Infection, Genetics and Evolution 11 (2011) 580–586

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Analysis of human P[4]G2 rotavirus strains isolated in Brazil reveals codon usage bias and strong compositional constraints

Mariela Martínez Gómez [a], Luis Fernando Lopez Tort [a], Eduardo de Mello Volotao [a], Ricardo Recarey [b], Gonzalo Moratorio [b,c], Héctor Musto [d], José Paulo G. Leite [a], Juan Cristina [b,*]

[a] Laboratório de Virologia Comparada e Ambiental, Instituto Oswaldo Cruz, FIOCRUZ, Av. Brasil 4365, Manguinhos, 21040-360 Rio de Janeiro, RJ, Brazil
[b] Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Iguá 4225, 11400 Montevideo, Uruguay
[c] Unidad de Biofísica de Proteínas, Instituto Pasteur-Montevideo, Mataojo 2020, 11400 Montevideo, Uruguay
[d] Laboratorio de Organización y Evolución del Genoma, Instituto de Biología, Facultad de Ciencias, Iguá 4225, 11400 Montevideo, Uruguay

ARTICLE INFO

Article history:
Received 14 October 2010
Received in revised form 11 January 2011
Accepted 12 January 2011
Available online 19 January 2011

Keywords:
Rotavirus A
Codon usage bias
Evolution

ABSTRACT

The *Rotavirus* genus belongs to the family *Reoviridae* and its genome consist of 11 segments of double-stranded RNA. Group A rotaviruses (RV-A) are the main etiological agent of acute viral gastroenteritis in infants and young children worldwide. Understanding the extent and causes of biases in codon usage is essential to the understanding of viral evolution. However, the factors shaping synonymous codon usage bias and nucleotide composition in human RV-A are currently unknown. In order to gain insight into these matters, we analyzed the codon usage and base composition constraints on the two genes that codify the two outer capsid proteins (VP4 [VP8*] and VP7) of 58 P[4]G2 RV-A strains isolated in Brazil and investigated the possible key evolutionary determinants of codon usage bias. The results of these studies revealed that the frequencies of codon usage in both RV-A proteins studied are significantly different than the ones used by human cells. In order to observe if similar trends of codon usage are found when RV-A complete genomes are considered, we compare these results with results found using a dataset of 10 reference strains for whom the complete codes of the 11 segments are known. Similar results were obtained using capsid proteins or complete genomes. The general correlations found between the position of each sequence on the first axis generated by correspondence analysis and the relative dinucleotide abundances indicate that codon usage in RV-A can also be strongly influenced by underlying biases in dinucleotide frequencies. CpG and GpC containing codons are markedly suppressed. Thus, the results of this study suggest that RV-A genomic biases are the result of the evolution of genome composition in relation to host adaptation and the ability to escape antiviral cell responses.

1. Introduction

Group A rotaviruses (RV-A) are the main etiological agent of acute viral gastroenteritis in infants and young children worldwide (Aoki et al., 2009; CDC, 2008). The *Rotavirus* genus belongs to the family *Reoviridae* and its genome consist of 11 double-stranded RNA (dsRNA) gene segments encoding six structural (VP) and six non-structural proteins (NSP) (Estes and Kapikian, 2007). Based on the two genes that codify the outer neutralizing capsid proteins, VP4 and VP7, a widely used binary classification system was established for RV-A that defined G (from VP7, glycoprotein) and P (from VP4, protease-cleaved protein) genotypes (Estes and Kapikian, 2007). To date, at least 25 G and 32 P genotypes have been identified (Matthijnssens et al., 2009, 2008; Collins et al., 2010; Abe et al., 2009; Ursu et al., 2009; Esona et al., 2010). Five RV-A G genotypes (G1–G4 and G9) and two P genotypes (P[8] and P[4]) are prevalent worldwide (Santos and Hoshino, 2005; Leite et al., 2008; Iturriza-Gómara et al., 2009). Different surveillance studies with RV-A-positive samples have shown that genotype P[4]G2 reemerges in Brazil in 2005, and since then has become predominant in this country (Carvalho-Costa et al., 2006; Gurgel et al., 2007; de Oliveira et al., 2008; Leite et al., 2008; Nakagomi et al., 2008; Mascarenhas et al., 2010).

Due to the degeneracy of the genetic code, most amino acids are coded by more than one codon. Synonymous codons are not used randomly, and in several organisms natural selection seems to bias codon usage toward a certain subset of optimal codons, mainly in highly expressed genes (Stoletzki and Eyre-Walker, 2007).

Two major models have been proposed to explain codon usage, the translation related model and the mutational model (Wong et al., 2010). Translational efficiency or translational accuracy bias may be due to the relationship between local tRNA abundance and

[*] Corresponding author. Tel.: +598 2 525 09 01; fax: +598 2 525 08 95.
E-mail address: cristina@cin.edu.uy (J. Cristina).

1567-1348 © 2011 Elsevier B.V. Open access under the Elsevier OA license.
doi:10.1016/j.meegid.2011.01.006

major codon preference, wherein a particular codon of an amino acid family pairs most optimally with the most abundant tRNA (Ikemura, 1982). The discrepancies of codon usage could also be due to genome compositional constraints and mutational biases (Sharp et al., 1986).

Understanding the extent and causes of biases in codon usage is essential to comprehend the interplay between viruses and the immune response (Shackelton et al., 2006). However, the factors shaping synonymous codon usage bias, like mutational pressure, nucleotide composition or translational selection are currently unknown for human RV-A.

In order to gain insight into these matters, we analyzed the codon usage and base composition constraints of VP4 [VP8*] and VP7 gene sequences of 72 P[4]G2 RV-A strains isolated in Brazil and investigated the possible key evolutionary determinants of codon usage bias. In order to observe if similar trends of codon usage are found when RV-A complete genomes are considered, we compared these results with the ones found using a dataset of reference strains from which the complete sequences of the 11 segments are known. The results of these studies revealed a significant codon usage bias and compositional constraints in the human RV-A strains studied.

2. Materials and methods

2.1. Fecal samples, viral RNA extraction and PCR amplification

A total of 72 diarrheic stool specimens were collected from 1996 to 2009 from children up to 5 years old hospitalized with acute diarrhea. These samples were obtained from children from the States of Acre (AC), Alagoas (AL), Bahia (BA), Espirito Santo (ES), Maranhão (MA), Mato Grosso do Sul (MS), Minas Gerais (MG), Pernambuco (PE), Rio de Janeiro (RJ), Rio Grande do Sul (RS) and Sergipe (SE), and were genotyped as P[4]G2 as previously described (Fischer et al., 2000; Das et al., 1994). The viral dsRNA was extracted by the glass powder method (Boom et al., 1990). The dsRNA was reverse transcribed (RT) and amplified by polymerase chain reaction (PCR) using a pair of consensus primers corresponding to a conserved nucleotide sequence of the VP7 (Gouvea et al., 1990; Das et al., 1994) or VP4 (VP8*) (Gentsch et al., 1992; Gómez et al., 2010) genes. Temperature and time conditions for PCR amplifications were performed as originally described (Gouvea et al., 1990; Gentsch et al., 1992). Distilled Milli-Q water was used as a negative control in all steps, and recommended manipulations for PCR procedures were carried out as a precaution to avoid false-positive results.

