Available online at www.sciencedirect.com
ScienceDirect
Virology 369 (2007) 431–442
VIROLOGY
www.elsevier.com/locate/yviro

Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses

Patrick C.Y. Woo [a,b,c], Beatrice H.L. Wong [c], Yi Huang [c], Susanna K.P. Lau [a,b,c], Kwok-Yung Yuen [a,b,c,*]

[a] State Key Laboratory of Emerging Infectious Diseases, Hong Kong
[b] Research Centre of Infection and Immunology, The University of Hong Kong, Hong Kong
[c] Department of Microbiology, The University of Hong Kong, Hong Kong

Received 12 July 2007; returned to author for revision 2 August 2007; accepted 7 August 2007
Available online 19 September 2007

Abstract

Using the complete genome sequences of 19 coronavirus genomes, we analyzed the codon usage bias, dinucleotide relative abundance and cytosine deamination in coronavirus genomes. Of the eight codons that contain CpG, six were markedly suppressed. The mean NNU/NNC ratio of the six amino acids using either NNC or NNU as codon is 3.262, suggesting cytosine deamination. Among the 16 dinucleotides, CpG was most markedly suppressed (mean relative abundance 0.509). No correlation was observed between CpG abundance and mean NNU/NNC ratio. Among the 19 coronaviruses, CoV-HKU1 showed the most extreme codon usage bias and extremely high NNU/NNC ratio of 8.835. Cytosine deamination and selection of CpG suppressed clones by the immune system are the two major independent biochemical and biological selective forces that shape codon usage bias in coronavirus genomes. The underlying mechanism for the extreme codon usage bias, cytosine deamination and G+C content in CoV-HKU1 warrants further studies.

Keywords: Coronavirus; Cytosine deamination; CpG suppression; Codon usage bias

Introduction

Codon usage bias is one of the most important indicators of the selective forces that shape genome evolution. In general, codon usage bias may be a result of mutation pressure and/or relative abundance of the corresponding acceptor tRNA molecules. For human RNA viruses, it has been observed in one study that codon usage bias was related to mutation pressure, G+C content, segmented nature of the genome and the route of transmission of the virus (Jenkins and Holmes, 2003). In other studies, it has been suggested that mutation pressure may result in bias in dinucleotide usage, such as CpG suppression, in small eukaryotic viruses (Karlin et al., 1994; Shackelton et al., 2006). Other factors, such as cytosine deamination, which results in C→U changes, have also been proposed to be responsible for shaping the G+C contents and GC skews of RNA viruses (Pyrc et al., 2004). Recently, it has been observed that codon usage is an important driving force in the evolution of astroviruses and small DNA viruses (Sewatanon et al., 2007; van Hemert et al., 2007). Despite all these fragmented observations, no study has integrated the various factors and been able to explain the basis for codon usage bias in viruses successfully.

Coronaviruses are positive sense, single-stranded RNA (ssRNA) viruses found in a wide range of animals in which they can cause respiratory, enteric, hepatic and neurological diseases of varying severity. The sizes of the genomes of coronaviruses are about 30 kb, the largest among RNA viruses. Based on genotypic and serological characterization, coronaviruses were divided into three distinct groups (Brian and Baric, 2005; Lai and Cavanagh, 1997; Ziebuhr, 2004). As a result of the low fidelity of the RNA-dependent–RNA polymerases, the mutation rates of RNA virus genomes are high, in the order of 1 per 10,000

* Corresponding author. State Key Laboratory of Emerging Infectious Diseases, Department of Microbiology, The University of Hong Kong, Room 423, University Pathology Building, Queen Mary Hospital Compound, Pokfulam, Hong Kong. Fax: +852 2855 1241.
E-mail address: hkumicro@hkucc.hku.hk (K.-Y. Yuen).

0042-6822/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2007.08.010

nucleotides replicated. Furthermore, the unique mechanism of viral replication has resulted in a high frequency of recombination in coronaviruses (Lai and Cavanagh, 1997; Woo et al., 2006b). Their tendency for recombination and high mutation rates have made their genomes highly plastic, allowed them to adapt to new hosts and ecological niches, and given them the potential to be good candidates for causing pandemics. These factors have made the study of coronavirus evolution particularly important, both biologically and for practical purposes (Grigoriev, 2004; Gu et al., 2004; Yap et al., 2003). However, the relative importance of the various selective forces that shape the codon usage bias in coronaviruses and their underlying biological and biochemical basis are still poorly understood.

The recent severe acute respiratory syndrome (SARS) epidemic, the discovery of SARS coronavirus (SARS-CoV) and identification of SARS-CoV-like viruses from Himalayan palm civets and a raccoon dog from wild live markets in China have led to a boost in interests in discovery of novel coronaviruses in both humans and animals (Guan et al., 2003; Marra et al., 2003; Peiris et al., 2003; Rota et al., 2003; Snijder et al., 2003; Woo et al., 2004). For human coronaviruses, in 2004, a novel group 1 human coronavirus, human coronavirus NL63 (HCoV-NL63), was reported (Fouchier et al., 2004; van der Hoek et al., 2004); and in 2005, we described the discovery, complete genome sequence and molecular diversity of another novel group 2 human coronavirus, coronavirus HKU1 (CoV-HKU1) (Lau et al., 2006; Woo et al., 2005a,b,c, 2006b). As for animal coronaviruses, six group 1 (Poon et al., 2005; Tang et al., 2006; Woo et al., 2006a; Lau et al., 2007), six group 2, including bat SARS coronavirus, sable antelope coronavirus, giraffe coronavirus, and two new subgroups of group 2 coronaviruses (Lau et al., 2005; Li et al., 2005; Woo et al., 2006a, 2007), 11 group 3 (Cavanagh et al., 2002; East et al., 2004; Jonassen et al., 2005; Liu et al., 2005; Hasoksuz et al., 2007) coronaviruses, and two unclassified coronaviruses from Asian leopard cats and Chinese ferret badgers (Dong et al., 2007) have recently been described. Since the number of coronavirus species with complete genomes available has increased from 9 in 2003 to 19 in 2007, this has provided a golden opportunity to study genome evolution in coronaviruses.

In this study, we analyzed the codon usage bias, dinucleotide relative abundance, cytosine deamination in coronavirus genomes and the codon usage bias in the hosts of the various coronaviruses. The relative importance of the various forces in shaping the codon usage bias in the various coronaviruses and the extreme codon usage bias and cytosine deamination in CoV-HKU1 were also discussed.

Results

Codon usage in coronavirus genomes

The mean (S.D.) effective number of codons (Nc) of the 19 coronaviruses is 45.448 (4.207) (Table 1). The codon usage fractions in the 19 coronavirus genomes are shown in Table 2. For all amino acids, the codon usage patterns of every individual coronavirus species are similar to the general codon usage patterns in coronaviruses. CoV-HKU1, HCoV-NL63, murine hepatitis virus (MHV) and bat coronavirus HKU5 (bat-CoV HKU5) are the four coronaviruses with relatively larger number of codons showing usage fractions outside the mean±2 S.D. usage fraction range of the corresponding codons, probably due to their relatively high (MHV and bat-CoV HKU5) or low (CoV-HKU1 and HCoV-NL63) G+C contents (Tables 1 and 2).

To study the possible effect of CpG suppression on codon usage bias, the usage fractions of the eight codons that contain CpG (CCG, GCG, UCG, ACG, CGC, CGG, CGU and CGA) were analyzed. Of these eight codons, six [CCG (mean 0.058), GCG (mean 0.060), UCG (mean 0.038), ACG (mean 0.070), CGG (mean 0.038) and CGA (mean 0.060)] were markedly suppressed. CGC is slightly suppressed (mean 0.122) whereas CGU is over-represented (mean 0.322).

To study the possible effect of cytosine deamination on codon usage bias, codons of amino acids that can use C or U in the codons were analyzed. For all amino acids that only use either NNU or NNC as codon (asparagine, histidine, aspartic acid, tyrosine, cysteine and phenylalanine), all NNU are markedly over represented with usage fractions of more than 0.700, whereas the usage fractions of all NNC are less than 0.300. For amino acids that use NNU, NNC or other codons (threonine, isoleucine, proline, leucine, alanine, glycine, valine and serine), the usage fractions of all NNU are at least three times more than those of the corresponding NNC. For leucine, UUA (mean 0.223) is used much more frequently than CUA (mean 0.081), and UUG (mean 0.261) is used much more frequently than CUG (mean 0.072).

To study the possible effect of A↔G transition on codon usage bias, codons of amino acids that can use A or G in the codons were analyzed. For amino acids that use either NNA or NNG as codons (lysine, glutamine and glutamic acid) and those that use NNA, NNG or other codons but excluding those codons with CpG (arginine, glycine and valine), the usage fractions of NNA are often higher than those of NNG, but the differences between the usage fractions of NNA and NNG are not as marked as those between the usage fractions of NNU and NNC.

