Gene 499 (2012) 37–40

Contents lists available at SciVerse ScienceDirect

Gene

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

High GC content of simple sequence repeats in *Herpes simplex virus type 1* genome ☆

Qingjian Ouyang [a], Xiangyan Zhao [a], Haiping Feng [a], You Tian [a], Dan Li [a], Mingfu Li [b], Zhongyang Tan [a,b,*]

[a] College of Biology, State Key Laboratory for Chemo/Biosensing and Chemometrics, Hunan University, Changsha 410082, China
[b] Chinese Academy of Inspection and Quarantine, Beijing 100029, China

ARTICLE INFO

Article history:
Accepted 24 February 2012
Available online 6 March 2012

Keywords:
Simple sequence repeat
Microsatellite
Herpes simplex virus type1

ABSTRACT

The presence, locations and composition of simple sequence repeats (SSRs) in *Herpes simplex virus type 1* (HSV-1) genome were extracted and analyzed by using the software Imperfect Microsatellite Extractor (IMEx). There were 663 mon-, 502 di-, 184 tri-, 20 tetra-, 4 penta- and 4 hexanucleotide SSRs that were observed in different distribution between coding and noncoding regions in the HSV-1 genome. G/C, GC/CG, and $(GGC)_n$ were predominant in mononucleotide, dinucletide, trinucleotide repeats respectively. Indeed, the results showed that GC content in simple sequence repeats was notably higher than that in entire HSV-1 genome. Our data might be helpful for studying the pathogenesis, genome structure and evolution of HSV-1.

1. Introduction

Herpes simplex virus type 1 (HSV-1) is widely distributed in the human population and is the leading cause of acute sporadic viral encephalitis and viral induced blindness (Karimi and MacLean, 2005). It is a neurotropic pathogen of humans that establishes a lifelong latent infection in the sensory ganglia innervating the site of primary infection. Although the mechanism(s) by which HSV establishes and reactivates from latency is an area of intensive study, much remains unknown (Perng et al., 1999). Previous studies have found that the latency-associated transcript (LAT) is the only viral gene that is abundantly transcribed during latency (Rock et al., 1987), while it was reported that the latency-associated transcript was related to the long repeat region of the viral genome (Perng et al., 1999). In addition, the HSV-1 genome sequence has one very special characteristic, it repeats at both ends. Obviously the repeated regions are an essential element of HSV-1 genome, and they may play an important part in entire genome. By analyzing the simple sequences we can learn more about the interrelationship between genome structure and pathogenesis in HSV-1.

Simple sequence repeats (SSRs) are synonym for microsatellites. They are sequences in which a short motif of 1–6 bases is tandemly repeated. The SSRs are prevalent in prokaryotes and eukaryotes, and they are not randomly distributed in the genomes (Hong et al., 2007; Li et al., 2004; Rajendrakumar et al., 2007; Toth et al., 2000). They are regarded as one of the most suitable markers for genome analysis (Pinto et al., 2006). Moreover, the SSRs may play a functional role in affecting gene expression and the polymorphism of SSR tracts is very important in the evolution of gene regulation (Gur-Arie et al., 2000; Kashi and King, 2006; Kashi and Soller, 1999; Kashi et al., 1997; Kunzler et al., 1995; Moxon and Wills, 1999; Rosenberg et al., 1994; Tonjum et al., 1998; Van Belkum, 1999). Their importance calls for intensive study in diverse organisms and recently we are interested in its investigation in viruses. Indeed, SSRs seem to be seldom noticed in viruses in the past few years (Chen et al., 2009).

In this study, the presence, locations and composition of SSR tracts in entire genome of HSV-1 were analyzed. The results show that the SSRs are differentially distributed between coding and noncoding regions. The SSRs in entire genome of HSV-1 also show the own distinct characteristics compared with that of other organisms. For example, G/C, GC/CG, and $(GGC)_n$ were predominant in mononucleotide, dinucletide, and trinucleotide repeats respectively. In fact, in all SSRs of HSV-1 the G and C were very abundant. The high GC content SSRs may be related to the pathogenesis of HSV-1. Besides, the nucleotide partiality of trinucleotide SSRs in coding regions is also closely related to the amino acid repeats, and it may affect the structure and function of encoded protein (Lawson and Zhang, 2006).

