EXPERIMENTAL and MOLECULAR MEDICINE, Vol. 41, No. 10, 746-756, October 2009

Comparative study of synonymous codon usage variations between the nucleocapsid and spike genes of coronavirus, and C-type lectin domain genes of human and mouse

Insung Ahn[1]*, Byeong-Jin Jeong[2,3]* and Hyeon Seok Son[2,3,4]

[1]Supercomputing Center
Korea Institute of Science and Technology Information
Daejon 305-806, Korea
[2]Laboratory of Computational Biology and Bioinformatics
Institute of Health and Environment
Graduate School of Public Health
Seoul National University
Seoul 110-799, Korea
[3]Interdisciplinary Graduate Program in Bioinformatics
College of Natural Science
Seoul National University
Seoul 151-742, Korea
[4]Corresponding author: Tel, 82-2-740-8864;
Fax, 82-2-762-9105; E-mail, hss2003@snu.ac.kr
*These authors contributed equally to this work.
DOI 10.3858/emm.2009.41.10.081

Accepted 16 June 2009

Abbreviations: CA, correspondence analysis; CoV, coronavirus; CTLD, C-type lectin domain; ENC, effective number of codons; RSCU, relative synonymous codon usage; SARS, severe acute respiratory syndrome

Abstract

Coronaviruses (CoVs) are single-stranded RNA viruses which contain the largest RNA genomes, and severe acute respiratory syndrome coronavirus (SARS-CoV), a newly found group 2 CoV, emerged as infectious disease with high mortality rate. In this study, we compared the synonymous codon usage patterns between the nucleocapsid and spike genes of CoVs, and C-type lectin domain (CTLD) genes of human and mouse on the codon basis. Findings indicate that the nucleocapsid genes of CoVs were affected from the synonymous codon usage bias than spike genes, and the CTLDs of human and mouse partially overlapped with the nucleocapsid genes of CoVs. In addition, we observed that CTLDs which showed the similar relative synonymous codon usage (RSCU) patterns with CoVs were commonly derived from the human chromosome 12, and mouse chromosome 6 and 12, suggesting that there might be a specific genomic region or chromosomes which show a more similar synonymous codon usage pattern with viral genes. Our findings contribute to developing the codon-optimization method in DNA vaccines, and further study is needed to determine a specific correlation between the codon usage patterns and the chromosomal locations in higher organisms.

Keywords: coronavirus; host-pathogen interactions; lectins, C-type

Introduction

Coronaviruses (CoVs) which are included in the family Coronaviridae are enveloped and contain the largest RNA genomes with some reaching almost 30,000 nucleotides (Dimmock *et al*., 2002). They primarily infect the upper respiratory and gastrointestinal tract of animals, and severe acute respiratory syndrome coronavirus (SARS-CoV), a newly emerged group 2 CoV, spread rapidly from Asia to North America and Europe with a high degree of transmissibility and mortality (Lew *et al*., 2003; Riley *et al*., 2003; Friman *et al*., 2008). In response to the SARS pandemic in 2003, many scientists have been interested in vaccine development against SARS-CoV. The phase I human study for a SARS DNA vaccine was reported by Martin and his colleagues showing immunogenicity with spike proteins of SARS-CoV in all subjects and neutralizing antibody responses in 8 of 10 subjects (Yang *et al*., 2005; Martin *et al*., 2008).

The DNA vaccination represents a new strategy for highly pathogenic and infectious diseases (Ramakrishna *et al*., 2004; Martin *et al*., 2006, 2007; Wang *et al*., 2006a, 2006b; Catanzaro *et al*., 2007), and it is usually produced in three successive steps. First, the primers specific to the target regions of viral genome are produced to generate cDNA fragments. Second, these cDNA fragments are inserted into a bacterial DNA vaccine plasmid such as *escherichia coli* plasmid, and lately, the prepared DNA vaccine is injected into the cells of the target organisms such as mouse, rabbit or human subjects to produce one or more specific proteins by mimicking viral replication and protein produc-

tion in the host. Because these proteins are recognised as foreign antigens in the target organisms, immune responses are triggered by them (Sin and Weiner, 2000; Donnelly *et al*., 2003). The phase I clinical trial of DNA vaccines against West Nile virus, Ebola virus and human immunodeficiency virus type 1 (HIV-1) in healthy adults have already been performed (Martin *et al*., 2006, 2007; Catanzaro *et al*., 2007). According to Wang *et al*. (2006a, 2006b) and Ramakrishna *et al*. (2004), the codon optimization of the Tat and envelope genes of HIV-1 as well as hemagglutinin genes of influenza A virus showed better antigen expression and immunogenicity in model animals such as mouse and rabbit. Each target gene of HIV-1 and influenza A virus was known to be changed to the preferred codons of the overall mammalian system to promote better expression of each encoded protein.

Synonymous codons usually encode common amino acids in protein synthesis, and they are not used randomly, with some codons being used more frequently than others (Moriyama and Hartl, 1993; McInerney, 1998; Duret, 2002; Lynn *et al*., 2002; Kawabe and Miyashita, 2003; Singer and Hickey, 2003). Codon usage bias has been known to mirror tRNA abundance in the early studies using *Bacillus subtilis* and *Caenorhabditis elegans* genes (Shields and Sharp, 1987; Stenico *et al*., 1994). In prokaryotes, such as thermophilic bacteria,

