Virus Research 196 (2015) 87–93

Contents lists available at ScienceDirect

Virus Research

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

Genome-wide analysis of codon usage bias in Ebolavirus

Juan Cristina[a,*], Pilar Moreno[a], Gonzalo Moratorio[a,b], Héctor Musto[c]

[a] Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Universidad de la República, Iguá 4225, 11400 Montevideo, Uruguay
[b] Viral Populations and Pathogenesis Laboratory, Institut Pasteur, CNRS UMR 3569, Paris, France
[c] Laboratorio de Organización y Evolución del Genoma, Instituto de Biología, Facultad de Ciencias, Universidad de la República, Iguá 4225, 11400 Montevideo, Uruguay

ARTICLE INFO

Article history:
Received 25 September 2014
Received in revised form 31 October 2014
Accepted 6 November 2014
Available online 14 November 2014

Keywords:
Ebola
Codon usage
Codon bias
Evolution

ABSTRACT

Ebola virus (EBOV) is a member of the family *Filoviridae* and its genome consists of a 19-kb, single-stranded, negative sense RNA. EBOV is subdivided into five distinct species with different pathogenicities, being *Zaire ebolavirus* (ZEBOV) the most lethal species. The interplay of codon usage among viruses and their hosts is expected to affect overall viral survival, fitness, evasion from host's immune system and evolution. In the present study, we performed comprehensive analyses of codon usage and composition of ZEBOV. Effective number of codons (ENC) indicates that the overall codon usage among ZEBOV strains is slightly biased. Different codon preferences in ZEBOV genes in relation to codon usage of human genes were found. Highly preferred codons are all A-ending triplets, which strongly suggests that mutational bias is a main force shaping codon usage in ZEBOV. Dinucleotide composition also plays a role in the overall pattern of ZEBOV codon usage. ZEBOV does not seem to use the most abundant tRNAs present in the human cells for most of their preferred codons.

1. Introduction

Ebola virus (EBOV) is a member of the family *Filoviridae* and is among the most deadly human pathogens, causing a severe hemorrhagic fever syndrome in both humans and non-human primates (Hoenen et al., 2006; Sanchez et al., 2007). EBOV is subdivided into five distinct species with different pathogenicity (Wauquier et al., 2010): *Zaire ebolavirus* (ZEBOV), the most lethal species with a case-fatality rate up to 90% (Khan et al., 1999), that caused numerous human outbreaks in Democratic Republic of Congo; *Sudan ebolavirus* (SEBOV), with a case-fatality rate of about 50%, which has caused outbreaks in Sudan and Uganda (WHO, 2004; CDC, 2001); *Cote d'Ivoire ebolavirus* (CIEBOV), who has been linked to a single non-fatal human case (Le Guenno et al., 1999), the newly discovered *Bundibugyo ebolavirus* (BEBOV) that caused an outbreak with a 25% case-fatality rate in 2007 in Uganda (Towner et al., 2008), and *Reston ebolavirus* (REBOV), which has caused outbreaks in non-human primates and swine in the Philippines and appears to be non-pathogenic for humans (Rollin et al., 1999).

EBOV is a filamentous enveloped virus containing a negative strand RNA genome of approximately 19 kb. The EBOV genome consists of eight major subgenomic mRNAs which in turn encode for seven structural proteins, namely a nucleoprotein (NP), two virion proteins (VP35 and VP40), a surface glycoprotein (GP), two additional viral proteins (VP30 and VP24), and a RNA-dependent RNA polymerase (L) and for one non-structural soluble protein (sGP) (Feldmann et al., 1993).

The disease caused by EBOV is characterized by the sudden onset of fever and malaise, accompanied by other nonspecific signs and symptoms such as myalgia, headache, vomiting, and diarrhea. EBOV initially replicates massively in macrophages and dendritic cells (DC), then spreads rapidly to all vital organs, infecting endothelial cells, epithelial cells, hepatocytes and other cell types (Stroher et al., 2001; Geisbert et al., 2003). Among EBOV patients, 30–50% experience hemorrhagic symptoms. A recent report by Rasmussen et al. (2014), suggested that the genetic factors play a significant role in determining hemorrhagic outcome in naïve individuals without prior exposure or immunologic priming by using mice model. Thence, in severe and fatal forms, multiorgan dysfunctions, including hepatic damage, renal failure, and central nervous system involvement occur, leading to shock and death (Dixon and Schafer, 2014). There is currently no vaccine and no specific treatment against EBOV (Becquart et al., 2014).

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

http://dx.doi.org/10.1016/j.virusres.2014.11.005
0168-1702/

Since 1976, there have been more than 20 EBOV outbreaks across Central Africa, with the majority caused by ZEBOV. No previous EBOV outbreak has been as large or persistent as the current epidemic, and none has spread beyond East and Central Africa (Dixon and Schafer, 2014; Gire et al., 2014). To date, more than 4555 people, including health care workers, have been killed by EBOV disease in 2014 (as of October 22, 2014), and the number of cases in the current outbreak now exceeds the number from all previous outbreaks combined (Frieden et al., 2014). Viral sequencing of specimens from the current epidemic were identified as ZEBOV (Baize et al., 2014).

