Infection, Genetics and Evolution 12 (2012) 1452–1457

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

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

Short communication

Differential distribution of compound microsatellites in various Human Immunodeficiency Virus Type 1 complete genomes

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

[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
[d] State Key Laboratory of Medical Genetics, Central South University, Changsha, Hunan 410078, China

ARTICLE INFO

Article history:
Received 23 August 2011
Received in revised form 4 May 2012
Accepted 12 May 2012
Available online 1 June 2012

Keywords:
Microsatellite
Simple sequence repeat
Human Immunodeficiency Virus Type 1
Compound microsatellite

ABSTRACT

Compound microsatellites consist of two or more individual microsatellites, and may originate from dynamic mutations or imperfection of microsatellites. Previous studies have found microsatellites were present in 81 completed Human Immunodeficiency Virus Type 1 (HIV-1) genomes, suggesting compound microsatellites may exist in viral genomes. However, up to now, compound microsatellites have not been analyzed in any viral genomes. We identified and characterized 238 compound microsatellites in 81 completed HIV-1 genomes. About 0–24.24% of all microsatellites could be categorized as compound microsatellites. Compound microsatellite distribution is very different in two aspects between diverse HIV-1 genomes. First, the number and motifs of compound microsatellites are variable between surveyed genomes. Second, the relative abundance and relative density of compound microsatellites exhibit very significant differences between these surveyed genomes, respectively. The relative abundance and relative density of compound microsatellites were weakly correlated with genome size and microsatellite density. We observed a more dynamic picture of compound microsatellites than previously reported in eukaryotes. This might be attributed to the lack of proofreading in HIV-1 genomes, as it has been demonstrated that the loss of polymerase proofreading activity can greatly enhance the mutation rate of microsatellites.

1. Introduction

Microsatellites, also known as simple sequence repeats (SSRs), are DNA/RNA stretches composed of a tandemly repeated unit with a length ⩽6 bp (Chen et al., 2009, 2010). Microsatellites are highly abundant and often exhibit length hypervariability in eukaryotic, prokaryotic and viral genomes (Bacolla et al., 2008; Chen et al., 2011a; Kim et al., 2008; Mrazek et al., 2007; Parisi et al., 2003; Power et al., 2009; Rajendrakumar et al., 2007). Owing to the polymorphisms observed in microsatellites, microsatellites have been extensively used as genetic markers today (Karaoglu et al., 2005). Dynamic mutations in microsatellites have been implicated to be associated with various diseases (Subramanian et al., 2003; Usdin, 2008). Many of these diseases are associated with the polymorphism of CAG repeat, including Huntington's disease, dentatorubro-pallidoluysian atrophy, hereditary ataxias, and spinobulbar muscular atrophy (Li et al., 2004; Usdin, 2008). Over the past years, microsatellites have been found and characterized within protein-coding and non-coding regions (Li et al., 2004; Madsen et al., 2008; Subramanian et al., 2003). These studies give useful information to explore possible microsatellite evolution. However, until now, we still not fully understood microsatellite evolution (Kofler et al., 2008). Debates over whether imperfection in microsatellites is responsible for microsatellite evolution are never-ending. It has been demonstrated that genomes harbor large numbers of imperfect microsatellites (Delgrange and Rivals, 2004; Mudunuri and Nagarajaram, 2007). Some authors have proposed imperfection of microsatellites has influence on the life cycle-'birth' and 'death' of microsatellites (Kofler et al., 2008). A tract composed of two or more microsatellites is called compound microsatellite which is expected to have higher polymorphism than single microsatellite (Chen et al., 2011b). Compound microsatellites are found in human genome (Bull et al., 1999; Weber, 1990). Compound microsatellites are investigated in eight eukaryotic genomes, and are proposed to be originated from imperfection in microsatellites (Kofler et al., 2008). Differently, Jakupciak and

Abbreviations: HIV-1, Human Immunodeficiency Virus Type 1; *E. coli*, *Escherichia coli*; SSRs, simple sequence repeats.

* Corresponding authors. Tel.: +86 731 88822754; fax: +86 731 88823701.
E-mail addresses: zhongyang@hnu.edu.cn (Z. Tan), zgming@hnu.edu.cn (G. Zeng).

