Infection, Genetics and Evolution 11 (2011) 1646–1654

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Microsatellite is an important component of complete *Hepatitis C virus* genomes

Ming Chen [a,b], Zhongyang Tan [c,*], Guangming Zeng [a,b,*]

[a] College of Environmental Science and Engineering, Hunan University, Changsha 410082, China
[b] Key Laboratory of Environmental Biology and Pollution Control, Hunan University, Ministry of Education, Changsha 410082, China
[c] College of Biology, State Key Laboratory for Chemo/Biosensing and Chemometrics, Hunan University, Changsha 410082, China

ARTICLE INFO

Article history:
Received 11 January 2011
Received in revised form 2 June 2011
Accepted 16 June 2011
Available online 23 June 2011

Keywords:
Microsatellite
Hepatitis C virus
Simple sequence repeat
Comparative genomics

ABSTRACT

Microsatellites are common and play diverse roles in eukaryotic and prokaryotic genomes. However, to our knowledge, microsatellite distribution remains largely enigmatic in viruses yet is crucial for understanding instability of viral genomes. We have therefore, examined microsatellite distribution in 54 complete genomes of *Hepatitis C virus* (HCV) from six genotypes, showing microsatellites were an important component of HCV genomes. Our results showed, in all analyzed HCV genomes, genome size and GC content had a weak influence on number, relative abundance and relative density of microsatellites, respectively. For each HCV genome, mono-, di- and trinucleotide repeats were very predominant, whereas other types of repeats rarely occurred. Our results revealed that the occurrence of microsatellites was significantly less than higher prokaryotes and eukaryotes and that all identified microsatellites were very short. The discovery of microsatellites in HCV genomes may become useful for population genetic, evolutionary analysis and strain (isolate) identification.

1. Introduction

Microsatellites or simple sequence repeats (SSRs) consist of mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats (Chen et al., 2011a, 2010), being highly polymorphic in eukaryotic (Ellegren, 2004; Li et al., 2004; Rajendrakumar et al., 2007) and prokaryotic genomes (Gur-Arie et al., 2000). The abundance of these six types of microsatellites varies between different species (Karaoglu et al., 2005). Microsatellites are found in diverse regions of genomes, including 3′-UTR, 5′-UTR, exon and intron (Li et al., 2004; Rajendrakumar et al., 2007; Toth et al., 2000). Triplet repeats are more common than non-triplet repeats in coding regions in eukaryotes, due to the fact that length changes in non-triplet repeats may lead to frameshift mutations in coding regions (Ellegren, 2004; Li et al., 2004). The most common microsatellite motifs may be different in various species. For example, in *Aspergillus nidulans*, the most common microsatellite motifs are AT/TA repeats, whereas AG/GA repeats are most abundant in *Fusarium graminearum* (Karaoglu et al., 2005). Genome size and GC content have been shown to have a certain influence on the occurrence of microsatellites in several species (Coenye and Vandamme, 2005; Dieringer and Schlotterer, 2003). Strand slippage and unequal recombination have been proposed to explain microsatellite instability (Toth et al., 2000). Intrinsic features of microsatellites (repeat number, length and motif size) have the strongest influence on the microsatellite mutability, whereas regional genomic factors have only minor effects (Kelkar et al., 2008). Mutability of microsatellites grows with the number of repeats, most likely because of an increase in the probability of slippage (Pearson et al., 2005). Imperfection in microsatellites is thought to influence replication slippage by limiting expansion of microsatellite size (Mudunuri and Nagarajaram, 2007). Because of their high instability, microsatellites are believed to serve a functional role in genome evolution (Tautz et al., 1986). It has been shown that microsatellites are associated with various genetic diseases (Usdin, 2008), including Huntington's disease and spinobulbar muscular atrophy (Li et al., 2004). Some microsatellites are related to bacterial pathogenesis and virulence, and can increase the antigenic variance to escape from the host immune response (Li et al., 2004; Mrazek et al., 2007).

However, despite widespread distribution and functional significance in genomes, little is known about distribution rules of microsatellites in viral genomes. Numerous polymorphic microsatellites were detected in *human cytomegalovirus* (HCMV), *herpes simplex virus type 1* (HSV-1), and *Ostreid herpesvirus 1* (OsHV-1) genomes (Davis et al., 1999; Deback et al., 2009; Segarra et al., 2010). Microsatellites have been also observed in genomes of epidemic human respiratory *adenovirus*, *influenza virus*, and *Sin Nombre virus* (Houng et al., 2009; Mudunuri et al., 2009). Our

Abbreviations: HCV, *Hepatitis C virus*; SSRs, simple sequence repeats; HIV-1, *Human Immunodeficiency Virus Type 1*.

* Corresponding authors.
E-mail addresses: zhongyang@hnu.cn (Z. Tan), zgming@hnu.cn (G. Zeng).

1567-1348/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2011.06.012

recent report comprehensively analyzed microsatellite distribution in viral pre-microRNAs, and found microsatellites were extensively presented in these small non-coding RNA sequences (Chen et al., 2010). In a previous study performed by us, *Human Immunodeficiency Virus Type 1* (HIV-1) was thought to be an excellent system to study evolution and roles of viral microsatellites, and this analysis indicated microsatellites were very short in length and were in low abundance (Chen et al., 2009). However, there remains much to be confirmed whether these features from HIV-1 genomes are suitable for other viruses. Moreover, until recently, there is some lack of knowledge about mononucleotide repeats and the correlation between genome features and microsatellite distribution in viral genomes. HCV has a positive sense RNA genome that is composed of a single open reading frame, mostly containing six genotypes (genotypes 1, 2, 3, 4, 5 and 6) (Kuiken et al., 2005). Genomic diversity of HCV can provide a very good opportunity to address abovementioned problems.

In the present study, we present a comprehensive analysis of the distribution of microsatellites over 6 nt in 54 complete HCV genomes which belong to six genotypes. We analyzed distribution of mononucleotide repeats and explored the correlation between genome features and microsatellite distribution using linear regression analyses for the first time. We also compared our results and other organisms, and discussed their similarity and difference.

2. Materials and methods

2.1. HCV genome sequences

We downloaded 54 complete HCV genomes from GenBank (http://www.ncbi.nlm.nih.gov). Analyzed sequences fall into six genotypes. The availability of complete HCV genomes from different genotypes is non-identical. The availably complete genomes from genotypes 1, 2, 3 and 6 are significantly more than those from genotypes 4 and 5. At the time of writing, only one complete genome was available for genotypes 4 and 5, respectively. Thus, the selected number of complete HCV genomes from various genotypes is different in this study. Detailed information on these genomes was given in Table 1.

2.2. Identification of microsatellites

We extracted imperfect mononucleotide repeats with lengths of 6 nt or more in each of surveyed HCV genomes using an IMEx program (Mudunuri and Nagarajaram, 2007). Perfect di-, tri-, tetra-, penta- and hexanucleotide repeats were detected by use of SSRIT (Temnykh et al., 2001); these microsatellites were repeating three times or more. These parameters were based on (i) Rajendrakumar et al. selected these significant threshold values for analyzing the microsatellite distribution in organellar genomes of rice (Rajendrakumar et al., 2007), (ii) we have made a survey of microsatellites with repeat number $\geqslant 3$ in 81 completed HIV-1 genomes (Chen et al., 2009), and (iii) most microsatellites were very short in viruses. Each sequence was analyzed separately. To differentiate coding regions, 3′-UTR and 5′-UTR, the existing annotation (the "CDS" features) were extracted from the corresponding GenBank files. The starting position of CDS in U89019 (S16) is significantly different from other 53 sequences. This may be a result of a correctly assigned start codon. Thus, the locations of microsatellites in S16 genome are not given. Note that a small number of microsatellites locate in overlap region of coding and non-coding regions in several HCV genomes, and these microsatellites are reported as non-coding microsatellites in the present study.

