1.56	30.83	1.75	24.00	1.44	28.00	1.62	37.00	2.04	32.00	1.78
	CGA	5.57	0.32	4.25	0.24	4.67	0.26	10.00	0.60	6.33	0.37	9.00	0.50	4.50	0.25
	CGG	13.96	0.80	13.25	0.74	16.17	0.92	7.00	0.42	8.33	0.48	8.00	0.44	15.50	0.86
Ser	AGU	13.52	0.67	13.75	0.64	13.83	0.67	10.00	0.46	16.00	0.73	17.00	0.78	14.50	0.69
	AGC	25.43	1.25	21.75	1.01	21.67	1.05	27.00	1.25	20.67	0.93	24.00	1.11	22.00	1.05
Arg	AGA	27.91	1.60	31.25	1.75	25.17	1.43	28.00	1.68	23.67	1.37	23.00	1.27	26.00	1.44
	AGG	15.09	0.86	14.50	0.81	14.83	0.84	9.00	0.54	15.33	0.89	15.00	0.83	15.00	0.83
Gly	GGU	43.22	1.10	35.75	0.91	34.83	0.89	36.00	0.89	39.00	0.98	42.00	1.08	39.00	0.99
	GGC	41.65	1.06	49.25	1.25	54.33	1.39	50.00	1.23	43.33	1.10	43.00	1.11	49.50	1.26
	GGA	39.17	1.00	39.00	0.99	34.17	0.87	42.00	1.04	42.33	1.07	45.00	1.16	36.50	0.93
	GGG	33.30	0.85	34.25	0.87	33.67	0.86	34.00	0.84	33.67	0.85	25.00	0.65	32.50	0.83

[a] Note: AA, amino acids; No., number of codons; RSCU, cumulative relative synonymous codon usage

Table 3 The values of ENC, GC_{3S} and ORF length for these 40 FMDV strains

SN	ENC	GC_{3S}	ORF length (bp)	SN	ENC	GC_{3S}	ORF length (bp)
1	50.59	0.64	6999	21	50.33	0.646	6999
2	52.79	0.626	6999	22	50.02	0.681	6969
3	52.74	0.629	6954	23	51.37	0.634	6999
4	51.3	0.632	6999	24	52.06	0.625	7002
5	51.25	0.633	6993	25	51.07	0.636	6999
6	51.36	0.633	6999	26	52.38	0.607	7011
7	50.87	0.638	6999	27	52.07	0.624	6987
8	50.59	0.672	6969	28	52.44	0.628	6984
9	51.32	0.634	6999	29	52.44	0.629	6984
10	51.34	0.633	6996	30	52.5	0.629	6984
11	51.85	0.628	6999	31	52.37	0.63	6984
12	51.8	0.632	6996	32	52.07	0.621	6990
13	51.76	0.632	6999	33	51.47	0.646	6993
14	51.5	0.637	6999	34	51.23	0.634	7008
15	51.17	0.673	6969	35	50.4	0.653	7008
16	51.09	0.675	6963	36	51.15	0.623	7020
17	51.46	0.672	6969	37	51.43	0.639	7008
18	51.23	0.636	6999	38	51.29	0.635	7008
19	51.31	0.635	6999	39	51.87	0.627	6999
20	51.49	0.633	6996	40	52.37	0.63	6984

indicate that most of the codon usage bias among these FMDV genomes is directly related to the nucleotide composition. Furthermore mutational bias is the major factor responsible for the variation of synonymous codon usage among ORFs in these virus genomes.

