



Research article

Codon usage similarity between viral and some host genes suggests a codon-specific translational regulation



Kunlakanya Jitobaom^a, Supinya Phakaratsakul^a, Thanyaporn Sirihongthong^a,
 Sasithorn Chotewutmontri^b, Prapat Suriyaphol^{c,d}, Ornpreeya Suptawiwat^b,
 Prasert Auewarakul^{a,*}

^a Department of Microbiology, Faculty of Medicine Siriraj Hospital, Mahidol University, Thailand

^b Faculty of Medicine and Public Health, HRH Princess Chulabhorn College of Medical Science, Chulabhorn Royal Academy, Bangkok, Thailand

^c Division of Bioinformatics and Data Management for Research, Department of Research and Development, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok, Thailand

^d Center of Excellence in Bioinformatics and Clinical Data Management, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

ARTICLE INFO

Keywords:

Microbiology
 Virology
 Viruses
 Viral genetics
 Genetics
 Gene expression
 Codon usage pattern
 RSCU
 RNA viruses
 Cell cycle
 Translation

ABSTRACT

The codon usage pattern is a specific characteristic of each species; however, the codon usage of all of the genes in a genome is not uniform. Intriguingly, most viruses have codon usage patterns that are vastly different from the optimal codon usage of their hosts. How viral genes with different codon usage patterns are efficiently expressed during a viral infection is unclear. An analysis of the similarity between viral codon usage and the codon usage of the individual genes of a host genome has never been performed. In this study, we demonstrated that the codon usage of human RNA viruses is similar to that of some human genes, especially those involved in the cell cycle. This finding was substantiated by its concordance with previous reports of an upregulation at the protein level of some of these biological processes. It therefore suggests that some suboptimal viral codon usage patterns may actually be compatible with cellular translational machineries in infected conditions.

1. Introduction

The genetic code is degenerate. There are 61 triplet codons coding for 20 amino acids and 3 stop codons [1]. Therefore, each amino acid is encoded by several codons, with the exception of two amino acids (methionine and tryptophan). This codon redundancy results in synonymous codon usage, whereby one amino acid is encoded by 2, 4, or 6 codons [2]. Several previous studies have revealed that synonymous codons are utilized with different frequencies and are not randomly used by different genomes or genes. This non-randomness is referred to as codon usage bias [3, 4]. Each species preferentially uses different synonymous codons [5]. This results in a species-specific codon usage bias.

Similarly, there are several tRNA species that carry the same amino acid. These tRNA species are called isoacceptors [6]. Codons and anti-codons in tRNA do not interact in a one-to-one fashion [7]. Base pairing at the third codon position is wobble; for example, G can pair with both cytosine (C) and uracil (U) [8]. It has been demonstrated that

tRNA modification directly affects tRNA and mRNA wobble base pairing [9, 10]. Both the available tRNA isoacceptors and tRNA modification change with the cell cycle, and they can be altered by cellular stresses [11, 12]. For efficient protein translation, the codon usage pattern should correlate with the population of available tRNA isoacceptors [13]. In the cellular stress-response, the alteration could enhance the expression of stress-response genes, with the codon usage patterns compatible with the changed tRNA modifications. Those genes shown to be regulated by this codon-specific manner are called Modification Tunable Transcripts (MoTTs) [12, 14].

Two major models have been proposed to explain the causes of codon usage bias: mutation pressure, and translational selection [15]. As to mutation pressure, it is believed that GC content is the major factor driving codon usage bias [16, 17]. The high mutation rates of some nucleotides or codons result in nucleotide substitution that might contribute to lower frequencies of some nucleotides and codons [15]. Mutation pressure has been suggested to be the most important factor determining

* Corresponding author.

E-mail address: prasert.aue@mahidol.ac.th (P. Auewarakul).

the codon usage bias in human RNA viruses [18, 19, 20]. However, there are correlations between the codon usage bias and other factors related to translation efficiency (such as available tRNAs, mRNA secondary structure, translation elongation rate, and the intragenic and intergenic codon bias) that cannot be explained by mutation pressure. This suggests that translational selection also influences codon usage bias [15]. The translational selection acts on codon usage bias to achieve efficient and accurate translation. The use of codons correlates with abundant tRNAs, resulting in a higher translation rate [21, 22, 23]. A correlation between codon usage bias and abundant tRNAs has been found in prokaryotes (such as *E. coli* [24]) and in some eukaryotes (such as *S. cerevisiae* [25], *C. elegans* [22], *Drosophila* [23], and human [26]). However, rare codons are preferred to encode some specific sets of genes or regions of genes, for instance, to enable protein oscillation in different phases of the cell cycle, slow down protein translation across the membrane, and reduce ribosome jamming and mRNA secondary structure at the 5' end of coding sequences [11, 27, 28, 29]. Therefore, the translational selection acting on the optimization of frequent and rare codon utilization is important in appropriate gene translation.

Viral replication is dependent on the cellular machineries of the host cells. Thus, one would intuitively think that the codon usage of a viral genome should match that of its host in order to be efficiently expressed. Surprisingly, however, most viruses have codon usage patterns that are different from the codon usage preference of their hosts [30, 31, 32, 33]. A previous study indicated an alteration in the cellular tRNA level after the infection of human immunodeficiency virus type 1 (HIV-1) [34]. In contrast, the cellular tRNA level was found to be unchanged following vaccinia and influenza A virus (IAV) infection, whereas an alteration in the polysome-associated tRNA population was observed, particularly the population of polysome tRNA isoacceptors correlated with viral codon usage [35]. These findings suggest that the codon usage pattern and the regulation of translational machineries may influence gene expression in some viruses.

In this study, we investigated the relationship between the codon usage bias of human genes and human RNA viruses. It is generally believed that the codon usage bias of viruses differs from that of human genes; however, various human genes possess various codon usage patterns [19, 26]. In addition, intragenic codon biases had previously been

reported in humans and mice [36]. A comparison of the codon usage at the genome level is therefore too generalized; a more precise comparison at the single-gene level may provide a better insight into the viral codon usage bias.

2. Results

2.1. Principal component analysis of the relative synonymous codon usage (PCA of RSCU)

A total of 20,190 major transcript variants of human protein coding sequences were recruited from the GENCODE database (version 26). The protein coding sequences of 77 human RNA viruses were downloaded from the NCBI database. The open reading frames (ORFs) of the protein coding sequences were rechecked by ORFfinder (NCBI) before performing the RSCU calculation. The RSCU is a simple parameter that represents the codon usage bias of synonymous codons in a coding sequence. In our analysis, the RSCU was calculated from the protein coding sequences of the human and RNA viruses. The RSCU of each gene consists of 59 values corresponding to 59 synonymous codons; thus, the PCA was performed to simplify the data to a smaller number of principal factors as a summary feature of the codon usage pattern of each gene. The PCA successfully reduced the 59 values of each RSCU into two significant components. The RSCUs of the human genes and RNA viruses were represented by the coordinates of principal component 1 (PC1, x-axis) and principal component 2 (PC2, y-axis) plotted on the PCA of an RSCU graph (Figure 1). The RSCU and PCA of the RSCUs of the human genes and human RNA viruses are shown in Supplementary File 1.

In Figure 1, the PCAs of the RSCUs of the human genes were represented with a transparent black dot; genes with a similar RSCU were located in the same area of the graph. The number of human genes located in each quadrant was counted: upper left, 6,020 genes; lower left, 4,227 genes; upper right, 4,664 genes; and lower right, 5,281 genes. Interestingly, many human genes were located densely in the right quadrants, specifically, between ($x = 0.95$ to 1.7) and ($y = -0.7$ to 0.6). Most RNA viruses were also located in the right quadrants. Additionally, negative sense-single strand RNA viruses (-ssRNA), ambisense RNA viruses (Ambi), and HIV-1 viruses were located in the area of the right quadrants

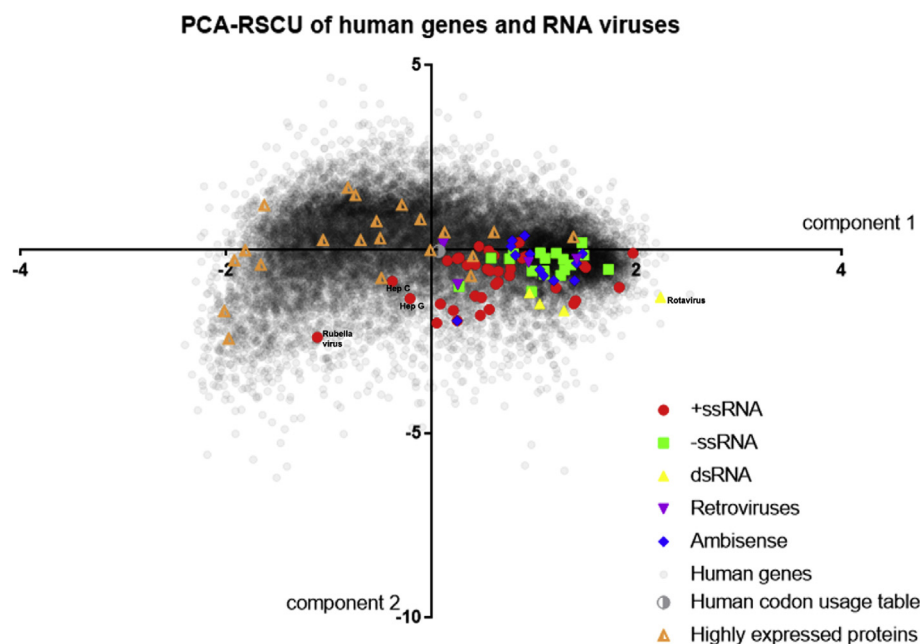


Figure 1. The PCA of the RSCU. The RSCU of human genes and RNA viruses were subjected to PCA. Then, the simplified RSCU values of human genes and RNA viruses were plotted on the graph as the coordinates of component 1 (x) and 2 (y). The color keys indicate the groups of RNA viruses. Human genes are represented using transparent black dots.

where the human genes were densely located. However, a great variation was observed in some groups of RNA viruses, especially positive-sense single strand RNA viruses (+ssRNA), double strand RNA viruses (dsRNA), and retroviruses (Retro).