The redundancy of the genetic code, in which most of the amino acids can be translated by more than one codon, offers evolution the opportunity to tune the efficiency and accuracy of protein production to various levels while maintaining the same amino acid sequence (Stoletzki and Eyre-Walker, 2007). The various codons that correspond to the same residue are often considered 'synonymous', yet their corresponding tRNAs might differ in their amounts in cells and thus also in the speed in which they will be recognized by the ribosome, and hence, influence the rate of translation and the accuracy in folding the encoded protein. While the non-random usage of synonymous codons is often correctly assumed to reflect the action of neutral drift, in an increasing number of cases it now turns out to reflect the result of natural selection, perhaps mainly for tuning the efficiency and accuracy of translation (Gingold and Pilpel, 2011). Studies on codon usage have determined several factors that could influence codon usage patterns, including mutational pressure, natural selection, secondary protein structure, replication and selective transcription among others (Butt et al., 2014).

The interplay of codon usage among viruses and their hosts is expected to affect overall viral survival, fitness, evasion from host's immune system and evolution (Burns et al., 2006; Mueller et al., 2006; Costafreda et al., 2014). Indeed, as is well known, synonymous triplets are generally not used randomly, and the main forces that drive this bias from equal usage are natural selection (which is mainly related to translation efficiency at two different levels: speed and accuracy) and mutational biases (for a review see Sharp et al., 2010). Therefore, the study of codon usage patterns in viruses can reveal important information about molecular evolution, regulation of viral gene expression and aid in vaccine design, where the efficient expression of viral proteins may be required to generate immunity (Butt et al., 2014).

In the present study, we performed comprehensive analyses of codon usage and composition of ZEBOV, including the recently isolated strains from the current epidemic of 2014, which represents all the complete genome sequences available in the databases, and investigated the possible key evolutionary determinants of the biases found.

2. Materials and methods

2.1. Sequences

Complete genome sequences for 25 Zaire Ebolavirus strains (ZEBOV) were obtained from DDBJ and GeneBank databases (available at: http://arsa.ddbj.nig.ac.jp and http://www.ncbi.nlm.nih.gov, respectively). For strain names and accession numbers see Supplementary Table S1. For each strain the ORFs were concatenated (NP+VP35+VP40+GP+VP30+VP24+L) and aligned using the MUSCLE program (Edgar, 2004). The alignment is available upon request. The data set comprised a total of 113,025 codons.

Supplementary Table S1 related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.virusres.2014.11.005.

2.2. Data analysis

Codon usage, dinucleotide frequencies, base composition, the relative synonymous codon usage (RSCU) (Sharp and Li, 1986), the effective number of codons (ENC) (Novembre, 2002), total G+C genomic content, as well as G+C content at first, second and third codon positions were calculated using the program CodonW (written by John Peden and available at http://sourceforge.net/projects/codonw). ENC index can vary between 20 and 61 and a low value indicates a strong bias in codon usage. To study codon usage preferences in EBOV in relation to the codon usage of human cells, we employed the codon adaptation index (CAI) (Sharp and Li, 1987). CAI was calculated using the approach of Puigbo et al. (2008a) (available at: http://genomes.urv.es/CAIcal) for EBOV and human cells. This method allows to compare a given codon usage (in our case, EBOV) to a predefined reference set (human). In order to show whether the EBOV genes are not better adapted to the codon usage of the reference set than the genes that define the reference dataset itself, as measured by CAI, we constructed a dataset composed of 322 human genes selected at random from Ensembl database (available at http://www.ensembl.org). A statistically significant difference among CAI values obtained was addressed by means of the use of a Student's *t*-test and a Wilcoxom & Mann–Whitney test (Wessa, 2012). In order to discern if the statistically significant differences in the CAI values arise from codon preferences, we used E-CAI (Puigbo et al., 2008b) to calculate the expected value of CAI (eCAI) at the 95% confident interval. A Kolmogorov–Smirnov test for the expected CAI was also performed (Puigbo et al., 2008b). The RSCU values of human cells were obtained from Kazusa database (available at: http://www.kazusa.or.jp/codon/). The frequencies of tRNAs in human cells were retrieved from the GtRNAdb database (Chan and Lowe, 2009).

2.3. Correspondence analysis

The relationship between variables and samples can be obtained using multivariate statistical analysis. Correspondence 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 different axes (Tao et al., 2009). COA was performed on the RSCU values using the CodonW program.

2.4. Correlation analysis

Correlation analysis was carried out using Spearman's rank correlation analysis method (Wessa, 2012; available at: www.wessa.net).

3. Results and discussion

3.1. General codon usage pattern in ZEBOV

In order to study the extent of codon usage bias in Zaire Ebolavirus strains (ZEBOV), the ENC's values were calculated for the 25 strains enrolled in this study. A mean value of 57.23 ± 0.51 was obtained. Due to the fact that all values obtained were >40, the results of these studies suggest that the overall codon usage among ZEBOV is similar and slightly biased.

This is in agreement with previous studies in other RNA viruses, like Chikungunya virus (ENC = 55.56) (Butt et al., 2014), bovine