Gene 545 (2014) 5–14

Contents lists available at ScienceDirect

Gene

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

Codon usage bias in human cytomegalovirus and its biological implication

Changyuan Hu [a], Jing Chen [b], Lulu Ye [c], Renpin Chen [d], Lifang Zhang [c], Xiangyang Xue [c,*]

[a] Department of General Surgery, The First Affiliated Hospital of Wenzhou Medical University, Ouhai District 325035, Wenzhou City, Zhejiang Province, China
[b] Department of Rheumatism and Immunology, The First Affiliated Hospital of Wenzhou Medical University, Ouhai District 325035, Wenzhou City, Zhejiang Province, China
[c] Department of Microbiology and Immunology, Institute of Molecular Virology and Immunology, Wenzhou Medical University, Ouhai District 325035, Wenzhou City, Zhejiang Province, China
[d] Department of Gastroenterology and Hepatology, The First Affiliated Hospital of Wenzhou Medical University, Ouhai District 325035, Wenzhou City, Zhejiang Province, China

ARTICLE INFO

Article history:
Received 14 December 2013
Received in revised form 2 May 2014
Accepted 6 May 2014
Available online 9 May 2014

Keywords:
Human herpesvirus 5
Codon usage bias
Host-relating protein

ABSTRACT

Human cytomegalovirus (HCMV) infection, a worldwide contagion, causes a serious disorder in infected individuals. Analysis of codon usage can reveal much molecular information about this virus. The effective number of codon (ENC) values, relative synonymous codon usage (RSCU) values, codon adaptation index (CAI), and nucleotide contents was investigated in approximately 160 coding sequences (CDS) among 17 human cytomegalovirus genomes using the software CodonW. Linear regression analysis and logistic regression were performed to explore the preliminary data. The results showed that, overall, HCMV genomes had low codon usage bias (mean ENC = 47.619). However, the ENC of individual CDS varied widely and was distributed unevenly between host-related genes and viral-self-function genes ($P = 0.002$, odds ratio (OR) = 3.194), as did the GC content ($P = 0.016$, OR = 2.178). The ENC values correlated with CAI, GC content, and the nucleotide composing at the 3rd codon position (GC3s) ($P < 0.001$). There was a significant variation in the codon preference that depended on the RSCU data. The predicted ENC curve suggested that mutational pressure, rather than natural selection, was one of the main factors that determined the codon usage bias in HCMV. Among 123 genes with known function, the genes related to viral self-replication and viral–host interaction showed different ENC and CAI values, and GC and GC3s contents. In conclusion, the detailed codon usage bias theoretically revealed information concerning HCMV evolution and could be a valuable additional parameter for HCMV gene function research.

1. Introduction

Human cytomegalovirus (HCMV), also named human herpesvirus 5, is a ubiquitous herpesvirus that infects a broad range of cell types in its human host. Depending on the tissue type and the host's immune state, HCMV shows different growth rates, which lead to three different infection modes: acute infections with highly productive growth, persistent infections with low levels of replication, and latent infections where no viral progeny is produced (Dunn et al., 2003). The persistence of HCMV in immunocompetent individuals, leading to asymptomatic infections, might increase the risk of age-related disorders such as frailty or immune aging (Brunner et al., 2011; Schmaltz et al., 2005). In immunodeficient or immunosuppressed patients, including acquired immune deficiency syndrome patients and patients undergoing transplantation, lytic or reactivated infections can cause fatal courses, such as hepatitis or pneumonitis (Griffiths, 1996; Smith et al., 2001). The mechanisms of HCMV infection, especially latent infection, in the various cell types of human host are largely unknown.

The HCMV genome, which encodes nearly 200 protein-coding genes, comprises the so-called unique long (UL) and unique short (US) domains, flanked at one end by terminal repeated sequences (TRL and TRS) and at the other end by internal repeats (IRL and IRS) (Davison et al., 2003; Murphy et al., 2003a,b). Approximately 45 genes are essential for viral replication in cultured fibroblasts and 117 genes are nonessential (Dunn et al., 2003). The expression patterns, transcript structure, and transcription characteristics of several HCMV genes have been determined (Ma et al., 2012). To understand the mechanisms of gene expression in HCMV proliferative and latent infection, it is necessary to identify all of the features of that genome that are subject to the action of natural selection or mutational pressure.