The relationship between the PCA of the RSCU analysis and the codon adaptive index (CAI) was investigated. CAI is a common parameter used to assess the codon usage bias of a gene. It is calculated from the frequency of the overall codons in a given protein coding sequence with respect to a reference set of genes [37]. In our analysis, the human codon usage table, which is the average codon frequency of a human genome, was used as the reference set. From the results (Figure 1), the PCA of the RSCU of the human codon usage table was plotted near the x and y intercepts, showing the average codon usage pattern of all human genes. A number of human genes in the PCA of the RSCU graph were selected and subjected to the CAI calculation. The graphs of PC1 and CAI, of PC2 and CAI, were plotted; the linear regression and Pearson correlation coefficient (PCC) were subsequently analyzed. From Figure 2A, it was found that the CAI of genes gradually decreased with an increase in PC1 ($R^2 = 0.7958$, $PCC = -0.8921$, p -value < 0.0001), while a positive correlation was observed between CAI and PC2 ($R^2 = 0.5824$, $PCC = 0.7631$, p -value < 0.0001). The percentages of the GC content at the third position of the codon (%G+C(3)) of the human genes were also determined. In a similar way to CAI (Figure 2B), %G+C(3) gradually decreased with an increase in PC1 ($R^2 = 0.9459$, $PCC = -0.9726$, p -value < 0.0001). A weak correlation between PC2 and %G+C(3) was observed ($R^2 = 0.4373$, $PCC = 0.6613$, p -value < 0.0001). Thus, the PCA of the RSCU analysis could be used to characterize the heterogeneity of the codon usage bias in the human genome, in which genes in the left-upper quadrant contain more optimal codon usage for high expression, whereas those in the right quadrants near or below the x-axis have less optimal codon usage.

The PCAs of RSCUs of the human genes that coded for highly expressed proteins were plotted on a graph. The highly expressed proteins of humans had been previously identified using the proteomic approach (Figure 1; see the gene list in Supplementary Table 1). [38]. From Figure 1, most of the highly expressed proteins were located in the left quadrants. This is in agreement with our analysis showing the relationship between PCA and CAI, and it supports the validity of using PCA to predict CAI.

2.2. Human genes with codon usage bias similar to RNA viruses

From the PCA of the RSCU graph (Figure 1), it was demonstrated that most of the human RNA viruses were located in the right quadrants. The degree of difference in the codon usage pattern varied among the groups of viruses. This feature was also observed intragroup. The highest PC1 (x) belonged to rotavirus, indicating a high degree of difference in codon usage pattern compared to human genes and other RNA viruses. In particular, there were a number of human genes with RSCUs similar to RNA viruses (Figure 1), especially -ssRNA, Ambi, and HIV viruses, which were located in the right quadrants, where human genes were also located densely. To investigate the kinds of human genes with codon usage patterns similar to RNA viruses and the contributions of those genes in important biological processes, the human genes plotted in the same area with each subgroup of RNA viruses were retrieved. The criteria for selection of the human genes with RSCUs similar to RNA viruses are described in the methods section.

Figure 3 represents the selected human genes with RSCUs similar to RNA viruses. These human genes were subjected to gene ontology (GO) enrichment analysis using GO-TermFinder to identify the over-represented GO terms in biological processes [39]. REVIGO was then

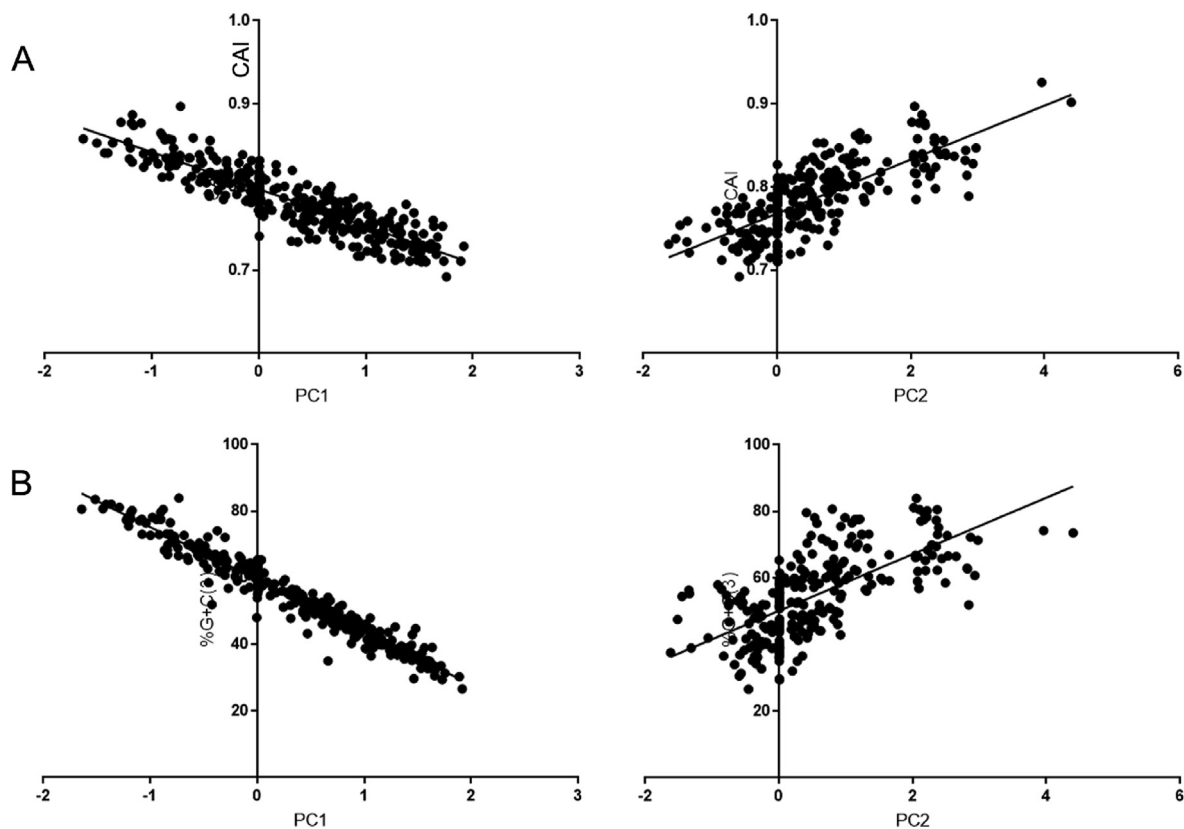


Figure 2. The correlation between principal components (PC1, PC2), codon adaptive index (CAI), and percentage of GC content at the third position of codon (%G+C(3)) were determined using simple linear regression and Pearson correlation coefficient. (A) CAI was plotted with either PC1 ($R^2 = 0.7958$, $PCC = -0.8921$, $p < 0.001$) or PC2 ($R^2 = 0.5824$, $PCC = 0.7631$, $p < 0.001$). (B) %G+C(3) was plotted with either PC1 ($R^2 = 0.9459$, $PCC = -0.9726$, $p < 0.001$) or PC2 ($R^2 = 0.4373$, $PCC = 0.6613$, $p < 0.001$).

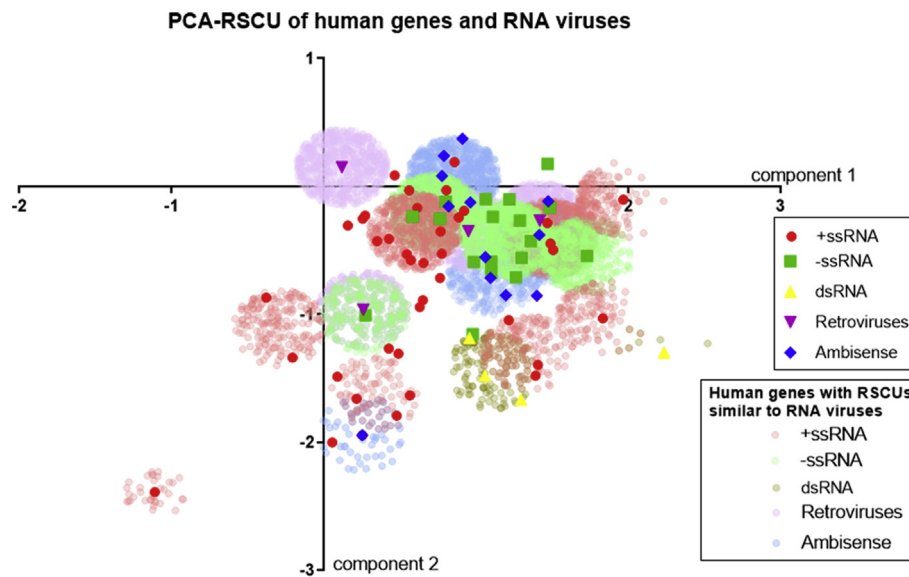


Figure 3. The human genes with RSCUs similar to RNA viruses in each subgroup are represented using transparent colored dots.

