FEBS Letters 583 (2009) 2959–2963

FEBS Letters
journal homepage: www.FEBSLetters.org

Similar distribution of simple sequence repeats in diverse completed *Human Immunodeficiency Virus Type 1* genomes

Ming Chen [a], Zhongyang Tan [a,b,*], Jianhui Jiang [b], Mingfu Li [c], Hongjun Chen [c], Guoli Shen [b], Ruqin Yu [b]

[a] *Institute of Life Sciences and Biotechnology, Hunan University, Changsha 410082, China*
[b] *State Key Laboratory for Chemo/Biosensing and Chemometrics, Hunan University, Changsha 410082, China*
[c] *Chinese Academy of Inspection and Quarantine, Beijing 100029, China*

ARTICLE INFO

Article history:
Received 17 June 2009
Revised 31 July 2009
Accepted 4 August 2009
Available online 11 August 2009

Edited by Takashi Gojobori

Keywords:
Simple sequence repeat
Microsatellite
Human Immunodeficiency Virus Type 1

ABSTRACT

The survey of simple sequence repeats (SSRs) has been extensively made in eukaryotes and prokaryotes. However, its still rare in viruses. Thus, we undertook a survey of SSRs in *Human Immunodeficiency Virus Type 1* (HIV-1) which is an excellent system to study evolution and roles of SSRs in viruses. Distribution of SSRs was examined in 81 completed HIV-1 genome sequences which come from 34 different countries or districts over 6 continents. In these surveyed sequences, although relative abundance and relative density exhibit very high similarity, some of these sequences show different preference for most common SSRs and longest SSRs. Our results suggest proportion of various repeat types might be related to genome stability.

1. Introduction

Simple sequence repeats (SSRs), also known as microsatellites, refer to the sequences that are 1–6 bp unit repeated in tandem in a genome. With the growing numbers of completed genome sequences in eukaryotic and prokaryotic organisms, SSRs have been extensively surveyed at the genome-wide level. Many characteristics of SSRs are found in eukaryotes and prokaryotes: (1) the distribution of SSRs in the genome was not random [1–4]; (2) most, but not all, of SSRs are ubiquitous and abundant in a genome [5–10], and SSR content and genome size have not always been shown positive relation [9]; (3) SSRs exist in 3′-UTR, 5′-UTR, exons, introns, i.e. protein-coding and non-coding regions [11–13]; (4) SSRs are highly variable and unusually polymorphic [1,14] and hence are extensively used as genetic markers [14–16]; (5) SSRs types are likely different in different taxa or different regions of the same taxon [6,9]; and (6) SSRs are thought to may influence transcriptional activity [17] and to play a functional role in the evolution of gene regulation [18].

Correspondingly, some hypotheses are developed to explain aforementioned phenomena. Its believed that formation of SSRs might be attributable to de novo genesis or adoptive genesis [1], and the instability and polymorphisms of SSRs are primarily due to slipped-strand mispairing errors during DNA replication [19]. Toth et al. thought the fixation of de novo generated SSRs was determined by the interplay of several factors including repeat type, the genomic position of the SSR, and the genetic–biochemical background of the cell [6].

Up to now, to our knowledge, SSRs still have not been studied in detail in viruses. Some SSRs have been known to be distributed in the HIV genomes. In particular, the ends of each strand of HIV RNA contain an RNA sequence called the long terminal repeat (LTR) which contains important regulatory regions, especially those for transcription initiation and polyadenylation. SLIP, a TTTTTT slippery site, is involved in the frameshift in the Gag-Pol reading frame required to make functional Pol [20]. Thus, HIV, a lentivirus (a member of the retrovirus family) that can lead to acquired immunodeficiency syndrome (AIDS), is an excellent system to study evolution and biological function of viral SSRs. Two strains of HIV (HIV-1 and HIV-2) are known to exist, of which HIV-1 is more virulent, relatively easily transmitted, and is globally pandemic [21].