Codon usage in CoV-HKU1

Among all the 19 coronaviruses, CoV-HKU1 showed the most extreme codon usage bias. CoV-HKU1 is the only coronavirus that showed Nc outside the mean±2 S.D. range. CoV-HKU1 also possessed the lowest G+C content, highest GC skew, lowest percentages of G and C and highest percentage of U among all coronavirus genomes (Table 1). For the six amino acids that only use either NNU or NNC as codon (asparagine, histidine, aspartic acid, tyrosine, cysteine and phenylalanine), amino acids that use NNU, NNC or other codons (threonine, isoleucine, proline, leucine, alanine, glycine, valine and serine), and for leucine that use UNN or CNN as codon, the average (S.D.) ratio of the usage fractions of the codons with U to those with C is 9.66 (2.49) (Table 2). For amino acids that use either NNA or NNG as codons (lysine, glutamine and glutamic acid) and those that use NNA, NNG or other codons but excluding those codons with CpG (arginine, glycine and valine), the average (S.D.) ratio of the usage fractions of the codons with A to those with G is 2.72 (0.57) (Table 2).

Table 1
Coronavirus genomes used in the present study

Coronavirus	Host	GenBank accession no.	Reference	Genome size (bases)	G+C content (%)	GC skew	Mononucleotide frequencies (%)				Nc
							G	A	U	C	
Group 1a											
TGEV	Pig	NC_002306	Almazan et al., 2000	28,586	37.5	0.097	20.6	29.5	32.9	17.0	44.737
FIPV	Cat	AY994055	Haijema et al., 2003	29,355	38.1	0.102	21.0	29.2	32.7	17.1	46.150
PRCV	Pig	DQ811787	Zhang et al., 2007	27,550	37.4	0.107	20.7	29.3	33.2	16.7	44.406
Group 1b											
HCoV-229E	Human	NC_002645	Thiel et al., 2001	27,317	38.2	0.129	21.6	27.2	34.6	16.7	44.281
HCoV-NL63	Human	NC_005831	van der Hoek et al., 2004	27,553	34.4	0.161	20.0	26.3	39.2	14.4	37.275
PEDV	Pig	NC_003436	Kocherhans et al., 2001	28,033	42.0	0.086	22.8	24.7	33.2	19.2	48.424
BtCoV	Bat	DQ648858	Tang et al., 2006	28,203	40.1	0.102	22.1	26.2	33.7	18.0	46.905
Bat-CoV HKU2	Bat	EF203064	Lau et al., 2007	27,164	38.9	0.140	22.2	24.9	35.1	16.8	43.342
Group 2a											
HCoV-OC43	Human	NC_005147	Vijgen et al., 2005	30,738	36.8	0.176	21.7	27.6	35.6	15.2	43.791
CoV-HKU1	Human	NC_006577	Woo et al., 2005b	29,926	32.0	0.188	19.0	27.8	40.1	13.0	35.671
BCoV	Cattle	NC_003045	Chouljenko et al., 2001	31,028	37.1	0.174	21.8	27.4	35.5	15.3	43.856
PHEV	Pig	NC_007732	Vijgen et al., 2006	30,480	37.2	0.164	21.7	27.3	35.4	15.6	44.380
MHV	Mouse	NC_001846	Leparc-Goffart et al., 1997	31,357	41.7	0.142	23.9	26.0	32.3	17.9	51.237
Group 2b											
SARS-CoV	Human	NC_004718	Marra et al., 2003	29,751	40.7	0.020	20.8	28.5	30.7	20.0	49.423
Bat-SARS-CoV HKU3	Bat	DQ022305	Lau et al., 2005	29,728	41.1	0.027	21.1	28.4	30.5	20.0	49.882
Group 2c											
Bat-CoV HKU4	Bat	EF065506	Woo et al., 2006b	30,286	37.8	0.093	20.7	27.6	34.6	17.1	44.585
Bat-CoV HKU5	Bat	EF065511	Woo et al., 2006b	30,488	42.9	0.004	21.6	26.6	30.4	21.4	53.230
Group 2d											
Bat-CoV HKU9	Bat	EF065513	Woo et al., 2006b	29,114	41.0	0.138	23.3	25.3	33.7	17.7	46.162
Group 3											
IBV	Chicken	NC_001451	Boursnell et al., 1987	27,608	37.9	0.144	21.7	28.9	33.2	16.2	45.777

Codon usage in hosts of coronaviruses

The codon usage fractions in the hosts of coronaviruses, including human, mouse, pig, cat and chicken, are shown in Table 3. To study the possible effect of CpG suppression on codon usage bias, the usage fractions of the eight codons that contain CpG (CCG, GCG, UCG, ACG, CGC, CGG, CGU and CGA) were analyzed. Among these eight codons, six (CCG, GCG, UCG, ACG, CGU and CGA) were suppressed, of which five were also suppressed in the coronavirus genomes. To study the possible effect of C↔U transition and A↔G transition on codon usage bias, codons of amino acids that can use C or U and those of amino acids that can use A or G in the codons were analyzed. No pattern of difference was observed between the use of NNU and NNC and between the use of NNA and NNG.

Dinucleotide relative abundance in coronavirus genomes

The relative abundance of the 16 dinucleotides in the 19 coronavirus genomes are shown in Table 4. Among the 16 dinucleotides, the relative abundance of CpG showed the most marked deviation from the "normal range" (mean±S.D.= 0.509±0.063, 0.271 less than 0.78), with all 19 genomes showing CpG under-representation. In addition, the relative abundance of UpG and CpA also showed slight deviation from the "normal range" (mean±S.D.=1.331±0.057 and 1.257± 0.070, respectively, both >1.23), with all 19 and 13 genomes showing UpG and CpA over-representation, respectively.

Correlations between CpG suppression and cytosine deamination in coronaviruses

The relationship between CpG suppression and cytosine deamination in the 19 coronavirus genomes is shown in Fig. 1. The mean (S.D.) of the NNU/NNC in the six amino acids that only use either NNC or NNU as the codons of the 19 coronavirus genomes is 3.262 (1.785). CoV-HKU1 showed extremely high NNU/NNC ratio of 8.835. No significant correlation was observed between CpG abundance and mean NNU/NNC ratio in the 19 coronavirus genomes ($r=-0.339$, $P=0.156$).

Discussion

Marked CpG suppression is observed in all coronavirus genomes. The discovery of Toll-like receptors (TLRs) that recognize pathogen-associated molecular patterns and the downstream molecular pathways was one of the biggest advances in the understanding of vertebrate innate immunity