2.2. Sequencing

DNA sequencing was performed with an ABI Prism Big Dye Terminator Cycle Sequencing Ready Reaction Kit® and an ABI Prism 3730 Genetic Analyzer (both from Applied Biosystems, Foster City, CA, USA). Sequences of the VP4 [VP8]* and VP7 genes were obtained by using the same set of primers utilized in the RT-PCR. For strain names and accession numbers, see Supplementary Material, Table 1. From the initial 72 stool samples, a total of 58 VP4 [VP8]* and 60 VP7 sequences, 818 and 978 nucleotides in-length, respectively, were obtained.

2.3. Codon usage analyses

The relative synonymous codon usage (RSCU) values of each codon in each gene (VP8* or VP7) were determined in order to measure the synonymous codon usage bias (Sharp and Li, 1986). This was done using the CodonW program (available at: http://mobyle.pasteur.fr). The RSCU of P[4]G2 RV-A VP8* and VP7 genes were compared with corresponding values of human cells (International Human Genome Sequencing Consortium, 2001). The effective number of codons (ENC) and the frequency of use of G+C at synonymous variable third positions of codons (GC_3S) (excluding Met, Trp, and termination codons) were also calculated with CodonW. ENC was used to quantify the codon usage bias of an ORF (Wrigth, 1990; Comeron and Aguade, 1998). Similarly, the fraction of the G+C nucleotides not involved in the GC_3S fraction (GC_{12}) was also calculated. All these indices were also calculated using CodonW. Dinucleotides relative frequencies were also calculated using this program as implemented in the Mobyle server (http://mobyle.pasteur.fr).

2.4. Correspondence analysis (COA)

The relationship between variables and samples can be obtained using multivariate statistical analysis. COA is a type of multivariate analysis that allows a geometrical representation of the sets of rows and columns in a dataset (Wong et al., 2010; Greenacre, 1984). Each ORF is represented as a 59-dimensional vector and each dimension correspond to the RSCU value of one codon (excluding AUG, UGG and stop codons). Major trends within a dataset can be determined using measures of relative inertia and genes ordered according to their position along the axis of major inertia (Tao et al., 2009). COA was performed on the RSCU values of the ORFs studied using the CodonW program.

2.5. Statistical analysis

Correlation analysis was carried out using Spearman's rank correlation analysis method (Wessa, 2010; available at: www.wessa.net).

2.6. Sequence alignment

Sequences were aligned using the MUSCLE program (Edgar, 2004).

2.7. Comparative analysis

In order to observe if the codon usage bias found in the outer capsid proteins of P[4]G2 RV-A strains isolated in Brazil, can also be found in other genome regions or considering complete genome codes of human RV-A strains of different genotypes and isolated elsewhere, a new dataset composed of 10 human RV-A reference strains for whom the complete codes of the 11 genome segments are known was constructed. For strain names, genotypes, accession numbers and genomic constellations see Supplementary Material Table 3.

3. Results

In order to study the extent of codon usage bias in P[4]G2 RV-A isolated in Brazil, the RSCU values of the codons in VP4 [VP8*] and VP7 ORFs were calculated, and the figures obtained for these genes, comprising a dataset of 58 and 60 sequences, respectively, are shown in Table 1.

Interestingly, the frequencies of codon usage in both VP4 [VP8*] and VP7 P[4]G2 RV-A ORFs are significantly different in relation to human cells. Particularly, extremely high biased frequencies were found for UUU (Phe), UUA (Leu), GUU and GUA (Val), UCA (Ser), CCA (Pro), GCU (Ala), UAU (Tyr), CAU (His), CAA (Gln), AAU (Asn), AAA (Lys), GAA (Glu), UGU (Cys), AGA (Arg) and GGA (Gly) in both ORFs (see Table 1). As can be seen, highly preferred codons are all U/A ending, which strongly suggests that mutational bias is the

Table 1
Codon usage in P[4]G2 RV-A strains, displayed as RSCU values.

AA	Cod	HC	VP4	VP7	AA	Cod	HC	VP4	VP7	AA	Cod	HC	VP4	VP7	AA	Cod	HC	VP4	VP7
Phe	**UUU**	**0.92**	**1.80**	**1.86**	Ser	UCU	1.14	1.07	1.00	Tyr	**UAU**	**0.88**	**1.96**	**1.50**	**Cys**	**UGU**	**0.92**	**1.97**	**1.24**
	UUC	1.08	0.20	0.14		UCC	1.32	0.40	0.27		UAC	1.12	0.04	0.50		UGC	1.08	0.03	0.76
Leu	**UUA**	**0.48**	**2.72**	**2.33**		**UCA**	**0.90**	**1.96**	**3.56**	TER	UAA	**	**	**	TER	UGA	**	**	**
	UUG	0.78	0.46	0.94		UCG	0.30	0.55	0.59		UAG	**	**	**	Trp	UGG	1.00	1.00	1.00
	CUU	0.78	1.14	0.39	Pro	CCU	1.16	0.31	0.36	His	**CAU**	**0.84**	**1.65**	**2.00**	Arg	CGU	0.48	0.04	0.64
	CUC	1.20	0.44	0.37		CCC	1.28	0.29	0.00		CAC	1.16	0.35	0.00		*CGC*	*1.08*	*0.00*	*0.00*
	CUA	**0.42**	**0.89**	**1.27**		**CCA**	**1.12**	**3.38**	**2.91**	Gln	**CAA**	**0.54**	**1.42**	**1.81**		CGA	0.66	0.02	1.78
	CUG	2.40	0.35	0.70		CCG	0.44	0.01	0.72		CAG	1.46	0.58	0.19		CGG	1.20	0.41	0.64
Ile	AUU	1.08	1.79	1.45	Thr	**ACU**	**1.00**	**1.98**	**1.51**	Asn	**AAU**	**0.94**	**1.85**	**1.77**	Ser	AGU	0.90	1.99	0.33
	AUC	1.41	0.31	0.25		ACC	1.44	0.18	0.25		AAC	1.06	0.15	0.23		AGC	1.44	0.03	0.26
	AUA	0.51	0.90	1.30		ACA	1.12	1.47	1.42	Lys	**AAA**	**0.86**	**1.65**	**1.90**	Arg	**AGA**	**1.26**	**4.70**	**2.59**
Met	AUG	1.00	1.00	1.00		ACG	0.44	0.37	0.82		AAG	1.14	0.35	0.10		AGG	1.26	0.84	0.35
Val	**GUU**	**0.72**	**1.31**	**1.97**	Ala	**GCU**	**1.08**	**1.74**	**2.49**	Asp	GAU	0.92	1.57	1.22	Gly	GGU	0.64	2.03	0.98
	GUC	0.96	0.75	0.01		GCC	1.60	0.02	0.13		GAC	1.08	0.43	0.78		GGC	1.36	0.26	0.24
	GUA	**0.48**	**1.65**	**1.57**		GCA	0.92	1.74	1.11	Glu	**GAA**	**0.84**	**1.41**	**1.57**		**GGA**	**1.00**	**1.66**	**2.51**
	GUG	1.84	0.28	0.45		GCG	0.44	0.50	0.28		GAG	1.16	0.59	0.43		GGG	1.00	0.05	0.26