In this study, we investigated the distribution and size variability of SSRs in different sequences of the same species (HIV-1). We intended to address the questions of whether the relative abundance and relative density are similar or not in those sequences from different countries or districts and how different are the frequencies of occurrence of various repeat types in these surveyed

Abbreviations: SSrs, simple sequence repeats; HIV-1, *Human Immunodeficiency Virus Type 1*

* Corresponding author. Address: Institute of Life Sciences and Biotechnology, Hunan University, Changsha 410082, China.
E-mail address: hudatzycm@163.com (Z. Tan).

0014-5793/$36.00
doi:10.1016/j.febslet.2009.08.004

completed HIV-1 genomes. We also expect our data to be able to provide insights into the distributional rules of SSRs in viruses.

2. Materials and methods

A total of 81 completed genomic sequences of HIV-1, from 34 different countries or districts over 6 continents, were collected from GenBank (http://www.ncbi.nlm.nih.gov). The list of genomic sequences, their attributed regions and sizes are shown in Table 1. All of the genomic sequences were scanned for various SSRs using SSRIT [16] which was employed to search each of five SSRs motifs (di-, tri-, tetra-, penta-, hexanucleotide repeats), with length of 6 bp or more. No mononucleotide repeats were surveyed, and no difference between the occurrence of repeats in coding and non-coding regions was made. To facilitate the comparison among genomic sequences of different sizes, we chose to use the "relative abundance", which is calculated by dividing the number of SSRs by kilo base pair (kb) of sequences, and decided to utilize the "relative density", which is a value by calculating the number of base pairs of sequence contributed by each SSR in the total sequences analyzed (kb).

3. Results and discussion

We analyzed perfect SSRs over 6 bp long, from 81 completely sequenced HIV-1 genomes, ranging from 8540 bp (AF042102) to 10 035 bp (AF133821), and these sequences were sampled from 34 different countries or districts over 6 continents (Table 1). Using computer software for a genome-wide scan of sequences of HIV-1, 22–48 SSRs were found in each of genomic sequences (Supplementary Table 1). Relative abundance of di-, tri-, tetra-, penta-, and hexanucleotide repeats and relative density of SSRs across the selected HIV-1genomes are presented in Tables 2 and 3, respectively. Results here indicate that the relative abundance and relative density are very similar in each of these surveyed sequences.

To date, some authors have reported the correlation of the SSR content with the genome size in fungal genomes [9], and other genomes [22,23]. Herein, similar work is finished, revealing that the total SSR contents in diverse HIV-1 genomes are not directly proportional to the genome size. For example, in a comparison of same-sized genomes, AF063224 have five excess SSRs as compared to that of AF193277 although they have the same genomic size (8961 bp) (Supplementary Table 1).

3.1. Total and relative abundance of SSRs

Compared with prokaryotic and eukaryotic genomes, HIV-1 genomes are very small. As expected, fewer SSRs were identified in each of these surveyed genomic sequences. Among the sequences examined, the highest number of SSR was 48 found in the sequence of EU448295, and the least one was 22 in the sequence of AY169815. The relative abundance of the various repeat

Table 1
List of analyzed completed HIV-1 genomes, their attributed regions and genome size.