1567-1348/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.meegid.2012.05.006

Wells thought that recombination between homologous microsatellites generated compound microsatellites (Jakupciak and Wells, 1999). Recently, we identified compound microsatellites in 22 complete *Escherichia coli* (*E. coli*) genomes (Chen et al., 2011b).

Several tools have been developed to detect compound microsatellites. Currently, IMEx (Mudunuri and Nagarajaram, 2007) and SciRoKo (Kofler et al., 2007) are two main computer aided approaches. However, until now, to our knowledge, the information on compound microsatellites is still blank in viral genomes. Thus, in the present study, we provide a survey of compound microsatellites in 81 completed HIV-1 genomes.

2. Materials and methods

2.1. Genome sequences

A total of 81 completed HIV-1 genome sequences were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/). Detailed information of these genomes can be obtained from data published by (Chen et al., 2009).

2.2. Analysis of compound microsatellites

The perfect microsatellites and compound microsatellites search was done with the software IMEx (Mudunuri and Nagarajaram, 2007). Previous analysis has shown that microsatellites are prevalent in 81 completed HIV-1 genome sequences and that there is a similar distribution pattern of microsatellites in these sequences (Chen et al., 2009). However, these conclusions are based on the fact that no mononucleotide repeats are considered. To address this problem, we assessed mononucleotide repeat distribution. The following parameters were used: Type of Repeat: perfect; Repeat Size: all; Minimum Repeat Number: 6, 3, 3, 3, 3, 3 (Rajendrakumar et al., 2007); all compound microsatellites were not standardized because of their low abundance and the fact that HIV-1 genome consists of single strand. Consistent with previous studies in eukaryotes and *E. coli* (Chen et al., 2011b; Kofler et al., 2008), in the present study, we used a dMAX (Max. distance allowed between any two SSRs) of 10 bp.

Statistical analyses were carried out using the software SPSS 11.5. The Pearson correlation coefficient was calculated to evaluate the influence of genome size and microsatellite density on the relative abundance and relative density of compound microsatellites.

3. Results

3.1. The effect of dMAX on occurrence of compound microsatellites

dMAX is the maximum distance between any two adjacent microsatellites to evaluate whether these microsatellites can be classified as a compound microsatellite (Kofler et al., 2008). When the distance separating any two microsatellites is less than or equivalent to dMAX, these microsatellites are accounted as a compound microsatellite. To determine the impact of dMAX, we selected six HIV-1 genomes for this analysis according to their geographical locations (AF042102 from Oceania; AF049337 from Asia; AF049494 from North America; AF061640 from Euroupe; AF063223 from Africa; and AY173956 from South America), and introduced percentage of individual microsatellites being part of a compound microsatellite (cSSRs-%) as the measure (Table 1). Noteworthy, the dMAX value can be only set between 0 and 50 for IMEx (Mudunuri and Nagarajaram, 2007). As a result, we observed that this percentage increased with the dMAX in each of these six genomes overall. This is easily expected, since all microsatellites being separated by less than the selected dMAX would be accounted as compound microsatellites in whole genomes. However, it must be noted that this increase is relatively fast in the cases of dMAX $\leqslant$30, but very slow in the cases of 30 < dMAX $\leqslant$ 50 (Supplementary Fig. S1).

3.2. Number, relative abundance and relative density of compound microsatellites

It must be noted that the compound microsatellite abundance and density were calculated on the basis of dMAX of 10 bp. Using computer software for a genome-wide scan of HIV-1 genomes, 0–8 compound microsatellites were found in each of surveyed sequences (Table 1 and Supplementary Table S1). Noteworthy, only EU000514 contained no compound microsatellites. In all these surveyed sequences, a total of 238 compound microsatellites were observed (Table 1). Most compound microsatellites were found in genic regions; these genes mainly included *gag, tat, rev* and *env* (Supplementary Table S2). These observations suggested some compound microsatellites were not likely to emerge by chance, and had evolutionary and functional significance in HIV-1 genomes. The percentage of an individual microsatellite being part of compound microsatellite (cSSRs-%) was very low in some completed HIV-1 genomes. For example, EF633445 had 50 microsatellites, but only 2 of them took part in the formation of compound microsatellites. AF042104 had the highest proportion of microsatellites being classified as compound microsatellites (24.24%), whereas EU000514 had the lowest (0.00%). It is estimated that about 10% of all *Homo sapiens* microsatellites have a compound motif (Weber, 1990). The relative density of compound microsatellites changes drastically in all selected HIV-1 genome sequences. The relative density of compound microsatellite ranged from 0.00 bp/kb in the sequence EU000514 to 12.14 bp/kb in the sequence AF042104 (Table 1).