2. Materials and methods

The DNA sequences of HSV-1 and Cercopithecine herpesvirus 9 genome from GenBank (http://www.ncbi.nlm.nih.gov) were analyzed for the purpose of generation of SSR data. The accession numbers of the sequences are X14112 and AF275348.3 respectively. SSRs were identified and localized by using the software IMEx (Imperfect Microsatellite Extractor) (Suresh et al., 2007), which identified perfect mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats. The mononucleotide repeats with a repeat length of 6 bp or more were identified; and the di-, tri-, tetra-, penta- and hexanucleotide SSRs repeated more than 3 times could be picked out from the

Abbreviations: SSRs, simple sequence repeats; HSV-1, *Herpes simplex virus type1*.

☆ Qingjian Ouyang and Xiangyan Zhao are the Co-First Author.

* Corresponding author at: College of biology, Hunan University, Changsha, 410082, China. Tel./fax: +86 731 88822606.
E-mail address: zhongyang@hnu.cn (Z. Tan).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2012.02.049

analyzed sequences by setting relevant parameters. The sequences in coding and noncoding regions were gained based on the sequence annotation information available in GenBank database. In order to better compare the content of SSRs in coding, noncoding regions and genome-wide sequence, the relative frequency was used. The relative frequency is the number of times that studied SSRs occurs in analyzed sequence per kb.

A former triplet classification system was used for analyzing the trinucleotide SSRs (Jurka and Pethiyagoda, 1995). In this classification the trinucleotide SSRs were divided into ten types (AAT, AAG, AAC, ATG, AGT, AGG, AGC, ACG, ACC, and GGC).

3. Results and discussion

3.1. The distribution and composition of SSRs in the HSV-1 genome

A total of 1377 SSR tracts were found in the HSV-1 genome, including 948 SSRs in coding regions (Supplementary Table 1) and 429 SSRs in noncoding regions (Supplementary Table 2). The relative frequency of SSRs in coding regions was lower than that in noncoding regions. Compared with the content of different length repeats, the results showed that in HSV-1 genome the SSRs gradually decreased with the increase of repeat length (Supplementary Table 3). Similar phenomenon was also found in the *human immunodeficiency virus type 1* genomes (Tonjum et al., 1998) and many eukaryotic genome sequences (Katti et al., 2001). Some authors have reported that this could be relevant to longer repeats with higher mutation rates, which make it unstable to exist (Wierl et al., 1997).

In mononucleotide SSRs of the HSV-1 genome the C/G repeats are predominant while it was just opposite observed in many other organism genomes. For example, the C/G repeats were very rare in *Drosophila*, *Arabidopsis*, *C. elegans*, yeast (Katti et al., 2001), and *Escherichia coli* (Gur-Arie et al., 2000). In addition, the C/G repeats were also abundant in both coding and noncoding regions of HSV-1 genome. The relative frequency of C/G repeats in coding regions was lower than that in noncoding regions.

The dinucleotide SSRs are ubiquitously distributed in both coding and noncoding regions with similar relative frequency (Table 1). The CG/GC repeats are observed to be overrepresented in entire genome (Table 1), but it is reversed in some other organisms. For example: the CG/GC dinucleotide SSRs are very rare in genome sequences of human, *Drosophila, Arabidopsis, Caenorhabditis, C. elegans*, yeast (Katti et al., 2001), fungi (Karaoglu et al., 2005; Kim et al., 2008) and some prokaryotes (Field and Wills, 1998).

The higher-order SSRs (tri-, tetra, penta or hexanucleotide repeats) were also found in the HSV-1 genome. The trinucleotide SSRs in coding regions were more than in noncoding regions (Table 1). There were totally 20 tetranucleotide SSRs in the noncoding regions of the HSV-1 genome, but 45% of those were GTGG SSRs. Only 4 penta- and 4 hexanucleotide SSRs were observed in HSV-1 genome. The 4 pentanucleotide repeats were evenly distributed in coding and noncoding regions, but the hexanucleotide SSRs were only found in coding regions. The longest SSR with a length of 96 bp (Supplementary Table 1) was found to be hexanucleotide repeats in coding regions. It was significantly longer than the other repeats in entire genome.

Table 1
Distribution of SSR tracts in HSV-1 genome.

Repeat motif	Genome-wide (152,261 bp)		Coding (121,248 bp)		Noncoding (31,013 bp)	
	Number	Relative frequency	Number	Relative frequency	Number	Relative frequency
Mono-						
A	33	0.22	17	0.14	16	0.52
T	30	0.20	16	0.13	14	0.45
C	290	1.90	168	1.39	122	3.93
G	310	2.04	180	1.48	130	4.19
Total	663	4.36	381	3.14	282	9.09
Di-						
AC/CA	58	0.38	34	0.28	24	0.77
AG/GA	24	0.16	16	0.13	8	0.26
AT/TA	25	0.16	11	0.09	14	0.45
CG/GC	325	2.13	291	2.40	34	1.10
CT/TC	19	0.12	11	0.13	8	0.26
GT/TG	51	0.33	31	0.26	20	0.64
Total	502	3.28	394	3.29	108	3.48
Tri-	184	1.21	167	1.38	17	0.55
Tetra-	20	0.13	0	0.00	20	0.64
Penta-	4	0.03	2	0.02	2	0.06
Hexa-	4	0.03	4	0.03	0	0.00