Codon usage and DNA base composition may be used to determine the rate of viral gene expression (Zhao et al., 2003). Studies of synonymous codon usage have not only provided knowledge about the molecular evolution of individual genes, but also identified potential modulations

Abbreviations: A3s, A content in the 3rd codon position; AA, amino acids; C3s, C content in the 3rd codon position; CAI, codon adaptation index; CDS, coding sequences; ENC, effective number of codon; G3s, G content in the 3rd codon position; GC3s, G + C content in the 3rd codon position; GCs, G + C content; HCMV, human cytomegalovirus; HIV, human immunodeficiency virus; HPV, human papillomavirus; IRS, internal repeat sequences; MHC, major histocompatibility complex; OR, odds ratio; pENC, predicted values of ENC; RSCU, relative synonymous codon usage; T3s, T content in the 3rd codon position; TRS, terminal repeated sequences; UL, unique long; US, unique short.

* Corresponding author.
E-mail address: wzxxy@aliyun.com (X. Xue).

http://dx.doi.org/10.1016/j.gene.2014.05.018
0378-1119/

of gene expression as the result of selection that affects translational efficiency (Heitzer et al., 2007; Roychoudhury and Mukherjee, 2010). Codon usage bias may be related to various biological factors, such as gene length, transfer RNA abundance, GC content, and gene expression level (Gouy and Gautier, 1982; Heitzer et al., 2007; Ikemura, 1981; Lobry and Gautier, 1994).

Codon usage bias has been analyzed in several viruses, for example in polioviruses, hepatitis virus, HIV, papillomavirus, and herpesvirus (Fu, 2010; Pandit and Sinha, 2011; J. Zhang et al., 2011; Y. Zhang et al., 2011; Zhao and Chen, 2011). These studies revealed that among 43 herpesviruses, most showed a low codon preference, with the exception of simplex viruses and some varicelloviruses (Supek and Vlahovicek, 2005). However, the results of HCMV obtained from a single strain, which has been replaced by a new sequence, remain to be explored. Therefore, we analyzed the whole coding sequence (CDS) of HCMV by synonymous codon usage frequency, and reported differences in codon usage bias between viral basic functional proteins and host–virus-related proteins in HCMV. Our findings provide a new approach for viral protein function prediction and related research in HCMV.

2. Materials and methods

2.1. Sequence data

Seventeen human cytomegalovirus genomes were downloaded from the National Center for Biotechnology Information GenBank (http://www.ncbi.nlm.nih.gov/). The strain name, gene length, effective number of codon (ENC), GC composition, and GenBank accession numbers for these strains are listed in Table 1. The clinical isolates of Fix and PH were not included in the study because the CDSs in their genome are not provided in the GenBank. The ENC and GCs values were obtained as described in the next paragraph. CDSs from the sequences were deduced using the criteria of Paul and Elizabeth (Sharp and Cowe, 1991): 1) the gene has a complete CDS; 2) the length of the encoding gene is more than 300 bp; and 3) the CDS begins with ATG and ends with a termination codon. These criteria identified 160 CDSs per genome, resulting in approximately 2700 sequences for analysis.

2.2. Nucleotide composition, ENC, codon adaptation index (CAI), and relative synonymous codon usage (RSCU) of HCMV and coding sequences

The nucleotide compositions of the HCMV genomes are shown in Table 1. The ENC is an estimator that measures the degree of departure from the equal usage of synonymous codons in a gene. Its value ranges from 20 (when only a single codon is used for each kind of amino acid, which means extreme preference) to 61 (when all available codons are used, signifying no bias). If the value is greater than 40, a gene is regarded as having a weak codon bias (J. Zhang et al., 2011). The whole HCMV genomes and their respective genes were analyzed for their mean ENC values, base content, total GC composition and the G + C content in the 3rd codon position (GC3s) are listed in Table 3. They could be divided into three groups by their ENC values. To determine the impact of GC content on codon usage, the relationship between the ENC and GC3s content of each gene was plotted. The predicted values of ENC were calculated as $pENC = 2 + GC3s + 29 / [GC3s^2 + (1 - GC3s)^2]$ (Wright, 1990).