Effect of other factors on codon usage

Generally, mutational bias and natural selection, such as, ORF length and the hydrophobicity of each protein are thought to be the factors accounting for the codon usage

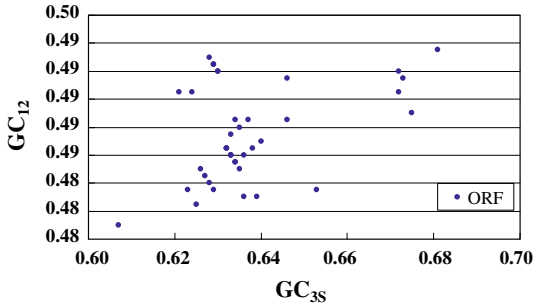

Fig. 1 Correlation between GC content at first and second codon positions (GC_{12}) with that at synonymous third codon positions (GC_{3S}). ORF, open reading frame

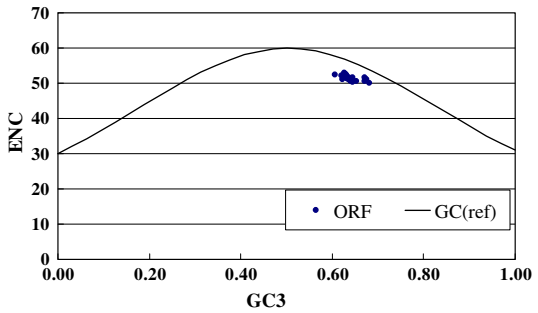

Fig. 2 Distribution of the codon usage index, ENC, and GC content at synonymous third codon positions (GC3S). The curve indicates the expected codon usage if GC compositional constraints alone account for codon usage bias. ENC: effective number of codons; ORF, open reading frame

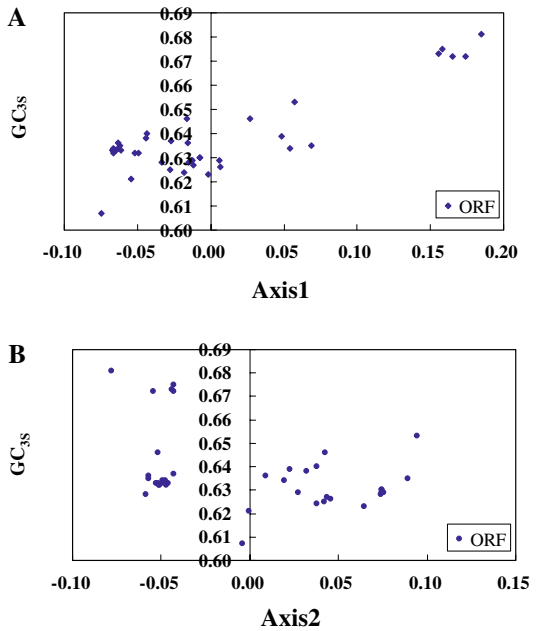

Fig. 3 Correlation between the first (**A**), second axis (**B**) values in COA and GC_{3S} values of each FMDV strains. COA: Correspondence analysis; GC_{3S}: the frequency of the nucleotide G + C at the synonymous third codon position (excluding Met, Trp and the termination codons); ORF, open reading frame

variation among ORFs in different organisms. However, whether there is any selection pressure that also contributes to the codon usage variation among these virus ORFs and which selection pressure determines the codon usage variation remained to be understood. Therefore, we performed a linear regression analysis on axis 1, axis 2 and axis 3 between the hydrophobicity of each protein and ORF length. It was found that axis 2 coordinates are also significantly correlated with the hydrophobicity of each protein ($r = 0.659$, $P < 0.01$), while axis 1 and axis 3 coordinates are also significantly correlated with the ORF length ($r = –0.564$, $P < 0.01$; $r = 0.610$, $P < 0.01$) respectively, indicating that the hydrophobicity of each protein and ORF length are also critical in affecting these viruses' codon usage, although they were less important than that of the mutational bias.

Cluster

Based on the RSCU variation of these 40 FMDV strains examined, a cluster tree was generated by using a hierarchical cluster method. As shown in Fig. 4, these 40 FMDV strains examined were divided into three main lineages (I, II and III).