3.3. Related parameters influencing compound microsatellite distribution

Differences in relative abundance and relative density of compound microsatellites in surveyed sequences can result from the parameters 'microsatellite density' (Kofler et al., 2008) and 'genome size' (Chen et al., 2011b). Noteworthy, the 'microsatellite density' mentioned by Kofler et al. (2008) indeed refers to the 'relative abundance of microsatellites' in this study. We tested the possible influence of these parameters on the relative abundance and relative density of compound microsatellite by use of SPSS 11.5. Microsatellite density has been observed to have a weakly positive influence on the relative abundance of compound microsatellites in *E. coli* genomes (Chen et al., 2011b). We observed the parameter 'microsatellite density' had a weakly positive influence on the relative abundance ($p < 0.01$, $r = 0.468$) and relative density ($p < 0.01$, $r = 0.457$) of compound microsatellites. Thus, our result was consistent with the observation in *E. coli* genomes (Chen et al., 2011b). It is surprising that genome size was negatively weakly correlated with the relative abundance ($p > 0.05$, $r = -0.111$) and relative density ($p > 0.05$, $r = -0.077$). For example, AF049337 had higher genome size than AF042102, but had less compound microsatellites and lower compound microsatellite density. It is generally assumed that the number and density of microsatellites/compound microsatellites increase with the increase of genome size. However, Morgante et al. (2002) found that genome size was inversely correlated with microsatellite abundance in plants. Thus, it is not necessary that genome size is positively correlated with the relative abundance and relative density of microsatellites or compound microsatellites in organisms.

Table 1
Overview of compound microsatellites in completed HIV-1 genome sequences.