Table 2
Codon usage fractions in coronaviruses

Amino acid	Codon[a]	Codon usage fraction																			
		HCoV-229E	HCoV-NL63	TGEV	PEDV	PRCV	FIPV	BtCoV	Bat-CoV HKU2	HCoV-OC43	CoV-HKU1	BCoV	MHV	PHEV	SARS-CoV	Bat-SARS-CoV HKU3	Bat-CoV HKU4	Bat-CoV HKU5	Bat-CoV HKU9	IBV	Mean ± S.D.
Lysine	AAA	0.561	0.590	0.610	0.347	0.610	0.584	0.464	0.440	0.511	0.695	0.509	0.392	0.513	0.532	0.541	0.557	0.474	0.394	0.518	0.518 ± 0.086
	AAG	0.439	0.410	0.390	0.653	0.390	0.416	0.536	0.560	0.489	0.305	0.491	0.608	0.487	0.468	0.459	0.443	0.526	0.606	0.482	0.482 ± 0.086
Asparagine	AAC	0.311	0.208	0.347	0.359	0.346	0.338	0.368	0.257	0.172	0.124	0.166	0.257	0.179	0.372	0.366	0.237	0.376	0.217	0.238	0.276 ± 0.083
	AAU	0.689	0.792	0.653	0.641	0.654	0.662	0.632	0.743	0.828	0.876	0.834	0.743	0.821	0.628	0.634	0.763	0.624	0.783	0.762	0.724 ± 0.083
Threonine	ACA	0.356	0.290	0.409	0.290	0.398	0.424	0.316	0.293	0.318	0.261	0.300	0.261	0.292	0.391	0.380	0.305	0.271	0.273	0.381	0.327 ± 0.054
	ACC	0.115	0.075	0.103	0.172	0.101	0.104	0.177	0.128	0.134	0.057	0.140	0.218	0.144	0.132	0.121	0.125	0.173	0.144	0.080	0.129 ± 0.039
	ACG	0.045	0.032	0.064	0.081	0.076	0.086	0.077	0.045	0.055	0.018	0.062	0.134	0.059	0.047	0.073	0.065	0.100	0.126	0.076	0.070 ± 0.029
	ACU	0.484	0.603	0.424	0.457	0.425	0.385	0.431	0.534	0.492	0.664	0.499	0.387	0.504	0.430	0.426	0.505	0.457	0.458	0.463	0.475 ± 0.069
Arginine	AGA	0.365	0.236	0.464	0.188	0.477	0.457	0.287	0.214	0.305	0.337	0.299	0.249	0.321	0.352	0.383	0.266	0.276	0.188	0.347	0.316 ± 0.087
	AGG	0.109	0.141	0.145	0.152	0.143	0.136	0.169	0.193	0.117	0.098	0.126	0.184	0.117	0.157	0.111	0.129	0.135	0.195	0.155	0.143 ± 0.028
	CGA	0.049	0.047	0.035	0.058	0.037	0.065	0.066	0.052	0.076	0.056	0.070	0.075	0.058	0.080	0.060	0.071	0.079	0.036	0.061	0.060 ± 0.014
	CGC	0.115	0.084	0.082	0.195	0.077	0.077	0.097	0.125	0.120	0.075	0.114	0.171	0.143	0.128	0.147	0.129	0.173	0.137	0.134	0.122 ± 0.035
	CGG	0.030	0.017	0.035	0.036	0.030	0.022	0.030	0.043	0.050	0.036	0.053	0.057	0.047	0.019	0.030	0.041	0.064	0.041	0.032	0.038 ± 0.013
	CGU	0.332	0.475	0.240	0.371	0.237	0.244	0.350	0.373	0.331	0.399	0.337	0.262	0.315	0.264	0.269	0.364	0.273	0.404	0.271	0.322 ± 0.066
Isoleucine	AUA	0.252	0.241	0.230	0.187	0.239	0.280	0.259	0.221	0.323	0.296	0.320	0.304	0.326	0.219	0.229	0.315	0.229	0.392	0.389	0.276 ± 0.058
	AUC	0.130	0.062	0.163	0.212	0.148	0.167	0.181	0.133	0.098	0.056	0.098	0.152	0.102	0.213	0.237	0.152	0.294	0.099	0.088	0.147 ± 0.062
	AUU	0.618	0.697	0.607	0.601	0.613	0.554	0.560	0.646	0.579	0.649	0.582	0.544	0.572	0.568	0.534	0.533	0.477	0.508	0.522	0.577 ± 0.054
Glutamine	CAA	0.652	0.656	0.638	0.436	0.620	0.613	0.551	0.403	0.541	0.690	0.531	0.464	0.517	0.604	0.568	0.602	0.550	0.440	0.623	0.563 ± 0.082
	CAG	0.348	0.344	0.362	0.564	0.380	0.387	0.449	0.597	0.459	0.310	0.469	0.536	0.483	0.396	0.432	0.398	0.450	0.560	0.377	0.437 ± 0.082
Histidine	CAC	0.284	0.206	0.250	0.320	0.243	0.287	0.389	0.266	0.213	0.105	0.212	0.274	0.221	0.359	0.327	0.221	0.361	0.233	0.320	0.268 ± 0.068
	CAU	0.716	0.794	0.750	0.680	0.757	0.713	0.611	0.734	0.787	0.895	0.788	0.726	0.779	0.641	0.673	0.779	0.639	0.767	0.680	0.732 ± 0.068
Proline	CCA	0.318	0.303	0.411	0.327	0.414	0.364	0.338	0.298	0.306	0.239	0.300	0.285	0.288	0.426	0.394	0.301	0.306	0.261	0.386	0.330 ± 0.054
	CCC	0.108	0.052	0.061	0.141	0.064	0.093	0.122	0.079	0.124	0.073	0.133	0.221	0.108	0.093	0.122	0.081	0.156	0.118	0.099	0.108 ± 0.040
	CCG	0.054	0.029	0.049	0.049	0.037	0.050	0.059	0.066	0.058	0.031	0.062	0.100	0.064	0.043	0.051	0.039	0.069	0.109	0.086	0.058 ± 0.021
	CCU	0.520	0.616	0.479	0.483	0.485	0.492	0.481	0.556	0.512	0.657	0.504	0.395	0.541	0.439	0.433	0.578	0.469	0.512	0.429	0.504 ± 0.065
Leucine	CUA	0.069	0.041	0.093	0.080	0.084	0.108	0.061	0.058	0.065	0.036	0.068	0.086	0.066	0.117	0.126	0.072	0.115	0.086	0.109	0.081 ± 0.025
	CUC	0.053	0.037	0.088	0.106	0.087	0.093	0.073	0.060	0.049	0.021	0.049	0.085	0.053	0.136	0.135	0.069	0.153	0.060	0.060	0.077 ± 0.035

	CUU	0.283	0.311	0.366	0.316	0.361	0.314	0.268	0.392	0.248	0.227	0.241	0.209	0.252	0.299	0.301	0.265	0.290	0.194	0.298	0.286 ± 0.052
	CUG	0.069	0.022	0.048	0.100	0.044	0.063	0.085	0.059	0.073	0.023	0.069	0.110	0.065	0.094	0.097	0.053	0.130	0.082	0.076	0.072 ± 0.028
	UUA	0.184	0.281	0.204	0.129	0.212	0.198	0.153	0.182	0.248	0.415	0.247	0.225	0.247	0.177	0.182	0.308	0.148	0.271	0.235	0.223 ± 0.066
	UUG	0.342	0.308	0.202	0.270	0.212	0.224	0.360	0.249	0.316	0.278	0.325	0.285	0.316	0.178	0.159	0.233	0.164	0.306	0.223	0.261 ± 0.061
Glutamic acid	GAA	0.676	0.662	0.740	0.462	0.732	0.684	0.599	0.467	0.591	0.684	0.577	0.449	0.627	0.531	0.552	0.627	0.567	0.496	0.604	0.596 ± 0.089
	GAG	0.324	0.338	0.260	0.538	0.268	0.316	0.401	0.533	0.409	0.316	0.423	0.551	0.373	0.469	0.448	0.373	0.433	0.504	0.396	0.404 ± 0.089
Aspartic acid	GAC	0.368	0.218	0.339	0.357	0.324	0.328	0.369	0.360	0.178	0.150	0.184	0.257	0.179	0.372	0.420	0.262	0.371	0.296	0.299	0.296 ± 0.081
	GAU	0.632	0.782	0.661	0.643	0.676	0.672	0.631	0.640	0.822	0.850	0.816	0.743	0.821	0.628	0.580	0.738	0.629	0.704	0.701	0.704 ± 0.081
Alanine	GCA	0.277	0.266	0.334	0.282	0.327	0.318	0.284	0.236	0.274	0.231	0.280	0.226	0.276	0.274	0.284	0.293	0.318	0.257	0.373	0.285 ± 0.037
	GCC	0.127	0.114	0.099	0.170	0.127	0.129	0.191	0.130	0.138	0.089	0.140	0.217	0.148	0.145	0.156	0.128	0.170	0.147	0.089	0.140 ± 0.032
	GCG	0.058	0.025	0.034	0.075	0.032	0.046	0.047	0.064	0.050	0.030	0.050	0.113	0.043	0.059	0.067	0.056	0.082	0.132	0.080	0.060 ± 0.028
	GCU	0.538	0.595	0.533	0.473	0.513	0.507	0.478	0.569	0.538	0.650	0.530	0.444	0.533	0.522	0.493	0.523	0.430	0.463	0.457	0.515 ± 0.054
Glycine	GGA	0.115	0.073	0.179	0.104	0.185	0.166	0.096	0.084	0.171	0.104	0.159	0.177	0.156	0.228	0.233	0.131	0.205	0.074	0.198	0.149 ± 0.051
	GGC	0.199	0.076	0.143	0.252	0.136	0.147	0.195	0.172	0.159	0.107	0.152	0.268	0.184	0.252	0.226	0.171	0.272	0.183	0.152	0.181 ± 0.054
	GGG	0.030	0.023	0.032	0.051	0.035	0.028	0.051	0.044	0.060	0.035	0.069	0.086	0.076	0.041	0.059	0.065	0.067	0.072	0.043	0.051 ± 0.018
	GGU	0.655	0.828	0.645	0.594	0.644	0.658	0.657	0.700	0.610	0.755	0.620	0.470	0.584	0.479	0.483	0.633	0.456	0.671	0.607	0.618 ± 0.096
Valine	GUA	0.120	0.106	0.194	0.121	0.177	0.216	0.122	0.097	0.174	0.196	0.163	0.136	0.169	0.215	0.199	0.226	0.158	0.210	0.224	0.170 ± 0.042
	GUC	0.119	0.090	0.167	0.182	0.160	0.145	0.151	0.125	0.082	0.056	0.084	0.140	0.089	0.171	0.175	0.103	0.244	0.107	0.101	0.131 ± 0.046
	GUG	0.190	0.078	0.162	0.193	0.167	0.181	0.187	0.146	0.191	0.060	0.197	0.285	0.202	0.189	0.196	0.157	0.197	0.221	0.172	0.177 ± 0.048
	GUU	0.571	0.727	0.477	0.504	0.496	0.458	0.539	0.632	0.553	0.688	0.556	0.440	0.540	0.424	0.431	0.514	0.401	0.463	0.502	0.522 ± 0.087
Tyrosine	UAC	0.326	0.190	0.391	0.335	0.376	0.403	0.389	0.275	0.168	0.118	0.199	0.226	0.204	0.439	0.428	0.222	0.436	0.255	0.266	0.297 ± 0.102
	UAU	0.674	0.810	0.609	0.665	0.624	0.597	0.611	0.725	0.832	0.882	0.801	0.774	0.796	0.561	0.572	0.778	0.564	0.745	0.734	0.703 ± 0.102
Serine	UCA	0.187	0.192	0.217	0.175	0.214	0.204	0.213	0.166	0.146	0.144	0.143	0.141	0.155	0.294	0.275	0.214	0.200	0.162	0.207	0.192 ± 0.042
	UCC	0.082	0.043	0.073	0.122	0.070	0.094	0.101	0.082	0.074	0.038	0.071	0.105	0.088	0.066	0.066	0.082	0.134	0.085	0.035	0.080 ± 0.026
	UCG	0.028	0.010	0.020	0.043	0.021	0.035	0.037	0.034	0.031	0.012	0.032	0.063	0.032	0.040	0.046	0.033	0.077	0.071	0.048	0.038 ± 0.018
	UCU	0.353	0.393	0.312	0.313	0.319	0.276	0.317	0.340	0.302	0.454	0.327	0.239	0.290	0.322	0.313	0.331	0.274	0.305	0.344	0.322 ± 0.046
	AGC	0.087	0.044	0.111	0.113	0.108	0.100	0.070	0.101	0.096	0.031	0.100	0.148	0.113	0.087	0.091	0.077	0.113	0.074	0.064	0.091 ± 0.027
	AGU	0.263	0.318	0.267	0.235	0.268	0.291	0.262	0.277	0.351	0.322	0.328	0.304	0.322	0.190	0.209	0.263	0.202	0.304	0.301	0.278 ± 0.045
Cysteine	UGC	0.282	0.091	0.316	0.326	0.290	0.271	0.259	0.267	0.246	0.082	0.225	0.340	0.246	0.375	0.344	0.197	0.373	0.270	0.180	0.262 ± 0.082
	UGU	0.718	0.909	0.684	0.674	0.710	0.729	0.741	0.733	0.754	0.918	0.775	0.660	0.754	0.625	0.656	0.803	0.627	0.730	0.820	0.738 ± 0.082
Phenylalanine	UUC	0.206	0.105	0.292	0.329	0.276	0.296	0.238	0.158	0.126	0.071	0.126	0.241	0.129	0.377	0.404	0.201	0.429	0.177	0.213	0.231 ± 0.104
	UUU	0.794	0.895	0.708	0.671	0.724	0.704	0.762	0.842	0.874	0.929	0.874	0.759	0.871	0.623	0.596	0.799	0.571	0.823	0.787	0.769 ± 0.104