RSCU, relative synonymous codon usage; AA, amino acid; Cod, codons; HC, human cells; TER, termination codon. More frequent codons in both VP4 [VP8*] and VP7 with respect to human cells are shown in bold. Codon CGC (Arg) not used in VP4 [VP8*] and VP7 P[4]G2 RV-A isolated in Brazil are shown in italics.

main force shaping codon usage in these two genes. It is interesting to note that CGC (Arg) is not used in both ORFs.

In order to investigate if these P[4]G2 RV-A strain sequences display similar composition features, the ENC values were calculated for VP8* and VP7 ORFs. These values range from 35.21 to 40.49 for VP8* and from 38.97 to 41.88 for VP7 (mean ENCs values are 37.36 and 40.56 for VP4 [VP8*] and VP7, respectively). For results obtained for Brazilian strains enrolled in these studies, see Supplementary Material Table 2. Due to the fact that almost all ENC values are <40, the results obtained for the two ORFs studied reveal that codon usage in P[4]G2 RV-A is biased.

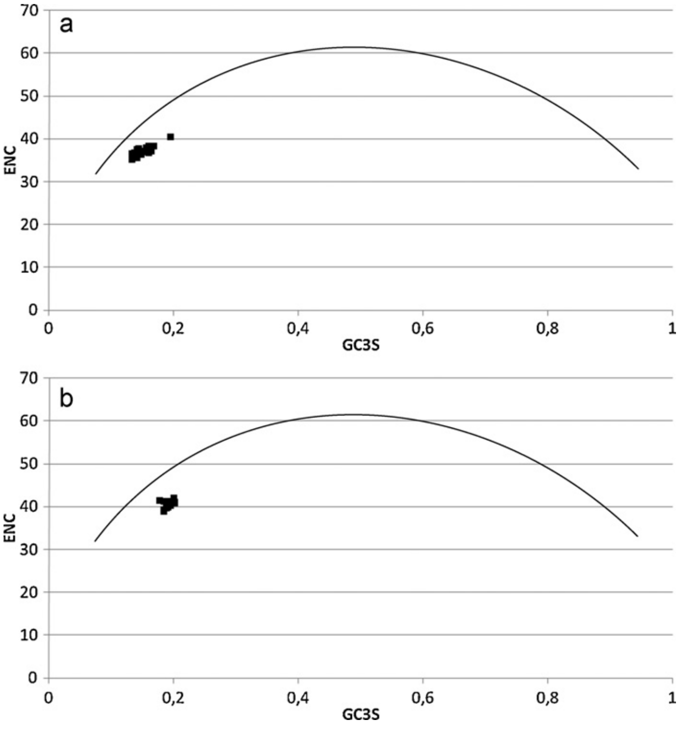

Fig. 1. Effective number of codons used in each ORF plotted against the GC_3S. The curve plots the relationship between GC_3S and ENC in absence of selection. Black square dots show the results obtained for RV-A strains. All of them lie below the expected curve. The results found for VP4 and VP7 are shown in (A) and (B), respectively.

An ENC–GC_3S plot (ENC plotted against GC_3S) can be used as a method that quantifies how far the codon usage of a gene departs from equal usage of synonymous codons (Wrigth, 1990). As shown in Fig. 1, the dotted continuous line in the plot represents a curve if codon usage is only determined by GC content at the third codon position. In other words, if GC_3S is the only determinant factor shaping the codon usage pattern, the values of ENC would fall on a continuous curve, which represents random codon usage (Jiang et al., 2007). If G+C compositional constraint influences the codon usage, then the GC_3S and ENC correlated spots would lie on or bellow the expected curve (Tsai et al., 2007). Otherwise, the codon usage bias of genes may be affected by other factors such as translational selection.

When the GC_3S values were calculated for VP4 [VP8*] and VP7 ORFs and the ENC–GC_3S plots constructed (for ENC and GC_3S values obtained for Brazilian strains enrolled in these studies, see Supplementary Material Table 2), all spots lie below and "parallel" in relation to the expected curve for both ORFs studied, indicating that the codon usage bias may be influenced by the G+C compositional constraints (see Fig. 1).

Since codon usage by its very nature is multivariate, it is necessary to analyze the data using multivariate statistical techniques (i.e. COA) in order to confirm these findings. COA is an ordination technique that identifies the major trends in the variation of the data and distributes genes along continuous axes in accordance with these trends. Moreover, it has the advantage that it does not assume that the data falls into discrete clusters and therefore can represent continuous variation accurately (Greenacre, 1984). COA creates a series of orthogonal axes to identify trends that explain the data variation, with each subsequent axis explaining a decreasing amount of the variation (Greenacre, 1984). The correlation between the position on the first axis generated by COA for each gene and the respective GC_3S values of each strain was analyzed for both VP4 [VP8*] and VP7 ORFs studied. We have found that the position of the sequences on the first axis from COA are highly correlated with the GC_3S values in both VP4 [VP8*] and VP7 ORFs ($r = 0.625$, $P < 0.0001$ and $r = -0.469$, $P < 0.001$ for VP4 [VP8] and VP7, respectively). Taking altogether, these results reveal that most of the codon usage bias is directly related to the nucleotide composition. Nevertheless, other factors may be also acting in shaping codon usage bias.

In order to analyze if the codon usage biases reported above can also be found using other genome regions or considering complete genome sequences, a new dataset was constructed composed of 10

Table 2
Codon usage in RV-A strains of different genotypes, expressed by RSCU values.