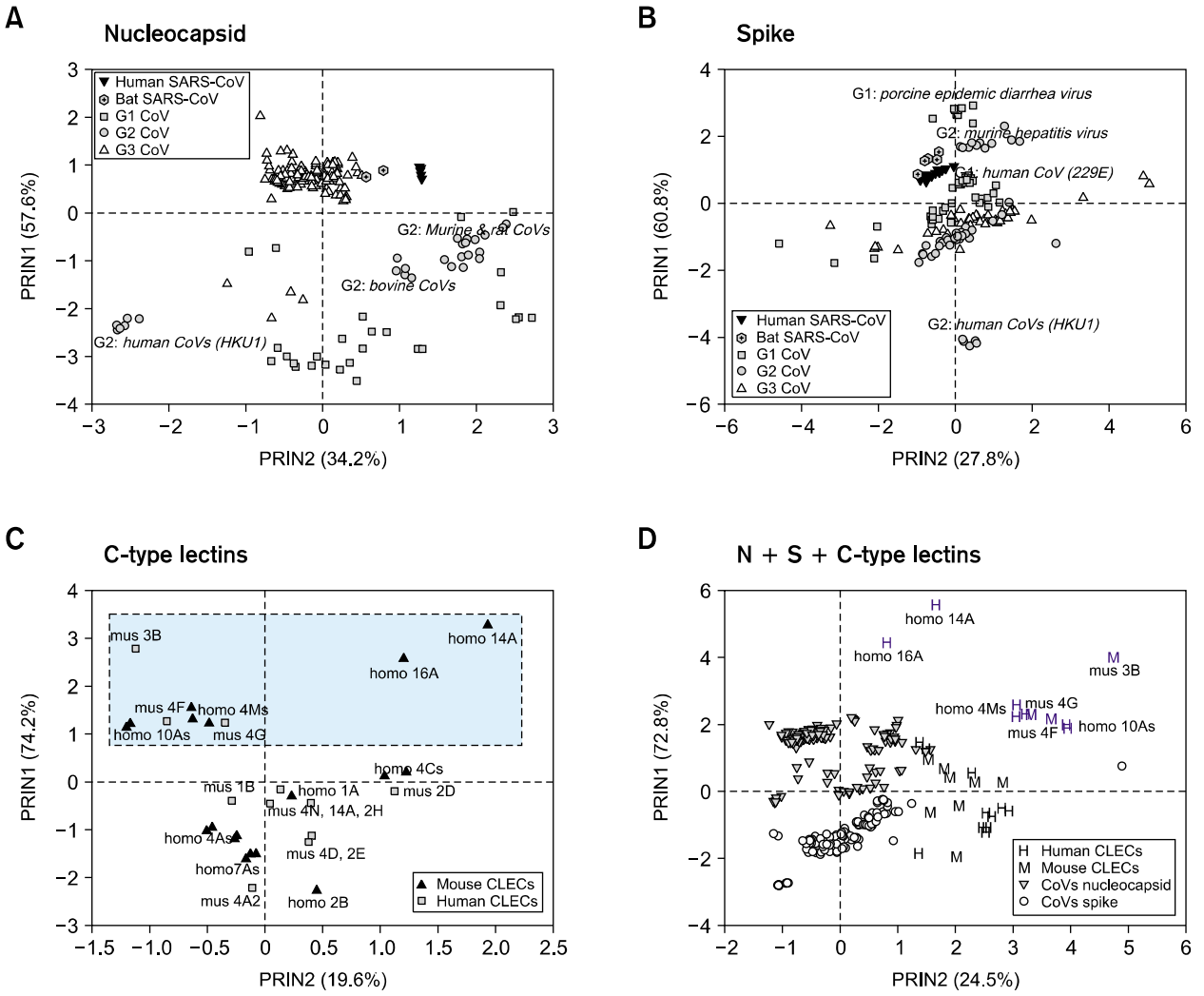

Figure 1. Principal component analysis of the % GC contents on the 1st, 2nd and 3rd codon position. The first two factors from the principal component analysis (PRIN1 and PRIN2) were presented with each eigenvalue proportion. Nucleocapsid (A) and spike (B) coding genes of *Coronavirus* genus were compared with CTLD genes of human (*homo sapiens*) and mouse (*mus musculus*) species (C, D). Family names of CTLDs were also presented with each plot. G1, Group 1 CoV; G2, Group 2 CoV; G3, Group 3 CoV; N, nucleocapsid gene of CoV; S, spike gene of CoV; homo, *homo sapiens; mus, mus musculus*.

highly expressed genes shift their codon usage toward a more restricted set of preferred synonymous codons compared to less highly expressed genes within the genome (Lynn *et al*., 2002; Singer and Hickey, 2003). As for the viral genomes, Gu *et al*. (2004) reported that the relative synonymous codon usage (RSCU) values of *Nidovirales* family including SARS-CoV are virus-specific, and translational selection and gene length may not affect the codon usage pattern in some viruses. However, Jenkins and Holmes (2003) who analyzed the extent of codon usage bias in the complete genomic coding region of 50 genetically and ecologically diverse human RNA viruses using the effective number of codon (ENC) as a parameter showed that the overall extent of codon usage bias was low and that there was little variation in bias between genes. More recently, Shackelton *et al*. (2006) reported that there was a striking difference in CpG content between DNA virus with large and small genomes as the majority of large genome viruses show the expected frequency of CpG, while most small genome viruses had CpG contents far below expected values. They suggested that the main reason for these differences might be due to the differences in the viral replication and repairing mechanisms, such as cellular or viral replicative machinery. In our previous study, synonymous codon usage patterns among RNA viruses such as influenza A viruses and HIV-1s were divided into each region, subtype, host or occurring-year group, with an expectation that there might be some correlations between the nucleotide patterns and the direction of viral variations on the codon basis (Ahn and Son, 2006, 2007; Ahn *et al*., 2006). Furthermore, van Hemert *et al*. (2007) reported that the recent evolution of astroviruses was associated with a switch in nucleotide composition and codon usage among non-human mammalian versus human/avian astroviruses. They suggested that evolutionary events within a virus family might be driven by forces operational at the level of synonymous substitutions, such as nucleotide composition, translational selection, and codon usage.

In this study, we hypothesized that the codon usage bias of viral genes might tend to mimic the specific genes and perform a key role during the initial immune responses, in their host species. C-type lectins, a supefamily of proteins containing C-type lectin domains (CTLDs), are a large group of extracellular Metazoan proteins with diverse functions (Zelensky and Gready, 2005). They usually provide Ca^{2+}-dependent sugar-recognition activity and initiate a various kinds of biological processes, such as adhesion, endocytosis, and pathogen neutralization (Drickamer and Dodd, 1999; Dodd and Drickamer, 2001). As a point of immune responses, C-type lectins are also known to perform an important function in dendritic cell (DC) immune regulations, which include the triggering of inflammatory cytokines, as well as delivering antigens to T cell to initiate the specific immune response (Cella *et al*., 1997, 1999). C-type lectin receptors in DCs have been determined to act as a capture of attachment factor for influenza A virus (H5N1 subtype) or HIV-1 (Lambert *et al*., 2008; Wang *et al*., 2008), and SARS-CoV infection is also known to induce a immune responses related with DC functions such as delaying an activation of alpha interferon (Spiegel *et al*., 2006). In this study, we compared the synonymous codon usage patterns of *Coronavirus* genus with the CTLD genes of human (*homo sapiens*) and mouse (*mus musculus*) to investigate the possible relations between microbes and their host species in codon basis.

Table 1. Eigenvectors and eigenvalues of the principal component analysis using the % GC contents on each codon position.