The CAI quantifies the relative adaptiveness of the codon usage of a gene towards the codon usage of highly expressed genes. The relative adaptiveness of each codon is the ratio of the usage of each codon (ranging from 0 to 1), to that of the most abundant codon for the same amino acid. We calculated every CAI for each CDS and listed them in Table 3. To investigate the characteristics of RSCU without the confusing influence of amino acid composing in all HCMV types, we analyzed the CDSs with high-frequency codon usage (ENC < 40). The RSCU values of the codons in each CDS were calculated by the formula:

$$RSCU = \frac{g_{ij}}{\frac{ni}{\sum_j g_{ij}}} n_i,$$

where g_{ij} is the observed number of the ith codon for jth amino acid, which has n_i type of synonymous codons (J. Zhang et al., 2011; Y. Zhang et al., 2011). This means that when the RSCU value of a codon is more than 1.0, the codon has high-frequency usage. When the RSCU is close to 1 or less than 1, the synonymous codon is chosen equally or has a positive codon usage bias. We summarized the means of the RSCUs of different CDSs among the strains in Table 2.

CodonW 1.4.2 (http://sourceforge.net/projects/codonw/) was used to calculate the genes' characteristics.

2.3. Statistical analysis

The statistical analysis software SPSS version 16.0 was used to carry out all statistical analyses. Most RSCU values fitted the Gaussian distribution; therefore, the independent sample t-test, the Mann–Whitney U test, and the Kruskal–Wallis H test were used to evaluate the differences in codon bias between different groups. The differences in ENC, CAI, GCs, and GC3s values, and the bias in different genes were analyzed by a chi-square test. Linear regression analysis was used to find the correlation between the gene analysis data using the Pearson correlation analysis or Spearman's rank correlation analysis. The effect of independent variables in functional genes was predicted by logistic regression.

Table 1
General features of human cytomegalovirus genome.

Strain	Length (bp)	A%	T%	C%	G%	CG%	CDSs	GenBank accession number
Human herpesvirus 5	235,646	21.446	21.072	28.507	28.975	57.482	166	NC_006273.2
3157	235,154	21.432	21.084	28.482	29.001	57.484	168	GQ221974.1
3301	235,703	21.440	21.051	28.481	29.029	57.510	169	GQ466044.1
AD169	231,781	21.444	21.005	28.526	29.025	57.511	160	FJ527563.1
AD169 substrain varUK	230,290	21.561	21.245	28.315	28.880	57.195	146	BK000394.5
AF1	235,937	21.460	21.073	28.502	28.965	57.467	165	GU179291.1
HAN13	236,219	21.452	21.050	28.503	28.996	57.499	165	GQ221973.1
HAN20	235,728	21.467	21.068	28.476	28.989	57.465	167	GQ396663.1
HAN38	236,112	21.407	21.007	28.544	29.042	57.586	165	GQ396662.1
JP	236,375	21.446	21.072	28.507	28.975	57.481	165	GQ221975.1
Merlin	235,646	21.486	20.983	28.609	28.922	57.482	166	AY446894.2
Toledo	235,398	21.425	21.072	28.473	29.030	57.531	165	GU937742.1
Towne	235,147	21.425	21.063	28.487	29.026	57.503	162	FJ616285.1
TR	235,681	21.440	21.080	28.486	28.994	57.480	170	KF021605.1
U8	235,709	21.468	21.109	28.472	28.951	57.512	166	GU179288.1
U11	234,732	21.458	21.092	28.450	28.999	57.423	166	GU179290.1
VR1814	235,233	21.446	21.072	28.507	28.975	57.449	167	GU179289.1