2.3. Calculation of the expected number of microsatellites

We compared the observed number of microsatellites (O) with the expected number of microsatellites (E) in the form of a ratio of O/E in order to evaluate whether microsatellites were over- or underrepresented in HCV genome sequences. To assess statistical significance of the microsatellite representation (O/E), we used Z-scores defined as $(O-E)/\sqrt{E}$ (Mrazek, 2006). The expected number of microsatellite composed of M_t (M is motif of the microsatellite with repeat number of t, and its length is L) in a genome of length G was calculated as given by (de wachter, 1981):

Table 1
List of analyzed HCV genomes.

No.	Acc. No.	Genotype	Size (nt)	GC%	No.	Acc. No.	Genotype	Size (nt)	GC%
S1	AB520610	1	9598	58.5	S28	AF238484	2	9416	57.6
S2	AF271632	1	9618	58.3	S29	AF238485	2	9416	57.7
S3	EF621489	1	9599	58.6	S30	AF238486	2	9416	56
S4	EU155241	1	9275	58.9	S31	D50409	2	9513	57.7
S5	EU256080	1	9329	58.5	S32	NC_009823	2	9711	56.9
S6	FJ024275	1	9269	59	S33	D17763	3	9456	55.6
S7	FJ390395	1	9264	58.9	S34	D28917	3	9454	55.8
S8	FN435993	1	9613	57.9	S35	D49374	3	9444	56
S9	L02836	1	9400	57.9	S36	D63821	3	9450	54.9
S10	M62321	1	9401	58.9	S37	NC_009825	4	9355	56.2
S11	M67463	1	9416	58.8	S38	NC_009826	5	9343	57.1
S12	M84754	1	9425	58.4	S39	AY878650	6	9388	56
S13	S62220	1	9440	58.9	S40	AY878651	6	9373	55.7
S14	U01214	1	9446	58.2	S41	DQ278891	6	9440	55.9
S15	U16362	1	9415	58.5	S42	DQ278893	6	9430	55.7
S16	U89019	1	9400	58.3	S43	DQ278894	6	9441	55.4
S17	AB030907	2	9654	56.1	S44	DQ314805	6	9468	55.5
S18	AB031663	2	9488	55.4	S45	DQ480519	6	9358	56.4
S19	AB047639	2	9678	58.3	S46	DQ480520	6	9358	56.5
S20	AF169002	2	9661	57.4	S47	DQ480522	6	9358	56.4
S21	AF169003	2	9693	57.2	S48	DQ480523	6	9358	56.2
S22	AF169004	2	9653	57.5	S49	DQ480524	6	9361	56.9
S23	AF169005	2	9700	57.3	S50	DQ835770	6	9447	55.9
S24	AF177036	2	9711	56.9	S51	EF424627	6	9450	57.2
S25	AF238481	2	9416	57.8	S52	EF424628	6	9453	56
S26	AF238482	2	9416	57.8	S53	EF424629	6	9459	56.3
S27	AF238483	2	9416	57.4	S54	NC_009827	6	9628	55.4

$$\mathrm{Exp}(M_t) = f(M)^t[1 - f(M)][G'(1 - f(M)) + 2L] \quad (1)$$

$$G' = G - tL - 2L + 1 \quad (2)$$

where $\mathrm{Exp}(M_t)$ is the expected number of M_t, and $f(M)$ is the probability of M.

2.4. Statistical analysis

We used SPSS 18.0 and EXCEL 2007 to perform all statistical analysis. Linear regression was used to reveal the correlation between the number, relative abundance, relative density of microsatellites and two genome features (genome size and GC content).

3. Results and discussion

Different studies used different parameters to search a genome for microsatellites. Power et al. showed the number of microsatellites significantly changed by increasing or decreasing the threshold value of repeat units (Power et al., 2009). Thus, it is very important to select an appropriate threshold value of repeat length. Previous studies have selected threshold repeat length of 6 nt in HIV-1 genomes whose genome sizes are very similar with HCV genomes (Chen et al., 2009). Likewise, in the present study, we also analyzed microsatellites over 6 nt in 54 completely sequenced HCV genomes. Until now, there are no studies which have systematically addressed and compared the distribution of mono-

Table 2
Occurrence of microsatellites among coding and non-coding regions for HCV genomes.

No.	Mononucleotide repeats				Microsatellites[2–6]			
	Coding	5′-UTR	3′-UTR	Genome-wide	Coding	5′-UTR	3′-UTR	Genome-wide
S1	20	1	4	25	19	0	2	21
S2	18	1	1	20	19	0	2	21
S3	13	1	1	15	24	0	2	26
S4	18	1	0	19	19	0	0	19
S5	20	1	0	21	17	0	0	17
S6	15	1	0	16	19	0	0	19
S7	19	1	0	20	25	0	0	25
S8	18	1	2	21	22	0	3	25
S9	19	1	0	20	26	0	0	26
S10	20	1	0	21	18	0	0	18
S11	14	1	0	15	23	0	0	23
S12	19	1	1	21	22	0	0	22
S13	18	1	1	20	20	0	0	20
S14	17	1	1	19	17	0	0	17
S15	25	1	0	26	24	0	0	24
S16	N/A	N/A	N/A	18	N/A	N/A	N/A	23
S17	18	1	2	21	20	0	2	22
S18	12	1	1	14	27	0	0	27
S19	16	1	1	18	26	0	3	29
S20	12	1	1	14	24	0	3	27
S21	17	1	1	19	26	0	4	30
S22	13	1	1	15	27	0	3	30
S23	14	1	1	16	18	0	3	21
S24	15	1	1	17	23	0	3	26
S25	16	1	0	17	22	0	0	22
S26	16	1	0	17	24	0	0	24
S27	14	1	0	15	27	0	0	27
S28	17	1	0	18	26	0	0	26
S29	18	1	0	19	26	0	0	26
S30	16	1	0	17	28	0	0	28
S31	17	1	1	19	25	0	0	25
S32	15	1	1	17	23	0	3	26
S33	13	1	2	16	24	0	0	24
S34	12	1	2	15	29	0	0	29
S35	11	1	0	12	23	0	0	23
S36	15	1	1	17	23	0	0	23
S37	9	1	1	11	20	0	0	20
S38	17	1	0	18	25	0	0	25
S39	8	1	0	9	21	0	0	21
S40	9	1	0	10	26	0	0	26
S41	8	1	1	10	21	0	0	21
S42	9	1	1	11	26	0	0	26
S43	8	1	1	10	26	0	0	26
S44	16	1	1	18	16	0	0	16
S45	13	1	0	14	24	0	0	24
S46	11	1	0	12	23	0	0	23
S47	8	1	0	9	24	0	0	24
S48	11	1	0	12	33	0	0	33
S49	12	1	0	13	21	0	0	21
S50	13	1	1	15	25	0	0	25
S51	11	1	0	12	28	0	0	28
S52	19	1	0	20	25	0	0	25
S53	10	1	1	12	22	0	0	22
S54	10	1	1	12	26	0	2	28

N/A, not available (see Section 2).

Table 3
Occurrence, relative abundance and relative density of mononucleotide repeats in analyzed HCV genomes.