No.	Acc. No.	Subtype (s)	Size (bp)	%GC	N[a]	RA[b]	RD[c]	P (%)[d]	c.c.[e]		
									2	3	4
S1	AF042102	B	8540	41.55	4	0.47	8.31	18.75	3	1	0
S2	AF042104	B	8733	42.46	8	0.92	12.14	24.24	8	0	0
S3	AF042105	B	8669	41.34	4	0.46	7.04	14.29	4	0	0
S4	AF042106	B	9096	42.19	4	0.44	7.37	15.25	3	1	0
S5	AF049337	04_cpx	9050	42.15	3	0.33	3.98	10.71	3	0	0
S6	AF049494	B	9193	41.88	2	0.22	2.94	7.14	2	0	0
S7	AF049495	B	9196	41.78	2	0.22	2.94	6.67	2	0	0
S8	AF061640	G	9048	41.95	3	0.33	4.31	11.32	3	0	0
S9	AF061641	G	9047	41.99	3	0.33	4.31	11.11	3	0	0
S10	AF061642	G	9074	41.94	3	0.33	4.08	11.54	3	0	0
S11	AF063223	02_AG	9002	42.09	1	0.11	1.44	4.08	1	0	0
S12	AF063224	02_AG	8961	41.86	2	0.22	2.68	7.69	2	0	0
S13	AF070521	B	9699	42.46	6	0.62	9.69	20.97	5	1	0
S14	AF075719	N/A	9493	41.86	4	0.42	6.11	12.90	4	0	0
S15	AF086817	B	9694	42.13	1	0.10	1.24	2.90	1	0	0
S16	AF110980	C	8931	41.54	3	0.34	4.14	11.32	3	0	0
S17	AF133821	D	10035	41.21	3	0.30	6.58	10.14	2	1	0
S18	AF193276	03_AB	8808	41.24	4	0.45	6.58	16.00	4	0	0
S19	AF193277	03_AB	8961	41.41	3	0.34	4.80	12.00	3	0	0
S20	AF197341	01_AE	9597	42.12	6	0.63	8.65	20.00	6	0	0
S21	AF286236	U	9060	41.91	4	0.44	6.07	14.81	4	0	0
S22	AF443110	C	9093	41.41	4	0.44	6.05	13.33	4	0	0
S23	AF492623	11_cpx	8819	41.87	1	0.11	2.49	6.25	0	1	0
S24	AY008714	01_AE	8859	41.17	5	0.56	7.22	19.23	5	0	0
S25	AY008717	08_BC	8784	41.04	2	0.23	3.07	8.16	2	0	0
S26	AY008718	01_AE	8806	40.82	5	0.57	7.27	18.18	5	0	0
S27	AY049711	C	9054	41.42	2	0.22	2.76	8.16	2	0	0
S28	AY093604	09_cpx	8777	41.46	5	0.57	6.95	17.24	5	0	0
S29	AY169814	N/A	9186	41.29	1	0.11	1.63	4.08	1	0	0
S30	AY169815	O	9186	42.17	1	0.11	1.63	5.13	1	0	0
S31	AY173951	B	8996	41.61	2	0.22	3.45	8.16	2	0	0
S32	AY173955	B	9027	41.29	1	0.11	1.66	4.17	1	0	0
S33	AY173956	B	8940	41.56	4	0.45	6.82	16.33	4	0	0
S34	AY180905	B	9010	41.86	2	0.22	3.11	8.70	2	0	0
S35	AY314060	B	8964	41.27	6	0.67	9.15	19.35	6	0	0
S36	AY314062	B	8997	41.44	5	0.56	7.78	15.63	5	0	0
S37	AY500393	A1	9159	42.14	3	0.33	4.37	10.34	3	0	0
S38	AY536236	BF1	8937	41.25	3	0.34	5.48	14.29	2	1	0
S39	AY536238	BF1	8760	41.28	1	0.11	1.37	4.35	1	0	0
S40	AY586549	G	9185	41.55	4	0.44	5.33	14.55	4	0	0
S41	AY682547	B	9089	41.26	4	0.44	6.60	14.55	4	0	0
S42	AY771589	BF1	9058	42.56	2	0.22	3.42	8.33	2	0	0
S43	AY818642	N/A	8820	40.82	3	0.34	4.88	11.32	3	0	0
S44	AY968312	BC	8834	41.62	3	0.34	4.75	12.00	3	0	0
S45	DQ011178	C	9119	41.72	3	0.33	4.06	11.54	3	0	0
S46	DQ011180	C	9076	41.62	2	0.22	2.53	7.27	2	0	0
S47	DQ020274	20_BG	8944	41.86	1	0.11	1.34	4.17	1	0	0
S48	DQ295193	B	9468	42.17	3	0.32	4.44	11.29	2	1	0
S49	DQ295195	B	9402	42.18	4	0.43	6.59	11.76	4	0	0
S50	DQ358808	B	8956	41.94	3	0.33	4.69	12.50	3	0	0
S51	DQ358809	B	9419	42.28	3	0.32	4.57	9.68	3	0	0
S52	DQ396382	C	9068	41.7	1	0.11	1.43	3.57	1	0	0
S53	DQ396395	C	9047	41.68	3	0.33	4.09	10.53	3	0	0
S54	DQ853444	B	8645	41.03	4	0.46	6.71	15.38	4	0	0
S55	DQ853463	B	9210	41.94	4	0.43	8.58	16.39	3	0	1
S56	DQ853465	N/A	8798	41.49	4	0.45	6.59	14.29	4	0	0
S57	DQ912823	A1D	8900	41.09	1	0.11	1.69	4.26	1	0	0
S58	EF057102	B	9674	42.62	3	0.31	4.13	10.34	3	0	0
S59	EF420987	N/A	9703	42.14	2	0.21	2.58	6.67	2	0	0
S60	EF469243	C	9830	42.68	2	0.20	3.36	9.26	1	1	0
S61	EF633445	D	8916	41.44	1	0.11	1.35	4.00	1	0	0
S62	EU000514	BC	9043	41.21	0	0.00	0.00	0.00	0	0	0
S63	EU031914	01B	9563	42.38	2	0.21	2.51	6.67	2	0	0
S64	EU031915	01B	8942	41.32	4	0.45	6.15	14.04	4	0	0
S65	EU110085	A1	9009	41.86	1	0.11	1.33	3.33	1	0	0
S66	EU110096	A1CG	8963	41.62	5	0.56	7.25	20.00	5	0	0
S67	EU220698	A1C	9457	41.52	2	0.21	2.64	7.41	2	0	0
S68	EU448295	45_cpx	9680	42.16	5	0.52	7.13	15.63	5	0	0
S69	EU448296	BCU	9684	41.92	2	0.21	2.89	6.45	2	0	0
S70	EU735539	40_BF	9105	41.94	2	0.22	3.40	7.55	2	0	0
S71	EU735540	40_BF	9079	41.9	1	0.11	1.32	3.39	1	0	0
S72	EU861977	A1	9781	42.32	3	0.31	3.78	9.68	3	0	0
S73	HIV1U23487	N/A	9655	42.29	3	0.31	4.04	9.52	3	0	0