[a] Codons with CpG are in red and codons of amino acids that use either NNC or NNU as the codon are in green. (For interpretation of the references to colour in this table legend, the reader is referred to the web version of this article.)

Table 3
Codon usage fractions in different hosts of coronaviruses

Amino acids	Codon[a]	Codon usage fraction				
		Human (*Homo sapiens*)	Mouse (*Mus musculus*)	Pig (*Sus scrofa*)	Cat (*Felis catus*)	Chicken (*Gallus gallus*)
Lysine	AAA	0.43	0.39	0.38	0.42	0.44
	AAG	0.57	0.61	0.62	0.58	0.56
Asparagine	AAC	0.53	0.57	0.61	0.59	0.57
	AAU	0.47	0.43	0.39	0.41	0.43
Threonine	ACA	0.28	0.29	0.23	0.24	0.30
	ACC	0.36	0.35	0.42	0.40	0.31
	ACG	0.11	0.11	0.14	0.15	0.14
	ACU	0.25	0.25	0.21	0.21	0.25
Arginine	AGA	0.21	0.21	0.19	0.22	0.23
	AGG	0.21	0.22	0.20	0.23	0.21
	CGA	0.11	0.12	0.10	0.09	0.10
	CGC	0.18	0.17	0.22	0.19	0.19
	CGG	0.20	0.19	0.21	0.20	0.18
	CGU	0.08	0.09	0.08	0.07	0.10
Isoleucine	AUA	0.17	0.16	0.14	0.15	0.18
	AUC	0.47	0.50	0.56	0.53	0.46
	AUU	0.36	0.34	0.30	0.31	0.35
Glutamine	CAA	0.26	0.25	0.22	0.27	0.27
	CAG	0.74	0.75	0.78	0.73	0.73
Histidine	CAC	0.58	0.60	0.65	0.63	0.60
	CAU	0.42	0.40	0.35	0.37	0.40
Proline	CCA	0.28	0.29	0.24	0.24	0.28
	CCC	0.32	0.30	0.36	0.37	0.30
	CCG	0.11	0.10	0.14	0.13	0.14
Leucine	CCU	0.29	0.31	0.26	0.26	0.28
	CUA	0.07	0.08	0.06	0.06	0.06
	CUC	0.20	0.20	0.22	0.21	0.18
	CUT	0.13	0.13	0.11	0.11	0.13
	CUG	0.40	0.39	0.45	0.44	0.41
	UUA	0.08	0.06	0.05	0.05	0.08
	UUG	0.13	0.13	0.11	0.13	0.14
Glutamic acid	GAA	0.42	0.40	0.37	0.42	0.43
	GAG	0.58	0.60	0.63	0.58	0.57
Aspartic acid	GAC	0.54	0.56	0.60	0.58	0.49
	GAU	0.46	0.44	0.40	0.42	0.51
Alanine	GCA	0.23	0.23	0.18	0.19	0.27
	GCC	0.40	0.38	0.45	0.45	0.32
	GCG	0.11	0.09	0.12	0.13	0.12
	GCU	0.27	0.29	0.24	0.24	0.29
Glycine	GGA	0.25	0.26	0.23	0.25	0.27
	GGC	0.34	0.33	0.37	0.35	0.31
	GGG	0.25	0.24	0.26	0.24	0.24
	GGU	0.16	0.18	0.14	0.15	0.18
Valine	GUA	0.12	0.12	0.09	0.10	0.13
	GUC	0.24	0.25	0.27	0.28	0.22
	GUG	0.46	0.46	0.50	0.47	0.45
	GUU	0.18	0.17	0.50	0.15	0.21
Tyrosine	UAC	0.56	0.57	0.64	0.61	0.60
	UAU	0.44	0.43	0.36	0.39	0.40
Serine	UCA	0.15	0.14	0.12	0.12	0.15
	UCC	0.22	0.22	0.25	0.25	0.20
	UCG	0.05	0.05	0.06	0.06	0.07
	UCU	0.19	0.20	0.17	0.19	0.18
	AGC	0.24	0.24	0.27	0.24	0.26
	AGU	0.15	0.15	0.13	0.13	0.14
Cysteine	UGC	0.55	0.52	0.61	0.57	0.60
	UGU	0.45	0.48	0.39	0.43	0.40
Phenylalanine	UUC	0.54	0.56	0.61	0.59	0.54
	UUU	0.46	0.44	0.39	0.41	0.46

[a] Codons with CpG are in red and codons of amino acids that use either NNC or NNU as the codon are in green. (For interpretation of the references to colour in this table legend, the reader is referred to the web version of this article.)

in recent years. Among the TLR that recognize viral components, TLR3, 7, 8 and 9 detect viral nucleic acids (Bowie and Haga, 2005). It has been shown that TLR9 bound to CpG of double-stranded DNA and elicited the downstream inflammatory response, and administration of CpG oligodeoxynucleotides has been shown to protect mice from herpes simplex virus 2

Table 4
Relative abundance of the 16 dinucleotides in the 19 coronavirus species with complete genomes available

Coronavirus	Relative abundance of the 16 dinucleotides[a]															
	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
TGEV	1.080	1.176	0.954	0.865	1.316	0.830	0.456	1.126	0.921	1.142	0.895	1.062	0.803	0.822	1.401	1.016
FIPV	1.067	1.221	0.945	0.869	1.301	0.923	0.501	1.108	0.913	1.113	0.884	1.092	0.848	0.786	1.383	1.000
PRCV	1.083	1.185	0.956	0.863	1.307	0.860	0.462	1.136	0.939	1.157	0.886	1.062	0.812	0.811	1.382	1.016
HCoV-229E	1.162	1.254	0.885	0.818	1.320	0.896	0.499	1.090	0.868	1.192	0.857	1.097	0.786	0.709	1.418	1.035
HCoV-NL63	1.171	1.293	0.874	0.834	1.293	1.012	0.416	1.098	0.817	1.111	0.924	1.122	0.872	0.744	1.339	1.008
PEDV	1.065	1.201	0.976	0.865	1.349	0.922	0.548	1.098	0.852	1.165	0.923	1.070	0.853	0.784	1.321	1.007
BtCoV	1.092	1.251	0.949	0.815	1.378	0.987	0.502	1.038	0.846	1.080	0.921	1.127	0.826	0.741	1.342	1.047
Bat-CoV HKU2	1.048	1.243	0.940	0.881	1.266	0.921	0.509	1.136	0.886	1.233	0.872	1.039	0.903	0.695	1.360	0.990
HCoV-OC43	1.050	1.024	0.985	0.946	1.239	1.168	0.485	1.053	0.884	1.303	0.891	1.022	0.926	0.720	1.281	1.002
CoV-HKU1	1.086	1.106	0.965	0.923	1.106	1.183	0.445	1.131	0.908	1.174	0.914	1.063	0.959	0.805	1.246	0.976
BCoV	1.052	1.049	0.987	0.945	1.216	1.196	0.479	1.067	0.887	1.289	0.904	1.020	0.935	0.699	1.292	0.999
PHEV	1.059	1.033	0.995	0.951	1.244	1.150	0.502	1.050	0.877	1.329	0.891	1.015	0.931	0.706	1.301	0.997
MHV	1.079	0.945	1.013	0.952	1.160	1.217	0.607	1.037	0.901	1.262	0.875	1.023	0.916	0.709	1.295	0.996
SARS-CoV	1.034	1.175	0.995	0.857	1.298	0.824	0.456	1.205	0.944	1.153	0.924	0.986	0.800	0.846	1.409	1.018
Bat SARS-CoV HKU3	1.053	1.179	1.001	0.842	1.267	0.849	0.497	1.196	0.967	1.137	0.920	1.010	0.808	0.852	1.382	1.010
Bat-CoV HKU4	1.037	1.122	0.997	0.911	1.186	1.025	0.508	1.132	0.857	1.214	0.886	1.075	0.952	0.777	1.298	0.960
Bat-CoV HKU5	1.088	1.071	0.974	0.902	1.229	0.873	0.605	1.152	0.922	1.146	0.900	1.035	0.828	0.922	1.355	0.952
Bat-CoV HKU9	0.968	1.138	1.000	0.938	1.161	1.117	0.678	1.039	0.780	1.236	0.902	1.095	1.079	0.637	1.235	0.959
IBV	1.053	1.132	1.068	0.833	1.238	0.990	0.512	1.115	0.877	1.194	0.913	1.082	0.906	0.762	1.249	1.034