No.	Repeat type				Total	MRA[a]	MRD[b]	O/E[c]	No.	Repeat type				Total	MRA[a]	MRD[b]	O/E[c]
	A	T	G	C						A	T	G	C				
S1	2	4	9	10	25	2.60	21.05	2.99	S28	0	1	7	10	18	1.91	13.17	2.26
S2	0	1	8	11	20	2.08	23.50	2.25	S29	0	2	10	7	19	2.02	12.96	2.20
S3	0	1	7	7	15	1.56	15.52	1.87	S30	0	0	7	10	17	1.81	12.00	2.58
S4	0	0	8	11	19	2.05	13.05	2.14	S31	0	1	10	8	19	2.00	15.24	2.35
S5	1	0	9	11	21	2.25	14.47	2.44	S32	0	1	7	9	17	1.75	23.89	2.65
S6	0	0	8	8	16	1.73	11.22	1.90	S33	2	3	7	4	16	1.69	13.01	2.08
S7	0	0	7	13	20	2.16	13.60	2.34	S34	0	4	5	6	15	1.59	10.79	3.19
S8	1	2	10	8	21	2.18	23.82	2.55	S35	0	1	3	8	12	1.27	9.00	2.71
S9	1	0	12	7	20	2.13	13.51	2.58	S36	2	2	6	7	17	1.80	12.70	2.53
S10	1	0	8	12	21	2.23	15.00	2.85	S37	1	1	5	4	11	1.18	8.55	1.95
S11	0	0	8	7	15	1.59	10.51	1.75	S38	3	0	5	10	18	1.93	12.31	2.51
S12	1	1	10	9	21	2.23	15.49	2.35	S39	0	0	3	6	9	0.96	6.07	1.48
S13	0	1	11	8	20	2.12	14.62	2.40	S40	0	0	3	7	10	1.07	6.93	1.58
S14	0	1	10	8	19	2.01	17.15	2.18	S41	0	1	3	6	10	1.06	8.90	1.63
S15	2	1	12	11	26	2.76	17.63	2.83	S42	0	1	3	7	11	1.17	8.59	1.73
S16	1	0	8	9	18	1.91	12.77	2.18	S43	0	1	6	3	10	1.06	9.53	1.58
S17	1	3	8	9	21	2.18	18.65	2.67	S44	0	2	9	7	18	1.90	15.00	2.78
S18	1	1	4	8	14	1.48	10.43	2.18	S45	1	1	4	8	14	1.50	9.94	1.74
S19	0	1	5	12	18	1.86	19.22	2.27	S46	1	1	5	5	12	1.28	8.66	1.49
S20	0	2	3	9	14	1.45	15.22	1.92	S47	1	0	2	6	9	0.96	6.73	1.21
S21	0	1	8	10	19	1.96	23.32	2.56	S48	1	1	4	6	12	1.28	8.55	1.49
S22	0	2	4	9	15	1.55	15.02	2.15	S49	1	1	3	8	13	1.39	9.40	1.61
S23	0	2	5	9	16	1.65	21.55	2.07	S50	0	2	5	8	15	1.59	12.70	1.84
S24	0	1	7	9	17	1.75	23.89	2.65	S51	0	2	7	3	12	1.27	11.11	1.71
S25	0	1	8	8	17	1.81	12.53	2.17	S52	1	1	7	11	20	2.12	16.29	2.88
S26	1	0	7	9	17	1.81	12.96	2.31	S53	0	2	7	3	12	1.27	11.63	2.16
S27	0	0	7	8	15	1.59	11.05	1.88	S54	0	1	3	8	12	1.25	15.68	2.07

[a] Relative abundance is the total mononucleotide repeats per kb of sequence analyzed.
[b] Relative density is defined as the total length (nt) contributed by each mononucleotide repeat per kb of sequence analyzed.
[c] Observed number of mononucleotide repeats/expected number of mononucleotide repeats.

nucleotide repeats in any viral genomes. Mononucleotide repeats are found to strongly affect the local mutation rate (Levinson and Gutman, 1987). The presence of mononucleotide repeats is thought to be an important determinant of stability (Ackermann and Chao, 2006). Thus, the analysis of imperfect mononucleotide repeats can give a good insight into the evolutionary and potential roles of microsatellites. In this study, we identified imperfect mononucleotide repeats over 6 nt. To compare with our previous results in HIV-1 genomes where only perfect di-, tri-, tetra-, penta- and hexanucleotide repeats are identified (Chen et al., 2009), we divided microsatellites into two categories: mononucleotide repeats and microsatellites[2–6] (refer to perfect di-, tri-, tetra-, penta- and hexanucleotide repeats). Among *Escherichia coli* (*E. coli*), microsatellites are richer in coding regions than in non-coding regions, because the bulk of the genome is composed of open reading frames (Chen et al., 2011b; Gur-Arie et al., 2000). Similarly, coding density of HCV genome is also very high (Kuiken et al., 2005). To assess whether microsatellites were also richer in coding regions for HCV, we detected microsatellites in coding regions, 3′-UTR and 5′-UTR. Our results clearly showed microsatellites were significantly more abundant in coding regions than in non-coding regions in HCV genomes (Table 2 and Supplementary Tables 1 and 2). As expected, this result was similar to that of *E. coli* (Gur-Arie et al., 2000).

3.1. Mononucleotide repeats

Mononucleotide repeats were observed in each HCV genome, showing relatively high occurrence. Obviously, poly (G/C) repeats were significantly more predominant than poly (A/T) repeats in each complete HCV genome (Table 3). It is generally assumed that the higher poly (G/C) frequencies in the genomes can be attributable to the high GC content of the genomes (Karaoglu et al., 2005). However, it must be noted that the GC content is only slightly higher than AT content in each of analyzed sequences, but leading to a significant difference between the occurrences of both poly (G/C) and poly (A/T) repeats (t test, $p < 0.001$). Thus, GC content had a weak influence on the occurrence of poly (G/C) repeats in HCV genomes. Each HCV genome contained 0–6 poly (A/T) repeats and 8–23 poly (G/C) repeats. Interestingly, in general, poly (A/T) tracts are more abundant than poly (G/C) tracts in eukaryotic and prokaryotic genomes (Gur-Arie et al., 2000; Karaoglu et al., 2005; Toth et al., 2000). For example, the genome of *Saccharomyces cerevisiae* showed a significant preference for poly (A/T) over poly (G/C) (99.5% vs. 0.5%) (Karaoglu et al., 2005). Mononucleotide repeats were consistently overrepresented in all surveyed HCV genomes (O/E ranged from 1.21 to 3.19) (Table 3 and Supplementary Table 3). The strongest overrepresentation of mononucleotide repeats was exhibited by the sequence D28917. The relative abundance of mononucleotide repeats was similar in analyzed HCV genomes overall, ranging from 0.96 to 2.76. The highest relative density of mononucleotide repeats was found in AF177036 and NC_009823 (23.89 nt/kb), followed by FN435993 (23.82 nt/kb), and the lowest one was in AY878650 (6.07 nt/kb). Some authors have assessed the relationship between microsatellite content and genome size (Chen et al., 2009; Coenye and Vandamme, 2005; Karaoglu et al., 2005), showing the total microsatellite contents in these organisms are not directly proportional to the genome sizes (Chen et al., 2009; Karaoglu et al., 2005), although it is generally inferred that the larger genome owns more microsatellites than do the smaller one (Hancock, 2002). Moreover, the correlation between GC content and distribution of mononucleotide repeats is analyzed as well in prokaryotes (Coenye and Vandamme, 2005). Similarly, we surveyed the correlation between distribution of mononucleotide repeats (number, relative abundance and relative density) and two genome features (genome size and GC content). Our results here indicated that genome size did not significantly correlate with number ($R^2 = 0.0257$, $p > 0.05$)