Table 1 (*continued*)

No.	Acc. No.	Subtype (s)	Size (bp)	%GC	N[a]	RA[b]	RD[c]	P (%)[d]	c.c.[e]		
									2	3	4
S74	HIVU51190	N/A	8999	41.73	3	0.33	4.11	11.11	3	0	0
S75	HIVU52953	N/A	8959	41.41	2	0.22	2.79	7.55	2	0	0
S76	HIVU86780	N/A	8990	42.19	4	0.44	6.12	16.00	4	0	0
S77	U88822	D	8975	41.2	3	0.33	5.01	12.50	3	0	0
S78	U88823	A1C	8992	41.38	5	0.56	6.78	18.52	5	0	0
S79	U88824	D	8952	41.39	1	0.11	2.01	3.77	1	0	0
S80	U88825	A1G	8966	41.79	3	0.33	4.02	11.11	3	0	0
S81	U88826	G	8987	41.7	1	0.11	1.45	3.70	1	0	0

The information on subtypes of HIV-1 is obtained from HIV Database (http://www.hiv.lanl.gov/content/index).
N/A, not available.
[a] Number of compound microsatellites.
[b] Relative abundance = number of compound microsatellites/genome size in kilo base (kb).
[c] Relative density is defined as the total length (bp) contributed by each compound microsatellite per kb of sequence analyzed.
[d] Percentage of individual microsatellites being part of a compound microsatellite (cSSRs-%).
[e] Compound microsatellite complexity (number of individual microsatellites in a compound microsatellite).

3.4. Motifs and complexity of compound microsatellites

Different species including eukaryotes, prokaryotes and HIV-1 showed different preferences for microsatellite types and motifs (Chen et al., 2009; Gur-Arie et al., 2000; Karaoglu et al., 2005). This may lead to a bias towards one or several motif abundance in the composition of compound microsatellites. Two or more individual microsatellites gather together, leading to the formation of a compound microsatellite. The numbers, repeat types (mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats) and repeat motifs of individual microsatellites among compound microsatellites might vary, giving rise to compound microsatellite complexity. Thus, the major difficulties in characterizing the differential distribution of compound microsatellites are the analyses of motifs and complexity of compound microsatellites compared to individual microsatellite. In the present study, we examined which microsatellites are the most frequent found in close proximity in variously completed HIV-1 genomes. In an attempt to understand this, an additional concept 'SSR-couples' was introduced. Kofler et al. (2008) defined SSR-couples having the form $[m1]_n$-x_n-$[m2]_n$ as SSR-couples of motif m1-m2. For instance, the SSR-couple $[A]_6$-x_8-$[AGC]_3$ has the motif A-AGC. Consistent with microsatellites, the average GC content of compound microsatellite motifs is generally lower than AT content in the analyzed HIV-1 genomes (Supplementary Table S1). Rare compound microsatellites only contained A/T. Among these surveyed sequences, the most frequent SSR-couple motif is GA-T (36 sequences contained this motif), followed by GA-A motif (33 sequences had this motif). In contrast, about half of SSR-couple motifs are unique, i.e. each of these motifs occurred once only in a surveyed sequence (Table 2).