[a] Numbers >1.23 and <0.78 are shown in red and green, respectively. (For interpretation of the references to colour in this table legend, the reader is referred to the web version of this article.)

infections (Ashkar et al., 2003; Lund et al., 2003). Furthermore, it has been shown that CpG is under-represented in the genomes of small DNA viruses, which could be related to their evasion of the host immune systems (Karlin et al., 1994; Shackelton et al., 2006). Although CpG suppression was also observed in RNA viruses, no known TLR has been shown to recognize CpG of ssRNA. However, recently it has been shown that ssRNA can stimulate human $CD14^{+}CD11c^{+}$ monocytes to produce large amounts of interleukin 12, but this activation of monocytes by CpG oligoribonucleotides was not mediated through TLR3, 7, 8 or 9 (Sugiyama et al., 2005). The results suggested that CpG oligoribonucleotides may stimulate monocytes through a novel mechanism distinct from previously known immunostimulatory nucleic acids. In the present study, we showed that the mean CpG relative abundance in the coronavirus genomes is markedly suppressed (Table 4). This concurs with the results observed in a study on di- and trinucleotide frequencies in nine coronaviruses 10 years ago (Tobler and Ackermann, 1998). The most logical way to avoid CpG is to mutate them to either UpG or CpA. This is in line with the observation that these two dinucleotides are over-represented in the coronavirus genomes, but their deviations from the upper limit of the "normal range" is not as remarkable as that of CpG from the lower limit of the "normal range", as the CpG suppression pressure is equally shared by UpG and CpA over-representation. Interestingly, only CpG containing codons in the context of purine-CpG (ACG and GCG), pyrimidine-CpG (UCG and CCG) and CpG-purine (CGA and CGG) are suppressed (Table 2), whereas CpG-pyrimidine (CGU and CGC) are not. However, when trinucleotide frequencies were analyzed in the 19 coronavirus genomes, all the eight trinucleotides with CpG were suppressed (Fig. 2). This indicates that there is probably another force that has led to an increase use of CGU and CGC as codons for arginine, but this force does not act on trinucleotides over the whole genome in general. This force is probably unrelated to the relative abundance of the corresponding tRNA molecules in the hosts of the coronaviruses, as the pattern of bias in the hosts is not the same as that in the coronaviruses.

In addition to CpG suppression, marked cytosine deamination is also observed in all coronavirus genomes. Although it has been recognized that deamination of cytosine is a significant source of spontaneous mutations for a few decades (Duncan and Miller, 1980), DNA-cytosine deaminases, which are able to attack cytosines in single-stranded DNA, have only been discovered in the recent few years (Bransteitter et al., 2003; Sohail et al., 2003). The discovery of the ability to edit human immunodeficiency virus DNA, and subsequently RNA as well, by the human cytidine deaminase APOBEC3G has allowed the speculation that APOBEC-mediated cytosine deamination may contribute to the sequence variation of RNA viruses that replicate without any DNA intermediates (Bishop et al., 2004). GC skew, which reflects cytosine deamination, has been studied in various coronaviruses, and it has been shown that the GC skews of coronavirus genomes become less pronounced in the one third of the genome that encodes the structural proteins (Grigoriev, 2004; Pyrc et al., 2004). In the present study, using the six amino acids that are only encoded by NNU or NNC, hence excluding most other pressures that may affect the relative abundance of cytosine and uracil, we showed that all these NNU and NNC had usage fractions of >0.700 and <0.300, respectively (Table 2). In fact, for all codons that encode the same amino acid and with either C or U in any position, the usage fraction of the codon that uses U is invariably higher than the one that uses C in all coronaviruses. Furthermore, the percentage of C showed strong inverse relationships with the percentage of U in coronavirus genomes ($r=-0.902$, $P<0.0001$) (Fig. 3). All these suggest that cytosine deamination is an important biochemical force in shaping coronavirus evolution.

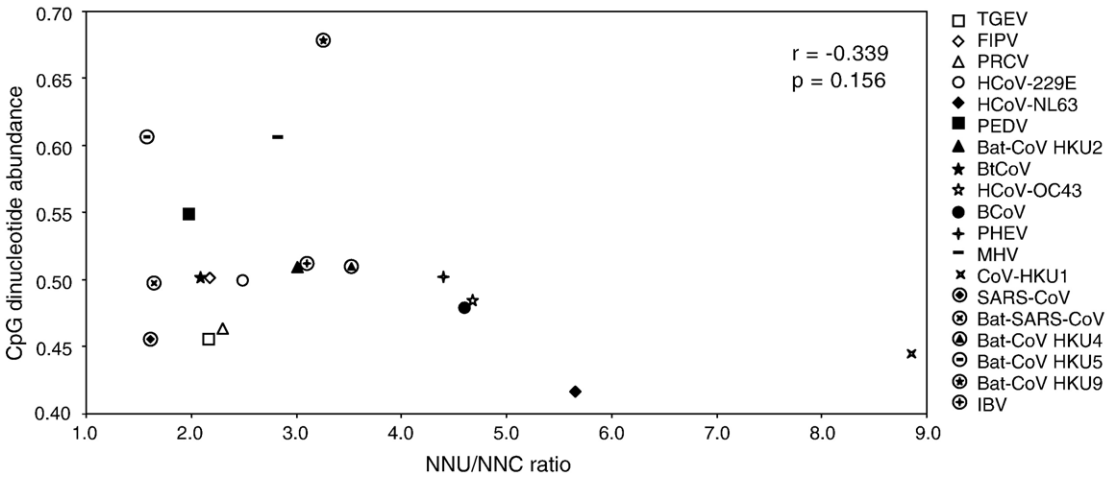

Fig. 1. Correlation between CpG dinucleotide abundance and NNU/NNC ratio in the 19 coronavirus genomes.

Cytosine deamination and selection of CpG suppressed clones by the immune system are the two major independent biochemical and biological selective forces that shape codon usage bias in coronavirus genomes. Codon usage bias in coronaviruses is unrelated to the relative abundance of the corresponding tRNA molecules, as the patterns of bias in codon usage fractions in the hosts are not the same as those in the coronaviruses (Tables 2 and 3). Although others have tried to explain variations in codon usage in coronaviruses by compositional constraints (Gu et al., 2004), we think that both codon usage bias and nucleotide composition of the coronavirus genomes, which are apparently related to each other, are both results of other biological and biochemical selective forces, rather than nucleotide composition as a cause of codon usage bias. On the other hand, most of the codon usage bias in the coronaviruses can be easily explained by CpG suppression and cytosine deamination (Table 2). For asparagine, isoleucine, histidine, aspartic acid, glycine, valine, tyrosine, cysteine and phenylalanine, NNU are used more frequently than NNC because of cytosine deamination. For lysine, glutamine and glutamic acid, NNA are used slightly more frequently than NNG because of cytosine deamination in the minus strand during RNA replication. For threonine, ACG is suppressed because of CpG suppression and ACU is used more frequently than ACC because of cytosine deamination. For arginine, CGA and CGG are suppressed because of CpG suppression and CGU is used more frequently than CGC because of cytosine deamination. AGA is used more frequently than AGG and CGA is used more frequently than CGG because of cytosine deamination in the minus strand during RNA replication. For proline, CCG is suppressed because of CpG suppression and CCU is used more frequently than CCC because of cytosine deamination. For leucine, CUU is used more frequently than CUC, UUA is used more frequently than CUA, and UUG is used more frequently than CUG because of cytosine deamination. For alanine, GCG is suppressed because of CpG suppression and GCU is used more frequently than GCC because of cytosine deamination. For serine, UCG is