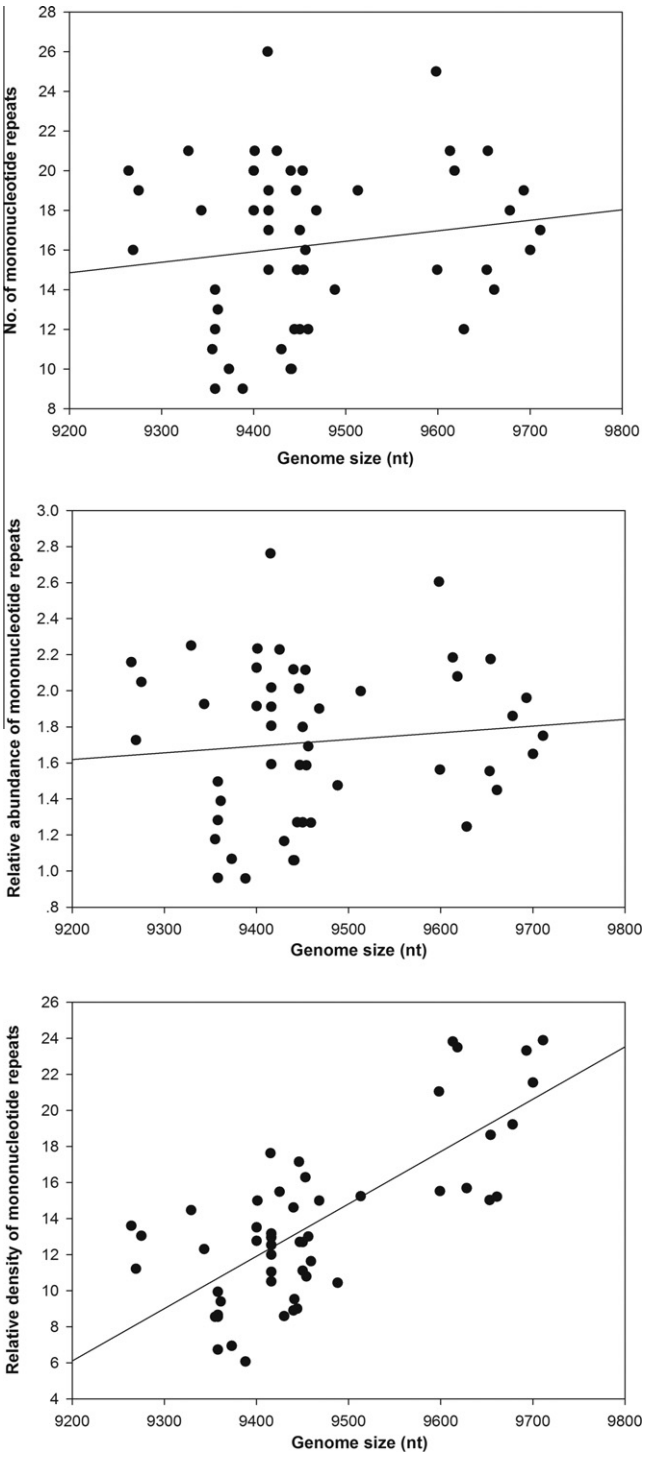

Fig. 1. Relationship between the genome size and the number, relative abundance and relative density of mononucleotide repeats in analyzed HCV genomes. Relative abundance is the total mononucleotide repeats per kb of sequence analyzed. Relative density is defined as the total length (nt) contributed by each mononucleotide repeat per kb of sequence analyzed.

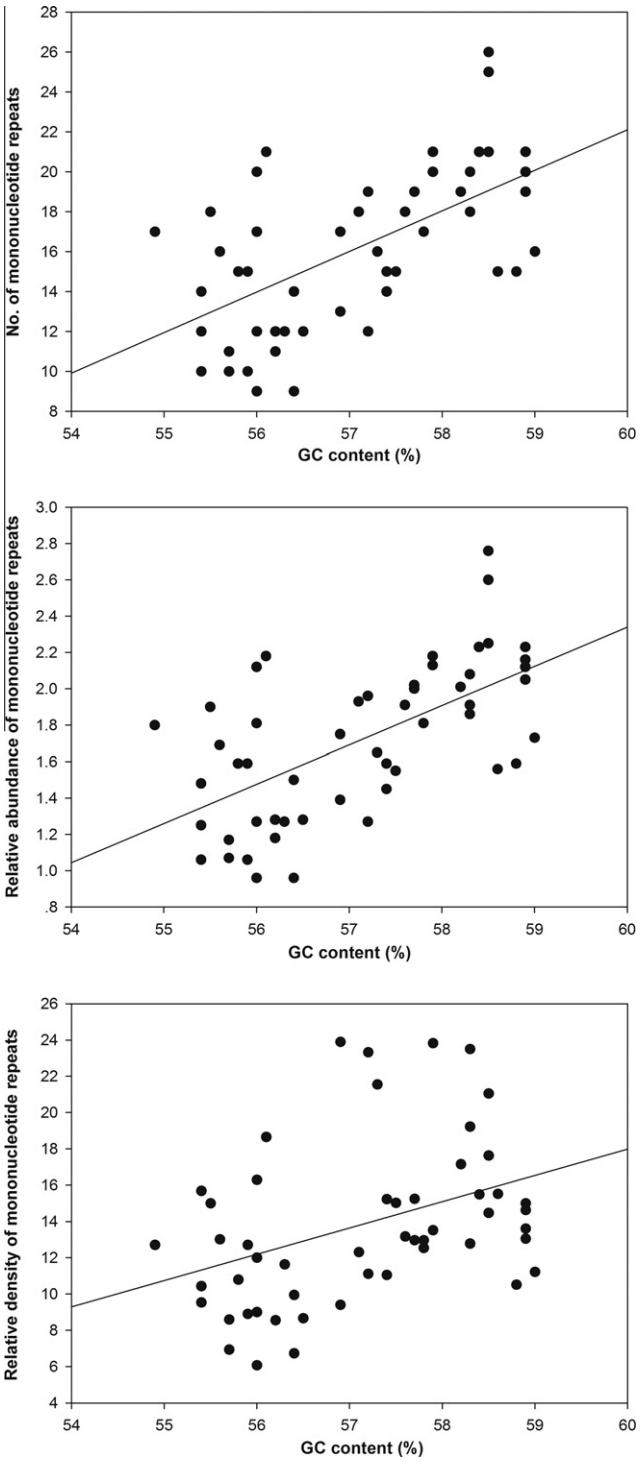

Fig. 2. Relationship between the GC content and the number, relative abundance and relative density of mononucleotide repeats in analyzed HCV genomes. See Fig. 1 legend.

and relative abundance ($R^2 = 0.0114$, $p > 0.05$) of mononucleotide repeats but significantly correlated with relative density ($R^2 = 0.5633$, $p < 0.05$) of mononucleotide repeats (Fig. 1). Compared with genome size, GC content was found to be weakly but significantly correlated with number ($R^2 = 0.3712$, $p < 0.05$) and relative abundance ($R^2 = 0.3818$, $p < 0.05$) of mononucleotide repeats, except for relative density of mononucleotide repeats ($R^2 = 0.1375$, $p > 0.05$) (Fig. 2).