used to categorize the redundant GO terms [40]. The number of human genes with RSCUs similar to RNA viruses are listed in Table 1. The results (Figure 4) show the significant GO terms in the biological processes of human genes with RSCUs similar to RNA viruses. Only human genes retrieved from six subgroups of RNA viruses resulted in significant enrichment, namely, +ssRNA (subgroups 6, 7), -ssRNA (subgroups 3, 4), Retro (HIV-1), and Ambi (subgroup 4), where the human genes recruited from +ssRNA (subgroups 6, 7), -ssRNA (subgroup 4), Retro (HIV-1), and Ambi (subgroup 4) were in the same or adjacent area. The human genes with RSCUs similar to +ssRNA (subgroup 6), -ssRNA (subgroup 4), Retro (HIV-1), and Ambi (subgroup 4) shared similar GO terms in biological processes, including the cell cycle, the regulation of the cell cycle process, cell division, microtubule cytoskeletal organization, chromosome segregation, DNA repair, macromolecule catabolism, and cellular localization (Fig. 4A, B, D, E, and F), while human genes with RSCUs similar to +ssRNA (subgroup 3) were related to RNA processing (Figure 4C). The list of enriched GO terms in biological processes is shown in Supplementary File 2.

2.3. Codon usage bias of cell cycle regulated genes is similar to that of RNA viruses

From the previous section, we demonstrated that human genes in the GO terms of the cell cycle and the regulation of the cell cycle process adopt codon usage patterns similar to those of +ssRNA (subgroups 6, 7), -ssRNA (subgroup 4), retrovirus (HIV-1), and ambisense (subgroup 4) viruses. To confirm that the codon usage patterns of these RNA viruses are similar to human genes in the cell cycle and the regulation of the cell cycle process, the cell cycle codon score (CCCS) was used to evaluate the similarity of the codon usage pattern between that of viral genes and a set of cell cycle-regulated human genes (top-600 set) [11]. The CCCS of human RNA viruses had been calculated and is detailed in Table 2. A positive CCCS indicates that a gene has a codon usage pattern similar to the top-600 set. The results revealed that most of the +ssRNA viruses had a negative CCCS score. Only some +ssRNA viruses had a positive CCCS score (such as dengue viruses [DENVs], MERS-coronavirus, SARS-coronavirus, human coronaviruses, human enterovirus 68, and the hepatitis A virus), whereas most of the -ssRNA, dsRNA, HIV-1, HIV-2, and ambisense viruses had a positive CCCS score. However, HTLV-1 and HTLV-2 in the retrovirus group gave a negative CCCS.

A previous study by Frenkel-Morgenstern *et al.* also demonstrated that the codon usage pattern of the cell cycle regulated genes (CCRs) influence

the cell cycle-dependent protein expression [11]. Thus, the similarity of the codon usage pattern of the CCR genes and RNA viruses was investigated. CCR genes cycling at the protein level and non-cycle regulated genes (NCCRs) that were found not to cycle at the protein level were selected from previous studies (Table 3). The RSCUs of the CCR and NCCR genes were plotted on a graph and compared with the RSCUs of RNA viruses (Figure 5). The results showed that the CCR genes were located with -ssRNA, dsRNA, and some +ssRNA viruses, indicating similar codon usage patterns, while the RSCUs of the NCCR genes were distributed all over the graph with no resemblance to RNA viruses. This result corresponds with the CCCS of RNA viruses (Table 2).

2.4. Human genes with similar codon usage patterns to RNA virus were upregulated in viral infections

From the previous section, we demonstrated that human genes with codon usage patterns similar to RNA viruses contributed to some important biological processes, such as the cell cycle, the regulation of the cell cycle process, cell division, microtubule cytoskeletal organization, chromosome segregation, DNA repair, macromolecule catabolism, cellular localization, and RNA processing (Figure 4). When coupled with the fact that some viral infections can manipulate host cellular pathways (especially the translation machineries), this finding suggests that human genes with codon usage patterns similar to RNA viruses may be upregulated during viral infection [41].

To substantiate this hypothesis, sets of proteomics data of RNA virus infection were reanalyzed. The lists of upregulated protein profiles upon HIV-1 [42], IAV [43], Zika virus (ZIKV) [44], and dengue virus serotype 2 (DENV-2) [45] infections were obtained from previous studies. The lists of upregulated proteins upon viral infection were submitted to GO-TermFinder (Supplementary File 3), and the enriched GO terms were compared to the enriched GO terms of human genes with codon usage patterns similar to RNA viruses from every subgroup. Several enriched GO terms of upregulated protein profiles during viral infections were found to be identical to the GO terms of human genes with codon usage patterns similar to RNA viruses (Figure 4). In the case of HIV-1 and ZIKV, the identical GO terms included the cell cycle, the regulation of the cell cycle process, the mitotic cell cycle, organelle organization, cell division, microtubule-based process, and cellular localization. The identical GO terms of IAV and DENV-2 included macromolecule metabolic processes, nucleic acid metabolic process, chromosome organization, cellular stress response and RNA processing.

Table 1. Categories of human RNA viruses. RNA viruses were categorized based on the nucleic acid types of their genomes. The viruses in each group were divided into subgroups based on their PCA of RSCU. In subgroups containing more than one virus, the mean virus RSCU (mean vRSCU [a, b]) was calculated. The number of human genes with RSCUs similar to RNA viruses were also represented.

RNA viruses	Subgroups	Members	PCA of RSCU component		No. of human genes with similar RSCU		
			1 (x)	2 (y)			
+ssRNA	1	Rubella virus	-1.11152	-2.38693	67		
		Hepatitis G virus	-0.20805	-1.33632			
	2	Hepatitis C virus	-0.38012	-0.86880	192		
		Mean vRSCU	-0.29409	-1.10256			
	3	Hepatitis E virus	0.05314	-1.99911	101		
		Ross river virus	0.08755	-1.48581			
		Chikungunya virus	0.21340	-1.65931			
		Sindbis virus	0.25123	-1.94371			
		Eastern equine encephalitis virus	0.48833	-1.30690			
		Western equine encephalitis virus	0.47744	-1.79137			
		Venezuelan equine encephalitis virus	0.42568	-1.26628			
		O'nyong nyong virus	0.56417	-1.63171			
		Mean vRSCU	0.32012	-1.63553			
		4	Enterovirus 71	0.64889		-0.88714	659
			Human coxsackievirus A9	0.54435		-0.52889	
	Human coxsackievirus B4		0.65238	-0.59692			
	Echoviruses		0.57004	-0.57533			
	Polioviruses		0.76484	-0.35123			
	Norwalk virus		0.77184	-0.52468			
	Sapporo virus		0.62754	-0.94239			
	Dengue virus type 1		0.91914	-0.19005			
	Dengue virus type 2		0.85638	0.19119			
	Dengue virus type 3		0.88466	-0.24134			
	Dengue virus type 4		0.80321	-0.02901			
	Japanese encephalitis virus		0.34886	-0.42656			
	Murray Valley encephalitis virus		0.76079	-0.71468			
	St. Louis encephalitis virus		0.61121	-0.16698			
	West Nile virus		0.42713	-0.40840			
	Yellow fever virus		0.55894	-0.03048			
	Zika virus		0.46368	0.08730			
	Kyasanur Forest disease virus		0.26948	-0.22710			
	Omsk hemorrhagic fever virus		0.25430	-0.25169			
Tick-borne encephalitis virus	0.15574	-0.30465					
Mean vRSCU	0.59467	-0.35595					
+ssRNA	5	SARS coronavirus	1.38814	-1.47854	169		
		MERS coronavirus	1.40490	-1.39321			
		Human astrovirus	1.21324	-1.04323			
		Mean vRSCU	1.33543	-1.30499			
	6	Rhinovirus A,B and C	1.46621	-0.28559	946		
		Human parechovirus	1.48761	-0.44531			
		Enterovirus 68	1.50423	-0.49488			
		Mean vRSCU	1.48602	-0.40859			
	7	Hepatitis A virus	1.96492	-0.10146	188		
	8	Human coronaviruses	1.83078	-1.03063	129		
-ssRNA	1	Borna virus	0.27188	-1.00708	233		
		Rabies virus	0.76128	-0.25367			
	2	Mokola virus	0.79901	-0.11880	837		
		Measles virus	0.58253	-0.23633			
		Mean vRSCU	0.71427	-0.20293			
		Vesicular stomatitis virus	1.10766	-0.23715			
	3	Influenza A virus H3N2	0.98353	-0.58886			
		Marburg virus	1.29734	-0.55782			
		Ebola viruses	1.09920	-0.67494			
		Influenza B virus	1.28533	-0.26627			
		Human parainfluenza virus type 1	1.09898	-0.58691			
		Human parainfluenza virus type 3	1.48386	-0.16452			

(continued on next page)

Table 1 (continued)

