Infection, Genetics and Evolution 11 (2011) 2098–2102

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

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

Short communication

The characteristic of codon usage pattern and its evolution of hepatitis C virus

Jin-song Hu [a,b,1], Qin-qin Wang [a,1], Jie Zhang [a,1], Hao-tai Chen [a], Zhi-wen Xu [b], Ling Zhu [b], Yao-zhong Ding [a], Li-na Ma [a], Kai Xu [b], Yuan-xing Gu [a], Yong-sheng Liu [a,*]

[a] State Key Laboratory of Veterinary Etiological Biology, Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Lanzhou 730046, China
[b] Animal Biological Technological Center, Sichuan Agricultural University, Ya'an 625014, China

ARTICLE INFO

Article history:
Received 20 May 2011
Received in revised form 22 August 2011
Accepted 24 August 2011
Available online 1 September 2011

Keywords:
Hepatitis C virus
Codon usage
Relative synonymous codon usage values
Effective number of codons values
Correlation analysis

ABSTRACT

To give a new perspective on the codon usage of the hepatitis C virus (HCV) and the factors accounting for shaping the codon usage pattern of the virus, the relative synonymous codon usage (RSCU) values, aromaticity and hydrophobicity of each polyprotein of the virus, effective number of codons (ENC) values and nucleotide contents were calculated to implement a comparative analysis to evaluate the dynamics of the virus evolution. The RSCU values of each codon of 144 HCV ORFs indicated that all abundant codons were C/G-ended codons. The plots of principal component analysis based on sub-genotype of HCV indicated that sub-genotype 1a and 1b separated clearly on the axis of *f2* suggesting that the codon usage bias between sub-genotype 1a and 1b strains was different. By comparing the codon usage between HCV and human cells, we found that the synonymous codon usage pattern of HCV was a mixture of coincidence and antagonism to that of host cells. The characteristics of the synonymous codon usage patterns and nucleotide contents of HCV, and the correlation analysis between GC_{3s}, $GC_{1,2s}$, GC% (ORF), GC% (5′-UTR), GC% (3′-UTR), aromaticity, hydrophobicity and ENC value, respectively, indicated that mutational pressure was the dominant factor accounting for the codon usage variation and selection pressure also accounted for HCV codon usage pattern.

1. Introduction

Studies of codon usage pattern can reveal the molecular evolution of organisms, and contribute to understand the interaction between RNA viruses and the immune response of the hosts (Shackelton et al., 2006). Selection pressure is a phenomenon that alters the behavior and fitness of living organisms within the host environment such as the the fine-tuning translation kinetics selection mechanism (Aragones et al., 2008, 2010). Mutation pressure is the change in some gene frequencies due to the repeated occurrence of the same mutations. The molecular bases of this genetic variability may be the high error rate of the viral RNA-dependent RNA polymerase and the absence of proofreading mechanisms. The mutation frequencies for a variety of RNA viruses range from 10^{-4} to 10^{-5} substitution per base per round of copying (Domingo, 1996). It was reported that compositional constraints and translational selection were the main factors that accounted for codon usage variation among genes/organisms (Lesnik et al., 2000; Ghosh et al., 2000). In some unicellular organisms such as *Escherichia coli* and *Bacillus subtilis*, the highly expressed genes had a strong selective preference for codons with a high concentration of the corresponding tRNA molecule, whereas the lowly expressed genes displayed a uniform pattern of codon usage (Ikemura, 1981, 1985; Sharp and Li, 1986; Lesnik et al., 2000). Additionally, in some prokaryotes with extremely high A + T or G + C contents, mutation bias was the major factor accounting for the variation in codon usage (Bulmer, 1988).