3.2. Microsatellite[2–6]

To compare with our previous results in HIV-1 genomes where only microsatellites[2–6] were surveyed (Chen et al., 2009) and to investigate whether the features of microsatellites[2–6] from HIV-1

Table 4
Occurrence, relative abundance, relative density and representation of microsatellites[2–6].

No.	Di/tri/tetra/penta/hexa[a]	Total	RA[b]	RD[c]	O/E[d]	No.	Di/tri/tetra/penta/hexa[a]	Total	RA[b]	RD[c]	O/E[d]
S1	17/4/0/0/0	21	2.19	14.38	0.88	S28	22/4/0/0/0	26	2.76	18.48	1.32
S2	16/5/0/0/0	21	2.18	15.08	1.23	S29	20/5/0/1/0	26	2.76	19.54	1.14
S3	21/5/0/0/0	26	2.71	18.02	1.25	S30	23/4/1/0/0	28	2.97	20.18	1.25
S4	15/4/0/0/0	19	2.05	14.23	1.13	S31	19/6/0/0/0	25	2.63	18.29	1.23
S5	13/4/0/0/0	17	1.82	12.43	1.10	S32	22/4/0/0/0	26	2.68	17.92	1.24
S6	15/4/0/0/0	19	2.05	13.59	0.89	S33	17/7/0/0/0	24	2.54	17.45	1.39
S7	19/6/0/0/0	25	2.70	18.35	1.32	S34	19/9/1/0/0	29	3.07	22.11	1.97
S8	18/6/1/0/0	25	2.60	18.31	1.28	S35	18/5/0/0/0	23	2.44	17.15	1.14
S9	19/7/0/0/0	26	2.77	19.26	1.45	S36	20/3/0/0/0	23	2.43	15.77	1.14
S10	13/5/0/0/0	18	1.91	13.51	1.25	S37	15/5/0/0/0	20	2.14	14.86	1.36
S11	18/5/0/0/0	23	2.44	16.46	1.22	S38	20/5/0/0/0	25	2.68	17.87	1.43
S12	16/6/0/0/0	22	2.33	16.13	1.15	S39	17/4/0/0/0	21	2.24	14.91	1.00
S13	17/3/0/0/0	20	2.12	13.88	1.03	S40	21/5/0/0/0	26	2.77	18.24	1.07
S14	15/2/0/0/0	17	1.80	11.86	1.18	S41	17/4/0/0/0	21	2.22	14.83	1.00
S15	18/6/0/0/0	24	2.55	17.63	1.41	S42	21/5/0/0/0	26	2.76	18.13	1.06
S16	16/6/1/0/0	23	2.45	17.45	1.48	S43	23/3/0/0/0	26	2.75	17.69	1.24
S17	17/4/1/0/0	22	2.28	15.95	1.27	S44	11/5/0/0/0	16	1.69	11.72	0.89
S18	23/4/0/0/0	27	2.85	18.76	1.31	S45	17/7/0/0/0	24	2.56	17.63	1.43
S19	24/5/0/0/0	29	3.00	20.15	1.23	S46	18/4/1/0/0	23	2.46	16.67	1.44
S20	22/5/0/0/0	27	2.79	18.94	1.32	S47	18/6/0/0/0	24	2.56	17.74	2.26
S21	23/7/0/0/0	30	3.10	21.15	1.23	S48	24/9/0/0/0	33	3.53	24.26	2.48
S22	25/5/0/0/0	30	3.11	20.41	1.49	S49	16/5/0/0/0	21	2.24	15.28	1.47
S23	19/2/0/0/0	21	2.16	14.23	0.99	S50	20/5/0/0/0	25	2.65	17.47	1.19
S24	22/4/0/0/0	26	2.68	17.92	1.24	S51	21/7/0/0/0	28	2.96	20.21	1.36
S25	17/4/1/0/0	22	2.34	16.36	1.06	S52	21/4/0/0/0	25	2.64	17.14	1.25
S26	20/4/0/0/0	24	2.55	16.99	1.19	S53	14/8/0/0/0	22	2.33	17.23	1.23
S27	22/4/0/1/0	27	2.87	19.86	1.40	S54	21/7/0/0/0	28	2.91	19.84	1.34

Microsatellites[2–6] indicates the di-, tri-, tetra-, penta- and hexanucleotide repeats.
[a] Di/tri/tetra/penta/hexa: number of dinucleotide repeats/trinucleotide repeats/tetranucleotide repeats/pentanucleotide repeats/hexanucleotide repeats, respectively.
[b] Relative abundance is the total microsatellites[2–6] per kb of sequence analyzed.
[c] Relative density is defined as the total length (nt) contributed by each microsatellite[2–6] per kb of sequence analyzed.
[d] Observed number of microsatellites[2–6]/expected number of microsatellites[2–6].

are consistent with other viruses, we investigated the presence of microsatellites[2–6] in 54 complete HCV genomes from six genotypes. Our results showed (i) microsatellites[2–6] were prevalently present in these surveyed sequences, (ii) with the repeat unit increasing, the number of repeats became less and less, and (iii) minor differences for total, relative abundance and relative density of microsatellites[2–6] could be seen between diverse HCV genomes, respectively. These features are very consistent with our previous results in HIV-1 genomes (Chen et al., 2009).

An overview of the occurrence, relative abundance and relative density of microsatellites[2–6] for HCV genomes was shown in Table 4. For these surveyed HCV genomes, the number of microsatellites[2–6] ranged from 16 to 33. Comparison between observed number of microsatellites[2–6] and expected number of microsatellites[2–6] based on the formula proposed by de wachter (1981) revealed the ratio of O/E was variable with the range from 0.88 to 2.48. DQ480523 had the highest relative abundance of microsatellites[2–6] (3.53 repeats/kb) whereas DQ314805 had the lowest (1.69 repeats/kb). The relative density of microsatellites[2–6] was nearly as equally represented across the 54 complete HCV genomes, regardless of whether the sequences are selected from different genotypes. The highest relative density was 24.26 nt/kb found in the sequence DQ480523 which is from genotype 6, and the lowest microsatellite[2–6] density was 11.72 nt/kb in the sequence DQ314805 which likewise belongs to genotype 6. The relative density of microsatellites[2–6] in most surveyed genomes was smaller than 20 nt/kb. In sharp contrast to this, the relative density of microsatellites[2–6] was 20 nt/kb or more in most analyzed HIV-1 genomes (Chen et al., 2009). The microsatellite[2–6] with the longest nucleotide stretch belonged to AF238483 and AF238485, consisting of $(GCTCT)_3$ motif of 15 nt. 11 complete HCV genomes investigated contained microsatellites of length $\geqslant$12 nt (Supplementary Table 4). For the longest microsatellite[2–6] motifs, only three types of repeats (tri-, tetra- and pentanucleotide repeat types) were found in all analyzed HCV genomes; most of longest microsatellite[2–6] motifs were 9 nt in length, and belonged to trinucleotide repeat type. This is drastically different from HIV-1 and fungi genomes in which the repeat types of the longest microsatellite motifs are diverse (Chen et al., 2009; Karaoglu et al., 2005). The plots for correlation between microsatellites[2–6] distribution (number, relative abundance and relative density) and genome features (genome size and GC content) were shown in Figs. 3 and 4. Clearly, number ($R^2 = 0.0939$, $p < 0.05$) of microsatellites[2–6] was weakly but significantly correlated with genome size, whereas relative abundance ($R^2 = 0.0506$, $p > 0.05$) and relative density ($R^2 = 0.0385$, $p > 0.05$) of microsatellites[2–6] were not significantly related to genome size. GC content did not show significant correlation with number ($R^2 = 0.0706$, $p > 0.05$) and relative density ($R^2 = 0.0582$, $p > 0.05$) of microsatellites[2–6], but had a significant relation to relative abundance ($R^2 = 0.0724$, $p < 0.05$) of microsatellites[2–6].