No.	Acc. no.	Size (bp)	Countries/districts	No.	Acc. no.	Size (bp)	Countries/districts
S1	AF110980	8931	Botswana	S42	AF061640	9048	Finland
S2	AF443110	9093	Botswana	S43	AF061641	9047	Finland
S3	AY169815	9186	Cameroon	S44	EU448296	9684	France
S4	AF492623	8819	Cameroon	S45	EU448295	9680	France
S5	AF197341	9597	Central African Republic	S46	EU861977	9781	Italy
S6	AF063223	9002	Djibouti	S47	HIV1U23487	9655	Manchester
S7	AF063224	8961	Djibouti	S48	AY682547	9089	Russia
S8	EU110096	8963	Kenya	S49	AF193277	8961	Russia
S9	EU110085	9009	Kenya	S50	AF193276	8808	Russia
S10	AF133821	10 035	Kenya	S51	AY500393	9159	Russia
S11	U88826	8987	Nigeria	S52	AF061642	9074	Sweden
S12	U88825	8966	Nigeria	S53	EU220698	9457	Canada
S13	AF286236	9060	Republic of the Congo	S54	DQ020274	8944	Cuba
S14	U88823	8992	Rwanda	S55	AY586549	9185	Cuba
S15	AY093604	8777	Senegal	S56	AF075719	9493	USA
S16	DQ396395	9047	South Africa	S57	DQ853444	8645	USA
S17	DQ011178	9119	South Africa	S58	AY314060	8964	USA
S18	DQ011180	9076	South Africa	S59	AF049495	9196	USA
S19	EF633445	8916	South Africa	S60	AY314062	8997	USA
S20	DQ396382	9068	South Africa	S61	DQ853463	9210	USA
S21	HIVU51190	8999	Uganda	S62	AF070521	9699	USA
S22	U88824	8952	Uganda	S63	AY173955	9027	USA
S23	U88822	8975	Zaire	S64	AY169814	9186	USA
S24	HIVU86780	8990	Zambia	S65	DQ853465	8798	USA
S25	AY008717	8784	China	S66	AF049494	9193	USA
S26	AY180905	9010	China	S67	AY536236	8937	Venezuela
S27	EF057102	9674	China	S68	AF042106	9096	Australia
S28	AY008714	8859	China	S69	AF042105	8669	Australia
S29	AY008718	8806	China	S70	AF042104	8733	Australia
S30	EF420987	9703	China	S71	AY818642	8820	Australia
S31	AF049337	9050	Cyprus	S72	AF042102	8540	Australia
S32	EF469243	9830	India	S73	AY968312	8834	Argentina
S33	AY049711	9054	India	S74	AY536238	8760	Argentina
S34	EU000514	9043	India	S75	DQ358808	8956	Brazil
S35	EU031914	9563	Malaysia	S76	DQ358809	9419	Brazil
S36	EU031915	8942	Malaysia	S77	AY173956	8940	Brazil
S37	DQ295193	9468	South Korea	S78	AY771589	9058	Brazil
S38	DQ295195	9402	South Korea	S79	EU735539	9105	Brazil
S39	AF086817	9694	Taiwan	S80	HIVU52953	8959	Brazil
S40	AY173951	8996	Thailand	S81	EU735540	9079	Brazil
S41	DQ912823	8900	Denmark				

Table 2
Relative abundance[a] of di, tri, tetra, penta, hexanucleotide repeats in HIV-1 genomes.

No.	Relative abundance	No.	Relative abundance	No.	Relative abundance	No.	Relative abundance
S1	3/0.8/0/0/0[b]	S22	3/0.8/0/0/0	S43	3/0.8/0/0/0	S64	3/0.3/0/0/0
S2	4/0.4/0/0/0	S23	2/0.9/0/0/0	S44	4/0.4/0/0/0	S65	3/0.9/0/0/0
S3	2/0.3/0/0/0	S24	3/0.8/0/0/0	S45	5/0.3/0/0/0	S66	3/0.7/0/0/0
S4	2/0.5/0/0/0	S25	2/0.6/0/0/0	S46	4/0.5/0/0/0	S67	2/1/0/0/0
S5	4/0.7/0/0/0	S26	2/0.7/0/0/0	S47	3/0.9/0/0/0	S68	3/1/0/0/0
S6	3/0.4/0/0/0	S27	3/0.7/0/0/0	S48	3/0.4/0/0/0	S69	3/0.8/0/0/0
S7	4/0.6/0/0/0	S28	3/0.6/0/0/0	S49	3/0.6/0/0/0	S70	4/0.7/0/0/0
S8	3/0.4/0/0/0	S29	3/0.8/0/0/0	S50	3/0.8/0/0/0	S71	3/0.7/0/0/0
S9	4/0.8/0/0/0	S30	3/0.7/0/0/0	S51	4/0.5/0/0/0	S72	3/1/0/0/0
S10	3/1/0/0/0	S31	4/0.8/0/0/0	S52	3/0.8/0/0/0	S73	3/0.9/0/0/0
S11	3/0.8/0/0/0	S32	3/0.5/0/0/0	S53	3/0.4/0/0/0	S74	2/0.6/0/0/0
S12	3/0.6/0/0/0	S33	3/0.6/0/0/0	S54	3/0.8/0/0/0	S75	3/0.8/0/0/0
S13	3/1/0/0/0	S34	3/0.6/0/0/0	S55	3/0.5/0/0/0	S76	4/0.6/0/0/0
S14	3/0.7/0/0/0	S35	3/0.7/0/0/0.1	S56	3/0.6/0/0/0	S77	2/0.7/0/0/0
S15	4/0.7/0/0/0	S36	3/0.9/0/0/0	S57	3/1/0/0/0	S78	3/1/0/0/0
S16	4/0.3/0/0/0	S37	4/0.7/0/0/0	S58	4/0.7/0/0/0	S79	3/1/0/0/0
S17	3/0.3/0/0/0	S38	4/1/0/0/0	S59	3/0.7/0/0/0	S80	3/0.8/0/0/0
S18	3/0.7/0/0/0	S39	4/0.8/0/0/0	S60	3/0.8/0/0/0	S81	4/0.7/0/0/0
S19	3/0.9/0/0/0	S40	3/0.8/0/0/0	S61	3/1/0/0/0		
S20	3/0.4/0.1/0/0	S41	3/0.6/0/0/0	S62	4/0.8/0/0/0		
S21	3/0.7/0/0/0	S42	3/0.7/0/0/0	S63	3/0.6/0/0/0		