AA	Cod	HC	OC	IM	IC	NSP	Full	AA	Cod	HC	OC	IM	IC	NSP	Full
Phe	**UUU**	**0.92**	**1.53**	**1.58**	**1.52**	**1.52**	**1.51**	Ser	UCU	1.14	1.30	1.02	1.13	1.66	1.30
	UUC	1.08	0.47	0.42	0.48	0.48	0.49		UCC	1.32	0.26	0.43	0.29	0.28	0.28
Leu	**UUA**	**0.48**	**2.74**	**1.41**	**2.59**	**1.89**	**2.33**		**UCA**	**0.90**	**2.84**	**3.07**	**3.26**	**2.29**	**2.93**
	UUG	0.78	0.99	1.25	1.14	1.32	1.17		UCG	0.30	0.54	0.55	0.62	0.54	0.57
	CUU	0.78	0.47	1.37	0.60	0.92	0.70	Pro	CCU	1.16	0.61	0.26	0.43	0.90	0.54
	CUC	1.20	0.29	0.32	0.18	0.39	0.25		CCC	1.28	0.27	0.02	0.12	0.20	0.15
	CUA	**0.42**	**1.09**	**1.18**	**1.04**	**1.03**	**1.07**		**CCA**	**1.12**	**2.68**	**3.34**	**2.85**	**2.55**	**2.81**
	CUG	2.40	0.42	0.48	0.46	0.45	0.48		CCG	0.44	0.44	0.38	0.60	0.35	0.51
Ile	AUU	1.08	1.13	1.80	1.02	1.80	1.24	Thr	**ACU**	**1.00**	**1.14**	**1.35**	**1.35**	**1.64**	**1.35**
	AUC	1.41	0.22	0.30	0.23	0.27	0.24		ACC	1.44	0.24	0.12	0.24	0.24	0.23
	AUA	0.51	1.65	0.90	1.75	0.93	1.51		ACA	1.12	1.73	2.11	1.63	1.52	1.66
Met	AUG	1.00	1.00	1.00	1.00	1.00	1.00		ACG	0.44	0.89	0.43	0.78	0.61	0.76
Val	**GUU**	**0.72**	**0.95**	**1.09**	**1.36**	**1.67**	**1.33**	Ala	**GCU**	**1.08**	**1.25**	**1.57**	**1.44**	**1.34**	**1.40**
	GUC	0.96	0.33	0.71	0.38	0.32	0.37		GCC	1.60	0.26	0.32	0.33	0.17	0.28
	GUA	**0.48**	**1.77**	**1.46**	**1.46**	**1.11**	**1.43**		GCA	0.92	1.76	1.36	1.62	2.03	1.69
	GUG	1.84	0.94	0.74	0.80	0.90	0.86		GCG	0.44	0.73	0.75	0.61	0.45	0.63
Tyr	**UAU**	**0.88**	**1.46**	**1.13**	**1.47**	**1.33**	**1.43**	**Cys**	**UGU**	**0.92**	**1.35**	**1.80**	**1.39**	**1.17**	**1.31**
	UAC	1.12	0.54	0.87	0.53	0.67	0.57		UGC	1.08	0.65	0.20	0.61	0.83	0.69
TER	UAA	**	**	**	**	**	**	TER	UGA	**	**	**	**	**	**
	UAG	**	**	**	**	**	**	Trp	UGG	1.00	1.00	1.00	1.00	1.00	1.00
His	**CAU**	**0.84**	**1.69**	**1.80**	**1.66**	**1.48**	**1.62**	Arg	CGU	0.48	0.12	0.46	0.60	0.74	0.53
	CAC	1.16	0.31	0.20	0.34	0.52	0.38		*CGC*	*1.08*	*0.22*	*0.12*	*0.23*	*0.06*	*0.18*
Gln	**CAA**	**0.54**	**1.20**	**1.28**	**1.28**	**1.38**	**1.29**		CGA	0.66	0.87	0.31	0.50	0.59	0.56
	CAG	1.46	0.80	0.72	0.72	0.62	0.71		*CGG*	*1.20*	*0.29*	*0.00*	*0.15*	*0.18*	*0.18*
Asn	**AAU**	**0.94**	**1.53**	**1.38**	**1.52**	**1.56**	**1.51**	Ser	AGU	0.90	0.85	0.55	0.58	0.92	0.71
	AAC	1.06	0.47	0.62	0.48	0.44	0.49		AGC	1.44	0.22	0.39	0.11	0.31	0.21
Lys	**AAA**	**0.86**	**1.52**	**1.66**	**1.49**	**1.48**	**1.49**	Arg	**AGA**	**1.26**	**3.75**	**4.98**	**3.86**	**3.50**	**3.85**
	AAG	1.14	0.48	0.34	0.51	0.52	0.57		AGG	1.26	0.75	0.12	0.67	0.92	0.70
Asp	GAU	0.92	1.38	1.54	1.43	1.63	1.47	Gly	GGU	0.64	1.44	1.03	1.31	1.45	1.34
	GAC	1.08	0.62	0.46	0.57	0.37	0.53		GGC	1.36	0.34	0.34	0.26	0.20	0.28
Glu	**GAA**	**0.84**	**1.30**	**1.46**	**1.46**	**1.47**	**1.43**		**GGA**	**1.00**	**1.90**	**2.40**	**2.02**	**2.04**	**2.01**
	GAG	1.16	0.70	0.54	0.54	0.53	0.57		GGG	1.00	0.32	0.23	0.41	0.31	0.37

RSCU, relative synonymous codon usage; AA, amino acid; Cod, codons; HC, human cells; OC, outer capsid shell proteins; IM, intermediate protein shell; IC, inner capsid shell proteins; NSP, non-structural proteins; Full, full genome; TER, termination codon. More frequent codons with respect to human cells found in all genome regions studied are shown in bold. Frequencies sharply reduced with respect to frequencies found in human cells are shown in italics.

human RV-A reference strains, for which the complete genomes of the 11 segments are known. For strains names, genotypes, accession numbers and genomic constellations, see Supplementary Material Table 3.

By concatenation of different genome ORF's sequences, the RSCU values of the different codons were calculated for different virus regions (outer capsid shell proteins, OC, VP4+VP7; intermediate protein shell, IM, VP6; inner capid shell proteins, IC, VP1+VP2+VP3; non-structural proteins, NSP, NSP1+NSP2+NSP3+NSP4+NSP5; and full genome, VP4+VP7+VP6+VP1+VP2+VP3+NSP1+NSP2+NSP3+NSP4+NSP5, which accounts for a total of 54,318 codons). The results of these studies are shown in Table 2.