RNA viruses	Subgroups	Members	PCA of RSCU component		No. of human genes with similar RSCU		
			1 (x)	2 (y)			
		Human parainfluenza virus type 2	1.25962	-0.71004			
		Mumps virus	0.97594	-1.15319			
		Nipah virus	1.21851	-0.10109			
		Hendra virus	1.06209	-0.09805			
		Respiratory syncytial virus	1.46890	0.17766			
		Metapneumovirus	1.35819	-0.42703			
		Mean vRSCU	1.20763	-0.41448	898		
		4	Influenza C virus	1.72634	-0.54316	586	
		dsRNA	1	Colorado tick fever virus	1.29334	-1.66493	
				Mammalian orthoreovirus	1.05591	-1.47831	
Human picobirnavirus	0.95531			-1.18070			
Mean vRSCU	1.10152			-1.44131	161		
Retro	2	Rotaviruses	2.23210	-1.29556	7		
		1	HTLV-1	0.25781	-0.96125	261	
		2	HTLV-2	0.11613	0.15063	790	
		3	HIV-1	1.41668	-0.26596	1037	
Ambi	4	HIV-2	0.94712	-0.34939	812		
		1	Sin nombre virus	0.25123	-1.94371	70	
		2	Lymphocytic choriomeningitis virus	0.77488	0.08308		
			Lassa virus	0.78719	0.24056		
			Machupo virus	0.90932	0.37322		
			Rift valley fever virus	0.81849	-0.15444		
			Junin virus	0.96106	-0.12456		
	Mean vRSCU	0.85019	0.08357	950			
	3	Guanarito virus	1.09370	-0.71405			
		La Crosse virus	1.19379	-0.84942			
		Crimean-Congo virus	1.05851	-0.55100			
	Mean vRSCU	1.11533	-0.70482	529			
	4	Bunyamwera virus	1.39643	-0.85192			
		Hantaan virus	1.47251	-0.11594			
Seoul virus		1.41416	-0.37791				
Mean vRSCU		1.42770	-0.44859	522			

3. Discussions

Synonymous codons are distributed unequally and in a non-random fashion, which is referred to as codon usage bias [46]. Moreover, there are significant variations of codon usage bias among different species, and even among genes in the same organism [5, 21]. Theoretically, two major factors shape the codon usage bias: mutation pressure and translational selection [18]. Mutation pressure can result in uneven frequencies of nucleotide content, which can in turn influence codon usage bias [15, 47]. As to translational optimization, the frequent codons are usually found correlated with the population of tRNA isoacceptors [13, 15]. Thus, the frequent or optimal codons would result in more rapid protein translation due to the greater availability of tRNAs corresponding to the frequently used codons [48].

Although the replication of viruses relies on the host cell machinery, several viruses possess different a codon usage pattern to the codon usage preferences of their host [31]. For instance, the HIV-1 genome has been found to be A-rich [49]. The G-to-A hypermutation in the HIV-1 genome has been attributed to viral reverse transcriptase (RT), which lacks 3' to 5' exonuclease proofreading activity, leading to the misincorporation of nucleotides [50, 51]. In addition, the function of host enzymes of the APOBEC3 (A3) family has been found to partially contribute to a G-to-A mutation [52, 53]. Furthermore, a difference in codon usage has been observed among individual genes of HIV-1 [51]. The HIV-1 gag gene, encoded for structural protein, adopts a great difference in codon usage pattern compared to human host cells. In contrast, the HIV-1 genes

involved in the regulation of the replication cycle, tat and rev genes, have been demonstrated to be more similar to human codon usage bias [54].

In this analysis, the PCA of RSCU represented codon usage patterns of human genes and RNA viruses as a coordinate of PC1 (x) and PC2 (y) on a graph. We demonstrated that the PCA of RSCU analysis is compatible with a well-established index, CAI. This suggested that the PCA of RSCU could be used in assessing codon usage bias and comparing the difference in codon usage pattern. In particular, the PCA of RSCU allowed a comparison to be made of individual genes in the whole genome scale. Human genes possess various, different codon usage patterns, as observed in each quadrant of a graph. As mentioned earlier, there are a number of human genes located densely in the right quadrants; these genes adopt a non-optimal codon usage pattern similar to CCR genes. Although most of the RNA viruses have a non-optimal codon usage pattern, a great variation in codon usage patterns was observed among the groups of +ssRNA, dsRNA, and retroviruses. The greatest difference in the codon usage patterns belonged to rotavirus, as seen in the graph. By comparison, the rubella virus exhibited a more similar codon usage pattern to humans. These results correspond with those of another study which demonstrated that the codon usage patterns of +ssRNA viruses are closer to human than other RNA viruses, and that the lowest CAI belongs to dsRNA viruses. In more detail, rubella virus (+ssRNA) had the highest CAI at 0.773, and rotavirus had the lowest CAI at 0.683 [55].

A number of human genes with codon usage patterns similar to that of RNA viruses were found by the present study (Figure 3). The human genes with RSCUs similar to RNA viruses were retrieved and subjected to

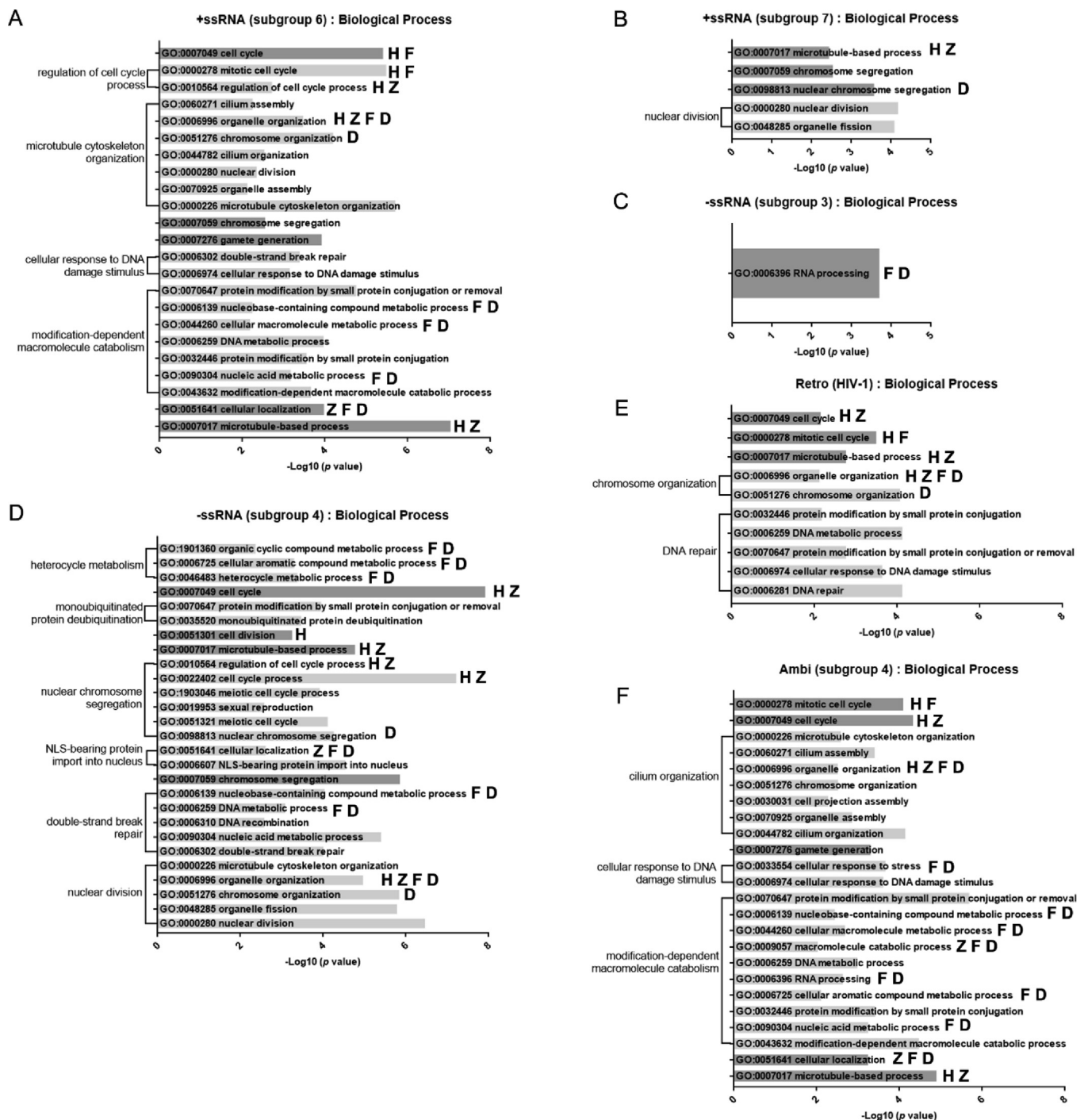


Figure 4. GO terms enrichment in the biological processes from human genes with RSCUs similar to RNA viruses. GO terms with $p \leq 0.01$ were taken as a significant enrichment and represented with $-\text{Log}_{10}(p\text{-value})$. The bold alphabets indicate GO terms that were found to be identical to the enriched GO terms of upregulated protein profiles in viral infections, determined by high-throughput quantitative proteomics; H (HIV-1), Z (Zika virus), F (influenza A virus), and D (dengue virus serotype 2).

gene ontology enrichment analysis. Interestingly, it was found that only human genes similar to groups of viruses in the right quadrant resulted in significant enrichment, namely, the cell cycle, the cell cycle regulation process, cell division, microtubule cytoskeletal organization, chromosome segregation, DNA repair, macromolecule catabolism, and cellular localization.