SSR relative frequency is the total repeats per kb of sequence analyzed. For example: the relative frequency of the mononucleotide repeats A in HSV-1 genome = (33/152,261) × 1000 ≈ 0.22; the relative frequency of the mononucleotide repeats A in coding regions = (17/121,248) × 1000 ≈ 0.14; the relative frequency of the mononucleotide repeats A in noncoding regions = (16/31,013) × 1000 ≈ 0.52.
The types of higher-order SSR (tri-, tetra, penta or hexanucleotide repeats) are not be shown one by one.

3.2. The different distribution of trinucleotide SSRs in HSV-1 genome

All trinucleotide repeat combinations were grouped into ten unique classes (Jurka and Pethiyagoda, 1995), namely, $(AAT)_n$, $(AAC)_n$ $(AAG)_n$, $(ATG)_n$, $(AGT)_n$, $(AGG)_n$, $(AGC)_n$, $(ACG)_n$, $(ACC)_n$ and $(GGC)_n$. The occurrence frequency of ten repeat types in HSV-1 was shown in Table 2. The type of $(GGC)_n$ was predominant in trinucleotide repeats, but the $(AAT)_n$ was absent. The deviation of the trinucleotide repeat type showed different features in diverse organisms. For example, human chromosomes 21 and 22 contain more $(AAT)_n$ and $(AAC)_n$ repeats, the $(AGC)_n$ repeats are predominant in *Drosophila* chromosomes, and the *Arabidopsis* and *C. elegans* chromosomes have comparatively higher frequencies of $(AAG)_n$ trinucleotide repeats. In contrast, the yeast genome contains more $(AAT)_n$, $(AAG)_n$, $(AAC)_n$, $(ATG)_n$ and $(AGC)_n$ repeats (Katti et al., 2001). Surprisingly, there is a common characteristic that in those studied organisms the type of $(GGC)_n$ is relatively few or none at all except for HSV-1. With the type $(GGC)_n$ repeats abound in the coding regions of HSV-1, it may have a special function for of HSV-1. In fact, it was reported that the $(CGC)_n$ repeats in gene Poly(A)-binding protein2 (PABP2) of human genome had adjusted the Oculopharyngeal muscular dystrophy function (Brais et al., 1998) and the $(GGC)_n$ repeats in 5′-UTR of gene Fragile X mental retardation-2 (FMR-2) could cause abnormal neuronal gene regulation (Cummings and Zoghbi, 2000). When the $(CGC)_n$ and $(GGC)_n$ repeats are expanded in human genome, they will cause the above diseases. Interestingly, the two repeats which have special functions in human genome abound in HSV-1 genome by chance, though we do not know its specific functions.

Table 2
Distribution of different types of trinucleotide SSRs in HSV-1 genome.

Type[a]	Repeat motifs						Total
T1	AAT(0)[b]	ATA(0)	TAA(0)	ATT(0)	TTA(0)	TAT(0)	0
T2	AAG(1)	AGA(0)	GAA(0)	CTT(1)	TTC(1)	TCT(1)	4
T3	AAC(0)	ACA(1)	CAA(1)	GTT(0)	TTG(0)	TGT(1)	3
T4	ATG(0)	TGA(0)	GAT(0)	CAT(0)	ATC(0)	TCA(1)	1
T5	AGT(0)	GTA(1)	TAG(0)	ACT(0)	CTA(0)	TAC(1)	2
T6	AGG(2)	GGA(2)	GAG(7)	CCT(1)	CTC(1)	TCC(4)	17
T7	AGC(2)	GCA(5)	CAG(1)	GCT(2)	CTG(1)	TGC(2)	13
T8	ACG(8)	CGA(7)	GAC(7)	CGT(8)	GTC(5)	TCG(3)	38
T9	ACC(3)	CCA(5)	CAC(4)	GGT(4)	GTG(4)	TGG(2)	22
T10	GGC(14)	GCG(16)	CGG(9)	GCC(6)	CCG(11)	CGC(18)	84

[a] The trinucleotide SSRs are divided into ten types(T1–T10), the method of classification was cited from Jurka and Pethiyagoda (1995).
[b] The numbers of repeat motifs in HSV-1 genome are listed.