Again, the frequencies of codon usage found in different genomic regions or considering complete genomes of RV-A are significantly different in relation to human cells (see Tables 1 and 2). Highly biased frequencies were also found for the same amino acids in all genomic regions or considering full genomes (Table 2) and in agreement with the previous results found using outer capsid proteins from P[4]G2 RV-A strains isolated in Brazil. The correlation between the position on the first axis generated by COA and the respective GC_3S values of each strain was analyzed for the complete genome dataset. A high and significant correlation among the position of the sequences on the first axis of COA and the GC_3S values ($r = -0.9879$, $P < 0.01$) was also found using full, complete genomes.

It has been suggested that dinucleotide biases can affect codon bias (Tao et al., 2009). To study this possibility, the relative abundances of the 16 dinucleotides in VP8* and/or VP7 ORFs was established. The results of these studies are shown in Table 3. As can be seen, the occurrences of dinucleotides are not random and no dinucleotides is present at the expected frequencies.

In the case of VP4 [VP8*] protein, the relative abundance of CpG and GpC showed a strong deviation from the expected frequencies (i.e. 1.0) (mean ± S.D. = 0.230 ± 0.035 and 0.282 ± 0.009, respectively) and were markedly underrepresented. On the other hand, ApU and ApA are markedly over-used (mean ± S.D. = 1.951 ± 0.033 and 1.979 ± 0.04, respectively) (Table 3). Among the 16 dinucleotides, 10 are correlated with the first axis value in COA (P values <0.01, Table 3). These observations indicated that the composition of dinucleotides also determines the variation in synonymous codon usage among P[4]G2 RV-A VP4 [VP8*] ORFs. To study the possible effects of CpG and GpC under-representation on codon usage bias of VP4 [VP8*] protein, the RSCU value of the 14 codons that contain CpG and/or GpC (CCG, GCG, UCG, ACG, CGC, CGG, CGU, CGA, GCU, GCC, GCA, UGC, AGC, GGC) were analyzed. Of these triplets, 12 [CCG (mean 0.01), GCG (mean 0.50), UCG (mean 0.35), ACG (mean 0.37), CGC (mean 0.00), CGG (mean 0.41) and CGU (mean 0.04), GCC (mean 0.02), CGA (mean 0.02), UGC (mean 0.03), AGC (mean 0.03) and GGC (mean 0.26)] were markedly suppressed.

In the case of VP7 protein, again, the relative abundance of CpG and GpC showed a strong deviation from the expected frequencies (mean ± S.D. = 0.397 ± 0.014 and 0.330 ± 0.018, respectively) and were underrepresented. Interestingly, the frequencies of ApU and ApA showed a sharp deviation from the expected frequencies and again we found a markedly over-use of these dinucleotides (mean ± S.D. = 2.056 ± 0.029 and 1.948 ± 0.038, respectively) (Table 3). Among the 16 dinucleotides, seven are correlated with the position of

Table 3
Relative abundance of dinucleotides in VP4 [VP8*] and VP7 proteins from P[4]G2 RV-A Brazilian strains and summary of COA.

VP4 [VP8]									
		UU	UC	UA	UG	CU	CC	CA	CG
Mean ± S.D.[a]		1.490 ± 0.035	0.823 ± 0.024	1.665 ± 0.021	0.846 ± 0.020	0.610 ± 0.022	0.381 ± 0.012	1.157 ± 0.023	0.230 ± 0.035
Axis 1[b]	*r*	−0.261	0.384	−0.427	0.038	−0.010	0.560	0.127	0.453
	P	0.040	<0.01	<0.001	0.764	0.928	<0.0001	0.317	<0.001

VP4 [VP8]									
		AU	AC	AA	AG	GU	GC	GA	GG
Mean ± S.D.		1.951 ± 0.033	0.869 ± 0.034	1.979 ± 0.04	1.177 ± 0.032	0.790 ± 0.025	0.282 ± 0.009	1.166 ± 0.019	0.561 ± 0.017
Axis 1	*r*	−0.223	0.104	−0.469	0.584	0.422	0.422	0.628	0.191
	P	0.080	0.412	<0.001	<0.001	0.001	0.001	<0.0001	0.133

VP7									
		UU	UC	UA	UG	CU	CC	CA	CG
Mean ± S.D.[a]		1.547 ± 0.020	0.640 ± 0.011	1.808 ± 0.023	1.142 ± 0.013	0.755 ± 0.018	0.250 ± 0.016	0.987 ± 0.024	0.397 ± 0.014
Axis 1[b]	*r*	−0.249	0.209	0.227	0.180	0.345	0.781	−0.341	0.126
	P	0.054	0.105	0.080	0.164	<0.01	<0.001	<0.01	0.327

VP7									
		AU	AC	AA	AG	GU	GC	GA	GG
Mean ± S.D.		2.056 ± 0.029	1.167 ± 0.020	1.948 ± 0.038	0.595 ± 0.037	0.793 ± 0.025	0.330 ± 0.018	1.013 ± 0.020	0.572 ± 0.024
Axis 1	*r*	−0.223	−0.327	−0.025	0.285	0.419	0.093	0.101	0.320
	P	0.085	0.011	0.841	0.027	<0.01	0.471	0.429	0.013

[a] Mean values of 58 P[4]G2 RV-A strains' relative dinucleotide ratios ± standard deviation.
[b] Correlation analysis between the first axis in COA and the sixteen dinucleotides frequencies in VP4 [VP8*] and VP7 proteins is shown.

Table 4
Position of codons in each of the four major axes of COA for RV-A VP4 [VP8*] and VP7 proteins.

	Axis 1			Axis 2			Axis 3			Axis 4		
	Codon	Value	Aminoacid	Codon	Value	Aminoacid	Codon	Value	Aminoacid	Codon	Value	Aminoacid
VP4	GCC	−4.183	Ala	UGC	−7.683	Cys	CCG	−6.889	Pro	GCC	−2.634	Ala
	UGC	0.660	Cys	CCG	0.898	Pro	GCC	2.071	Ala	GGG	1.967	Gly
VP7	GUC	0.610	Val	GGG	0.212	Gly	GUC	−0.960	Val	GUC	−0.100	Val
	AGG	1.488	Arg	GUC	4.447	Val	UUC	0.337	Phe	UUC	0.279	Phe

the sequences along the first axis in COA (*P* values <0.01, Table 3). These results indicate that the composition of dinucleotides also determines the variation in synonymous codon usage among P[4]G2 RV-A VP7 ORFs. The RSCU value for the VP7 protein of the 14 codons that contain CpG and GpC (see above) revealed that six [GCG (mean 0.28), CGC (mean 0.00), GCC (mean 0.13), GCG (mean 0.28), AGC (mean 0.26) and GGC (mean 0.24)] were markedly suppressed and five [CCG (mean 0.73), UCG (mean 0.59), CGG (mean 0.64), CGU (mean 0.64) and UGC (mean 0.75)] were slightly suppressed.