Species	Variances	Eigenvectors	
		PRIN1	PRIN2
Coronavirus (Nucleocapsid)	GC_{1st}	0.700134	0.188262
	GC_{2nd}	0.710481	-0.087886
	GC_{3rd}	-0.070915	**0.978179**
	Eigenvalue %	57.6	34.2
Coronavirus (Spike)	GC_{1st}	0.628461	- 0.383325
	GC_{2nd}	0.419316	**0.899838**
	GC_{3rd}	0.655142	- 0.208216
	Eigenvalue %	60.8	27.8
Human + Mouse (C-type lectin domain genes)	GC_{1st}	0.629533	- 0.110096
	GC_{2nd}	0.525960	**0.789002**
	GC_{3rd}	0.571887	- 0.604445
	Eigenvalue %	74.2	19.6
Coronavirus+ Human + Mouse (Nucleocapsid, Spike+C-type lectin domain genes)	GC_{1st}	0.645441	-0.254191
	GC_{2nd}	0.630553	-0.354859
	GC_{3rd}	0.431057	**0.899701**
	Eigenvalue %	72.8	24.5

Results

Principal component analysis using the % GC contents on the 1st, 2nd and 3rd codon positions

The first two principal factors of the % GC contents on each codon position from the nucleocapsid and spike genes of CoVs as well as the CTLDs of human and mouse were investigated using the principal component analysis (Figure 1). Eigen-

vectors of each principal factor (PRIN1 and PRIN2) and eigenvalue proportions (%) were presented in Table 1. Among the CoV genes, the first two principal components of nucleocapsd genes accounted for 57.6% and 34.2%, whereas those of spike genes accounted for 60.8% and 27.8% of the total variance of the data set, respectively (Figure 1A and 1B). Eigenvector compositions of those two genes, however, showed different patterns. The % GC contents on the third codon position (GC_{3rd}) among nucleocapsid genes showed highly positive correlations (0.978) along with PRIN2-axis, whereas

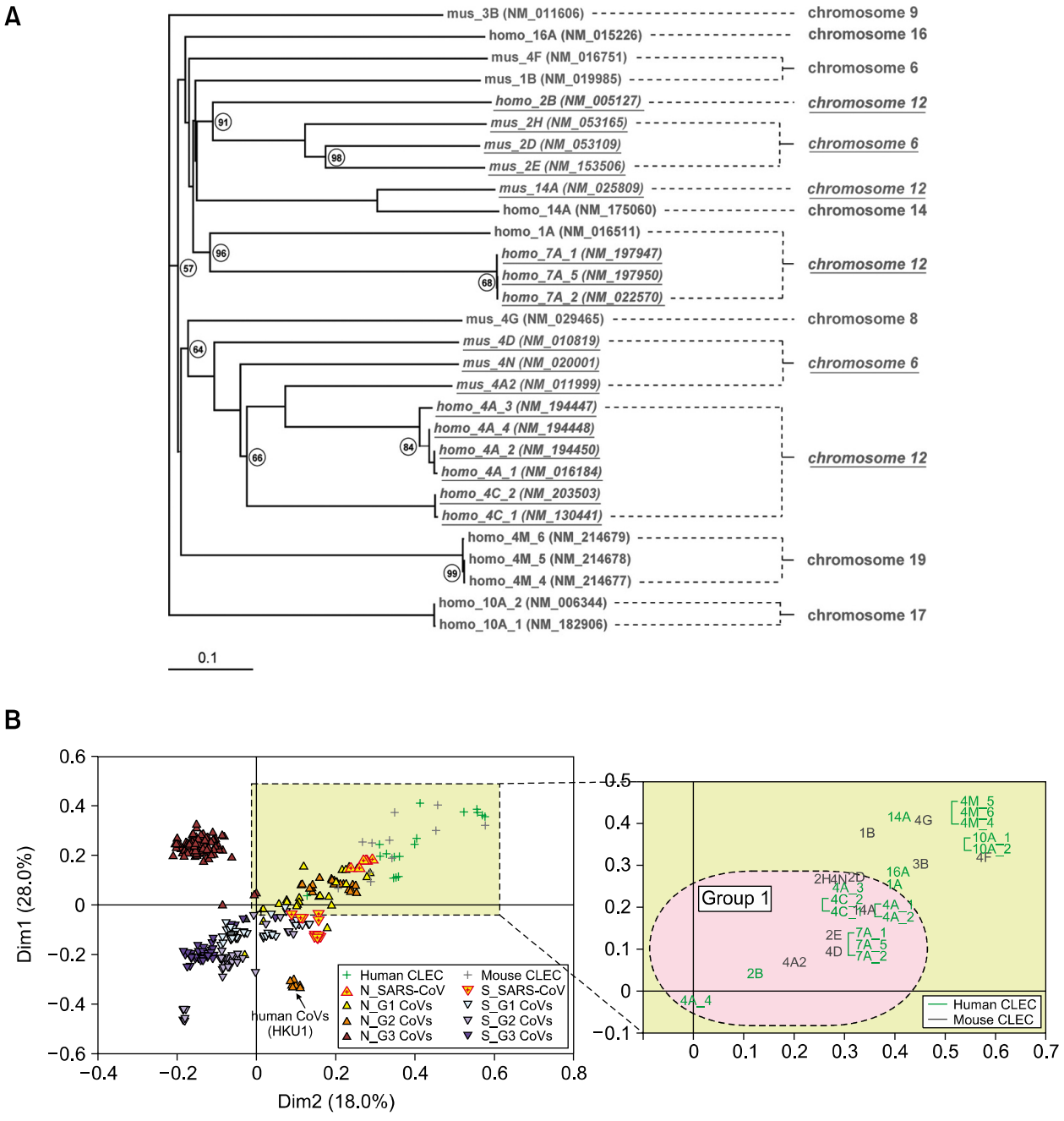

Figure 2. The results of phylogenetic analysis using the CTLD genes of human (*homo sapiens*) and mouse (*mus musculus*) species (A), and the scatter plots of the correspondence analysis using the relative synonymous codon usage values of the nucleocapsid and spike genes of CoVs as well as the CTLDs of human and mouse (B). Phylogram was derived by Neighbor-Joining method with bootstrap analysis of 1000 iterations, and bootstrap values (%) that are not 100% are represented as circulated numbers in each node. Each chromosome source of CTLD was also presented on the right column of tree. G1, Group 1 CoV; G2, Group 2 CoV; G3, Group 3 CoV; N, nucleocapsid gene of CoV; S, spike gene of CoV; CLEC, C-type lectin domain gene.

PRIN2 of spike genes was strongly dependent on the GC_{2nd} (0.899) (Table 1). The former pattern was also appeared when two genes from CoVs, and CTLDs from human and mouse were analyzed together (Figure 1D, Table 1), whereas CTLD gene itself revealed very similar eigenvector patterns with those of spike genes of CoVs (Figure 1C, Table 1). The eigenvectors of PRIN1s in all the cases commonly showed positive correlation with on the GC_{1st} and GC_{2nd}, with showing that PRIN1s were mainly dependent on the non-synonymous codon usage patterns of each gene.