In the present study, we divided dinucleotide repeats into six types: AG/GA, GT/TG, AC/CA, CT/TC, AT/TA and CG/GC. Our results clearly indicated six types of dinucleotide repeats were variable in number in different HCV genomes, respectively (Supplementary Table 5). Our previous observation in all 81 complete HIV-1 genomes showed AG/GA repeats were most predominant among dinucleotide repeats (Chen et al., 2009). However, GT/TG repeats were the most abundant dinucleotide repeat types in more than half of surveyed HCV genomes, followed by AC/CA repeats (Supplementary Table 5). Moreover, our results also revealed an additional difference in the two species: overall relative density of microsatellites[2–6] in complete HCV genomes was lower than that in HIV-1 genomes (11.72–24.26 nt/kb vs. 16–35 nt/kb; Chen et al., 2009). An interesting result was that CG/GC repeats were very predominant in HCV genomes and even could be the most common dinucleotide repeats in some sequences such as in

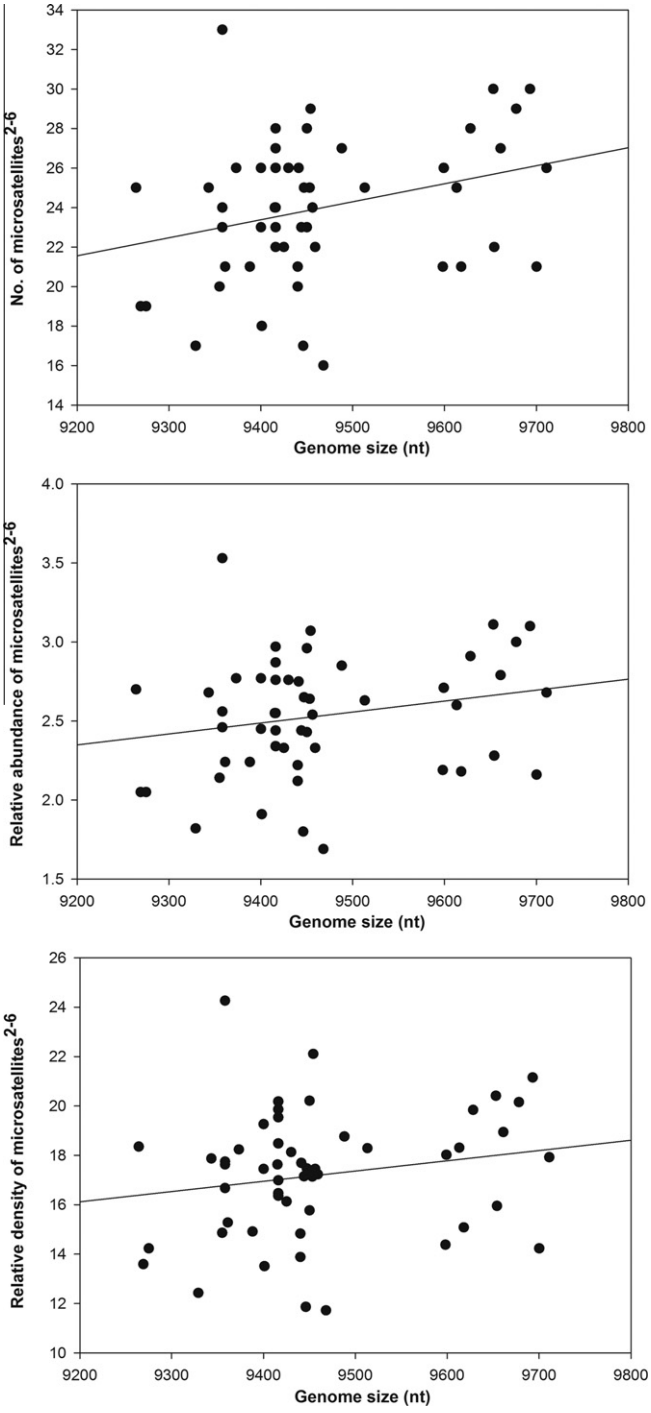

Fig. 3. Relationship between the genome size and the number, relative abundance and relative density of microsatellites$^{2-6}$ in analyzed HCV genomes. Relative abundance is the total microsatellites$^{2-6}$ per kb of sequence analyzed. Relative density is defined as the total length (nt) contributed by each microsatellite$^{2-6}$ per kb of sequence analyzed.

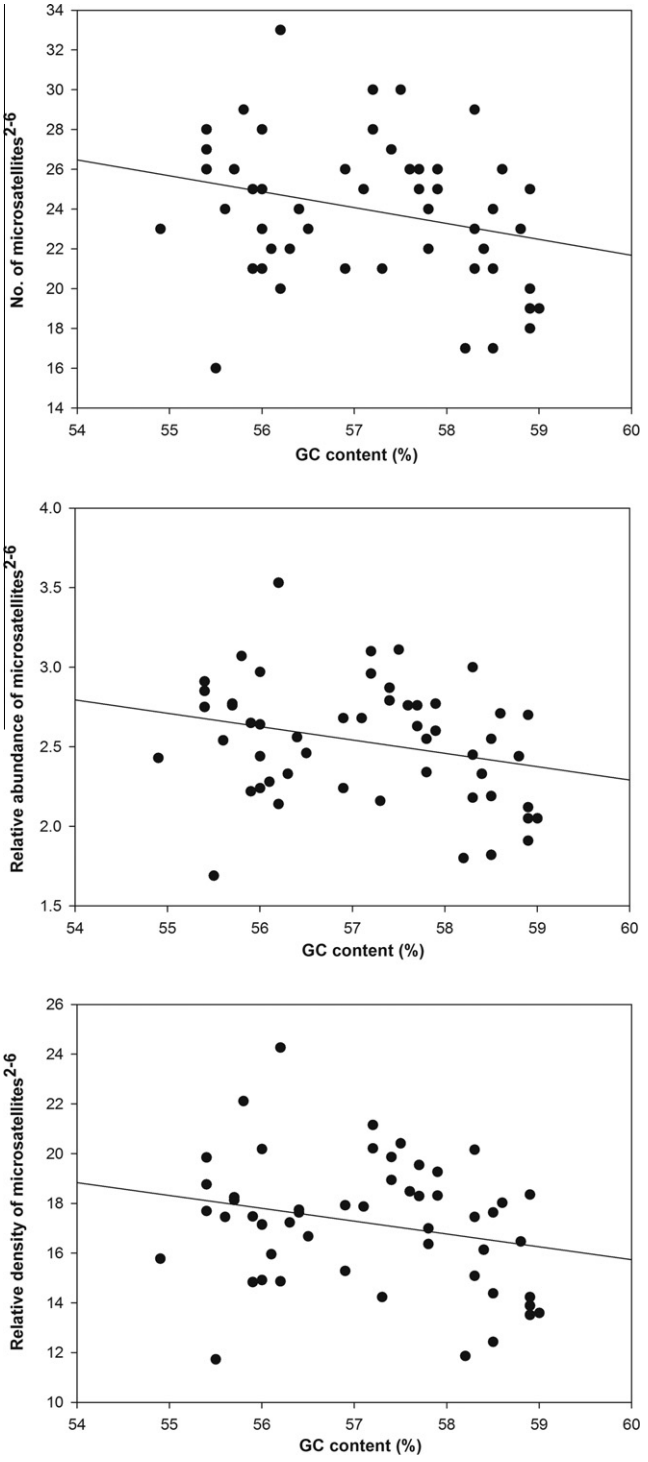

Fig. 4. Relationship between the GC content and the number, relative abundance and relative density of microsatellites$^{2-6}$ in analyzed HCV genomes. See Fig. 3 legend.