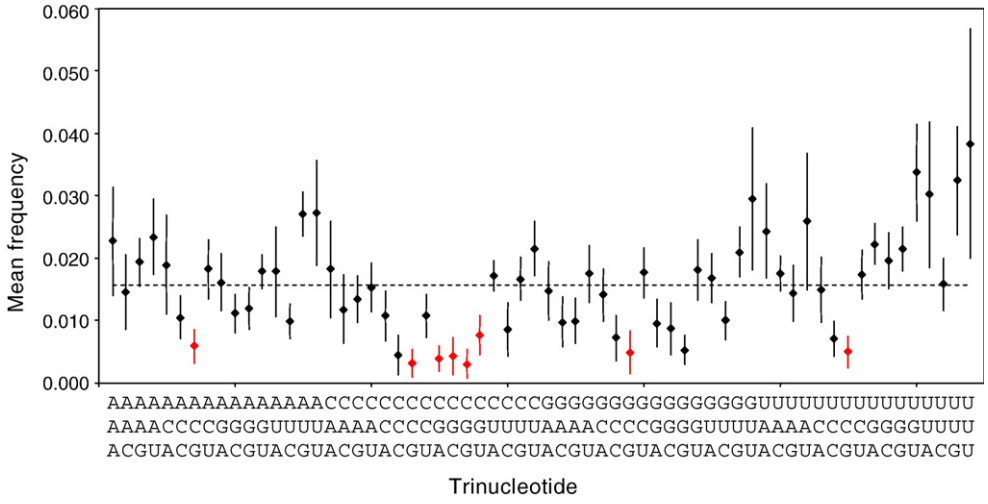

Fig. 2. Mean frequencies of 64 trinucleotides in the 19 coronavirus genomes. The dots and the bars represent the mean frequencies and the 95% confidence intervals of the trinucleotides. The dotted line represents the frequency of each trinucleotide (1/64=0.015625) if the bases are distributed in random. The CpG containing trinucleotides are in red.

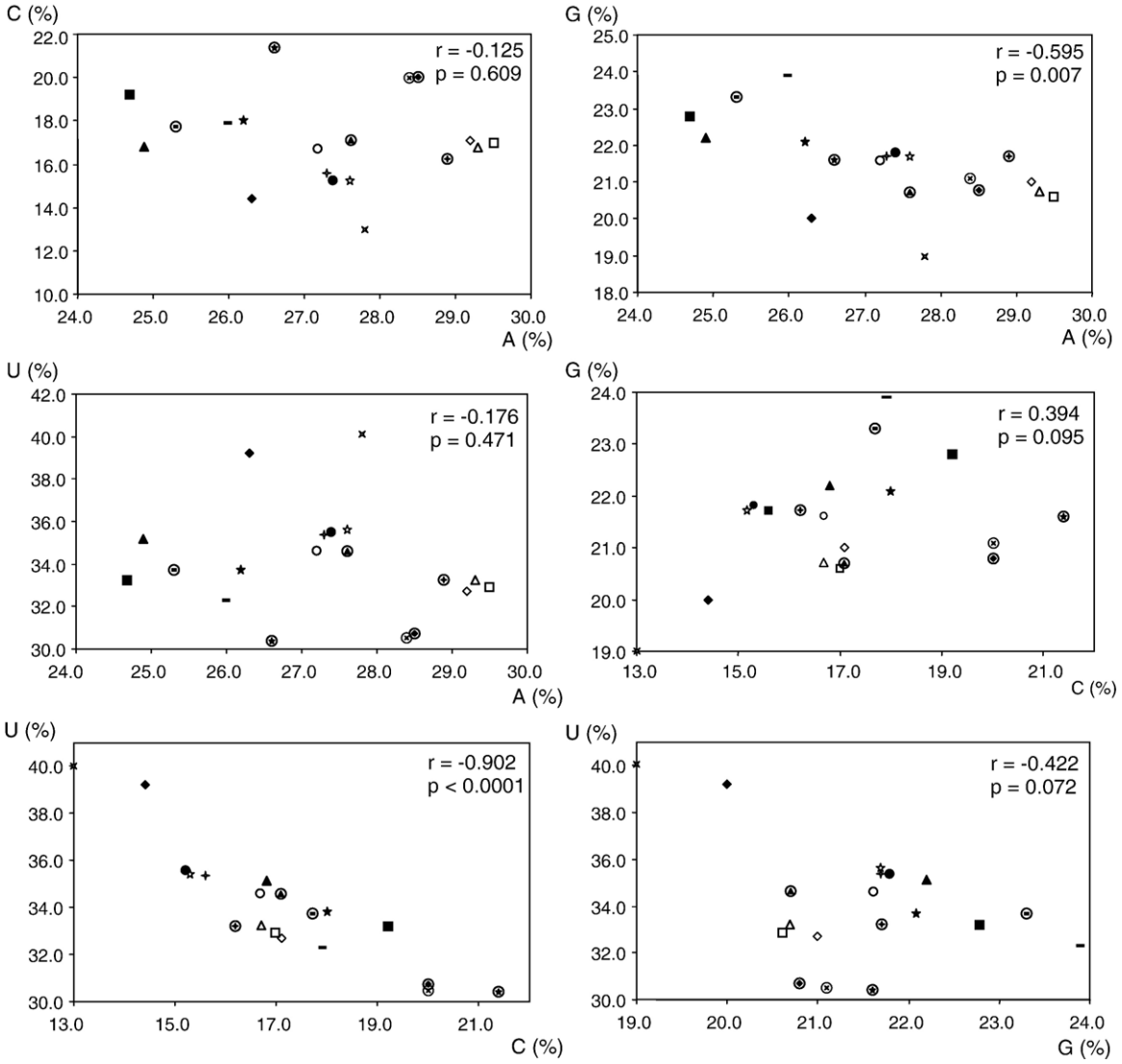

Fig. 3. Correlations among mononucleotide frequencies in the 19 coronavirus genomes. The symbols for the various coronaviruses are the same as those used in Fig. 1.

suppressed because of CpG suppression and UCU is used more frequently than UCC while ACU is used more frequently than ACC because of cytosine deamination. In addition to showing that CpG suppression and cytosine deamination are probably the two most important biological/biochemical forces that shape codon usage bias, we also demonstrated that these two forces are independent (Fig. 1), although cytosine deamination and subsequent selection of CpG suppressed clones by the immune system may be one of the mechanisms that has led to the resultant CpG suppression. Furthermore, we speculate that the species-specific number of CpG containing codons may not simply be the result of mutation pressure to avoid CpG, but an equilibrium between the immune pressure and the required number of CpG containing codons to serve biological functions such as to maintain RNA structure stability. Such an additional factor could explain the mere correlation between the NNU/NNC ratio and CpG dinucleotide abundance.

The underlying mechanism for the extreme codon usage bias, cytosine deamination and G+C content in CoV-HKU1 is enigmatic. The contribution of cytosine deamination to genome evolution varies from very low to very high among the 19 coronavirus genomes. For bat-CoV HKU5, SARS-CoV and bat-SARS-CoV, the mean NNU/NNC ratios are less than 1.7 (Fig. 1). Codon usage bias in these coronaviruses is relatively mild (Nc of 53.23, 49.423 and 49.882, respectively; Table 1), and is mainly due to CpG suppression (Table 2). On the other hand, for CoV-HKU1, the mean NNU/NNC ratio is more than 8.8 (Fig. 1), which is likely a result of rapid cytosine deamination. Although the biochemical basis for this extreme cytosine deamination is not known, this is probably the explanation for the extremely strong codon usage bias in CoV-HKU1 (Nc of 35.671) and its lowest G+C content of 32% among all coronavirus genomes (Table 1).

Materials and methods

Coronavirus and host genomes

One genome sequence of each of the 19 coronavirus species with complete genome sequence available was downloaded

from the GenBank database (Table 1). The genomes of the hosts of the coronaviruses, including those of human, mouse, pig, cat and chicken, were also downloaded.

Codon usage

Codon usage bias was calculated according to the method described by Wright (1990). Using this method, when only one codon is used for each amino acid, Nc for the virus would be 20, and when all codons are used equally, the Nc for the virus would be 61. The codon usage fraction of a particular codon in a genome is calculated by the ratio of the number of that codon to the number of the amino acid that codon and other synonymous codons encode for in the protein coding sequence of the genome. The method for calculating codon usage bias accounting for background nucleotide composition (Nc′) (Novembre, 2002) was not used because it had been proposed to suffer from methodology problems, although those problems did not affect the conclusions which had been drawn by using Nc of this study (Fuglsang, 2006).

Dinucleotide relative abundance in coronavirus genomes

The relative abundance of the dinucleotides in the coronavirus genomes was assessed using the method described by Karlin and Burge (1995). The odds ratio $\rho_{xy}=f_{xy}/f_x f_y$, where f_x denotes the frequency of the nucleotide X and f_{xy} the frequency of the dinucleotide XY, etc., for each dinucleotide were calculated. From data simulations and statistical theory, $\rho_{xy}\leq 0.78$ (extreme under-representation) or $\rho_{xy}\geq 1.23$ (extreme over-representation) occurs for sufficiently long ($\geq$20 kb) random sequences with the probability at most 0.001 for virtually any base composition.

Correlations between CpG suppression and cytosine deamination in coronaviruses

To study possible correlations between CpG suppression and cytosine deamination in coronaviruses, the relative abundance of CpG and the mean ratio of NNC to NNU in the six amino acids (asparagine, histidine, aspartic acid, tyrosine, cysteine and phenylalanine) that only use either NNC or NNU as the codons (NNU/NNC ratio, representing contribution of cytosine deamination) were calculated for all 19 coronavirus genomes. Analysis of correlation between CpG deamination and NNU/NNC ratio was performed using Pearson's correlation (SPSS version 11.0).