The number of human genes with RSCUs similar to RNA viruses retrieved from each subgroup varied from only seven to a thousand genes (Table 1). However, the number of retrieved genes did not affect the significance level or the number of enriched GO terms in Figure 4; for

instance, 898 genes for the $-\text{ssRNA}$ subgroup 3 resulted in only 1 enriched GO term, whereas 188 genes for the $+\text{ssRNA}$ subgroup 7 resulted in 5 GO terms. This suggests that the enriched GO terms did not result from a bias from the different numbers of retrieved genes among the virus groups. It is also possible that some virus groups with a limited number of retrieved genes might not have had sufficient statistical power to enable the detection of enriched GO terms. Among the retrieved human genes, only the genes that adopted a non-optimal codon usage pattern retrieved from the area where they were located densely in the right quadrants gave significant overrepresented GO terms. This

Table 2. Cell-Cycle Codon Score (CCCS) of human RNA viruses.

Groups	Gene isoform or viruses	PCA components		CCCS
		PC1 (x)	PC2 (y)	
+ssRNA	Rubella virus	-1.11152	-2.38693	-0.041109
	Hepatitis G virus	-0.20805	-1.33632	-0.012674
	Hepatitis C virus	-0.38012	-0.8688	-0.015136
	Hepatitis E virus	0.05314	-1.99911	-0.006142
	Ross river virus	0.08755	-1.48581	-0.011565
	Chikungunya virus	0.2134	-1.65931	-0.008180
	Sindbis virus	0.25123	-1.94371	-0.009176
	Eastern equine encephalitis virus	0.48833	-1.3069	-0.002974
	Western equine encephalitis virus	0.47744	-1.79137	-0.003173
	Venezuelan equine encephalitis virus	0.42568	-1.26628	-0.003554
	O'nyong nyong virus	0.56417	-1.63171	-0.001119
	Enterovirus 71	0.64889	-0.88714	-0.000267
	Human coxsackievirus A9	0.54435	-0.52889	-0.002991
	Human coxsackievirus B4	0.65238	-0.59692	-0.000998
	Echoviruses	0.57004	-0.57533	-0.002082
	Polioviruses	0.76484	-0.35123	0.001193
	Norwalk virus	0.77184	-0.52468	0.001137
	Sapporo virus	0.62754	-0.94239	0.004792
	Dengue virus type 1	0.91914	-0.19005	0.005112
	Dengue virus type 2	0.85638	0.19119	0.005403
	Dengue virus type 3	0.88466	-0.24134	0.004578
	Dengue virus type 4	0.80321	-0.02901	0.003912
	Japanese encephalitis virus	0.34886	-0.42656	-0.003955
	Murray Valley encephalitis virus	0.76079	-0.71468	0.002322
	St.Louis encephalitis virus	0.61121	-0.16698	0.000160
	West Nile virus	0.42713	-0.4084	-0.003391
	Yellow fever virus	0.55894	-0.03048	-0.000690
	Zika virus	0.46368	0.0873	-0.002752
	Kyasanur Forest disease virus	0.26948	-0.2271	-0.006086
	Omsk hemorrhagic fever virus	0.2543	-0.25169	-0.005482
	Tick-borne encephalitis virus	0.15574	-0.30465	-0.007427
	SARS coronavirus	1.38814	-1.47854	0.027135
	MERS coronavirus	1.4049	-1.39321	0.031220
	Human astrovirus	1.21324	-1.04323	0.014968
	Rhinovirus A,B and C	1.46621	-0.28559	0.013209
	Human parechovirus	1.48761	-0.44531	0.013904
	Enterovirus 68	1.50423	-0.49488	0.013644
	Hepatitis A virus	1.96492	-0.10146	0.022827
	Human coronaviruses	1.83078	-1.03063	0.024501
	-ssRNA	Borna virus	0.27188	-1.00708
Rabies virus		0.76128	-0.25367	0.000205
Mokola virus		0.79901	-0.1188	0.001436
Measles virus		0.58253	-0.23633	0.000810
Vesicular stomatitis virus		1.10766	-0.23715	0.006431
Influenza A virus H3N2		0.98353	-0.58886	0.008199
Marburg virus		1.29734	-0.55782	0.000675
Ebola viruses		1.0992	-0.67494	0.008390
Influenza B virus		1.28533	-0.26627	0.014146
Human parainfluenza virus type 1		1.09898	-0.58691	0.015018
Human parainfluenza virus type 3		1.48386	-0.16452	0.020038
Human parainfluenza virus type 2		1.25962	-0.71004	0.016267
Mumps virus		0.97594	-1.15319	0.009273
Nipah virus		1.21851	-0.10109	0.011326
Hendra virus		1.06209	-0.09805	0.009457
Respiratory syncytial virus		1.4689	0.17766	0.020368
Metapneumovirus		1.35819	-0.42703	0.016291
Influenza C virus		1.72634	-0.54316	0.021730

(continued on next page)

Table 2 (continued)

Groups	Gene isoform or viruses	PCA components		CCCS
		PC1 (x)	PC2 (y)	
dsRNA	Colorado tick fever virus	1.29334	-1.66493	-0.005517
	Mammalian orthoreovirus	1.05591	-1.47831	0.006214
	Human picobirnavirus	0.95531	-1.1807	0.003846
	Rotaviruses	2.2321	-1.29556	0.028856
Retro	HTLV-1	0.25781	-0.96125	-0.006770
	HTLV-2	0.11613	0.15063	-0.007187
	HIV-1	1.41668	-0.26596	0.008663
	HIV-2	0.94712	-0.34939	0.007142
Ambi	Sin nombre virus	0.25123	-1.94371	0.017737
	Lymphocytic choriomeningitis virus	0.77488	0.08308	0.006150
	Lassa virus	0.78719	0.24056	0.007772
	Machupo virus	0.90932	0.37322	0.008089
	Rift valley fever virus	0.81849	-0.15444	0.003661
	Junin virus	0.96106	-0.12456	0.009001
	Guanarito virus	1.0937	-0.71405	0.009082
	La Crosse virus	1.19379	-0.84942	0.007769
	Crimean-Congo virus	1.05851	-0.551	0.006304
	Bunyamwera virus	1.39643	-0.85192	0.018548
	Hantaan virus	1.47251	-0.11594	0.015166
	Seoul virus	1.41416	-0.37791	0.015762

Table 3. List of CCR and NCCR genes.

	Gene symbols	Descriptions	References
CCR	TRA2B	transformer 2 beta homolog	[70, 71, 72, 73]
	TOP1	topoisomerase	
	E2F5	E2F transcription factor 5	
	H2AFV	H2A histone family member V	
	ANP32E	acidic nuclear phosphoprotein 32 family member E	
	STAG1	stromal antigen 1	
	USP7	ubiquitin specific peptidase 7	
	EZH2	enhancer of zeste 2 polycomb repressive complex 2 subunit	
	RBBP7	RB binding protein 7, chromatin remodeling factor	
	DDX5	DEAD-box helicase 5	
	GTF2F2	general transcription factor IIF subunit 2	
	GARS	glycyl-tRNA synthetase	[11]
	TARS	threonyl-tRNA synthetase	
EPRS	glutamyl-prolyl-tRNA synthetase		
NCCR	EFHD2	EF-hand domain family member D2	[74]
	ZNF433	zinc finger protein 433	
	STAG3	stromal antigen 3	
	LMNA	lamin A/C	
	E2F4	E2F transcription factor 4	
	HMGA1	high mobility group AT-hook 1	
	YPEL1	yippee like 1	
	SET	SET nuclear proto-oncogene	
	DDX46	DEAD-box helicase 46	
	EZH1	enhancer of zeste 1 polycomb repressive complex 2 subunit	
	HMGA2	high mobility group AT-hook 2	
	SAE1	SUMO1 activating enzyme subunit 1	
	FHAD1	forkhead associated phosphopeptide binding domain 1	
	GAPDH	glyceraldehyde-3-phosphate dehydrogenase	[11]
	WARS	tryptophanyl-tRNA synthetase	

suggested that the difference in codon usage bias in human genes might have specific functions. The contribution of non-optimal codon usage bias in human genes on the regulation of protein expression has been investigated in previous research [11, 56, 57]. One study on CCR genes

revealed that the non-optimal codon usage pattern generates the oscillation in protein expression during cell cycle progression [11]. We demonstrated that the RSCUs of RNA viruses were similar to the RSCUs of CCR genes, using both the CCCS calculation and the PCA of RSCU. The