L02836 and D28917. However, the CG/GC repeats are very low in genomes of human, *Drosophila*, *Caenorhabditis elegans*, yeast, fungi, and HIV-1 (Chen et al., 2009; Karaoglu et al., 2005; Katti et al., 2001). Trinucleotide repeats were the second abundant repeats in surveyed HCV genomes. Consistent with dinucleotide repeats, our analysis also show trinucleotide repeat types were variable within and between different HCV genotypes. The $(ATG)_3$ repeat was the most prevalent trinucleotide repeat in HCV genomes except for AB030907, AF238486 and EF424628 (Supplementary Table 6). Tetranucleotide repeats $(TTCT)_3$ and $(AGGG)_3$ consisted of a string of the same character and another different character. Hence, only a single mutation would be required for a transformation of one mononucleotide repeat motif into the tetranucleoitde repeat motif. It is naturally assumed that the two tetranucleotide repeats $(TTCT)_3$ and $(AGGG)_3$ can originate from the mutations of $(T)_n$ and $(G)_n$, respectively. Seven sequences (FN435993, U89019,

AB030907, AF238481, AF238486, D28917 and DQ480520) were shown to contain tetranucleotide repeats (Supplementary Table 7), and only two sequences (AF238483 and AF238485) had pentanucleotide repeats (Supplementary Table 8). Among all surveyed genomes, no hexanucleotide repeats were found. There is evidence that different taxa show different preference for microsatellite types (Karaoglu et al., 2005; Toth et al., 2000). For example, the $(GT)_n$ is the most abundant repeat motif in animals and invertebrates (Stallings et al., 1991). However, rare work is done to prove whether this preference is present or not in different complete genomes from the same species. To date, only a related study is finished, showing that the most common microsatellite motifs may be different in diverse completed HIV-1 genomes (Chen et al., 2009). Clearly, our study also showed the most common microsatellite types changed between analyzed HCV genomes.

3.3. Identification of microsatellite polymorphisms

To compute triplet repeat length polymorphism in the human transcriptome, Molla et al. first identified triplet repeat blocks composed of triplet repeats and their corresponding flanking sequences in human reference genome (assembly: NCBI36), and then detected these repeat blocks in genome of James Watson (Molla et al., 2009). Similar method was used to detect tandem repeat variation in protein-coding regions of human genes (O'Dushlaine et al., 2005). Recently, a polymorphic microsatellite database has been constructed for prokaryotes by comparing the genome sequences of different strains from the same species (Kumar et al., 2011). Clearly, these studies employed genome alignment methods to identify microsatellite polymorphisms. According to the above methods, we used the sequence S1 (AB520610) as the template for microsatellite detection in the present study. Other 53 sequences (S2-S54) were used to construct the database '53sequences.fna'. In addition, 10 bp of flanking sequence on both sides of each microsatellite was also extracted. Microsatellite lacking 10 bp of flanking sequence on its both sides was omitted. Microsatellites detected in sequence S1 were defined as the reference microsatellites. We searched the database for the similar microsatellite regions with the reference microsatellites using their 10 bp of flanking sequences. We defined a reference microsatellite as having polymorphism if the length of the reference repeat block is non-identical with that of the other sequences in the database, and this length difference must be a multiple of the repeat unit. The schematic representation of our method was shown in Supplementary Fig. 1. The prerequisite for the detection of microsatellite variations by genome alignment is the conservation of the flanking sequences. However, comparison of flanking sequences from each microsatellite site in HCV genomes clearly indicated one or more insertions, deletions and substitutions existed between these sequences. Low conservation of the flanking sequences from microsatellite sites in HCV genomes hindered the use of the above method in the present study. Manual methods might help to identify microsatellite polymorphism in HCV genomes. Mononucleotide repeats were used as the representative for manually detecting polymorphisms. The distinction (base composition, coding density, or other genome features) between the sequences from different genotypes should be more significant than that between the sequences from the same genotypes. To avoid errors resulting from this distinction, we selected the sequences from genotype 1 for the purpose of this work. However, the starting position of Coding Sequence (CDS) of U89019 is significantly different from that of other sequences (Supplementary Table 1). This may be a result of incorrect annotation in GenBank. Thus, this sequence is not considered for this analysis and a total of 15 sequences from genotype 1 are used for this purpose. Each mononucleotide repeat in a genome is manually evaluated with regards to its flanking genome sequences and its position relative to starting position of CDS, whether its counterpart is present or not in the other complete genome sequences and whether there is any variation in the microsatellite between these genome sequences. Then, we observed some microsatellite polymorphisms in HCV genomes (Supplementary Table 9). Clearly, manually estimation showed microsatellite polymorphisms were present in HCV genomes. However, it must be noted that manual methods were not rigorous and rough, and could not completely correctly identify all microsatellite polymorphisms.

In conclusion, the study of microsatellites in 54 complete HCV genomes is the first step towards a better understanding the nature, evolution and function of viral microsatellites. Similar study in all sequenced complete viral genomes is in process to investigate whether microsatellites make up an important proportion in all viral genomes and whether they have any functional significance and evolutionary dynamics. Our study showed microsatellites were an important component of complete HCV genomes and some microsatellites were significantly overrepresented, suggesting they may play important roles in genome organization. Genome features are weakly correlated with the number, relative abundance and relative density of microsatellites in these surveyed genomes. Consistent with HIV-1, we observed a similar distribution pattern of microsatellites[2–6] based on relative abundance and relative density. However, it must be noted that the repeat motifs varied between HCV genomes. In the present study, all identified microsatellites are very short. This may be because (i) longer microsatellites may be more unstable than shorter microsatellites due to the fact that longer microsatellites have more opportunities to undergo slipped-strand misparing (Wierdl et al., 1997), and (ii) longer microsatellites exhibit the downward mutation bias and short existence time (Harr and Schlotterer, 2000; Karaoglu et al., 2005). Because of high mutability, microsatellites may be involved in generating genomic diversity and take part in genome evolving in eukaryotes as well as in prokaryotes (Ellegren, 2004; Li et al., 2004). Our analysis showed microsatellites existed extensively in complete HCV genomes, and microsatellite distribution varied in these sequences, suggesting microsatellites have a potential for generating HCV genomic diversity and phenotypic changes (Li et al., 2004). Other mechanisms including neutral and adaptive evolution are also demonstrated to play important roles in the diversification of HCV, which can provide genetic variants for fast adaption to new selection pressures (Simmonds, 2004). Microsatellite variation may be a useful resource of HCV genome evolution, possibly helping HCV genome quickly adapt to environmental changes and counteract the human immune response (Li et al., 2004; Mrazek et al., 2007).