Acknowledgments

We are grateful to the generous support of Mr. Hui Hoy and Mr. Hui Ming in the genomic sequencing platform. This work was partly supported by the Research Grant Council Grant; University Development Fund, Outstanding Young Researcher Award, HKU Special Research Achievement Award and The Croucher Senior Medical Research Fellowship, The University of Hong Kong; The Tung Wah Group of Hospitals Fund for Research in Infectious Diseases; the HKSAR Research Fund for the Control of Infectious Diseases of the Health, Welfare and Food Bureau; and the Providence Foundation Limited in memory of the late Dr. Lui Hac Minh.

References

Almazan, F., Gonzalez, J.M., Penzes, Z., Izeta, A., Calvo, E., Plana-Duran, J., Enjuanes, L., 2000. Engineering the largest RNA virus genome as an infectious bacterial artificial chromosome. Proc. Natl. Acad. Sci. U.S.A. 97, 5516–5521.

Ashkar, A.A., Bauer, S., Mitchell, W.J., Vieira, J., Rosenthal, K.L., 2003. Local delivery of CpG oligodeoxynucleotides induces rapid changes in the genital mucosa and inhibits replication, but not entry, of herpes simplex virus type 2. J. Virol. 77, 8948–8956.

Bishop, K.N., Holmes, R.K., Sheehy, A.M., Malim, M.H., 2004. APOBEC-mediated editing of viral RNA. Science 305, 645.

Boursnell, M.E., Brown, T.D., Foulds, I.J., Green, P.F., Tomley, F.M., Binns, M. M., 1987. Completion of the sequence of the genome of the coronavirus avian infectious bronchitis virus. J. Gen. Virol. 68, 57–77.

Bowie, A.G., Haga, I.R., 2005. The role of Toll-like receptors in the host response to viruses. Mol. Immunol. 42, 859–867.

Bransteitter, R., Pham, P., Scharff, M.D., Goodman, M.F., 2003. Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. Proc. Natl. Acad. Sci. U.S.A. 100, 4102–4107.

Brian, D.A., Baric, R.S., 2005. Coronavirus genome structure and replication. Curr. Top. Microbiol. Immunol. 287, 1–30.

Cavanagh, D., Mawditt, K., Welchman Dde, B., Britton, P., Gough, R.E., 2002. Coronaviruses from pheasants (*Phasianus colchicus*) are genetically closely related to coronaviruses of domestic fowl (infectious bronchitis virus) and turkeys. Avian Pathol. 31, 81–93.

Chouljenko, V.N., Lin, X.Q., Storz, J., Kousoulas, K.G., Gorbalenya, A.E., 2001. Comparison of genomic and predicted amino acid sequences of respiratory and enteric bovine coronaviruses isolated from the same animal with fatal shipping pneumonia. J. Gen. Virol. 82, 2927–2933.

Dong, B.Q., Liu, W., Fan, X.H., Vijaykrishna, D., Tang, X.C., Gao, F., Li, L.F., Li, G.J., Zhang, J.X., Yang, L.Q., Poon, L.L., Zhang, S.Y., Peiris, J.S., Smith, G.J., Chen, H., Guan, Y., 2007. Detection of a novel and highly divergent coronavirus from Asian leopard cats and Chinese ferret badgers in southern china. J. Virol. 81, 6920–6926.

Duncan, B.K., Miller, J.H., 1980. Mutagenic deamination of cytosine residues in DNA. Nature 287, 560–561.

East, M.L., Moestl, K., Benetka, V., Pitra, C., Honer, O.P., Wachter, B., Hofer, H., 2004. Coronavirus infection of spotted hyenas in the Serengeti ecosystem. Vet. Microbiol. 102, 1–9.

Fouchier, R.A., Hartwig, N.G., Bestebroer, T.M., Niemeyer, B., de Jong, J.C., Simon, J.H., Osterhaus, A.D., 2004. A previously undescribed coronavirus associated with respiratory disease in humans. Proc. Natl. Acad. Sci. U.S.A. 101, 6212–6216.

Fuglsang, A., 2006. Accounting for background nucleotide composition when measuring codon usage bias: brilliant idea, difficult in practice. Mol. Biol. Evol. 23, 1345–1347.

Grigoriev, A., 2004. Mutational patterns correlate with genome organization in SARS and other coronaviruses. Trends Genet. 20, 131–135.

Gu, W., Zhou, T., Ma, J., Sun, X., Lu, Z., 2004. Analysis of synonymous codon usage in SARS coronavirus and other virus in the Nidovirales. Virus Res. 101, 155–161.

Guan, Y., Zheng, B.J., He, Y.Q., Liu, X.L., Zhuang, Z.X., Cheung, C.L., Luo, S.W., Li, P.H., Zhang, L.J., Guan, Y.J., Butt, K.M., Wong, K.L., Chan, K.W., Lim, W., Shortridge, K.F., Yuen, K.Y., Peiris, J.S., Poon, L.L., 2003. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 302, 276–278.

Haijema, B.J., Volders, H., Rottier, P.J., 2003. Switching species tropism: an effective way to manipulate the feline coronavirus genome. J. Virol. 77, 4528–4538.

Hasoksuz, M., Alekseev, K., Vlasova, A., Zhang, X., Spiro, D., Halpin, R., Wang, S., Ghedin, E., Saif, L.J., 2007. Biologic, antigenic, and full-length genomic characterization of a bovine-like coronavirus isolated from a giraffe. J, Virol. 81, 4981–4990.

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

Jonassen, C.M., Kofstad, T., Larsen, I.L., Lovland, A., Handeland, K., Follestad, A., Lillehaug, A., 2005. Molecular identification and characterization of novel coronaviruses infecting graylag geese (*Anser anser*), feral pigeons (*Columbia livia*) and mallards (*Anas platyrhynchos*). J. Gen. Virol. 86, 1597–1607.

Karlin, S., Burge, C., 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11, 283–290.

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

Kocherhans, R., Bridgen, A., Ackermann, M., Tobler, K., 2001. Completion of the porcine epidemic diarrhoea coronavirus (PEDV) genome sequence. Virus Genes 23, 137–144.

Lai, M.M., Cavanagh, D., 1997. The molecular biology of coronaviruses. Adv. Virus. Res. 48, 1–100.

Lau, S.K., Woo, P.C., Li, K.S., Huang, Y., Tsoi, H.W., Wong, B.H., Wong, S.S., Leung, S.Y., Chan, K.H., Yuen, K.Y., 2005. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. Proc. Natl. Acad. Sci. U.S.A. 102, 14040–14045.

Lau, S.K., Woo, P.C., Yip, C.C., Tse, H., Tsoi, H.W., Cheng, V.C., Lee, P., Tang, B.S., Cheung, C.H., Lee, R.A., So, L.Y., Lau, Y.L., Chan, K.H., Yuen, K.Y., 2006. Coronavirus HKU1 and other coronavirus infections in Hong Kong. J. Clin. Microbiol. 44, 2063–2071.

Lau, S.K., Woo, P.C., Li, K.S., Huang, Y., Wang, M., Lam, C.S., Xu, H., Guo, R., Chan, K.H., Zheng, B.J., Yuen, K.Y., 2007. Complete genome sequence of bat coronavirus HKU2 from Chinese horseshoe bats revealed a much smaller spike gene with a different evolutionary lineage from the rest of the genome. Virology [Epub ahead of print].

Leparc-Goffart, I., Hingley, S.T., Chua, M.M., Jiang, X., Lavi, E., Weiss, S.R., 1997. Altered pathogenesis of a mutant of the murine coronavirus MHV-A59 is associated with a Q159L amino acid substitution in the spike protein. Virology 239, 1–10.

Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J.H., Wang, H., Crameri, G., Hu, Z., Zhang, H., Zhang, J., McEachern, J., Field, H., Daszak, P., Eaton, B. T., Zhang, S., Wang, L.F., 2005. Bats are natural reservoirs of SARS-like coronaviruses. Science 310, 676–679.

Liu, S., Chen, J., Chen, J., Kong, X., Shao, Y., Han, Z., Feng, L., Cai, X., Gu, S., Liu, M., 2005. Isolation of avian infectious bronchitis coronavirus from domestic peafowl (*Pavo cristatus*) and teal (*Anas*). J. Gen. Virol. 86, 719–725.

Lund, J., Sato, A., Akira, S., Medzhitov, R., Iwasaki, A., 2003. Toll-like receptor 9 mediated recognition of herpes simplex virus-2 by plasmacytoid dendritic cells. J. Exp. Med. 198, 513–520.