The compound microsatellite distribution appears unrelated to geographical locations and subtypes of HIV-1. For example, the genomes EU735539 and EU735540 have the same geographical location (South America) and subtype (40_BF), but their motifs of compound microsatellites are completely different (Supplementary Table S2). It has been speculated that compound microsatellites are from imperfection of microsatellites (Kofler et al., 2008). Differential distribution of compound microsatellites suggests the dynamic picture of microsatellite evolution is different between various HIV-1 genomes in two aspects. First, the preexisting microsatellite whose imperfection forms a compound microsatellite may be non-identical in various HIV-1 genomes. Second, the mutations occurring at microsatellites for the generation of compound microsatellites are also different. For example, AC-AT compound microsatellite may be derived from base substitution in a microsatellite, whether AGA-GA compound microsatellite is more likely produced from insertion or deletion of base "A" in a microsatellite.

Compound microsatellite complexity is revealed relying on the number of individual microsatellites. Compound microsatellites composed of two individual microsatellites are called 'di-SSR' compound microsatellites, whereas those consisting of 3 individual microsatellites are named 'tri-SSR' compound microsatellites (Kofler et al., 2008). Among all surveyed sequences, 'di-SSR' compound microsatellites are the most abundant compound microsatellites. Only with a few exceptions the genomes contained more than two individual microsatellites (Table 1). Interestingly, in the present study, we found the number of compound microsatellites decreased gradually with an increasing complexity. This observation is consistent with that of eukaryotes or *E. coli* (Chen et al., 2011b; Kofler et al., 2008). We found the largest compound microsatellites in DQ853463, having 4 individual microsatellites (Table 1).

4. Discussion

It has been clearly demonstrated that eukaryotes and prokaryotes have a high proportion of microsatellites categorized as compound microsatellites (Chen et al., 2011b; Kofler et al., 2008). These compound microsatellites may serve a functional role in the gene regulation and relate to protein function in some species (Kashi and King, 2006). However, the detailed survey of compound microsatellites in viral genomes remains blank. The most important assignment of this study is therefore that it seeks to reveal the frequency, distribution and potential roles of compound microsatellites in viral genomes. Our work might be useful in three aspects in scientific or engineering fields. First, our work provided a general approach for survey of compound microsatellites in any viral genomes. Second, our study laid a good foundation for further exploring the origin of compound microsatellites. Third, compound microsatellites might be as an effective tool for evolutionary analysis and identification of viral isolate.

We presented the first survey of compound microsatellites in variously completed genomes of HIV-1 which is an excellent system to study evolution and roles of compound microsatellites in viruses. These sequences were sampled from 34 different countries or districts over six continents with genome sizes ranging from 8 540 to 10 035 bp. Although mononucleotide repeats were considered and distribution of mononucleotide repeats was different (Supplementary Fig. S2), relative abundance and relative density of microsatellites are still very similar between surveyed genome sequences (Supplementary Table S1). This observation suggested the similar trend for microsatellite distribution on the basis of relative abundance and relative density between variously completed

Table 2
Characterization of SSR-couples in completed HIV-1 genome sequences.

Motif	Occurrence[a]
A-AG	7
AAG-AGG	1
A-AGC	11
A-AGC-AGC	2
AAG-GA-A	1
AC-A	2
A-CA	2
A-CAG	1
A-CAG-A	1
AC-AT	7
A-G	1
AG-A	23
AGA-AAT	3
AGA-CAG	1
AGA-GA	6
AG-AT-CA	1
AG-CA	5
A-G-G	1
AG-GA	2
AG-T	6
AGT-A	2
AG-TA	1
AGT-AT	1
A-T	1
AT-CA	1
ATG-AG	1
CA-A	1
CAA-AAG	1
CT-A	1
CTA-ACT-TAA-A	1
CT-GT	1
CTT-TTG	1
G-A	2
GA-A	33
GAA-AGA	1
GA-AAT	1
GA-AG	1
GAA-GA	16
GA-CA	2
GA-T	36
G-AT	1
G-CA	2
GGA-GA	4
GT-CAA	2
TA-A	1
TA-ATA	7
TAC-TAG	1
TAG-AAT-TG	1
TCA-GA	1
TCC-TCC-CCT	1
TCC-TCT	1
T-CT	1
TG-A	1
TG-AT	2
TG-CA	1
TG-GA	18
TG-TA	3
T-TC	1