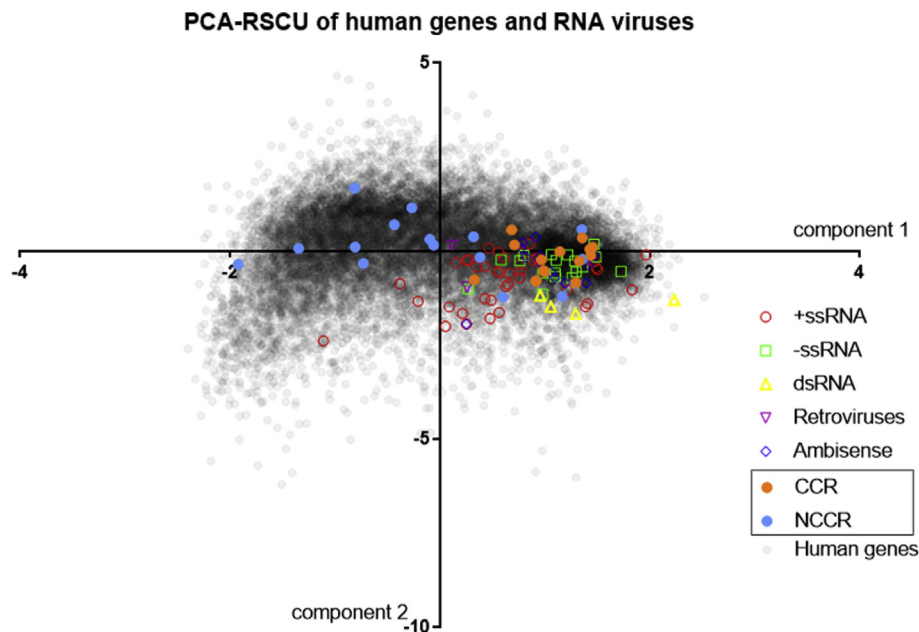


Figure 5. The graph shows the RSCUs of human cell cycle regulated (CCR) genes and non-cell cycle regulated (NCCR) genes, compared with the RSCUs of RNA viruses.

results showed that $-ssRNA$, $dsRNA$, HIV-1, HIV-2, ambisense viruses, and a few viruses in the $+ssRNA$ group exhibited non-optimal codon usage similar to that of the CCR genes. This suggests that despite having a non-optimal codon usage bias, viral genes might be efficiently expressed during the specific phase of the cell cycle correlated with the available tRNA population in that period.

The tRNA population is tissue specific and varies with cellular conditions [58]. Alteration of the tRNA population depends on the level of aminoacyl tRNA synthetase and cellular ATP concentration [59]. In yeast cells, oscillation of the aminoacyl tRNA synthetase and ATP during the cell cycle has been found to result in an increase in tRNA levels in the G2/M phase, but a low tRNA level was observed toward the end of the G1 phase [11]. Therefore, with a low-charged tRNA concentration, the genes expressed during G1 prefer optimal codon usage bias, whereas the genes with non-optimal codon usage bias are highly expressed in the other phases of the cell cycle with a high-charged tRNA concentration [11]. The availability of charged tRNAs during the cell cycle may regulate protein translation in a codon usage-specific manner. Several studies have revealed viral subversion of the cell cycle by arresting via various mechanisms to generate the resources and favorable environment for viral replication and viral protein production [60]. Cell cycle arrest has been observed in both DNA and RNA virus infections. In some RNA viruses, cell cycle arrest at a specific phase may lead to an increase in viral protein translation [61]. During the G2/M phase, the expression of many proteins has been found to fluctuate by arresting at G2/M, viruses may use this mechanism to regulate protein expression [62]. Another study found that HIV-1 was more transcriptionally active during the G2 phase, and that arresting of the cell cycle may help limit the host immune response [63]. Moreover, an HIV-1 infection also causes an alteration at the cellular tRNA level [34]. As to avian coronavirus infections, G2/M arrest has been found with an increased viral protein expression [64]. Furthermore, rotavirus infection arrests the cell cycle in the S/G2 phase, favoring viral protein expression [61], while influenza A virus infection arrests the cell cycle in the G0/G1 phase, resulting in increased viral protein expression [65, 66]. These findings suggest that viruses may manipulate the cell cycle and cellular translation machinery to create available tRNA population favoring the viral codon usage pattern.

Several GO terms of human genes with codon usage patterns similar to RNA viruses have been found by previous studies to be identical to the

GO terms of upregulated protein profiles in viral infections. Global proteomic and phosphoproteomic changes in HIV-1 infected CD4+ T cells revealed that HIV-1 affected transcriptional and translational regulation, and targeted RNA or protein degradation in order to modulate biological processes (including signal transduction, cell cycle, metabolic processes, and the immune system) [42]. In addition, a study of ZIKV infected human neurospheres also found an upregulation profile of proteins involving cell cycle arrest. This resulted in an alteration of the cell cycle in order to regulate the transcription and translation of the host cells [44]. IAV infection targeted several cellular pathways to favor its replication, including aminoacyl-tRNA biosynthesis, glycolysis, fatty acid biosynthesis, and spliceosome [43]. Moreover, regulated proteins and phosphoproteins in DENV-infected cells were related to cellular macromolecule biosynthesis, RNA splicing, chromatin modification, and cell stress response, and these regulations help facilitate viral protein expression [45]. This suggests that human genes with codon usage patterns similar to that of viruses may be upregulated at the translational level in a viral infection. It is unclear whether viruses evolved to regulate the translational machinery in order to accommodate the codon usage pattern that had already been shaped by mutational pressure, or whether they adapted their codon usage pattern to match the cellular translational machinery condition in infected cells. These two possibilities are not mutually exclusive. Further studies are required to gain more insight into this new aspect of the virus-host interaction.

4. Materials and methods

4.1. Protein coding sequences of human genomes

The protein coding sequences of human genomes were downloaded from the GENCODE database (version 26) in FASTA format [67]. The data set provided the nucleotide sequences of the coding transcripts on the reference chromosomes, including multiple transcript variants for each gene. Thus, only the major transcript variants were selected for analysis. The major transcript variant is the longest transcript variant of the gene with a complete ORF. For each gene, one representative as a major transcript variant with the longest sequence length was selected using a custom python script (Supplementary File 4). The ORFs of the protein coding sequences were rechecked by the ORFfinder tool at the

website <https://www.ncbi.nlm.nih.gov/orffinder> (NCBI) before performing further analysis. The transcript variants with an incomplete ORF were excluded and substituted by alternative variants. A list of the selected major transcripts variants is given in Supplementary File 1.

4.2. Sequences of human RNA virus genomes

The data set for the sequences of the human RNA virus genome utilized by a previous study was used [55]. A total of 77 RNA viruses that can cause diseases in humans were selected. The protein coding sequences of those RNA viruses were downloaded in FASTA format from the Nucleotide database (RefSeq, NCBI). For each virus species, the protein coding sequences from different isolations of the same virus species were downloaded. The sequences were selected based on their availability in the database (some viral protein coding sequences were only available in a few numbers). Any coding sequence with unidentified nucleotides that could not be translated or with an incomplete ORF was excluded. The selected human RNA viruses were categorized by their family, genus, and genome polarity. They comprised 39 positive-sense single strand RNA (+ssRNA) viruses; 18 negative-sense single strand RNA (-ssRNA) viruses; 4 double strand RNA (dsRNA) viruses; 4 retroviruses (retro); and 12 ambisense (+/-, ambi) viruses. Supplementary File 5 presents a list of the human RNA viruses used in this study, the accession numbers of the sequences, and the number of coding sequences of each virus.

4.3. The RSCU analysis

The RSCU is the ratio of the observed frequency of the codon in a gene to the expected frequency of the codon under the condition that all the synonymous codons are equally used. The three stop codons (TAA, TAG, TGA), Met (ATG), and Trp (TGG) were excluded from the analysis. The observed frequency of the codons in the genes was counted. The FASTA sequences were parsed, and the codons of the coding sequences for each transcript variant were counted by a python script and the Biopython library (Python version 3.5.2, with Biopython version 1.66; for the scripts, see Supplementary File 6). Then, the RSCU was calculated as follows:

$$RSCU_i = \frac{X_i}{\frac{1}{n} \sum_{i=1}^n X_i}$$

where n is the number of synonymous codons ($1 \leq n \leq 6$) for the amino acid, and X_i is the number of occurrences of codon i . The synonymous codons with RSCU values greater than 1.0 had a positive codon usage bias and were defined as abundant codons, while those with RSCU values less than 1.0 had a negative codon usage bias and were defined as less-abundant codons. In the case of RSCU values that were exactly 1.0, it meant that there was no codon usage bias, and the codons were chosen equally [68]. The RSCU of human RNA viruses were calculated using the CAIcal server, which is available at <http://genomes.urv.es/CAIcal> [37]. The multiple protein coding sequences from different isolations of the virus in the same species were submitted to the CAIcal server for the RSCU calculation. After that, the average RSCU was calculated to represent the RSCU of each RNA virus species. The RSCUs of the human genes and RNA viruses are provided in Supplementary File 1.

4.4. The calculation of CAI

The CAI of a specific gene was calculated using the CAI calculator on the CAIcal server [37]. The reference human codon usage table was obtained from the Codon Usage Database (<http://www.kazusa.or.jp/codon/>) [69].

4.5. PCA of RSCUs

The RSCUs of the protein coding sequences of 20,190 human genes and 77 human RNA viruses were input to the PCA. The PCA was performed using PASW Statistics for Windows, version 18 (SPSS Inc., Chicago, Ill., USA). The Kaiser–Meyer–Olkin Measure of Sampling Adequacy test (KMO–MSA) was also analyzed. The overall KMO–MSA was 0.799, which was greater than a cut off of 0.5, indicating that the sample size was adequate. The principle components were successfully extracted using covariance matrix and Quartimax rotation, which reduced the high dimensions of the dataset to a smaller number of dimensions. The selection of the significant components was based on a scree plot and the proportions of variance. The scree plot was a plot of the component numbers and eigenvalues; only the first two components had eigenvalues greater than 1, and they accounted for 45.5% of the total variance. The RSCU of the gene was represented by the coordinate of PC1 and PC2 (x , y) on the graph. The PCAs of the RSCU graphs were plotted and analyzed using GraphPad Prism 7 (GraphPad Software, Inc., CA).