Lineage I mainly contained strains of A, O, C and Asia 1, and branched to give eight sublineages (I1—I8). Five C [27–30, 39] strains were clustered into sublineage I1 and

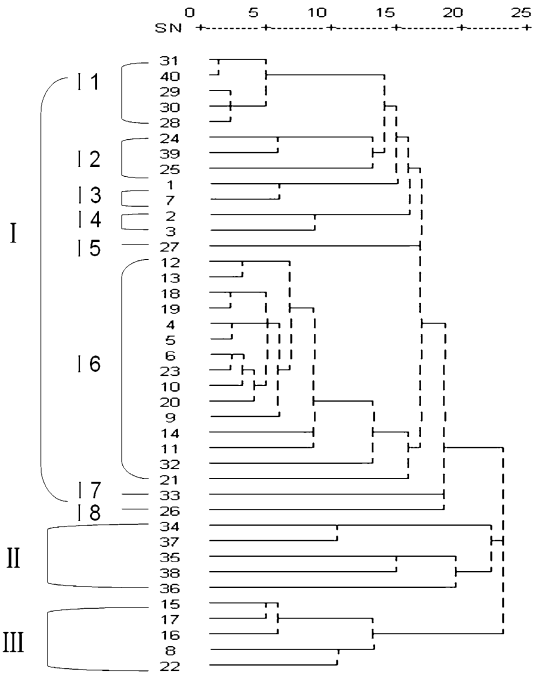

Fig. 4 Cluster tree based on the relative synonymous codon usage (RSCU) values of 40 FMDV strains examined. The cluster tree was generated by using Hierarchical cluster method

another C [26] was clustered into sublineage I5 separately. Sublineage I2 was composed of the three strains A [23, 24, 38], but another A [25] was clustered into sublineage I8 separately. Two O [1, 6] strains were clustered into sublineage I3. Sublineage I4 was composed of the two O strains [2], while 14 A strains [3–5, 8–13, 17–20, 22] and one Asia 1 strain [31] were clustered into sublineage I6. In the end, other Asia 1strain [32] was clustered into sublineage I7 separately.

Lineage II comprised five strains [33–36, 37] from SAT1, SAT2 and SAT3.

Lineage III included five strains [7, 14–16, 21] from O.

Discussion

Codon usage bias in FMDV was investigated in the present study. Up to now, it is unclear how FMDV serotypes might affect codon choice in FMD viruses and we used RSCU [42], ENC [32], COA [43, 44] and GC_{3S}, to measure the synonymous codon usage bias in order to minimize the effects of serotypes on codon bias, which have been successfully used to analyze the variation of codon usage among different viruses species [11, 13, 14, 22, 45, 46]. The analysis revealed that codon usage bias is low in most cases. As a case in point, the values of ENC vary from 50.02 to 52.79 (S.D. = 0.67) and the GC_{3S} values range from 61.00 to 68.00% (S.D. of 0.02). The average ENC value of 51.53 among 40 strains can be compared to those seen in other organisms such as H5N1 virus , severe acute respiratory syndrome *Coronavirus* (SARSCoV) and *Porcine adenovirus* where mean values of 50.91, 48.99, and 38.97, respectively, have been reported [14, 22, 47]. In the case of human RNA viruses, the average ENC value also probably lies close to 45 since the distribution of values for [13] individual virus ranges uniformly from just under 38.5 to 58.3 [13]. One possible explanation about why FMDV had had a lower codon usage bias than other RNA viruses [13, 14, 22, 47] is that a low bias is advantageous to viruses that need to replicate efficiently in vertebrate cells, with potentially distinct codon preferences.

A general mutational bias, which affects the whole genome would, certainly account for the majority of the codon usage variation. The genome base compositions affected the codon usage in *Entamoeba histolytica* genome [48]. Although the *Chlamydomonas reinhardtii* genome [49] had high GC contents, there was a little evidence that the genome composition shaped the codon usages in this genome. *C. elegans* showed a weak, but statistically significantly negative correlation between 'G + C' content and gene expression levels [50]. In human RNA viruses, H5N1 virus and SARS *Coronavirus*, mutation pressure rather than natural (translational) selection is the most important determinant of the codon bias. In this study, the general association between codon usage bias and base composition suggests that mutational pressure, rather than natural (translational) selection is supported by the highly significant correlation between GC_{12} and GC_{3S} ($r = 0.432$, $P < 0.05$), and the result of ENC-plot (Fig. 2). The fact that GC content varies in a similar way at all codon positions is usually assumed to be the result of mutational bias. A general mutational bias, which affects the whole genome would certainly account for the majority of the codon usage variation. A similar pattern of codon usage has been reported amongst some RNA viruses [13]. Since mutation rates in RNA viruses are much higher than those in DNA viruses [40], it is understandable that mutation pressure is the determinant source of codon usage bias in the 40 FMDV strains included in this study. Therefore, mutational bias is the major factor responsible for the variation of synonymous codon usage among ORFs in these virus genomes.