[a] SSR relative abundance is the total repeats per kb of sequence analyzed. For example: DQ396382 had 29 dinucleotide repeats, 4 trinucleotide repeats, 1 tetranucleotide repeat and the genome was 9068 bases in length. Thus, relative abundance of dinucleotide repeats = $(29/9068) \times 1000 \approx 3$; relative abundance of trinucleotide repeats = $(4/9068) \times 1000 \approx 0.4$; and relative abundance of tetranucleotide repeats = $(1/9068) \times 1000 \approx 0.1$.
[b] Relative abundance of dinucleotide repeats/trinucleotide repeats/tetranucleotide repeats/pentanucleotide repeats/hexanucleotide repeats, respectively.

Table 3
Relative density[a] of SSRs in HIV-1 genomes.

No.	Relative density	No.	Relative density	No.	Relative density	No.	Relative density
S1	27	S22	25	S43	26	S64	19
S2	28	S23	23	S44	29	S65	29
S3	16	S24	25	S45	32	S66	25
S4	20	S25	20	S46	28	S67	24
S5	29	S26	22	S47	30	S68	32
S6	24	S27	27	S48	24	S69	26
S7	27	S28	24	S49	24	S70	33
S8	24	S29	27	S50	26	S71	23
S9	29	S30	28	S51	28	S72	26
S10	32	S31	31	S52	24	S73	24
S11	24	S32	24	S53	25	S74	20
S12	25	S33	21	S54	27	S75	24
S13	28	S34	22	S55	24	S76	29
S14	25	S35	29	S56	27	S77	22
S15	31	S36	26	S57	26	S78	26
S16	25	S37	29	S58	29	S79	28
S17	22	S38	35	S59	27	S80	24
S18	24	S39	33	S60	30	S81	29
S19	27	S40	25	S61	33		
S20	26	S41	22	S62	32		
S21	26	S42	25	S63	22		

[a] SSR relative density is defined as the total length (bp) contributed by each SSR per kb of sequence analyzed. For example: a total of 34 repeats were found in DQ396382 genome with the length of 9068 bases, and the total length of these repeats is 234 bp. Thus, relative density of repeats = $(234/9068) \times 1000 \approx 26$.

types was similar across genomes. For example, most, but not all, of relative abundance of dinucleotide repeats of each sequence was 3 or 4. It was even more evident for relative abundance of other repeat types, as the relative abundance of trinucleotide repeats is less than one or equivalent to 1 and the relative abundance of tetra-, penta- and hexanucleotide repeats was mostly 0 in each of surveyed HIV-1 genomic sequences. Its suggested that strand-slippage theories alone are insufficient to explain characteristic SSR distributions that differential abundance of repeats in different genomes of HIV-1 [10].