In Figure 1, we categorized the nucleocapsid and spike genes into each CoV groups such as group 1, 2 and 3, which were presented as G1, G2 and G3, respectively. The nucleocapsid genes of both human and bat SARS-CoVs were located closely to the G3 CoVs such as infectious bronchitis virus along the PRIN1, but human SARS-CoVs displayed similar patterns with the G2 CoVs such as bovine, murine and rat CoVs along with the PRIN2 (Figure 1A). On the other hand, spike genes of SARS-CoVs were located near the G2 murine hepatitis virus CoVs as well as G3 human CoV 229E (Figure 1B). Human CoV (HKU1) in G2 CoVs were distinctly located from other G2 CoVs in both nucleocapsid and spike genes. As for the CTLD genes, mouse and human genes spread broadly across the biplots, and they were not clearly separated each other (Figure 1C). On the basis of the PRIN1-axis, the family 4Ms, 10As, 16A and 14A of human CTLDs as well as 3B, 4F and 4G of mouse CTLDs were positively biased, and those genes also showed different % GC contents from CoV genes in Figure 1D.

Phylogenetic relationships among CTLD genes of human and mouse species

To compare the phylogenetic relationships among human and mouse CTLDs with the synonymous codon usage patterns, we constructed a phylogenetic tree using the Neighbour-Joining method with 1000 times bootstrapping test. All the genes were well grouped into each CTLD family, and human 7As, 1A, 14A, 2B and 16A CTLDs were separately located with other human CTLD genes, showing closer relationships with mouse genes (Figure 2A). Among the mouse genes, family 4A2, 4N, 4D and 4G CTLDs showed close relationships with human family 4As, 4Cs, and 4Ms. Family 10As of human genes were distinctly located from other families. On the basis of the chromosomes which each gene was transcribed from, mouse CTLDs were encoded from chromosome 6, 8, 9 and 12, and human genes were from chromosome 12, 14, 16, 17 and 19.

Synonymous codon usage analysis using the CA method

To investigate the synonymous codon usage patterns, we parsed each nucleotide sequence into each synonymous codon groups first, then, calculated the RSCU values per each sequence. After that, we assigned each kind of gene or species as rows, and RSCU values of 59 codons as columns for the CA. All the target sequences of CoVs, human and mouse species were analyzed together to compare the overall synonymous codon usage patterns (Figure 2B). First of all, the nucleocapsid and spike genes of CoVs showed opposite patterns along with the first dimensional factor (Dim1) of CA plots, and human and mouse CTLDs were located on the same side with nucleocapsid genes. Secondly, we also performed linear regression analysis to identify which codon usage parameters affect the Dim1 and Dim2 of CA result most (Table 2). The Dim1 showed the significant correlations with all the codon usage parameters such as GC_{1st}, GC_{2nd} and GC_{3rd} and ENCs. Among the % GC contents, Dim1 was strongly dependent on the GC_{1st} and GC_{2nd}, showing R^2 values of 0.781 and 0.671, respectively. As for the Dim2, however, only GC_{3rd} showed positive correlations (R^2=0.500) among all the % GC contents.

In Figure 2B, we also presented the enlarged region of CTLDs of both human and mouse species with each family, member and transcript variant (if exists) name 1 (Figure 2B right). The CTLD genes which were located within or near the CA plots of CoVs were clustered as 'group 1', and they were presented as the underlined italic characters in phylogenetic tree (Figure 2A). Interestingly, the group 1 CTLDs of human species were derived from the chromosome 12, when those of mouse were from the chromosome 6 and 12.

The CTLDs can be divided into seven groups based on their domain architecture, and seven new groups were added in his revised article in 2002 (Drickamer, 1993; Drickamer and Fadden, 2002; Zelensky and Gready, 2005). Among CTLD genes, human clec4C_1 and mouse clec14A showed very close relationships with the nucleocapsid genes of SARS-CoVs, whereas human clec4A_4, 14A and 10A_2 as well as mouse clec4A2, 4G and 4F were located far from SARS-CoVs on the basis of both Dim1 and Dim2.

Comparison of RSCU values between SARS-CoVs and the most similar CTLD genes of human and mouse

In the CA result, human clec4C_1 and mouse clec14A showed very close relationships with the

Table 2. The results of the regression analysis between each dimensional factor of correspondence analysis using RSCU values of CoVs and each codon pattern parameters.

Variances	DIM1[a]		DIM2[b]	
	R-Square[c]	Parameter estimate[d]	*R-Square*	Parameter estimate
GC_{1st}	0.7809*	0.02896	0.0016	-0.00105
GC_{2nd}	0.6710*	0.03903	0.0135	-0.00440
GC_{3rd}	0.3497*	0.01701	0.5003*	0.01620
ENC[e]	0.5677*	0.03538	0.1070*	0.01168

[a]First dimensional factor of the correspondence analysis, [b]Second dimensional factor of the correspondence analysis, [c]R^2 value of each linear regression analysis, [d]Parameter estimate which was resulted from linear regression analysis, [e]Effective number of codons. *$P < 0.0001$.

nucleocapsid genes of SARS-CoV (Figure 2B). So we compared the RSCU profiles of each gene to analyze the different patterns more intensively (Figure 3). The nucleocapsid and spike genes of human and bat SARS-CoVs (Figure 3A and 3B) were compared with the two types of CTLD genes such as human clec4C_1 and mouse clec14A which were included in 'group 1' in Figure 2B, and human clec10A_2 and mouse clec4F which were not included in group 1 (Figure 3C and 3D). First of all, nucleocapsid genes of human and bat SARS-CoVs did not use two synonymous codons for cysteine (CYS) as well as ACG for threonine (THR), and showed similar patterns with the spike genes in alanine (ALA), asparagine (ASN), glutamine (GLN), glycine (GLY), proline (PRO), serine (SER) and threonine encoding codon groups (Figure 3A and 3B). In the phenylalanine (PHE) and the first three codons of leucine (LEU) encoding codons, however, nucleocapsid and spike genes showed somewhat opposite patterns from each other, and spike showed more biased patterns in CCU and UCU which encode proline and serine, respectively. Secondly, human and mouse CTLD genes in Figure 3C and 3D showed common RSCU profiles with SARS-CoVs in the alanine and proline encoding codon groups, but used different patterns from SARS-CoVs in glycine and phenylalanine codon groups. As for the human clec4C_1 and mouse clec14A, they showed similar RSCU profiles with spike genes in the arginine and serine coding groups, and human clec4C_1 alone showed the same patterns with nucleocapsid genes in the isoleucine (ILE) and serine encoding groups. The human clec10A_2 and mouse clec4F showed more biased patterns in leucine and serine encoding codons than those of the group 1 CLECs (Figure 3D).

Discussion

Codon usage bias has been studied in various organisms ranging from virus to eukaryote, and optimized codon usages in the target viral genes also have been made to improve the efficacy of DNA vaccines development (Ramakrishna *et al*., 2004; Shackelton *et al*., 2006; van Hemert *et al*., 2007; Wang *et al*., 2006a, 2006b). Based on our previous studies, synonymous codon usage itself among RNA viruses such as influenza A viruses and HIV-1s revealed specific bias on the basis of each region, subtype, host or occurring-year group, suggesting that there might be some correlations between the codon usage patterns and viral variations in the codon basis (Ahn and Son, 2006, 2007; Ahn *et al*., 2006).