[a] Indicates the times of occurrence of compound microsatellites in all 81 completed HIV-1 genomes.

HIV-1 genomes was not changed by the participation of mononucleotide repeats. This can be expected, since differently completed HIV-1 genomes can have the same organizations of genes, similar genome sizes and base composition. A lower complexity of compound microsatellites was observed in HIV-1 genomes and *E. coli* genomes than in eukaryotic genomes can be explained by their smaller genome sizes and higher coding density (Chen et al., 2011b). Consistent with eukaryotes and prokaryotes, the cSSRs-% increased with the dMAX in HIV-1 genomes (Chen et al., 2011b; Kofler et al., 2008). Interestingly, cSSRs-% varied between HIV-1 genome, eukaryotes and *E. coli* genomes: 0–24.24% in HIV-1 genomes; 4–25% in eight eukaryotic genomes (Kofler et al., 2008) and 1.75–2.85% in *E. coli* genomes (Chen et al., 2011b). Microsatellite density and genome size have a weak influence on relative abundance and relative density of compound microsatellites in HIV-1 genomes. This observation is similar to that of *E. coli* but different from that of eukaryote (Chen et al., 2011b; Kofler et al., 2008).

Microsatellites have 10^3–10^6-fold higher mutation rates than does the eukaryotic DNA sequences (10^{-3}–10^{-6} vs. 10^{-9}) (Xu et al., 2000), and hence are believed to play an important role in genome evolution by providing quantitative genetic variation (Tautz et al., 1986). Compound microsatellites composed of two or more microsatellites are expected to have higher mutation rates than single microsatellite. Thus, compound microsatellites may be a more excellent resource for genome evolution. In this study, a certain number of compound microsatellites are found in RNA viruses (HIV-1), suggesting compound microsatellites may contribute to genetic variability of RNA viruses to some extent. However, it must be noted that genetic RNA recombination is a major factor responsible for the emergence of new viral strains or species (Tan et al., 2005). Because of higher polymorphism, compound microsatellites are capable of faster leading to changes of gene function by expansions or contractions than single microsatellite. In particular, compound microsatellites variations found in genic regions is likely to affect the evolution, transmission and pathogenicity of acquired immunodeficiency syndrome (AIDS). Moreover, compound microsatellites could be used as molecular strain markers (Houng et al., 2009).

A microsatellite was interrupted by a nucleotide base, becoming compound in *Human respiratory adenovirus* (Houng et al., 2009). Previous study found individual microsatellites of compound microsatellites consisted of very similar motifs, and proposed that imperfection in microsatellites were responsible for generation of compound microsatellites in eukaryotes (Kofler et al., 2008). However, compound microsatellite motifs are different overall in each of surveyed HIV-1 genome sequences. We observed less than 50% of all compound microsatellite motifs contained very similar motifs. Instead, many SSR-couples consist of very distinct motifs differing by two or more single mutations in more than 50% of cases. This might be because HIV-1 genomes have an exceptionally high mutation rate due to lack of proofreading function, leading to great enhancement of the rate of microsatellite tract alterations (Levinson and Gutman, 1987; Wells, 1996). Moreover, we observed an interesting phenomenon that most of the microsatellites identified are very short, in contrast to the much longer arrays normally studied in eukaryotes (Supplementary Fig. S3). The same was true for *Hepatitis C virus* (HCV) genomes in which all identified microsatellites were also short in length (Chen et al., 2011a). Such dominance of short microsatellites over long microsatellites in HIV-1 and HCV genomes may be explained based on the fact that longer microsatellites more easily mutated than shorter microsatellites (Chen et al., 2011a; Wierdl et al., 1997).