In *Drosophila* [51] genome, longer genes had lower codon usage bias. But, the longer genes had higher expression level and higher codon usage bias in *S.penumoniaes* genome [52]. In some virus, such as nucleopolyhedroviruses [11], H5N1 virus [14], SARS Coronavirus [22], adenoviruses [47], ORF length has no effect on the variations of synonymous codon usage. Those indicated that different genomes had different ORF lengths which accommodated their particular genome's best requirements, and there were not universal rules about ORF length and codon usage in all genomes. In this study, the ORF length had played a critical role in affecting FMDV codon usage. The mechanisms that lead this is not clear, which is needed a more comprehensive analysis.

It was reported that codon choices were influenced the hydropathy level of each protein in *Chlamydia trachomatis*, and *Thermotoga maritime* [53, 54]. In this study, codon usage is significantly positively correlated with the hydrophobicity of each FMDV. The link with hydropathy and codon usage may be caused by the fact that the expressed sequences are hydrophilic just because they accomplish their function in the aqueous media of the cell.

Up to date, phylogenetic analyses have been performed largely on FMDV sequences from the 1D coding regions. These analyses have permitted the discrimination among serotypically related FMDV strains [55]. Another analysis of the complete FMDV genomes indicated phylogenetic incongruities between different genomic regions which were suggestive of interserotypic recombination [56]. In this study, phylogenetic analyses based on the RSCU values of the 40 FMDV strains examined were carried out using a hierarchical cluster method. The result indicated complex phylogenetic relationships also exist between different FMDV isolates as determined by Carrillo et al.

2005[56]. For instance, Five O strains (HKN/2002, Tau-Yuan TW97 (Taiwan, 1997), Chu-Pei (Taiwan) (pig strain), Yunlin/Taiwan/97, and lz were clustered into Lineage III; One C strain (Argentina/85) was clustered into sublineageI5 and one Asia 1 strain (YNBS/China/58) was clustered into sublineageI6. Although they all belong to EURO-SA topotype [55], five C in sublineage I1 are sub-serotype C1 and one C in sublineage I5 is sub-serotype C3. The three A strains in Sublineage I2 was EURO-SA topotype and A strains sublineageI8 is Asia topotype, O stains in sublineageI3 is EURO-SA topotype, while 14 O strains in sublineage I6 belong to ME-SA topotype, PanAsia stains, were clustered into. O strains in Lineage III are Cathay topotype. All five SAT strains were clustered into Lineage II. Taken together, just as stated by other researchers [30, 56–58], these results suggest that FMDV sequences may undergo intertypic recombination, which conceivably undergo complex recombination events and the result of phylogenetic analyses based on the RSCU values fail to display serotype-specific phylogenetic relationships. These observations raise interesting questions about FMDV genome evolution in nature and the relative contribution of recombination to the generation of FMDV genetic and population diversity.

As we know FMD is highly contagious, affects all cloven-hoofed animals, and is caused by FMDV that exists as antigenically diverse serotypes and intra-typical variants (subtypes); Some published results has shown that the overall extent of codon usage bias in RNA viruses is low and there is a little variation in bias between genes or genomes [13, 14, 21, 22]. Our analysis revealed that although there are a few, variations in codon usage bias among different FMD viruses, codon usage bias in FMDV is low. Due to lack of data and politic factors, in this article, it is impossible to obtain information on the virus isolation background, vaccination etc, but clearly, a more comprehensive analysis is needed to reveal more information about codon usage bias variation within and among FMD viruses and what other factors are responsible, including the influence of factors, such as cell tropism, principal host species, method of transmission, and viral genetic structure. Such information would then allow us to more precisely judge the relative importance of mutation pressure versus natural selection in determining base composition and codon usage in these pathogens.