Besides, the position of each codon in each of the four major axes of COA was determined for both proteins studied. For VP4 [VP8*] ORFs, the first major axis accounted for the 28.67% of the observed variation, while the second, third and fourth axis accounted for the 21.57%, 18.56% and 12.39%, respectively. For VP7 ORFs, the first major axis accounted for the 66.00% of the observed variation; the second, third and fourth major axis accounted for the 14.82%, 8.09% and 2.40% of the observed variation, respectively. Table 4 shows the codons for which the maximum and minimum values were obtained for each of the axes studied (i.e. the most divergent codons values), indicating a strong bias in their use by both VP4 and VP7 proteins. As can be seen, the most divergent triplets tend to be GC-rich (considering the two ORFs, G+C explains 19/24 positions of these codons). Again, this can be explained in terms of a strong mutational bias.

In order to observe if the same results found using outer capsid proteins of P[4]G2 RV-A strains can be found using complete genomes, the same studies were repeated using a dataset of full complete genomes (for strains, accession numbers and genomic constellations, see Supplementary Material Table 3). The results of these studies are shown in Supplementary Material Table 4. Again, the relative abundance of CpG and GpC showed a strong deviation from the expected frequencies (i.e. 1.0) (mean ± S.D. = 0.360 ± 0.021) and (0.468 ± 0.038, respectively) and were markedly underrepresented. The frequencies of ApU and ApA also showed a sharp deviation from the expected frequencies and were markedly over-used (mean ± S.D. = 1.907 ± 0.069 and 2.089 ± 0.048, respectively). Among the 16 dinucleotides, seven are correlated with the position of the sequences along the first axis in COA (*P* values <0.01, Supplementary Material Table 4). Taking all these results together, it is possible to observe that the composition of dinucleotides also determines the variation in synonymous codon usage in the complete sequences of human RV-A.

4. Discussion

The results of these studies revealed that codon usage for VP4 [VP8*] and VP7 in P[4]G2 RV-A is quite different from that of human genes (see Table 1). Moreover, this is also observed considering all different genome regions or complete, full genome codes (see Table 2). This is in agreement with results found for other viruses such as human immunodeficiency virus 1 (HIV-1) (Grantham and Perrin, 1986; Kypr and Mrazek, 1987) and hepatitis A virus (Aragones et al., 2008). In other RNA viruses, like poliovirus or foot-and-mouth disease virus (FMDV) the codon usage is very similar to that of their hosts, implying competence for tRNAs among virus and host (Sanchez et al., 2003). In these cases, competition is avoided by the induction of cellular shutoff of protein synthesis through carboxy cleavage of translation initiation factor 4G (eIF4G) by 2A and L proteases, respectively (Racaniello, 2001).

Early during the infection process RV-A also takes over the host translation machinery of the cell, causing a shutoff of cell protein synthesis, although by a different mechanism of picornaviruses. After RV-A infection, the translation initiation factor 2α (eIF2α) becomes phosphorylated and remains in this state throughout the virus replication cycle, leading to a further inhibition of cell protein synthesis (Montero et al., 2008). However, recent studies have shown that under these restrictive conditions, the viral proteins and some cellular proteins are efficiently translated (Montero et al., 2008). Whether this extremely different strategy in codon usage among RV-A and human cells is related to this fact is currently unknown, but might allow RV-A to compete successfully for translation of viral RNAs.

We analyzed synonymous codon usage and nucleotide compositional constraints in VP4 [VP8*] and VP7 genes of P[4]G2 RV-A and compare the results found with a dataset of RV-A reference strains from which the complete sequences for the 11 segments were previously known. Interestingly, in contrary to previous results found for other viruses such H5N1 influenza A Virus (mean ENC = 50.91) (Ahn et al., 2006; Zhou et al., 2005); SARS (mean ENC = 48.99) (Zhao et al., 2008); FMDV (mean ENC = 51.42) (Zhong et al., 2007); classical swine fever virus (mean ENC = 51.7) (Tao et al., 2009) and duck enteritis virus (mean ENC = 52.17) (Jia et al., 2009), the ENC values found for human P[4]G2 RV-A are comparatively low (mean ENC values of 37.36 and 40.56 for VP8* and VP7, respectively). Moreover, when the complete genomes are studied (accounting for 54,318 codons), the mean ENC value obtained is 41.60. This indicates that the overall extent of codon usage bias in RV-A genomes is significant.

We observed a general correlation between codon usage bias and base composition was observed, since all spots in the ENC–GC_3S plot lie below the curve of the predicted values (Fig. 1). Highly significant correlations between the first axis of COA and GC_3S values were obtained for both outer surface protein shells. Moreover, concatenation of complete sequences of the 11 segments of 10 reference human RV-A strains also show this significant correlation. All these results strongly suggest that mutational pressure is an important factor in determining codon usage bias in human RV-A. Nevertheless, we cannot completely discard other factors that may also account for codon usage bias.

The frequencies of dinucleotides were not random and no dinucleotides was present at the expected frequencies for both ORFs studied (VP8* and VP7, see Table 3). The same results are found using the complete genome dataset (Supplementary Material Table 4). CpG and GpC containing codons are markedly suppressed (see Tables 1 and 2). Marked CpG deficiency has been also observed in Coronaviruses (Woo et al., 2007), vertebrate-infecting members of the family *Flaviviridae* (Lobo et al., 2009), poliovirus (Rothberg and Wimmer, 1981) and other RNA viruses (Karlin et al., 1994). The CpG deficiency was proposed to be related to the immunostimulatory properties of unmethylated CpG, which were recognized by the host's innate immune system as a pathogen signature (Shackelton et al., 2006; Woo et al., 2007). This is now known to be triggered by the intracellular Pattern Recognition Receptor (PRR) Toll-like 9 (TLR9), which recognizes CpG-unmethylated DNA, and triggers several immune response pathways (Dorn and Kippenberger, 2008). Since the vertebrate immune system relies on unmethylated CpG recognition in DNA molecules as a sign of infection, and CpG under-representation in RNA viruses is exclusively observed in vertebrate viruses (Lobo et al., 2009), it is reasonable to suggest that a TLR9-like mechanism exists in the vertebrate immune system which recognizes CpG when in RNA context (such as in the genomes of RNA viruses) and triggers immune responses (Lobo et al., 2009). Moreover, recent studies on influenza A viruses, which have originated from an avian reservoir and have been infecting human hosts since 1918, were selected under strong pressure to reduce the frequency of CpG in its genome (Greenbaum et al., 2008).

The results of this work provide a basic knowledge of the mechanisms that give rise to codon usage bias in human RV-A and are also useful in understanding the processes involved in RV-A evolution. Further studies will be needed to reveal more about RV-A viral genome.