In this paper, we determined whether % GC contents on the first (GC_{1st}), second (GC_{2nd}), and third (GC_{3rd}) codon positions showed similar patterns among the same genes of viral species as well as the CTLDs of human and mouse (Figure 1). Among the two genes of CoVs, the nucleocapsid genes showed highly positive eigenvectors of GC_{3rd} (0.978) along with the PRIN2, and this pattern was also observed when we compared all the target genes from CoVs, human and mouse together using the principal factor analysis (Table 1). Traditionally, spike protein is known to define the viral tropism by its receptor specificity and also by its membrane fusion activity during virus entry into cells, so it has been the major target of neutralizing antibodies in vaccine development (Gallagher and Buchmeier, 2001). Recently, however, the nucleocapsid also has been studied as a new viral target protein in vaccine industry because of its good immunogenicity (Bode *et al*., 2003; Ye *et al*., 2007). As for hepatitic C virus (HCV), the nucleocapsid protein is known to play an important role in immune evasion, including the inhibition of IFN-α-induced tyrosine phosphorylation, and activation of STAT1 in hepatic cells (Bode *et al*., 2003). The nucleocapsid protein of SARS-CoV itself has become a potential candidate for DNA vaccine production because it revealed a critical role in viral infection process (Zhu *et al*., 2004; Zhao *et al*., 2007; Mark *et al*., 2008; Schulze *et al*., 2008). Ye *et al*. reported that the nucleocapsid gene of mouse hepatitis virus A59, a group 2 CoV, circumvented the effects of the type I interferon (2007). Furthermore, Okada and his colleagues reported that mice vaccinated with the nucleocapsid protein of SARS-CoV showed T-cell immune responses, and Gao resulted that SARS DNA vaccine encoding nucleocapsid protein generated INF-γ producing T-cells in rhesus monkeys (Gao *et al*., 2003; Okada *et al*., 2005).

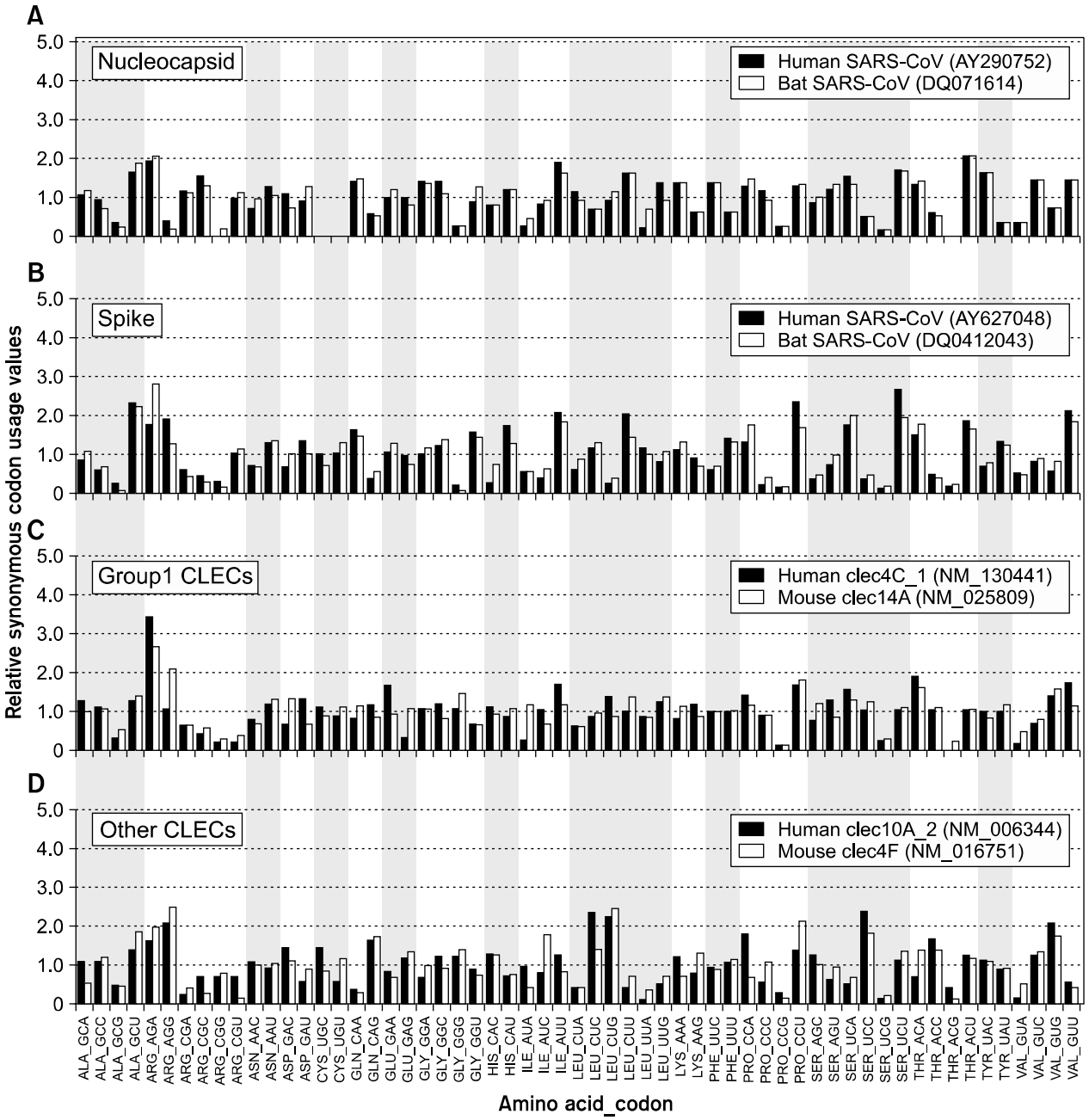

Figure 3. The profiles of the relative synonymous codon usage were shown as the vertical bar graph. The nucleocapaid (A) and spike (B) genes of human and mouse SARS-CoVs, as well as the human and mouse CTLD genes which were located near the nucleocapsid genes of SARS-CoVs (C) and were located far from those of SARS-CoVs (D) in the correspondence analysis in Figure 2B are presented. Genbank accession numbers are presented in legends. G1, Group 1 CoV; G2, Group 2 CoV; G3, Group 3 CoV; N, nucleocapsid gene of CoV; S, spike gene of CoV; CLEC, C-type lectin domain gene.

Our finding demonstrated that GC_{3rd} of nucleocapsid genes revealed highly positive relationships along with the RPIN2 among CoV species, whereas spikes did not show any specific patterns related to GC_{3rd}. This result implicates that the nucleocapsid genes of CoVs might be more heavily affected by the synonymous codon usage bias which is usually determined by the nucleotide on the third codon position than spike genes (Figure 1A and 1B).