4.6. The CCCS calculation

The CCCS of a gene evaluates the similarity of the codon usage between that of a specific gene and a set of cell cycle regulated human genes. The calculation has been previously described [11]. Briefly, the CCCS is the sum of the codon preference (CP) values of the cell cycle regulated human genes (top-600 set) over all codons in the coding sequence of a gene, normalized by the length of the cDNA. The CCCS of a specific gene was calculated as follows:

$$CCCS(g) = \frac{\sum_{\text{codon}(g)} CP^{\text{top-600}}(\text{codon})}{\text{length}(g)}$$

where g is every codon of a gene, and $CP^{\text{top-600}}(\text{codon})$ is the CP in the top-600 gene set (see cited reference, Table 1).

4.7. GO enrichment

On the PCA of the RSCU graph, the human genes located in the same area of RNA viruses were taken as human genes with RSCUs similar to RNA viruses. To select these human genes, the average RSCU of each group of RNA viruses was initially calculated. However, a great variation of RSCU was observed in each group of RNA virus categorized by the nucleic acid types of their genomes. Hence, the viruses in each group were divided into subgroups by their RSCU before calculation of the average PC1 (x) and PC2 (y); these RNA-virus subgroups are listed in Table 1. Subsequently, the average RSCU of the RNA viruses in each subgroup (mean vRSCU, coordinate $[a, b]$) were set as the circle center of a circle with a radius of 0.3 units (Figure 3). The radius was calculated to be minimal, to not exceed the standard deviation of the distance between the circle center and the human genes, and to cover most of the viruses in each subgroup. The human genes located within the circle were taken as the human genes with RSCUs similar to RNA viruses. The distance (r) from the circle center (a, b) was measured as follows:

$$(a-x)^2 + (b-y)^2 = r^2$$

where a and b are the coordinates of the mean vRSCU of each subgroup of RNA viruses, while x and y are the coordinates of the RSCUs of the human genes.

The human genes with an RSCU similar to each group of RNA viruses were analyzed for GO enrichment using Go-TermFinder, and REVIGO was used to categorize the redundant GO terms [39, 40]. The whole genome of *Homo sapiens* was used as a background list. The over-represented GO terms in the biological process were investigated in both analyses, with $p \leq 0.01$ taken as a significant enrichment.

4.8. Statistical analysis

The simple linear regression analysis and Pearson correlation coefficient (PCC) were determined using GraphPad Prism 7 with $p < 0.05$ was taken as significant.

Declarations

Author contribution statement

Kunlakanya Jitobaom: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Supinya Phakaratsakul: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Thanayaporn Sirihongthong: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data.

Sasithorn Chotewutmontri, Prapat Suriyaphol: Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Ornpreya Suptawiwat: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Prasert Auewarakul: Conceived and designed the experiments; Analyzed and interpreted the data.

Funding statement

This work was supported by Postdoctoral Fellowship, Mahidol University, Thailand (Grant no. R01612001, 2017), The Royal Golden Jubilee Ph.D. Program, Thailand (Grant No. PHD/0030/2556), Siriraj Graduate Scholarship, Mahidol University, and Thailand Research Fund (Grant No. IRN60W0002).

Competing interest statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at <https://doi.org/10.1016/j.heliyon.2020.e03915>.

References

- F.H.C. Crick, L. Barnett, S. Brenner, R.J. Watts-Tobin, General nature of the genetic code for proteins, *Nature* 192 (1961) 1227–1232.
- J.S. Trupin, F.M. Rottman, R.L. Brimacombe, P. Leder, M.R. Bernfield, M.W. Nirenberg, RNA codewords and protein synthesis VI. on the nucleotide sequences of degenerate codeword sets for isoleucine, tyrosine, asparagine and lysine, *Proc. Natl. Acad. Sci. U. S. A.* 53 (1965) 807–811.
- R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Pavé, Codon catalog usage and the genome hypothesis, *Nucleic Acids Res.* 8 (1980) r49–r62.
- R. Grantham, C. Gautier, M. Gouy, Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type, *Nucleic Acids Res.* 8 (1980) 1893–1912.
- Y. Nakamura, T. Gojobori, T. Ikemura, Codon usage tabulated from international DNA sequence databases: status for the year 2000, *Nucleic Acids Res.* 28 (2000) 292, 292.
- M. Staehelin, Isoacceptor tRNAs, in: E.K.F. Bautz, P. K. H. Kersten (Eds.), *Regulation of Transcription and Translation in Eukaryotes*, 24, Springer, Berlin, Heidelberg, 1973, pp. 26–28.
- F.H.C. Crick, Codon—anticodon pairing: the wobble hypothesis, *J. Mol. Biol.* 19 (1966) 548–555.
- G. Varani, W.H. McClain, The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems, *EMBO Rep.* 1 (2000) 18–23.
- E. Sochacka, E. Lodyga-Chruscinska, J. Pawlak, M. Cypryk, P. Bartos, K. Ebenryter-Olbinska, et al., C5-substituents of uridines and 2-thiouridines present at the wobble position of tRNA determine the formation of their keto-enol or zwitterionic forms - a factor important for accuracy of reading of guanosine at the 3'-end of the mRNA codons, *Nucleic Acids Res.* 45 (2017) 4825–4836.
- C. Yarian, H. Townsend, W. Czestkowski, E. Sochacka, A.J. Malkiewicz, R. Guenther, et al., Accurate translation of the genetic code depends on tRNA modified nucleosides, *J. Biol. Chem.* 277 (2002) 16391–16395.
- M. Frenkel-Morgenstern, T. Danon, T. Christian, T. Igarashi, L. Cohen, Y.-M. Hou, et al., Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels, *Mol. Syst. Biol.* 8 (2012) 572, 572.
- C. Gu, T.J. Begley, P.C. Dedon, tRNA modifications regulate translation during cellular stress, *FEBS Lett.* 588 (2014) 4287–4296.
- T. Ikemura, Codon usage and tRNA content in unicellular and multicellular organisms, *Mol. Biol. Evol.* 2 (1985) 13–34.
- L. Endres, P.C. Dedon, T.J. Begley, Codon-biased translation can be regulated by wobble-base tRNA modification systems during cellular stress responses, *RNA Biol.* 12 (2015) 603–614.
- E.F. Quax Tessa, J. Claassens Nico, D. Söll, J. van der Oost, Codon bias as a means to fine-tune gene expression, *Mol. Cell* 59 (2015) 149–161.
- S.L. Chen, W. Lee, A.K. Hottes, L. Shapiro, H.H. McAdams, Codon usage between genomes is constrained by genome-wide mutational processes, *Proc. Natl. Acad. Sci. U. S. A.* 101 (2004) 3480–3485.
- R.D. Knight, S.J. Freeland, L.F. Landweber, A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes, *Genome Biol.* 2 (2001).
- I.S. Belalov, A.N. Lukashov, Causes and implications of codon usage bias in RNA viruses, *PLoS One* 8 (2013), e56642.
- G.M. Jenkins, E.C. Holmes, The extent of codon usage bias in human RNA viruses and its evolutionary origin, *Virus Res.* 92 (2003) 1–7.
- V.V. Khurstalev, T.A. Khurstaleva, N. Sharma, R. Giri, Mutational pressure in Zika virus: local ADAR-editing areas associated with pauses in translation and replication, *Front. Cell Infect. Microbiol.* 7 (2017) 44, 44.
- C.G. Kurland, Codon bias and gene expression, *FEBS Lett.* 285 (1991) 165–169.
- G. Marais, L. Duret, Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*, *J. Mol. Evol.* 52 (2001) 275–280.
- J.R. Powell, K. Dion, Effects of codon usage on gene expression: empirical studies on *Drosophila*, *J. Mol. Evol.* 80 (2015) 219–226.
- M.A. Sørensen, C.G. Kurland, S. Pedersen, Codon usage determines translation rate in *Escherichia coli*, *J. Mol. Biol.* 207 (1989) 365–377.
- T. Ikemura, Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoacceptor transfer RNAs, *J. Mol. Biol.* 158 (1982) 573–597.
- D. Kotlar, Y. Lavner, The action of selection on codon bias in the human genome is related to frequency, complexity, and chronology of amino acids, *BMC Genom.* 7 (2006) 67, 67.
- D.B. Goodman, G.M. Church, S. Kosuri, Causes and effects of N-terminal codon bias in bacterial genes, *Science* 342 (2013) 475–479.
- T. Tuller, Y.Y. Waldman, M. Kupiec, E. Ruppin, Translation efficiency is determined by both codon bias and folding energy, *Proc. Natl. Acad. Sci. U. S. A.* 107 (2010) 3645–3650.
- S. Pechmann, J.W. Chartron, J. Frydman, Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo, *Nat. Struct. Mol. Biol.* 21 (2014) 1100–1105.
- T. Zhou, W. Gu, J. Ma, X. Sun, Z. Lu, Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses, *Biosystems* 81 (2005) 77–86.
- F. van Hemert, A.C. van der Kuyl, B. Berkhout, Impact of the biased nucleotide composition of viral RNA genomes on RNA structure and codon usage, *J. Gen. Virol.* 97 (2016) 2608–2619.
- A. Kofman, M. Graf, A. Bojak, L. Deml, K. Bieler, A. Kharazova, et al., HIV-1 gag expression is quantitatively dependent on the ratio of native and optimized codons, *Tsitologiya* 45 (2003) 86–93.
- J. Kypr, J.A.N. Mrázek, Unusual codon usage of HIV, *Nature* 327 (1987) 20, 20.
- A. van Weringh, M. Ragonnet-Cronin, E. Pranckeviciene, M. Pavon-Eternod, L. Kleiman, X. Xia, HIV-1 modulates the tRNA pool to improve translation efficiency, *Mol. Biol. Evol.* 28 (2011) 1827–1834.
- M. Pavon-Eternod, A. David, K. Dittmar, P. Berglund, T. Pan, J.R. Bennink, et al., Vaccinia and influenza A viruses select rather than adjust tRNAs to optimize translation, *Nucleic Acids Res.* 41 (2012) 1914–1921.
- D. Desai, K. Zhang, S. Barik, A. Srivastava, M.E. Bolander, G. Sarkar, Intragenic codon bias in a set of mouse and human genes, *J. Theor. Biol.* 230 (2004) 215–225.
- P. Puigbò, I.G. Bravo, S. Garcia-Vallve, CAIcal: a combined set of tools to assess codon usage adaptation, *Biol. Direct* 3 (2008) 38.
- T. Geiger, A. Wehner, C. Schaab, J. Cox, M. Mann, Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins, *Mol. Cell. Proteomics* 11 (2012).
- E.I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J.M. Cherry, et al., GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes, *Bioinformatics* 20 (2004) 3710–3715.
- F. Supek, M. Bošnjak, N. Skunca, T. Šmuc, REVIGO summarizes and visualizes long lists of gene ontology terms, *PLoS One* 6 (2011), e21800 e21800.
- S. Albers, A. Czech, Exploiting tRNAs to boost virulence, *Life (Basel)*. 6 (2016) 4.
- M. Golumbeanu, S. Desfarges, C. Hernandez, M. Quadroni, S. Rato, P. Mohammadi, et al., Proteo-transcriptomic dynamics of cellular response to HIV-1 infection, *Sci. Rep.* 9 (2019) 213.
- H.S. Smallwood, S. Duan, M. Morfouace, S. Rezinciuc, B.L. Shulkin, A. Shelat, et al., Targeting metabolic reprogramming by influenza infection for therapeutic intervention, *Cell Rep.* 19 (2017) 1640–1653.