Acknowledgements

We acknowledge support by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES, Brazil) and Universidad de la República (Uruguay) through Project CAPES/UDELAR 006/08. We also acknowledge support by PEDECIBA (Uruguay), Agencia Nacional de Investigación e Innovación (ANII) (Uruguay) through a Ph.D. and a M.Sc. fellowship to GM and RR, respectively, and project Fondo Clemente Estable FCE2007_722. We acknowledge support by the National Council for Scientific and Technological Development (CNPq, Brazil). We thank anonymous reviewers of previous versions of this work for important comments.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.meegid.2011.01.006.

References

Abe, M., Ito, N., Morikawa, S., Takasu, M., Murase, T., Kawashima, T., Kawai, Y., Kohara, J., Sugiyama, M., 2009. Molecular epidemiology of rotaviruses among healthy calves in Japan: isolation of a novel bovine rotavirus bearing new P and G genotypes. Virus Res. 144, 250–257.

Ahn, I., Jeong, B.J., Bae, S.E., Jung, J., Son, H.S., 2006. Genomic analysis of influenza A viruses, including avian flu (H5N1) strains. Eur. J. Epidemiol. 21, 511–519.

Aoki, S.T., Settembre, E.C., Trask, S.D., Greenberg, H.B., Harrison, S.C., Dormitzer, P.R., 2009. Structure of rotavirus outer-layer protein VP7 bound with a neutralizing Fab. Science 324, 1444–1447.

Aragones, L., Bosch, A., Pinto, R.M., 2008. Hepatitis A virus spectra under the selective pressure of monoclonal antibodies: codon usage constraints limit capsid variability. J. Virol. 82, 1688–1700.

Boom, R., Sol, C.J., Salimans, M.M., Jansen, C.L., Wertheim-van Dillen, P.M., van der Noordaa, J., 1990. Rapid and simple method for purification of nucleic acids. J. Clin. Microbiol. 28, 495–503.

Carvalho-Costa, F.A., Assis, R.M., Fialho, A.M., Bóia, M.N., Alves, D.P.D., Martins, C.M.M.A., Leite, J.P.G., 2006. Detection and molecular characterization of group A rotavirus from hospitalized children in Rio de Janeiro, Brazil, 2004. Mem. Inst. Oswaldo Cruz 101, 291–294.

Centers for Disease Control, 2008. Rotavirus surveillance—worldwide, 2001–2008. Morb. Mortal. Wkly. Rep. 57, 1255–12557.

Collins, P.J., Martella, V., Buonavoglia, C., O' Shea, H., 2010. Identification of a G2-like porcine rotavirus bearing a novel VP4 type, P[32]. Vet. Res. 41, 73.

Comeron, J.M., Aguade, M., 1998. An evaluation of measures of synonymous codon usage bias. J. Mol. Evol. 47, 268–274.

Greenacre, M., 1984. Theory and Applications of Correspondence Analysis. Academic Press, London.

Das, B.K., Gentsch, J.R., Cicirello, H.G., Woods, P.A., Gupta, A., Ramachandran, M., Kumar, R., Bhan, M.K., Glass, R.I., 1994. Characterization of rotavirus strains from newborns in New Delhi, India. J. Clin. Microbiol. 32, 1820–1822.

de Oliveira, A.S.L., Mascarenhas, J.D.P., Soares, L.S., Guerra, S.F.S., Gabbay, Y.B., Sanchez, N.O., Linhares, A.C., 2008. Reemergence of G2 rotavirus serotypes in Northern Brazil reflects a natural changing pattern over time. In: The 8th Rotavirus International Symposium. Istanbul, Turkey Abstracts pp. 60–61.

Dorn, A., Kippenberger, S., 2008. Clinical application of CpG-, non-CpG, and antisense oligodeoxynucleotides as immunomodulators. Curr. Opin. Mol. Ther. 10, 10–20.

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113.

Esona, M.D., Mijatovic-Rustempasic, S., Conrardy, C., Tong, S., Kuzmin, I.V., Agwanda, B., Breiman, R.F., Banyai, K., Niezgoda, M., Rupprecht, C.E., Gentsch, J.R., Bowen, M.D., 2010. Reassortant group A rotavirus from straw-colored fruit bat (*Eidolon helvum*). Emerg. Infect. Dis. 16 (12), 1844–1852.

Estes, M.K., Kapikian, A.Z., 2007. Rotaviruses. In: Knipe, D.M., Howley, P.M., Griffin, D.E. (Eds.), Fields Virology. 5th ed. Lippincott Williams & Wilkins, Philadelphia, pp. 1917–1975.

Fischer, T.K., Steinsland, H., Molbak, K., Ca, R., Gentsch, J.R., Valentiner-Branth, P., Aaby, P., Sommerfelt, H., 2000. Genotype profiles of rotavirus strains from children in a suburban community in Guinea-Bissau, Western Africa. J. Clin. Microbiol. 38, 264–267.

Gentsch, J.R., Glass, R.I., Woods, P., Gouvea, V., Gorziglia, M., Flores, J., Das, B.K., Bhan, M.K., 1992. Identification of group A rotavirus gene 4 types by polymerase chain reaction. J. Clin. Microbiol. 30, 1365–1373.

Gómez, M.M., Volotão, E.M., Lima de Mendonça, M.C., Tort, L.F.L., da Silva, M.F.M., Leite, J.P.G., 2010. Detection of uncommon rotavirus A strains P[8]G8 and P[4]G8 in the city of Rio de Janeiro, 2002. J. Med. Virol. 82, 1272–1276.

Gouvea, V., Glass, R.I., Woods, P., Taniguchi, K., Clark, F.H., Forrester, B., Fang, Z.Y., 1990. Polymerase chain reaction amplification and typing of rotavirus nucleic acid from stool specimens. J. Clin. Microbiol. 28, 276–282.

Grantham, P., Perrin, P., 1986. AIDS virus and HTLV-I differ in codon choices. Nature 319, 727–728.

Greenbaum, B.D., Levine, A.J., Bhanot, G., Rabadan, R., 2008. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog. 4, e1000079.

Gurgel, R.Q., Cuevas, L.E., Vieira, S.C., Barros, V.C., Fontes, P.B., Salustino, E.F., Nakagomi, O., Nakagomi, T., Dove, W., Cunliffe, N., Hart, C.A., 2007. Predominance of rotavirus P[4]G2 in a vaccinated population, Brazil. Emerg. Infect. Dis. 13, 1571–1573.