The recent European epizootic of FMD has made us aware of the great economic losses to be endured because no effective preventive and control measures are available for FMD [59]. Up to our knowledge, our work is the first report of the codon usage analysis on FMDV, which has provided a basic understanding of the mechanisms for codon usage bias and the processes governing the evolution of FMDV

Acknowledgments The work was supported by China Scholarship Council and the Department for Environment, Food and Rural Affairs (DEFRA), UK.

References

1. R. Grantham, C. Gautier, M. Gouy, Nucl Acids Res. **8**, 1893–1912 (1980)
2. M. Bulmer, J. Evol. Biol. **1**, 15–26 (1988)
3. D.C. Shields, P.M. Sharp, D.G. Higgins, F. Wright, Mol. Biol. Evol. **5**, 704–716 (1988)
4. M. Stenico, A.T. Lloyd, P.M. Sharp, Nucleic Acids Res. **22**, 2437–2446 (1994)
5. S. Karlin, J. Mrazek, J. Mol. Biol. **262**, 459–472 (1996)
6. S.L. Fennoy, J. Bailey-Serres, Nucleic Acids Res. **21**, 5294–5300 (1993)
7. H. Chiapello F. Lisacek M. Caboche A. Henaut, Gene **209**, GC1–GC38 (1998)
8. J. Ma, T. Zhou, W. Gu, X. Sun, Z. Lu, Biosystems **65**, 199–207 (2002)
9. S.K. Gupta, S. Majumdar, T.K. Bhattacharya, T.C. Ghosh, Biochem. Biophys. Res. Commun. **269**, 692–696 (2000)
10. K.C. Chou, C.T. Zhang, AIDS Res. Hum. Retrovirus. **8**, 1967–1976 (1992)
11. D.B. Levin, B. Whittome, J. Gen. Virol. **81**, 2313–2325 (2000)
12. S.M. Leisner, D.A. Neher, J. Theor. Biol. **217**, 195–201 (2002)
13. G.M. Jenkins, E.C. Holmes, Virus. Res. **92**, 1–7 (2003)
14. T. Zhou, W. Gu, J. Ma, X. Sun, Z. Lu, Biosystems **81**, 77–86 (2005)
15. G. Sanchez, A. Bosch, R.M. Pinto, J. Virol. **77**, 452–459 (2003)
16. S. Karlin, W. Doerfler, L.R. Cardon, J. Virol. **68**, 2889–2897 (1994)
17. S. Karlin, B.E. Blaisdell, G.A. Schachtel, J. Virol. **64**, 4264–4273 (1990)
18. B. Berkhout, A. Grigoriev, M. Bakker, V.V. Lukashov, AIDS Res. Hum. Retrovirus. **18**, 133–141 (2002)
19. D. Haydon, N. Knowles, J. McCauley, Virus Genes. **16**, 253–266 (1998)
20. C.R. Stephens, H. Waelbroeck, J. Mol. Evol. **48**, 390–397 (1999)
21. G.M. Jenkins, M. Pagel, E.A. Gould, P.M.D. Zanotto, E.C. Holmes, J. Mol. Evol. **52**, 383–390 (2001)
22. W.J. Gu, T. Zhou, J.M. Ma, X. Sun, Z.H. Lu, Virus Res. **101**, 155–161 (2004)
23. C. Tsai, C. Lin, C. Chang, Virus Res. **126**, 196–208 (2007)
24. S. Hassard, G. Ward, Biotechniques. **18**, 396–398, 400, (1995)
25. A.R. Samuel, N.J. Knowles, Trends Genet. **17**, 421–424 (2001)
26. E. Domingo, C. Escarmis, E. Baranowski, Ruiz-C.M. Jarabo, E. Carrillo, J.I. Nunez, F. Sobrino, Virus Res. **91**, 47–63 (2003)
27. M. Mittal, C. Tosh, D. Hemadri, A. Sanyal, S.K. Bandyopadhyay, Arch. Virol. **150**, 911–928 (2005)
28. Q. Feng, H. Yu, Y. Liu, C. He, J. Hu, H. Sang, N. Ding, M. Ding, Y.W. Fung, L.T. Lau, A.C. Yu, J. Chen, Biochem. Biophys. Res. Commun. **323**, 254–263 (2004)
29. P.W. Mason, J.M. Pacheco, Q.Z. Zhao, N.J. Knowles, J.Gene. Virol. **84**, 1583–1593 (2003)
30. C. Carrillo, E.R. Tulman, G. Delhon, Z. Lu, A. Carreno, A. Vagnozzi, G.F. Kutish, D.L. Rock, Dev Biol (Basel). 126, 23–30; discussion 323, (2006)
31. M.S. Paul, L. Wen-Hsiung, J. Mol. Evol. **V24**, 28–38 (1986)
32. F. Wright, Gene **87**, 23–29 (1990)
33. B.R. Morton, Proc. Natl. Acad Sci. U S A. **96**, 5123–5128 (1999)
34. H. Musto, W. Cruveiller, G. D'Onofrio, H. Romero, G. Bernardi, Mol. Biol. Evol. **18**, 1703–1707 (2001)
35. R.J. Grocock, P.M. Sharp, Int. J. Parasitol. **31**, 402–412 (2001)