- [44] P.P. Garcez, J.M. Nascimento, J.M. de Vasconcelos, R. Madeiro da Costa, R. Delvecchio, P. Trindade, et al., Zika virus disrupts molecular fingerprinting of human neurospheres, *Sci. Rep.* 7 (2017) 40780.
- [45] M. Miao, F. Yu, D. Wang, Y. Tong, L. Yang, J. Xu, et al., Proteomics profiling of host cell response via protein expression and phosphorylation upon dengue virus infection, *Virol. Sin.* 34 (2019) 549–562.
- [46] T. Ikemura, Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system, *J. Mol. Biol.* 151 (1981) 389–409.
- [47] G.A. Palidwor, T.J. Perkins, X. Xia, A general model of codon bias due to GC mutational bias, *PLoS One* 5 (2010), e13431.
- [48] J.F. Curran, M. Yarus, Rates of aminoacyl-tRNA selection at 29 sense codons in vivo, *J. Mol. Biol.* 209 (1989) 65–77.
- [49] F.J. van Hemert, A.C. van der Kuyl, B. Berkhout, The A-nucleotide preference of HIV-1 in the context of its structured RNA genome, *RNA Biol.* 10 (2013) 211–215.
- [50] B. Preston, B. Poiesz, L. Loeb, Fidelity of HIV-1 reverse transcriptase, *Science* 242 (1988) 1168–1171.
- [51] A.C. van der Kuyl, B. Berkhout, The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus, *Retrovirology* 9 (2012) 92.
- [52] P. Jern, R.A. Russell, V.K. Pathak, J.M. Coffin, Likely role of APOBEC3G-mediated G-to-A mutations in HIV-1 evolution and drug resistance, *PLoS Pathog.* 5 (2009), e1000367.
- [53] H. Zhang, B. Yang, R.J. Pomerantz, C. Zhang, S.C. Arunachalam, L. Gao, The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA, *Nature* 424 (2003) 94–98.
- [54] S. Phakratsakul, T. Sirihongthong, C. Boonarkart, O. Suptawiwat, P. Auewarakul, Codon usage of HIV regulatory genes is not determined by nucleotide composition, *Arch. Virol.* 163 (2018) 337–348.
- [55] S. Phakratsakul, T. Sirihongthong, C. Boonarkart, O. Suptawiwat, P. Auewarakul, Genome polarity of RNA viruses reflects the different evolutionary pressures shaping codon usage, *Arch. Virol.* 163 (2018) 2883–2888.
- [56] M. Zhou, J. Guo, J. Cha, M. Chae, S. Chen, J.M. Barral, et al., Non-optimal codon usage affects expression, structure and function of clock protein FRQ, *Nature* 495 (2013) 111–115.
- [57] F. Zhao, C-h Yu, Y. Liu, Codon usage regulates protein structure and function by affecting translation elongation speed in *Drosophila* cells, *Nucleic Acids Res.* 45 (2017) 8484–8492.
- [58] K.A. Dittmar, J.M. Goodenbour, T. Pan, Tissue-specific differences in human transfer RNA expression, *PLoS Genet.* 2 (2006) e221.
- [59] G. Orfanoudakis, M. Baltzinger, D. Meyer, N. Befort, J.P. Ebel, J.J. Befort, et al., Cell cycle variations of dinucleoside polyphosphates in synchronized cultures of mammalian cells, *Mol. Cell Biol.* 7 (1987) 2444–2450.
- [60] S. Bagga, M.J. Bouchard, Cell cycle regulation during viral infection, in: E. Noguchi, M.C. Gadaleta (Eds.), *Cell Cycle Control: Mechanisms and Protocols*, Springer New York, New York, NY, 2014, pp. 165–227.
- [61] S. Glück, A. Buttafuoco, A.F. Meier, F. Arnoldi, B. Vogt, E.M. Schraner, et al., Rotavirus replication is correlated with S/G2 interphase arrest of the host cell cycle, *PLoS One* 12 (2017), e0179607 e0179607.
- [62] C. Davy, J. Doorbar, G2/M cell cycle arrest in the life cycle of viruses, *Virology* 368 (2007) 219–226.
- [63] E.S. Zimmerman, M.P. Sherman, J.L. Blackett, J.A. Neidleman, C. Kreis, P. Mundt, et al., Human immunodeficiency virus type 1 Vpr induces DNA replication stress in vitro and in vivo, *J. Virol.* 80 (2006) 10407–10418.
- [64] B. Dove, G. Brooks, K. Bicknell, T. Wurm, J.A. Hiscox, Cell cycle perturbations induced by infection with the coronavirus infectious bronchitis virus and their effect on virus replication, *J. Virol.* 80 (2006) 4147–4156.
- [65] Y. He, K. Xu, B. Keiner, J. Zhou, V. Czudai, T. Li, et al., Influenza A virus replication induces cell cycle arrest in G0/G1 phase, *J. Virol.* 84 (2010) 12832–12840.
- [66] W. Jiang, Q. Wang, S. Chen, S. Gao, L. Song, P. Liu, et al., Influenza A virus NS1 induces G0/G1 cell cycle arrest by inhibiting the expression and activity of RhoA protein, *J. Virol.* 87 (2013) 3039–3052.
- [67] J. Harrow, A. Frankish, J.M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, et al., GENCODE: the reference human genome annotation for the ENCODE Project, *Genome Res.* 22 (2012) 1760–1774.
- [68] P.M. Sharp, W.-H. Li, An evolutionary perspective on synonymous codon usage in unicellular organisms, *J. Mol. Evol.* 24 (1986) 28–38.
- [69] J. Athey, A. Alexaki, E. Osipova, A. Rostovtsev, L.V. Santana-Quintero, U. Katneni, et al., A new and updated resource for codon usage tables, *BMC Bioinf.* 18 (2017) 391.
- [70] A. Sigal, R. Milo, A. Cohen, N. Geva-Zatorsky, Y. Klein, I. Alaluf, et al., Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins, *Nat. Methods* 3 (2006) 525–531.
- [71] A. Sigal, R. Milo, A. Cohen, N. Geva-Zatorsky, Y. Klein, Y. Liron, et al., Variability and memory of protein levels in human cells, *Nature* 444 (2006) 643–646.
- [72] N.P. Gauthier, L.J. Jensen, R. Wernersson, S. Brunak, T.S. Jensen, Cyclebase.org: version 2.0, an updated comprehensive, multi-species repository of cell cycle experiments and derived analysis results, *Nucleic Acids Res.* 38 (2010) D699–D702.
- [73] N.P. Gauthier, M.E. Larsen, R. Wernersson, U. de Lichtenberg, L.J. Jensen, S. Brunak, et al., Cyclebase.org—a comprehensive multi-organism online database of cell-cycle experiments, *Nucleic Acids Res.* 36 (2007) D854–D859.
- [74] M.L. Whitfield, G. Sherlock, A.J. Saldanha, J.I. Murray, C.A. Ball, K.E. Alexander, et al., Identification of genes periodically expressed in the human cell cycle and their expression in tumors, *Mol. Biol. Cell* 13 (2002) 1977–2000.