Marra, M.A., Jones, S.J., Astell, C.R., Holt, R.A., Brooks-Wilson, A., Butterfield, Y.S., Khattra, J., Asano, J.K., Barber, S.A., Chan, S.Y., Cloutier, A., Coughlin, S.M., Freeman, D., Girn, N., Griffith, O.L., Leach, S.R., Mayo, M., McDonald, H., Montgomery, S.B., Pandoh, P.K., Petrescu, A.S., Robertson, A.G., Schein, J.E., Siddiqui, A., Smailus, D.E., Stott, J.M., Yang, G.S., Plummer, F., Andonov, A., Artsob, H., Bastien, N., Bernard, K., Booth, T.F., Bowness, D., Czub, M., Drebot, M., Fernando, L., Flick, R., Garbutt, M., Gray, M., Grolla, A., Jones, S., Feldmann, H., Meyers, A., Kabani, A., Li, Y., Normand, S., Stroher, U., Tipples, G.A., Tyler, S., Vogrig, R., Ward, D., Watson, B., Brunham, R.C., Krajden, M., Petric, M., Skowronski, D.M., Upton, C., Roper, R.L., 2003. The genome sequence of the SARS-associated coronavirus. Science 300, 1399–1404.

Novembre, J.A., 2002. Accounting for background nucleotide composition when measuring codon usage bias. Mol. Biol. Evol. 8, 1390–1394.

Peiris, J.S., Lai, S.T., Poon, L.L., Guan, Y., Yam, L.Y., Lim, W., Nicholls, J., Yee, W.K., Yan, W.W., Cheung, M.T., Cheng, V.C., Chan, K.H., Tsang, D. N., Yung, R.W., Ng, T.K., Yuen, K.Y., 2003. Coronavirus as a possible cause of severe acute respiratory syndrome. Lancet 361, 1319–1325.

Poon, L.L., Chu, D.K., Chan, K.H., Wong, O.K., Ellis, T.M., Leung, Y.H., Lau, S.K., Woo, P.C., Suen, K.Y., Yuen, K.Y., Guan, Y., Peiris, J.S., 2005. Identification of a novel coronavirus in bats. J. Virol. 79, 2001–2009.

Pyrc, K., Jebbink, M.F., Berkhout, B., van der Hoek, L., 2004. Genome structure and transcriptional regulation of human coronavirus NL63. Virol. J. 17, 1–7.

Rota, P.A., Oberste, M.S., Monroe, S.S., Nix, W.A., Campagnoli, R., Icenogle, J.P., Penaranda, S., Bankamp, B., Maher, K., Chen, M.H., Tong, S., Tamin, A., Lowe, L., Frace, M., DeRisi, J.L., Chen, Q., Wang, D., Erdman, D.D., Peret, T.C., Burns, C., Ksiazek, T.G., Rollin, P.E., Sanchez, A., Liffick, S., Holloway, B., Limor, J., McCaustland, K., Olsen-Rasmussen, M., Fouchier, R., Gunther, S., Osterhaus, A.D., Drosten, C., Pallansch, M.A., Anderson, L.J., Bellini, W.J., 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science 300, 1394–1399.

Sewatanon, J., Srichatrapimuk, S., Auewarakul, P., 2007. Compositional bias and size of genomes of human DNA viruses. Intervirology 50, 123–132.

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

Snijder, E.J., Bredenbeek, P.J., Dobbe, J.C., Thiel, V., Ziebuhr, J., Poon, L.L., Guan, Y., Rozanov, M., Spaan, W.J., Gorbalenya, A.E., 2003. Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. J. Mol. Biol. 331, 991–1004.

Sohail, A., Klapacz, J., Samaranayake, M., Ullah, A., Bhagwat, A.S., 2003. Human activation-induced cytidine deaminase causes transcription-dependent, strand-biased C to U deaminations. Nucleic Acids Res. 31, 2990–2994.

Sugiyama, T., Gursel, M., Takeshita, F., Coban, C., Conover, J., Kaisho, T., Akira, S., Klinman, D.M., Ishii, K.J., 2005. CpG RNA: identification of novel single-stranded RNA that stimulates human $CD14^{+}CD11c^{+}$ monocytes. J. Immunol. 174, 2273–2279.

Tang, X.C., Zhang, J.X., Zhang, S.Y., Wang, P., Fan, X.H., Li, L.F., Li, G., Dong, B.Q., Liu, W., Cheung, C.L., Xu, K.M., Song, W.J., Vijaykrishna, D., Poon, L.L., Peiris, J.S., Smith, G.J., Chen, H., Guan, Y., 2006. Prevalence and genetic diversity of coronaviruses in bats from China. J. Virol. 80, 7481–7490.

Thiel, V., Herold, J., Schelle, B., Siddell, S.G., 2001. Infectious RNA transcribed in vitro from a cDNA copy of the human coronavirus genome cloned in vaccinia virus. J. Gen. Virol. 82, 1273–1281.

Tobler, K., Ackermann, M., 1998. Comparison of the di- and trinucleotide frequencies from the genomes of nine different coronaviruses. Adv. Exp. Med. Biol. 440, 801–804.

van der Hoek, L., Pyrc, K., Jebbink, M.F., Vermeulen-Oost, W., Berkhout, R.J., Wolthers, K.C., Wertheim-van Dillen, P.M., Kaandorp, J., Spaargaren, J., Berkhout, B., 2004. Identification of a new human coronavirus. Nat. Med. 10, 368–373.

van Hemert, F.J., Berkhout, B., Lukashov, V.V., 2007. Host-related nucleotide composition and codon usage as driving forces in the recent evolution of the *Astroviridae*. Virology 361, 447–454.

Vijgen, L., Keyaerts, E., Moes, E., Thoelen, I., Wollants, E., Lemey, P., Vandamme, A.M., Van Ranst, M., 2005. Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. J. Virol. 79, 1595–1604.

Vijgen, L., Keyaerts, E., Lemey, P., Maes, P., Van Reeth, K., Nauwynck, H., Pensaert, M., Van Ranst, M., 2006. Evolutionary history of the closely related group 2 coronaviruses: porcine hemagglutinating encephalomyelitis virus, bovine coronavirus, and human coronavirus OC43. J. Virol. 80, 7270–7274.

Woo, P.C., Lau, S.K., Tsoi, H.W., Chan, K.H., Wong, B.H., Che, X.Y., Tam, V.K., Tam, S.C., Cheng, V.C., Hung, I.F., Wong, S.S., Zheng, B.J., Guan, Y., Yuen, K.Y., 2004. Relative rates of non-pneumonic SARS coronavirus infection and SARS coronavirus pneumonia. Lancet 363, 841–845.

Woo, P.C., Huang, Y., Lau, S.K., Tsoi, H.W., Yuen, K.Y., 2005a. In silico analysis of ORF1ab in coronavirus HKU1 genome reveals a unique putative cleavage site of coronavirus HKU1 3C-like protease. Microbiol. Immunol. 49, 899–908.

Woo, P.C., Lau, S.K., Chu, C.M., Chan, K.H., Tsoi, H.W., Huang, Y., Wong, B.H., Poon, R.W., Cai, J.J., Luk, W.K., Poon, L.L., Wong, S.S., Guan, Y., Peiris, J.S., Yuen, K.Y., 2005b. Characterization and complete genome

sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. J. Virol. 79, 884–895.

Woo, P.C., Lau, S.K., Tsoi, H.W., Huang, Y., Poon, R.W., Chu, C.M., Lee, R.A., Luk, W.K., Wong, G.K., Wong, B.H., Cheng, V.C., Tang, B.S., Wu, A.K., Yung, R.W., Chen, H., Guan, Y., Chan, K.H., Yuen, K.Y., 2005c. Clinical and molecular epidemiological features of coronavirus HKU1-associated community-acquired pneumonia. J. Infect. Dis. 192, 1898–1907.

Woo, P.C., Lau, S.K., Li, K.S., Poon, R.W., Wong, B.H., Tsoi, H.W., Yip, B.C., Huang, Y., Chan, K.H., Yuen, K.Y., 2006a. Molecular diversity of coronaviruses in bats. Virology 351, 180–187.

Woo, P.C., Lau, S.K., Yip, C.C., Huang, Y., Tsoi, H.W., Chan, K.H., Yuen, K.Y., 2006b. Comparative analysis of 22 coronavirus HKU1 genomes reveals a novel genotype and evidence of natural recombination in coronavirus HKU1. J. Virol. 80, 7136–7145.

Woo, P.C., Wang, M., Lau, S.K., Xu, H., Poon, R.W., Guo, R., Wong, B.H., Gao, K., Tsoi, H.W., Huang, Y., Li, K.S., Lam, C.S., Chan, K.H., Zheng, B.J., Yuen, K.Y., 2007. Comparative analysis of 12 genomes of three novel group 2c and group 2d coronaviruses reveals unique group and subgroup features. J. Virol. 81, 1574–1585.

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

Yap, Y.L., Zhang, X.W., Danchin, A., 2003. Relationship of SARS-CoV to other pathogenic RNA viruses explored by tetranucleotide usage profiling. BMC Bioinform. 4, 43.

Zhang, X., Hasoksuz, M., Spiro, D., Halpin, R., Wang, S., Stollar, S., Janies, D., Hadya, N., Tang, Y., Ghedin, E., Saif, L., 2007. Complete genomic sequences, a key residue in the spike protein and deletions in nonstructural protein 3b of US strains of the virulent and attenuated coronaviruses, transmissible gastroenteritis virus and porcine respiratory coronavirus. Virology 358, 424–435.

Ziebuhr, J., 2004. Molecular biology of severe acute respiratory syndrome coronavirus. Curr. Opin. Microbiol. 7, 412–419.