In order to compare the synonymous codon usage patterns among two genes of CoVs as well as CTLDs of human and mouse more intensively, we calculated the RSCU values of all the target

genes from CoVs, human and mouse, and then, analyzed the Euclidean distances using the CA (Figure 2B). As a result, the nucleocapsid genes of SARS-CoVs from both human and bat showed the most biased patterns (0.292) among CoVs along with Dim2, which showed the significant correlations with the GC_{3rd} (R^2=0.50, $P < 0.0001$) of those genes in linear regression test (Table 2), and CTLDs of both human and mouse were broadly distributed on the first quadrant (Figure 2B). Interestingly, the group 1 CTLDs of human species were derived from the chromosome 12, and those of mouse were from the chromosome 6 and 12, whereas other CTLDs were from chromosome 14, 16, 17 or 19 for human, and 6, 8 or 9 for mouse. Our finding suggests clue that there might be a specific genomic region or chromosomes, which show a more similar synonymous codon usage pattern with antigenic viral genes. Recently, DNA vaccine has become a more and more important part of vaccine development against many infectious viruses (Martin *et al*., 2006, 2007; Catanzaro *et al*., 2007), and the codon-optimization method which switches the synonymous codons of viruses to those of their host organisms has been reported to improve the immunogenicity of HIV-1 and influenza A virus (Ramakrishna *et al*., 2004; Wang *et al*., 2006a, 2006b). For now, the preferred codons of the overall mammalian system are used in the codon-optimization process, but we observed that there were various synonymous codon biases even among CTLD genes of both human and mouse species. Although those differences might be due to the chromosomal region which each gene was transcribed from, or other factors, one thing is clear that the preferred codons of host organisms are more various than we thought. In the case of CTLDs of human and mouse host, the group 1 genes were commonly transcribed from the chromosome 6 or 12.

On the other hand, human CoV (HKU1) which is included in group 2 CoVs showed the most distinct synonymous codon usage biases in both % GC contents and RSCU patterns (Figure 1, 2), which agrees with the results from Woo *et al*. (Woo *et al*., 2005). Woo suggested that it might be because human CoV (HKU1) may have originated from a major recombination event and numerous minor recombination events among group 2 CoVs. In this study, the nucleocapsid genes of human CoV (HKU2) were found on the opposite side from other group 2 CoVs along with the PRIN2 (Figure 1A), which showed high relationships with GC_{3rd} in Table 1, and they also revealed the opposite RSCU patterns from other group 2 CoVs on the basis of Dim1 in Figure 2B.

In Figure 3, we compared the RSCU profiles of both nucleocapsid and spike genes of human SARS-CoVs (AY290752, AY627048) with other genes such as the nucleocapsid gene of bat SARS-CoVs (DQ071614, DQ0412043), the group 1 CLECs of human (NM_130441) and mouse (NM_025809), which were most closely located with human SARS-CoV, and other CLECs of human (NM_006344) and mouse (NM_016751), which showed distinct patterns from CoVs. As a result, the nucleocapsid genes of both human and bat SARS-CoVs did not use two synonymous codons for cysteine as well as ACG for threonine at all (Figure 3A), whereas other CoVs used them (data not shown). Among SARS-CoVs, the RSCU profile showed somewhat different patterns between the nucleocapsid and spike genes, especially in the phenylalanine and leucine encoding codons, and spike showed more biased patterns in U-ended codons such as CCU (RSCU = 2.34), and UCU (RSCU = 2.67) for profile and serine, respectively. In general, the RSCU value would be 1.00 if there is no codon usage bias. As for the human clec4C_1 and mouse clec14A, they showed very similar profiles with spike genes, especially with bat SARS-CoV, in the arginine coding groups, showing the high RSCU values over 2.50 in AGA. The human clec10A_2 and mouse clec4F showed more biased patterns in GC-ending codons such as CUC, CUG and UCC for leucine and serine encoding codons than those of the group 1 CLECs (Figure 3D).

Consequently, our study demonstrated that the nucleocapsid genes of CoVs might be more heavily affected by the synonymous codon usage than spike genes, and the CTLDs of human and mouse were partially overlapped with the nucleocapsid genes of CoVs. Furthermore, we showed that the group 1 CTLDs of human species were commonly derived from the chromosome 12, and those of mouse were from the chromosome 6 and 12. This suggests that there might be a specific genomic region or chromosomes which show a more similar synonymous codon usage pattern with viral genes. We also found the similar results between CoV genes and other human or mouse genes in our preliminary stage (data not shown). Our findings might be helpful for developing the codon-optimization method in DNA vaccines, and further study is necessary to determine a specific correlation between the codon usage patterns of coding sequences and the chromosomal locations where they are transcribed from in higher organisms.

Methods

Nucleotide sequences

The nucleocapsid (251 sequences) and spike (284 sequences) genes of *Coronavirus* genus including SARS-CoVs were collected from the NCBI Taxonomy Browser (www.ncbi.nlm.nih.gov/Taxonomy/) in GenBank format, and then, all the GenBank flat files were parsed into each category such as accession number, species name, gene name and sequence length using JAVA codes to construct a local database to facilitate the further computational works. As for the human and mouse species, we collected the coding sequences of human (*homo sapiens*) and mouse (*mus musculus*) from the genome section of NCBI's FTP site (ftp.ncbi.nih.gov/genomes/), and also parsed and constructed a local database. All the gene names, abbreviations, sequence lengths and their GenBank accession numbers of CTLD genes used in this study are shown in Supplemental Data Table S1. Abnormal sequences which include unknown characters except for A, G, C or U were not divided by three - maximum length of a codon unit - were removed. MySQL database management system was used to construct all the local databases on Linux operating system.

Principal component analysis of % guanine-cytosine contents data

Principal component analysis was performed using the % guanine-cytosine (GC) contents of the first (GC_{1st}), second (GC_{2nd}) and third (GC_{3rd}) position of each codon, which were calculated for the nucleocapsid and spike coding genes of *Coronavirus* genus as well as the CTLD genes of human and mouse species. All the screened target sequences were extracted from our local database first, then, each coding sequence was parsed into each codon unit. From the pool of codon units for each sequence, we calculated the % GC contents on the first, second and third codon position. JAVA was used in all calculation processes, and the SAS 9.1 statistical program (Cary, 2004) was used for the principal analysis.

Phylogenetic analysis

Twenty nine sequences of the CTLD genes from human and mouse species were used for the multiple sequence alignments using the ClustalW ver. 1.83 program (Thompson *et al*., 1997) with default parameters that set the DNA weight matrix as the IUB matrix, and values of gap opening and gap extension penalties as 15.0 and 6.66, respectively. The Neighbor-Joining method with 1000 times bootstrapping process were performed using PAUP* ver. 4.0b program (Swofford, 1999).