As for some RNA viruses, compared with translational selection, mutation pressure plays a more important role in the synonymous codon usage pattern (Jenkins and Holmes, 2003; Levin and Whittome, 2000). Although it is known that compositional constraints and translation selection are the generally accepted factors accounting for codon usage bias, other selection forces such as the fine-tuning translation kinetics selection, as well as escape from cellular antiviral responses have also been reported (Aragones et al., 2008, 2010; Karlin et al., 1994; Sugiyama et al., 2005). Hepatitis C virus (HCV), a small enveloped RNA virus, is the pathogen of chronic hepatitis C. It belongs to the hepacivirus genus of the *Flaviviridae* family (Suzuki et al., 2007; Alter, 1995). It consists of a single-stranded positive-sense RNA of about 9.6 kb, which contains an open reading frame (ORF) encoding a polyprotein precursor of approximately 3000 residues, flanked by untranslated regions (UTR) at both ends (Choo et al., 1989; Suzuki et al., 2007). An important feature of the HCV genome is its high degree of genetic variability (Pawlotsky, 2006; Suzuki et al., 2007). HCV is classified into at least seven principal genotypes that

* Corresponding author. Tel.: +86 931 8342771; fax: +86 931 8340977.
E-mail address: liuyongshengvip8@163.com (Y.-s. Liu).
[1] These authors contributed equally to this work.

1567-1348/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2011.08.025

differ in their nucleotide sequences by 31–34% and in their amino acid sequences by ~30%. These genotypes (1, 2, 3, 4, 5, 6 and 7) show differences with regard to their worldwide distribution, transmission and disease progression, and have been further classified into different sub-genotypes (a, b, c, d, etc) (Kim et al., 2010; Sharma, 2010; Duarte et al., 2010). HCV, which is most commonly spread by direct contact with the infected blood and the blood products, has been recognized as a major cause of chronic liver disease. Chronic infection eventually causes cirrhosis leading to hepatocellular carcinoma (HCC) and ultimately death. Currently, there is no vaccine to prevent hepatitis C (Suzuki et al., 2007; Sharma, 2010). Abundant genome sequences of HCV have been published and lots of studies have been performed in recent years (Suzuki et al., 2007; Duarte et al., 2010), but little codon usage information about HCV is available. In order to better understand the characteristics of the viral genome evolution, the codon usage and the potential factors accounting for codon usage pattern of HCV was analyzed in the study.

2. Materials and methods

2.1. HCV genome sequences

In this study, a total of 144 HCV complete genomic sequences representing seven genotypes and 35 sub-genotypes were selected from NCBI (http://www.ncbi.nlm.nih.gov/) according to the availability of collection time, sub-genotypes, sampling country, sampling year and patient information as much as possible. The serial number, GenBank accession numbers, sub-genotypes, and other detail information were listed in Supplementary Table 1.

2.2. Compositional properties measures

To examine compositional properties of 144 HCV sequences, the GC_{3s} (the frequencies of nucleotide G + C at the third codon position) and $GC_{1,2s}$ (the mean frequencies of nucleotide G + C at the first and the second position) of each ORF were calculated. The GC content of the ORF, 5′-UTR and 3′-UTR of HCV samples were also calculated, respectively.

2.3. Analysis of codon usage

The relative synonymous codon usage (RSCU) values of each codon (excluding AUG, UGG and the termination codons) of 144 complete coding sequences of HCV were calculated according the previously reported method (Sharp and Li, 1986; Sharp et al., 1986). The codons with RSCU values >1.0 have positive codon usage bias (abundant codons), while those with RSCU values <1.0 have negative codon usage bias (less-abundant codons), and when the RSCU values is 1.0, it means that these codons are chosen equally or randomly (Sharp and Li, 1986). To examine the synonymous codon usage bias of the whole coding sequences, the 'Effective Number of Codons' (ENC) values (Wright, 1990; Schubert and Putonti, 2010) of 144 HCV strains were calculated. The ENC value is the best overall estimator of one gene/ORF absolute synonymous codon usage bias (Comeron and Aguade, 1998). The ENC values are always between 20 (when only one codon is used for each amino acid) and 61 (when all codons are used equally) (Sharp and Li, 1986). A plot was drawn to show the distribution of the GC3s and ENC values among 144 HCV strains. Codon adaptation index (CAI) of all HCV ORFs was also calculated. The CAI value ranges from 0 to 1. The CAI was proposed as a quantitative way of predicting the expression level of a gene based on its codon sequence. The most frequent codons simply have the highest relative adaptiveness values, and sequences with higher CAIs are preferred over those with lower CAIs (Kadam and Ghosh, 2008). A rare codon was defined as one whose frequency was less than 30% that of its most abundant synonym in each of the codon usage tables (Gavrilin et al., 2000; Sanchez et al., 2003), and a comparative analysis of the codon usage was implemented between HCV and human cells.