5. Conclusions

Our study has helped to identify compound microsatellites in 81 completed HIV-1 genomes. To our knowledge, this is a first analysis of the viral genomes for such purposes and represents a general approach for analysis of compound microsatellites in other viral genomes. Taking our computational and statistical data together, we conclude that the distribution of compound microsatellites is differential between variously completed HIV-1 genomes. Information on the abundance, density and motifs of compound

microsatellites may give us some clues to the evolution of microsatellites such as 'birth' and 'death' of microsatellites (Kofler et al., 2008).

Acknowledgements

The authors sincerely thank Editor and anonymous reviewer for suggestions on improving the paper. 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 (2009ZX 08015-003A), 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) and the Hunan Key Scientific Research Project (2009FJ1010).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2012.05.006.

References

Bacolla, A., Larson, J.E., Collins, J.R., Li, J., Milosavljevic, A., Stenson, P.D., Cooper, D.N., Wells, R.D., 2008. Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties. Genome Res. 18, 1545–1553.

Bull, L.N., Pabon-Pena, C.R., Freimer, N.B., 1999. Compound microsatellite repeats: practical and theoretical features. Genome Res. 9, 830–838.

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., Peng, J., 2010. Comprehensive Analysis of Simple Sequence Repeats in Pre-miRNAs. Mol. Biol. Evol. 27, 2227–2232.

Chen, M., Tan, Z., Zeng, G., 2011a. Microsatellite is an important component of complete Hepatitis C virus genomes. Infect. Genet. Evol. 11, 1646–1654.

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. 585, 1072–1076.

Delgrange, O., Rivals, E., 2004. STAR: an algorithm to search for tandem approximate repeats. Bioinformatics 20, 2812–2820.

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.

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.

Jakupciak, J.P., Wells, R.D., 1999. Genetic instabilities in (CTG.CAG) repeats occur by recombination. J. Biol. Chem. 274, 23468–23479.

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

Kashi, Y., King, D.G., 2006. Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22, 253–259.

Kim, T.S., Booth, J.G., Gauch Jr., H.G., Sun, Q., Park, J., Lee, Y.H., Lee, K., 2008. Simple sequence repeats in *Neurospora crassa*: distribution, polymorphism and evolutionary inference. BMC Genomics 9, 31.

Kofler, R., Schlotterer, C., Lelley, T., 2007. SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics 23, 1683–1685.

Kofler, R., Schlotterer, C., Luschutzky, E., Lelley, T., 2008. Survey of microsatellite clustering in eight fully sequenced species sheds light on the origin of compound microsatellites. BMC Genomics 9, 612.

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.

Madsen, B.E., Villesen, P., Wiuf, C., 2008. Short tandem repeats in human exons: a target for disease mutations. BMC Genomics 9, 410.

Morgante, M., Hanafey, M., Powell, W., 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat. Genet. 30, 194–200.

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.

Parisi, V., De Fonzo, V., Aluffi-Pentini, F., 2003. STRING: finding tandem repeats in DNA sequences. Bioinformatics 19, 1733–1738.

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.

Subramanian, S., Madgula, V.M., George, R., Mishra, R.K., Pandit, M.W., Kumar, C.S., Singh, L., 2003. Triplet repeats in human genome: distribution and their association with genes and other genomic regions. Bioinformatics 19, 549–552.

Tan, Z., Gibbs, A.J., Tomitaka, Y., Sanchez, F., Ponz, F., Ohshima, K., 2005. Mutations in Turnip mosaic virus genomes that have adapted to *Raphanus sativus*. J. Gen. Virol. 86, 501–510.

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

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

Weber, J.L., 1990. Informativeness of human (dC-dA)n. (dG-dT)n polymorphisms. Genomics 7, 524–530.

Wells, R.D., 1996. Molecular basis of genetic instability of triplet repeats. J. Biol. Chem. 271, 2875–2878.

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

Xu, X., Peng, M., Fang, Z., 2000. The direction of microsatellite mutations is dependent upon allele length. Nat. Genet. 24, 396–399.