Acknowledgements

The authors thank the editors and reviewers for very helpful comments and suggestions. The study was financially supported by Production, Education and Research guiding project, Guangdong Province (2010B090400439), Great program for GMO, Ministry of Agriculture of the people Republic of China (2009ZX08015-003A), the National Natural Science Foundation of China (Nos. 50608029, 50978088, 50808073, 51039001), Hunan Provincial Innovation Foundation for Postgraduate, the National Basic Research Program (973 Program) (No. 2005CB724203), Program for Changjiang Scholars and Innovative Research Team in University (IRT0719), the Hunan Provincial Natural Science Foundation of China (10JJ7005), the Hunan Key Scientific Research Project (2009FJ1010), and Hunan Provincial Innovation Foundation For Postgraduate (CX2010B157).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.meegid.2011.06.012.

References

Ackermann, M., Chao, L., 2006. DNA sequences shaped by selection for stability. PLoS Genet. 2, e22.

Chen, M., Tan, Z., Jiang, J., Li, M., Chen, H., Shen, G., Yu, R., 2009. Similar distribution of simple sequence repeats in diverse completed Human Immunodeficiency Virus Type 1 genomes. FEBS Lett. 583, 2959–2963.

Chen, M., Tan, Z., Zeng, G., 2011a. MfSAT: detect simple sequence repeats in viral genomes. Bioinformation 6, 171–172.

Chen, M., Tan, Z., Zeng, G., Peng, J., 2010. Comprehensive analysis of simple sequence repeats in pre-miRNAs. Mol. Biol. Evol. 27, 2227–2232.

Chen, M., Zeng, G., Tan, Z., Jiang, M., Zhang, J., Zhang, C., Lu, L., Lin, Y., Peng, J., 2011b. Compound microsatellites in complete *Escherichia coli* genomes. FEBS Lett.. doi:10.1016/j.febslet.2011.03.005.

Coenye, T., Vandamme, P., 2005. Characterization of mononucleotide repeats in sequenced prokaryotic genomes. DNA Res. 12, 221–233.

Davis, C.L., Field, D., Metzgar, D., Saiz, R., Morin, P.A., Smith, I.L., Spector, S.A., Wills, C., 1999. Numerous length polymorphisms at short tandem repeats in human cytomegalovirus. J. Virol. 73, 6265–6270.

de wachter, R., 1981. The number of repeats expected in random nucleic acid sequences and found in genes. J. Theor. Biol. 91, 71–98.

Deback, C., Boutolleau, D., Depienne, C., Luyt, C.E., Bonnafous, P., Gautheret-Dejean, A., Garrigue, I., Agut, H., 2009. Utilization of microsatellite polymorphism for differentiating herpes simplex virus type 1 strains. J. Clin. Microbiol. 47, 533–540.

Dieringer, D., Schlotterer, C., 2003. Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res. 13, 2242–2251.

Ellegren, H., 2004. Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. 5, 435–445.

Gur-Arie, R., Cohen, C.J., Eitan, Y., Shelef, L., Hallerman, E.M., Kashi, Y., 2000. Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism. Genome Res. 10, 62–71.

Hancock, J.M., 2002. Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. Genetica 115, 93–103.

Harr, B., Schlotterer, C., 2000. Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. Genetics 155, 1213–1220.

Houng, H.S., Lott, L., Gong, H., Kuschner, R.A., Lynch, J.A., Metzgar, D., 2009. Adenovirus microsatellite reveals dynamics of transmission during a recent epidemic of human adenovirus serotype 14 infection. J. Clin. Microbiol. 47, 2243–2248.

Karaoglu, H., Lee, C.M., Meyer, W., 2005. Survey of simple sequence repeats in completed fungal genomes. Mol. Biol. Evol. 22, 639–649.

Katti, M.V., Ranjekar, P.K., Gupta, V.S., 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol. Biol. Evol. 18, 1161–1167.

Kelkar, Y.D., Tyekucheva, S., Chiaromonte, F., Makova, K.D., 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res. 18, 30–38.

Kuiken, C., Yusim, K., Boykin, L., Richardson, R., 2005. The Los Alamos hepatitis C sequence database. Bioinformatics 21, 379–384.

Kumar, P., Chaitanya, P.S., Nagarajaram, H.A., 2011. PSSRdb: a relational database of polymorphic simple sequence repeats extracted from prokaryotic genomes. Nucleic Acids Res. 39, D601–D605.

Levinson, G., Gutman, G.A., 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. 4, 203–221.

Li, Y.C., Korol, A.B., Fahima, T., Nevo, E., 2004. Microsatellites within genes: structure, function, and evolution. Mol. Biol. Evol. 21, 991–1007.

Molla, M., Delcher, A., Sunyaev, S., Cantor, C., Kasif, S., 2009. Triplet repeat length bias and variation in the human transcriptome. Proc. Natl. Acad. Sci. USA 106, 17095–17100.

Mrazek, J., 2006. Analysis of distribution indicates diverse functions of simple sequence repeats in Mycoplasma genomes. Mol. Biol. Evol. 23, 1370–1385.

Mrazek, J., Guo, X., Shah, A., 2007. Simple sequence repeats in prokaryotic genomes. Proc. Natl. Acad. Sci. USA 104, 8472–8477.

Mudunuri, S.B., Nagarajaram, H.A., 2007. IMEx: imperfect microsatellite extractor. Bioinformatics 23, 1181–1187.

Mudunuri, S.B., Rao, A.A., Pal lamsetty, S., Mishra, P., Nagarajaram, H.A., 2009. VMD: viral microsatellite database – a comprehensive resource for all viral microsatellites. J. Comput. Sci. Syst. Biol. 2, 283–286.

O'Dushlaine, C.T., Edwards, R.J., Park, S.D., Shields, D.C., 2005. Tandem repeat copy-number variation in protein-coding regions of human genes. Genome Biol. 6, R69.

Pearson, C.E., Nichol Edamura, K., Cleary, J.D., 2005. Repeat instability: mechanisms of dynamic mutations. Nat. Rev. Genet. 6, 729–742.

Power, P.M., Sweetman, W.A., Gallacher, N.J., Woodhall, M.R., Kumar, G.A., Moxon, E.R., Hood, D.W., 2009. Simple sequence repeats in *Haemophilus influenzae*. Infect. Genet. Evol. 9, 216–228.

Rajendrakumar, P., Biswal, A.K., Balachandran, S.M., Srinivasarao, K., Sundaram, R.M., 2007. Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions. Bioinformatics 23, 1–4.

Segarra, A., Pepin, J.F., Arzul, I., Morga, B., Faury, N., Renault, T., 2010. Detection and description of a particular Ostreid herpesvirus 1 genotype associated with massive mortality outbreaks of Pacific oysters, *Crassostrea gigas*, in France in 2008. Virus Res. 153, 92–99.

Simmonds, P., 2004. Genetic diversity and evolution of hepatitis C virus – 15 years on. J. Gen. Virol. 85, 3173–3188.

Stallings, R.L., Ford, A.F., Nelson, D., Torney, D.C., Hildebrand, C.E., Moyzis, R.K., 1991. Evolution and distribution of $(GT)_n$ repetitive sequences in mammalian genomes. Genomics 10, 807–815.

Tautz, D., Trick, M., Dover, G.A., 1986. Cryptic simplicity in DNA is a major source of genetic variation. Nature 322, 652–656.

Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S., McCouch, S., 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 11, 1441–1452.

Toth, G., Gaspari, Z., Jurka, J., 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10, 967–981.

Usdin, K., 2008. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res. 18, 1011–1019.

Wierdl, M., Dominska, M., Petes, T.D., 1997. Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics 146, 769–779.