36. G.A.C. Singer, D.A. Hickey, Gene **317**, 39–47 (2003)
37. L. Peixoto, A. Zavala, H. Romero, H. Musto, Gene **320**, 109–116 (2003)
38. H. Romero, A. Zavala, W. Musto, G. Bernardi, Gene **317**, 141–147 (2003)
39. M.J. Adams, J.F. Antoniw, Arch. Virol. **149**, 113–135 (2004)
40. J.W. Drake, J.J. Holland, Proc. Natl. Acad. Sci. U S A **96**, 13910–13913 (1999)
41. K.N. Zhao, W.J. Liu, I.H. Frazer, Virus Res. **98**, 95–104 (2003)
42. P.M. Sharp, W.H. Li, Nucleic Acids Res. **14**, 7737–7749 (1986)
43. S.K. Gupta, T.C. Ghosh, Gene **273**, 63–70 (2001)
44. W. Gu, T. Zhou, J. Ma, X. Sun, Z. Lu, Biosystems **73**, 89–97 (2004)
45. S.N. Sudha, S. Krishnaswamy, V. Sekar, Curr. Sci. **63**, 573–575 (1992)
46. I. Ahn, H.S. Son, Exp. Mol. Med. **38**, 643–651 (2006)
47. S. Das, S. Paul, C. Dutta, Virus Res. **117**, 227–236 (2006)
48. H. Romero, A. Zavala, H. Musto, Gene **242**, 307–311 (2000)
49. H. Naya, H. Romero, N. Carels, A. Zavala, H. Musto, FEBS Lett. **501**, 127–130 (2001)
50. G. Marais, L. Duret, J. Mol. Evol. **52**, 275–280 (2001)
51. J.M. Comeron, M. Aguade, J. Mol. Evol. **47**, 268–274 (1998)
52. Y.N. HouZC, YiChuan XueBao. **29**, 747–752 (2002)
53. H. Romero, A. Zavala, H. Musto, Nucleic Acids Res. **28**, 2084–2090 (2000)
54. A. Zavala, H. Naya, H. Romero, H. Musto, J. Mol. Evol. **54**, 563–568 (2002)
55. N.J. Knowles, A.R. Samuel, Virus Res. **91**, 65–80 (2003)
56. C. Carrillo, E.R. Tulman, G. Delhon, Z. Lu, A. Carreno, A. Vagnozzi, G.F. Kutish, D.L. Rock, J. Virol. **79**, 6487–6504 (2005)
57. O. Krebs, O. Marquardt, J. Gene. Virol. **73**, 613–619 (1992)
58. H. van Rensburg, D. Haydon, F. Joubert, A. Bastos, L. Heath, L. Nel, Gene **289**, 19–29 (2002)
59. E. Domingo, E. Baranowski, C. Escarmis, F. Sobrino, Comp. Immunol. Microbiol. Infect. Dis. **25**, 297–308 (2002)