Ikemura, T., 1982. 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 isoaccepting transfer RNAs. J. Mol. Biol. 158, 573–597.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Iturriza-Gómara, M., Dallman, T., Bányai, K., Böttiger, B., Buesa, J., Diedrich, S., Fiore, L., Johansen, K., Korsun, N., Kroneman, A., Lappalainen, M., László, B., Maunula, L., Matthinjnssens, J., Midgley, S., Mladenova, Z., Poljsak-Prijatelj, M., Pothier, P., Ruggeri, F.M., Sanchez-Fauquier, A., Schreier, E., Steyer, A., Sidaraviciute, I., Tran, A.N., Usonis, V., Van Ranst, M., de Rougemont, A., Gray, J., 2009. Rotavirus surveillance in Europe, 2005–2008: web-enabled reporting and real-time analysis of genotyping and epidemiological data. J. Infect. Dis. 200, S215–S221.

Jia, R., Cheng, A., Wang, M., Xin, H., Guo, Y., Zhu, D., Qi, X., Zhao, L., Ge, H., Chen, X., 2009. Analysis of synonymous codon usage in the UL24 gene of duck enteritis virus. Virus Genes 38, 96–103.

Jiang, P., Sun, X., Lu, Z., 2007. Analysis of synonymous codon usage in *Aeropyrum pernix* K1 and other Crenarchaeota microorganisms. J. Genet. Genomics 34, 275–284.

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

Kypr, J., Mrazek, J., 1987. Unusual codon usage in HIV. Nature 327, 20.

Leite, J.P., Carvalho-Costa, F.A., Linhares, A.C., 2008. Group A rotavirus genotypes and the ongoing Brazilian experience—a review. Mem. Inst. Oswaldo Cruz 103, 745–753.

Lobo, F.P., Mota, B.E.F., Pena, S.D.J., Azevedo, V., Macedo, A.M., Tauch, A., Machado, C.R., Franco, G.R., 2009. Virus-host coevolution: common patterns of nucleotide motif usage in *Flaviviridae* and their hosts. PLoS One 4, e6282.

Mascarenhas, J.D., Lima, C.S., de Oliveira, D.S., Guerra Sde, F., Maestri, R.P., Gabbay, Y.B., de Lima, I.C., de Menezes, E.M., Linhares Ada, C., Bensabath, G., 2010. Identification of two sublineages of genotype G2 rotavirus among diarrheic children in Parauapebas, southern Pará state. Brazil. J. Med. Virol. 82, 712–719.

Matthijnssens, J., Bilcke, J., Ciarlet, M., Martella, V., Banyai, K., Rahman, M., Zeller, M., Beutels, P., Van Damme, P., Van Ranst, M., 2009. Future Microbiol. 4, 1303–1316.

Matthijnssens, J., Ciarlet, M., Heiman, E., Arijs, I., Delbeke, T., McDonald, S.M., Palombo, E.A., Iturriza-Gómara, M., Maes, P., Patton, J.T., Rahman, M., Van Ranst, M., 2008. Full genome-based classification of rotaviruses reveals a common origin between human Wa-like and porcine rotavirus strains and human DS-1-like and bovine rotavirus strains. J. Virol. 82, 3204–3219.

Montero, H., Rojas, M., Arias, C.F., Lopez, S., 2008. Rotavirus infection induces the phosphorylation of eIF2α but prevents the formation of stress granules. J. Virol. 82, 1496–1504.

Nakagomi, T., Cuevas, L.E., Gurgel, R.G., Elrokhsi, S.H., Belkhir, Y.A., Abugalia, M., Dove, W., Montenegro, F.M., Correia, J.B., Nakagomi, O., Cunliffe, N.A., Hart, C.A., 2008. Apparent extinction of non-G2 rotavirus strains from circulation in Recife, Brazil, after the introduction of rotavirus vaccine. Arch. Virol. 153, 591–593.

Racaniello, V.R., 2001. Picornaviridae: the viruses and their replication. In: Knipe, D.M., Howley, P.M., Griffin, D.E., Lamb, R.A., Martin, M.A., Roizman, B., Straus, S.E. (Eds.), Fields Virology, 4th ed., vol. 1. Lippincott Williams & Wilkins, Philadelphia, PA, pp. 685–722.

Rothberg, P.G., Wimmer, E., 1981. Mononucleotide and dinucleotide frequencies, and codon usage in poliovirus RNA. Nucleic Acids Res. 9, 6221–6229.

Sanchez, G., Bosch, A., Pinto, R.M., 2003. Genome variability and capsid structural constraints of hepatitis A virus. J. Virol. 77, 452–459.

Santos, N., Hoshino, Y., 2005. Global distribution of rotavirus serotypes/genotypes and its implication for the development and implementation of an effective rotavirus vaccine. Rev. Med. Virol. 15, 29–56.

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

Sharp, P.M., Tuohy, T.M., Mosurski, K.R., 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14, 5126–5143.

Sharp, P.M., Li, W.H., 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24, 28–38.

Stoletzki, N., Eyre-Walker, A., 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. Mol. Biol. Evol. 24, 374–381.

Tao, P., Dai, L., Luo, M., Tang, F., Tien, P., Pan, Z., 2009. Analysis of synonymous codon usage in classical swine fever virus. Virus Genes 38, 104–112.

Tsai, C.T., Lin, C.H., Chang, C.Y., 2007. Analysis of codon usage bias and base compositional constraints in iridovirus genomes. Virus Res. 126, 196–206.

Ursu, K., Kisfali, P., Rigo, D., Ivanics, E., Erdélyi, K., Dan, A., Melegh, B., Martella, V., Bányai, K., 2009. Molecular analysis of the VP7 gene of pheasant rotaviruses identifies a new genotype, designated G23. Arch. Virol. 154, 1365–1369.

Wessa, P., 2010. Free Statistics Software. Office for Research Development and Education version 1. 1.23-r5. URL: http://www.wessa.net.

Woo, P.C.Y., Wong, B.H.L., Huang, Y., Lau, S.K.P., Yuen, K., 2007. Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in Coronaviruses. Virology 369, 431–442.

Wong, E., Smith, D.K., Rabadan, R., Peiris, M., Poon, L.L.M., 2010. Codon usage bias and the evolution of influenza A virus. Codon usage biases of Influenza virus. BMC Evol. Biol. 10, 253.

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

Zhao, S., Zhang, Q., Liu, X., Wang, X., Zhang, H., Wu, Y., Jiang, F., 2008. Analysis of synonymous codon usage in 11 human bocavirus isolates. Biosystems 92, 207–214.

Zhong, J., Li, Y., Zhao, S., Liu, S., Zhang, Z., 2007. Mutation pressures shapes codon usage in the GC-rich genome of foot-and-mouth disease virus. Virus Genes 35, 767–776.

Zhou, T., Gu, W., Ma, J., Sun, X., Lu, Z., 2005. Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses. Biosystems 81, 77–86.