Infection, Genetics and Evolution 28 (2014) 95–100

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Characterization of the porcine epidemic diarrhea virus codon usage bias

Ye Chen [1], Yuzhen Shi [1], Hongjuan Deng, Ting Gu, Jian Xu, Jinxin Ou, Zhiguo Jiang, Yiren Jiao, Tan Zou, Chong Wang *

Guangdong Provincial Key Lab of Agro-animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou, Guangdong 510642, People's Republic of China

ARTICLE INFO

Article history:
Received 26 July 2014
Received in revised form 26 August 2014
Accepted 1 September 2014
Available online 18 September 2014

Keywords:
Codon usage bias
Porcine epidemic diarrhea virus
Mutation bias
Natural selection

ABSTRACT

Porcine epidemic diarrhea virus (PEDV) has been responsible for several recent outbreaks of porcine epidemic diarrhea (PED) and has caused great economic loss in the swine-raising industry. Considering the significance of PEDV, a systemic analysis was performed to study its codon usage patterns. The relative synonymous codon usage value of each codon revealed that codon usage bias exists and that PEDV tends to use codons that end in T. The mean ENC value of 47.91 indicates that the codon usage bias is low. However, we still wanted to identify the cause of this codon usage bias. A correlation analysis between the codon compositions (A3s, T3s, G3s, C3s, and GC3s), the ENC values, and the nucleotide contents (A%, T%, G%, C%, and GC%) indicated that mutational bias plays role in shaping the PEDV codon usage bias. This was further confirmed by a principal component analysis between the codon compositions and the axis values. Using the Gravy, Aroma, and CAI values, a role of natural selection in the PEDV codon usage pattern was also identified. Neutral analysis indicated that natural selection pressure plays a more important role than mutational bias in codon usage bias. Natural selection also plays an increasingly significant role during PEDV evolution. Additionally, gene function and geographic distribution also influence the codon usage bias to a degree.

1. Introduction

Porcine epidemic diarrhea (PED) was first identified in Belgium and the United Kingdom in 1971 (Rosenberg et al., 1977). PED is a great threat to swine health and production, and it has caused devastating diseases with a severe impact on the swine-raising industry (Huang et al., 2013; Lee and Lee, 2014; Puranaveja et al., 2009; Wang et al., 2014). PED is characterized by vomiting, watery diarrhea and dehydration and causes a high death rate among suckling piglets. PED has been reported in many countries since 1971, including China (Chen et al., 2008; Tian et al., 2013), Thailand (Puranaveja et al., 2009; Temeeyasen et al., 2014), South Korea (Lee and Lee, 2014; Park et al., 2013), and the United States (Huang et al., 2013; Wang et al., 2014). Outbreaks of PED were first identified in the United States between 2013 and 2014, and in less than one year, 2692 PED cases were found in 23 US states, which led to great economic losses (Huang et al., 2013; Wang et al., 2014). The genome of the causative agent was identified as being closely related to a strain from China (Wang et al., 2014). During 2013 and 2014, outbreaks of PED also occurred in South Korea, and genomic sequences of the South Korean isolated virus were shown to be similar to the US PED strains (Lee and Lee, 2014). The recent frequency in outbreaks and the great economic losses associated with PED indicate the importance of understanding this virus.

PED is caused by the porcine epidemic diarrhea virus (PEDV). The first PEDV isolate was named CV777 (Egberink et al., 1988; Huang et al., 2013; Rosenberg et al., 1977), and sequencing of its viral genome indicated that PEDV and the bat coronavirus have a common evolutionary precursor (Huang et al., 2013). The PEDV genome is composed of seven open reading frames (ORFs) (Egberink et al., 1988; Kocherhans et al., 2001). ORF1a and ORF1b encode the replicase proteins. The next five ORFs encode the viral proteins, which include the spike protein (S), the ORF3 protein (ORF3), the small membrane proteins (E), the membrane proteins (M), and the nucleocapsid protein (N).

Synonymous codons are not used randomly in the genomes of organisms. The same codon may be present in different genes in a single genome as well as in different parts of one specific gene (Fu, 2010; Li et al., 2012; Liu et al., 2010a; Zhou et al., 2010). This phenomenon is called codon usage bias. Codon usage bias has been confirmed in many viruses (Auewarakul et al., 2009; Fu, 2010;

* Corresponding author.
E-mail address: betty@scau.edu.cn (C. Wang).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.meegid.2014.09.004
1567-1348/

D'Andrea et al., 2011; Li et al., 2012; Liu et al., 2010a; Shi et al., 2013; Wang et al., 2011; Wong et al., 2010; Zhou et al., 2010). Synonymous codon usage bias is strong in some viruses, such as the hepatitis A virus (D'Andrea et al., 2011). Other viruses have weak codon usage bias, such as SARS (Zhao et al., 2008), classical swine fever virus (Tao et al., 2009), and H9N2 influenza A virus (Liu et al., 2010a). Codon usage bias is related to mutation bias, natural selection, gene function, gene length, tRNA abundance, and RNA structure. Codon usage bias is generally influenced by one of the two following biological pressures: (1) pressure from mutational bias and/or (2) pressure from natural selection (Fu, 2010). In some RNA and DNA viruses, mutational bias has been identified as playing a more important role, compared with natural selection, such as in foot-and-mouth virus (Zhou et al., 2010)and herpes virus (Fu, 2010). Natural selection has also been confirmed as being the dominant factor in some viruses such as *Parvoviridae* (Shi et al., 2013). The nucleotide content in the third codon position (A3s, T3s, G3s, C3s, GC3s, AT3s) is based on the codon usage bias (Wong et al., 2010). Many analytic methods for analyzing codon usage bias have been developed and are based on examining the nucleotide content of the third codon position. These methods have been shown to be effective in uncovering the codon usage bias of the genome and other significant information, such as the effective number of codons (Wright, 1990).