Correspondence analysis

The correspondence analysis (CA) method was used to compare the RSCU values for the 59 codons described above using the SAS 9.1 statistical program (Cary, 2004). The RSCU value is the number of times that a particular codon is observed relative to the number of times that the codon would be observed in the absence of any codon usage bias. If there is no codon usage bias, the RSCU value is 1.00. The RSCU was calculated as

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}}$$

where X_{ij} is the frequency of occurrence of the *j*th codon for the ith amino acid, and n_i is the number of codons for the *i*th amino acid. Each gene is represented as a 59-dimensional vector excluding the start and stop codons and UGG, which codes for tryptophan which has no synonyms (Sharp and Li, 1986). We assigned each kind of gene or species as rows, and RSCU values of 59 codons as columns in an input data set for CA. The biplot graph from a CA includes the best two dimensional representations of the data, along with the coordinates of the plotted points, and a measure of the amount of information retained in each dimension. CA uses *chi*-square to standardize the frequency values, so the distance between two coordinates with the same row or column value indicates the *chi*-square distance (Hair *et al*., 1998). If this distance is long enough to have statistical meaning, the coordinates of the output plots along with each column or row direction will be located far from the origin, and they usually exist on the opposite side of each coordinate axis. The distance between each row or each column reveals the Euclidean distance (Gu *et al*., 2004; Perrière and Thioulouse, 2002), but there is no meaning between the row and column coordinates.

Other statistical analysis

Linear regression analysis was conducted to determine the correlations between the first two dimensional factors (Dim1 and Dim2) of the CA results, and the % GC contents on each codon position, effective number of codons (ENC) and the average hydrophilicities of encoded proteins. ENC values were often used to measure the magnitude of codon bias, which yields values raging from 20, when one codon is used for each amino acid, to 61, when all synonymous codons are used in equal frequency (Wright, 1990). We calculated each ENC value per each nucleotide sequence using JAVA codes, and all these analyses were performed using the SAS 9.1 statistical program (Cary, 2004).

Supplemental data

Supplemental Data include a Table and can be found with this article online at http://e-emm.or.kr/ article/article_files/ SP-41-10-07.pdf.

Acknowledgements

This study was supported by the Korea Institute of Science and Technology Information. We also acknowledge the invaluable contribution of the researchers who have made their data publicly available.

References

Ahn I, Jeong BJ, Bae SE, Jung J, Son HS. Genomic analysis of influenza A viruses, including avian flu (H5N1) strains. Eur J Epidemiol 2006;21:511-9

Ahn I, Son HS. Epidemiological comparisons of codon usage patterns among HIV-1 isolates from Asia, Europe, Africa and the Americas. Exp Mol Med 2006;38:643-51

Ahn I, Son HS. Comparative study of the hemagglutinin and neuraminidase genes of influenza A virus H3N2, H9N2, and H5N1 subtypes using bioinformatics techniques. Can J Microbiol 2007;53:830-9

Bode JG, Ludwig S, Ehrhardt C, Erhardt A, Albercht U, Schaper F, Heinrich PC, Häussinger D. IFN-α antagonistic activity of HCV core protein involves induction of suppressor of cytokine signaling-3. FASEB J 2003;17:488-90

Cary NC. SAS 9.1.2 Qualification Tolls User's Guide, 2004, SAS Institute Inc. NC

Catanzaro AT, Roederer M, Koup RA, Bailer RT, Enama ME, Nason MC, Martin JE, Rucker S, Andrews CA, Gomez PL, Mascola JR, Nabel GJ, Graham BS, VRC 007 Study Team. Phase I clinical evaluation of a six-plasmid multiclade HIV-1 DNA candidate vaccine. Vaccine 2007;25:4085-92

Cella M, Sallusto F, Lanzavecchia A. Origin, maturation and antigen presenting function of dendritic cells. Curr Opin Immunol 1997;9:10-6

Cella M, Salio M, Sakakibara Y, Langen H, Julkunen I, Lanzavecchia A. Maturation, activation, and protection of dendritic cells induced by double-stranded RNA. J Exp Med 1999;189:821-9

Dimmock NJ, Easton AJ, Leppard KN. Coronaviruses, p141. Introduction to modern virology, 5th ed, 2002, Blackwell Publishing, Malden, MA.

Dodd RB, Drickamer K. Lectin-like proteins in model organisms: implications for evolution of carbohydrate-binding acivity. Glycobiol 2001;11:71R-79R

Donnelly J, Berry K, Ulmer JB. Technical and regulatory hurdles for DNA vaccines. Int J Parasitol 2003;33:457-67

Drickamer K. Evolutional of Ca^{2+}-dependent animal lectins. Prog Nucleic Acid Res Mol 1993;45:207-32

Drickamer K, Dodd RB. C-type lectin-like domains in *Caenorhabditis elegans*: predictions from the complete genome sequence. Glycobiol 1999;9:1357-69

Drickamer K, Fadden AJ. Genomic analysis of C-type lectins. Biochem Soc Symp 2002;69:59-72

Duret L. Evolution of synonymous codon usage in metazoans. Curr Opin Gene & Dev 2002;12:640-9

Frieman M, Heise M, Baric R. SARS coronavirus and innate immunity. Virus Res 2008;113:101-12

Gallagher TM, Buchmeier MJ. Coronavirus spike proteins in viral entry and pathogenesis. Virol 2001;279:371-4

Gao W, Tamin A, Soloff A, D'Aiuto L, Nwanegbo E, Robbins PD, Bellini WJ, Barratt-Boyes S, Gambotto A. Effects of a SARS-associated coronavirus vaccine in monkey. Lancet 2003;362:1895-6

Gu W, Zhou T, Ma J, Sun X, Lu Z. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. Virus Res 2004;101:155-61

Hair JF Jr, Anderson RE, Tatham RL, Black WC. Multivariate Data Analysis: fifth edition, 1998, Prentice-Hall International, Inc., USA

Jenkins GM, Holmes EC. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res 2003;92:1-7

Kawabe A, Miyashita NT. Patterns of codon usage bias in three dicot and four monocot plant species. Genes Genet Syst 2003;78:343-52

Lambert AA, Gilber CG, Richard M, Beaulieu AD, Tremblay MJ. The C-type lectin surface receptor DCIR acts as a new attachment factor for HIV-1 in dendritic cells and contributes to *trans*- and *cis*-infection pathways. Blood 2008;112:1299-307

Lew TWK., Kwek TK, Tai D, Earnest A, Loo S, Singh K, Kwan KM, Chan Y, Yim CF, Bek SL, Kor AC, Yap WS, Chelliah YR, Lai YC, Goh SK. Acute respiratory distress syndrome in critically ill patients with severe acute respiratory syndrome. JAMA 2003;290:374-80

Lynn DJ, Singer GAC, Hickey DA. Synonymous codon usage in subject to selection in thermophilic bacteria. Nucleic Acids Res 2002;30:4272-7