2.4. Analysis of influencing factors of codon usage pattern

Correlation analysis between GC_{3s}, $GC_{1,2s}$, GC% (ORF), GC% (5′-UTR), GC% (3′-UTR), aromaticity, hydrophobicity of corresponding encoding polyprotein of each gene and ENC value among 144 samples were carried out, respectively, using the Pearson's rank correlation analysis method (Ewens and Grant, 2001).

Principal component analysis (PCA) was used to investigate the major trend in codon usage variation among ORFs (Jolliffe, 2002; Mardia et al., 1979; Liu et al., 2011). In order to minimize the effect of amino acid composition on codon usage, each ORF is represented as a 59-dimensional vector. Each dimension corresponds to the RSCU value of one sense codon (excluding AUG, UGG and the termination codons).

These indices mentioned above were calculated by the program Codon W and all statistical processes were done by statistical software SPSS17.0.

3. Results

3.1. Compositional properties

Among the 144 samples, the GC_{3s} values ranged from 56.5% to 69.1% with a mean value 65.9% and a standard error 0.03, the $GC_{1,2s}$ values ranged from 53.3% to 55.1% with a mean value 54.4% and a standard error 0.003, and the GC content of ORF ranged from 54.7% to 59.4% with a mean value 58.2% and a standard error 0.012 (Supplementary Table 2). These indicated that HCV is a GC abundant virus.

3.2. Codon usage

The RSCU values of each codon of 144 HCV complete coding sequences indicated that all abundant codons were C/G-ended codons and the overwhelming majorities of less-abundant codons were A/U-end codons. Although first nucleotide position is a synonymous position in Leu (UUA-CUA, UUG-CUG) and Arg (CGA-AGA, CGG-AGG), the all G-ended codons (UUG, CUG, CGG and AGG) are used more often than the A-ended codons (UUA, CUA, CGA and AGA) when the coding sequence of HCV is being translated (Table 1). The phenomenon suggested that the codon usage bias of HCV was related to the G/C bias of the coding sequences.

The CAI values among the 144 HCV ORFs ranged from 0.167 to 0.199 with a mean value 0.179 and a standard error 0.007. It was low, implying that the codon usage bias and the expression level of HCV were low. Furthermore, the ENC values among the 144 HCV ORFs ranged from 50.68 to 56.99 with a mean value 52.62 and a standard error 1.707 (Supplementary Table 2). Based on the comparative analysis with the published data of codon usage bias among other RNA viruses such as bovine viral diarrhea virus (ENC = 51.43), classical swine fever virus (ENC = 51.7), hepatitis A virus (ENC = 39.78) and hepatitis E virus (ENC = 48.2) (Wang et al., 2010; Tao et al., 2009; Sanchez et al., 2003; D' Andrea et al., 2011; Jenkins and Holmes, 2003), we could conclude that codon usage bias of HCV whole coding sequence was lower. In addition, a tendency that ENC values of genotype 1 strains were lower than that of the other genotypes (2, 3, 4, 5, 6 and 7) was observed (Fig. 1). Furthermore, the differences of the ENC values between