Considering the recent increase in the prevalence of PEDV and its great threat to the swine industry, we chose to examine the codon usage bias of PEDV to better understand the codon usage pattern of the PEDV genome, the differences in codon usage between the PEDV and pig genomes, and the evolutionary pattern of PEDV codon usage.

2. Materials and methods

2.1. Sequence data

The complete genome sequences of the PEDV isolates were retrieved from the GenBank database (http://www.ncbi.nlm.nih.gov). To better understand PEDV codon usage bias, only the viruses with complete genomic information were included in our study. Detailed information about the 43 PEDV isolates, including their accession number, the time when they were isolated, and the country where they were isolated, is listed in the supplied materials (Table S1).To perform the phylogenetic analysis, the sequence data were compiled and edited using the DNASTAR software package (Madison, WI, USA). The edited data were then aligned using the BioEdit (version 7.0.9.0) sequence analysis program and the ClustalW method. The unrooted phylogenetic tree was constructed with the MEGA 4.0 software with the evolutionary pattern of the 43 PEDV isolates using the pairwise deletion model and calculated based on 1,000 replicates.

2.2. Nucleotide composition

The nucleotide content (A%, T%, G%, and C%) of each PEDV strain was analyzed using the MEGA 4.0 biosoftware for windows. The nucleotide composition of the third synonymous codon position of each codon (A3s, T3s, G3s, and C3s) was calculated using the CodonW program (version 1.4.2) (http://codonw.sourceforge.net//).

2.3. Codon usage indices

Relative synonymous codon usage (RSCU) values were first proposed in 1986 (Sharp and Li, 1986). The RSCU value is independent of the amino acid composition and has been used widely to estimate the codon usage bias between genes. A higher RSCU value means that the codon is used more frequently or has a higher codon usage bias. If the RSCU value of a specific codon is higher than 1.0, it is considered to have a positive codon usage bias. When the RSCU value is less than 1.0, it is considered to have a negative codon usage bias.

The effective number of codons (ENC) value is not influenced by the amino acid composition or the gene length. In the ENC analysis, an ENC value is given to each codon. The ENC value ranges from 20 to 61. In contrast to the RSCU value, a higher ENC value correlates to a weaker codon usage bias. If the codon of one gene is completely randomly and unbiased, then the expected ENC value (ENC^*) is calculated from GC3s (Sharp and Li, 1986):

$$ENC = 2 + s + \frac{29}{s^2 + (1-s)^2}$$

The s value is the GC3s content of each codon. When the expected ENC value is plotted against the GC3s value, an expected curve is formed. A dot located on the curve is regarded as unbiased.

2.4. Principal component analysis

Principal component analysis (PCA) is a common statistical method used to explain the codon usage of a specific gene. In the analysis, the RSCU value of each codon is explained by a 59-dimension space and transformed into unrelated factors. In this model, PCA can determine the major variation from the RSCU value of each codon. Using both the PCA and correlation analysis, the factors influencing the codon usage bias can be effectively determined.

2.5. Codon adaptation index

Codon adaptation index (CAI) is one of the most widespread methods for analyzing codon usage bias due to the natural selection pressure. It represents the adaption of the virus to the host. The value of the CAI value is between 0 and 1. A higher CAI value indicates the stronger adaptation to the host. The codon usage pattern of the *Sus scrofa* is obtained from an online website (http://www.kazusa.or.jp/codon/) (Puigbo et al., 2008). To estimate the codon adaption of the PEDV to *Sus scrofa*, the CAI value is calculated using the CAIcal biosoftware (http://genomes.urv.es/CAIcal). In the analysis, the synonymous codon usage pattern of the viral host is deposited as reference and the CAI value of the PEDV is calculated after comparison with the reference from *Sus scrofa*. ORFs 1a/b is not included in the analysis because of the nucleotide length restrictions of the online tool.

2.6. Hydropathicity and aromaticity indices

The hydropathicity and aromaticity of a single gene product are thought to be the result of translation selection and according to the natural selection (Lobry and Gautier, 1994). In our study, the Gravy and Aroma score of each gene product was obtained using the CodonW program (version 1.4.2) to reflect the hydropathicity and aromaticity, respectively. A higher Gravy or Aroma score means a more hydrophobic or aromatic amino acid product, respectively.

2.7. Neutral evolution analysis

The neutral analysis is used to estimate the varying role of mutational pressure and natural selection in the PEDV. In the analysis, the *P12* (GC12s) value of the synonymous codon is plotted against its *P3* (GC3s) value (Sueoka, 1988). To study the evolutional