Mark J, Li X, Cyr T, Fournier S, Jaentschke B, Hefford MA. SARS coronavirus: unusual lability of the nucleocapsid protein. Biochem Biophys Res Comm 2008;377:429-33

Martin JE, Sullivan NJ, Enama ME, Gordon IJ, Roederer M, Koup RA, Bailer RT, Chakrabarti BK, Bailey MA, Gomez PL, Andrews CA, Moodie Z, Gu L, Stein JA, Nabel GJ, Graham BS, VRC 204 Study Team. A DNA vaccine for Ebola virus is safe and immunogenic in a phase I clinical trial. Clin Vaccine Immunol 2006;13:1267-77

Martin JE, Pierson TC, Hubka S, Rucker S, Gordon IJ, Enama ME, Andrews CA, Xu Q, Davis BS, Nason MC, Fay MP, Koup RA, Roederer M, Bailer RT, Gomez PL, Mascola JR, Chang GJ, Nabel GJ, Graham BS. A West Nile virus DNA vaccine induces neutralizing antibody in healthy adults during a phase 1 clinical trial. J Infect Dis 2007;196:1732-40

Martin JE, Louder MK, Holman LA, Gordon IJ, Enama ME, Larkin BE, Andrews CA, Vogel L, Koup RA, Roederer M, Bailer RT, Gomez PL, Nason M, Mascola JR, Nabel GJ, Graham BS, VRC 301 Study Team. A SARS DNA vaccine induces neutralizing antibody and cellular immune responses in healthy adults in a phase I clinical trial. Vaccine 2008; 26:6338-43

McInerney JO. Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. Proc Natl Acad Sci 1998;95:10698-703

Moriyama EN, Hartl DL. Codon Usage Bias and Base Composition of Nuclear Genes in Drosophila. Genetics 1993;134:847-58

Okada M, Takemono Y, Okuno Y, Hashimoto S, Yoshida S, Fukunaga Y, Tanaka T, Kita Y, Kuwayama S, Muraki Y,

Kanamaru N, Takai H, Okada C, Sakaguchi Y, Furukawa I, Yamada K, Matsumoto M, Kase T, deMello DE, Peiris JSM, Chen PJ, Yamamoto N, Yoshinaka Y, Nomura T, Ishida I, Morikawa S, Tashiro M, Sakatani M. The development of vaccines against SARS corona virus in mice and SCID-PBL/hu mice. Vaccine 2005;23:2269-72

Perrière G, Thioulouse J. Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Res 2002;30:4548-55

Ramakrishna L, Anand KK, Mohankumar KM, Ranga U. Codon optimization of the Tat antigen of human immunodeficiency virus type 1 generates strong immune responses in mice following genetic immunization. J Virol 2004; 78:9174-89

Riley S, Fraser C, Donnely CA, Ghani AC, Abu-Raddad LJ, Hedley AJ, Leung GM, Ho LM, Lam TH, Thach TQ, Chau P, Chan KP, Lo SV, Leung PY, Tsang T, Ho W, Lee KH, Lau EMC, Ferguson NM, Anderson RM. Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. Science 2003;300:1961-6

Schulze K, Staib C, Schätzl HM, Ebensen T, Erfle V, Guzman CA. A prime-boost vaccination protocol optimizes immune responses against the nucleocapsid protein of the SARS coronavirus. Vaccine 2008;26:6678-84

Shackelton LA, Parrish CR, Holmes EC. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. J Mol Evol 2006;62:551-63

Sharp PM, Li WH. Codon usage in regulatory genes in Escherichia coli does not reflect for 'rare' codons. Nucleic Acids Res 1986;14:7737-49

Shields DC, Sharp PM. Synonymous codon usage in Bacillus stbtilis reflects both translational selection and mutational biases. Nucleic Acids Res 1987;15:8023-40

Sin JI, Weiner DB. Improving DNA vaccines targeting viral infection. Intervirology 2000;43:233-46

Singer GAC, Hickey DA. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. Gene 2003;317:39-47

Spiegel M, Schneider K, Weber F, Weidmann M, Hufert FT. Interaction of severe acute respiratory syndrome-associated coronavirus with dendritic cells. J Gen Virol 2006;87: 1953-60

Stenico M, Lloyd AT, Sharp PM. Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. Nucleic Acids Res 1994;22:2437-46

Swofford DL. PAUP*. Phylogenetic analysis using parsimony (* and other methods). ver. 4, 1999, Sinauer, Sunderland, Mass

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. CLUSTLAX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997;25:4876-82

van Hemert FJ, Berkhout B, Lukashov VV. Host-related nucleotide composition and codon usage as driving forces in the recent evolution of the Astroviridae. Virol 2007;361: 447-54

Wang S, Farfan-Arribas D, Shen S, Chou TW, Hirsch AL, He F, Lu S. Relative contributions of codon usage, promoter efficiency and leader sequence to the antigen expression and immunogenicity of HIV-1 and Env DNA vaccine. Vaccine 2006a;24:4531-40

Wang S, Tafffe J, Parker C, Solórzano A, Cao H, García-Sastre A, Lu S. Hemagglutinin (HA) proteins from H1 and H3 serotypes of influenza A viruses require different antigen designs for the induction of optimal protective antibody responses as studied by codon-optimized HA DNA vaccines. J Virol 2006b;80:11628-37

Wang SF, Huang JC, Lee YM, Liu SJ, Chan YJ, Chau YP, Chong P, Chen MA. DC-SIGN mediates avian H5N1 influenza virus infection in cis and in trans. Biochem. Biophys. Res Comm 2008;373:561-6

Woo PCY, Lau SKP, Huang Y, Tsoi HW, Chan KH, Yuen KY. Phylogenetic and recombination analysis of coronavirus HKU1, a novel coronavirus from patients with pneumonia. Arch Virol 2005;150:2299-311

Wright F. The 'effective number of codons' used in a gene. Gene 1990;87:23-9

Yang ZY, Kong WP, Huang Y, Roberts A, Murphy BR, Subbarao K, Nabel GJ. A DNA vaccine induces SARS coronavirus neutralization and protective immunity in mice. Nature 2005;428:561-4

Ye Y, Hauns K, Langland JO, Jacobs BL, Hogue BG. Mouse hepatitis coronavirus A59 nucleocapsid protein is a type I interferon antagonist. J Virol 2007;81:2554-63

Zelensky AN, Gready JE. The C-type lectin-like domain superfamily. FEBS J 2005;272:6179-217

Zhao J, Huang Q, Wang H, Zhang Y, Lv P, Gao XM. Identification and characterization of dominant helper T-cell epitopes in the nucleocapsid protein of severe acute respiratory syndrome coronavirus. J Virol 2007;81:6079-88

Zhu MS, Pan Y, Chen HQ, Shen Y, Wang XC, Sun YJ, Tao KH. Induction of SARS-nucleoprotein-specific immune response by use of DNA vaccine. Immunol Lett 2004;92: 237-43