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/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
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