In each of these surveyed genomic sequences, as expected, the analysis of motif patterns revealed the smaller repeat motifs were overrepresented. Among dinucleotide repeats, AG/GA repeats were predominant, while CG/GC repeats were relatively rare (Supplementary Figs. S1–S81). This is especially interesting because the content of CG/GC repeats is also very low in some chromosome or genomes of human, *Drosophila*, *Arabidopsis*, *Caenorhabditis elegans*, yeast [10], fungi [1,9] and some prokaryotes [24]. Trinucleotide repeats were the second most abundant repeats, and their repeat types might be different in various completed HIV-1 genomes (Supplementary Table 2). Nearly no tetra-, penta-, and hexanucleotide repeats were identified in all of the genomes. However, there are a few notable exceptions, e.g., a $(GGAA)_3$ tetranucleotide repeat which existed in the sequence of DQ396382 and an $(AAGAGG)_3$ hexanucleotide repeat which was presented in the sequence of EU031914.

3.2. Relative density of SSRs

The relative density of SSRs is nearly as equally represented across the 81 completed HIV-1 genomic sequences, regardless of whether the sequences were selected from the same countries or districts. The highest SSR density was 35 bp/kb found in the sequence of DQ295195 which was from South Korea of Asia, and the lowest SSR density was 16 bp/kb in the sequence of AY169815, which was originated from Cameroon of Africa, but the relative density of most of these sequences was 20 bp/kp or more.

In recent years it has been demonstrated that trinucleotide repeats are more abundant than other repeat types in coding regions of some eukaryotic and prokaryotic genomes [1,2,22], and dynamic mutations in trinucleotide repeats occasionally associated with diseases [25] and other important functions [26]. Here, we focused on SSRs which locate in completed HIV-1 genomes, and investigated whether the density of SSRs in these sequences was the same or not. Our results show that dinucleotide repeats are extremely common compared to trinucleotide repeats in completed sequences. Dinucleotide repeats were thought to allow more possible slippage events per unit length of DNA and hence be more unstable because it own highest slippage rate than trinucleotide repeats and tetranucleotide repeats [10]. The observations appeared to indicate that the difference of proportion of repeat types might have impact on the organization and stability of completed HIV-1 genomes.

3.3. Most common SSR motifs

There is evidence that the most common SSR motifs are different in various organisms. For instance, it has been reported that $(GT)_n$ is the most common SSR motif in animals and invertebrates [27], whereas in plants the repeats $(AT)_n$ are the most common [28] and in insects $(CT)_n$ are the most common one [29]. However, it seems still rare to be reported that whether the most common SSR motifs were different in various completed sequences of the same species. Our results indicate the most common SSR motifs also might be different in the same species. Among 58 sequences of these surveyed sequences in the study, the most common SSR motifs are $(GA)_n$ which can occur between 4 and 13 times in single sequence, while in other sequences the common SSR motifs can be $(AG)_n$, $(CA)_n$, $(AT)_n$, $(GT)_n$ or $(TA)_n$ (Supplementary Table 3). Moreover, it is also noted that a sequence might harbor most common SSR motifs one or more and the total occurrences of most common SSR motifs might be different in a variety of sequences.

3.4. Longest SSRs

Its inferred that longer repeats have higher mutation rates and hence are unstable [30], which lead to the frequency of SSRs decrease gradually with repeat length [10]. According to the hypothesis, longest SSR might be one of the most unstable SSRs in a sequence. In this study, the longest SSR is a hexanucleotide repeat, which is $(AAGAGG)_3$ with length of 18 nt in all analyzed genomes. However, each sequence has its own longest SSR with length of 9 nt, 10 nt or 12 nt and so on (Supplementary Table 4). Besides hexanucleotide repeat type, di-, tri-, tetranucleotide repeat types were also observed in longest SSRs. Generally, in a sequence, no more than two SSRs with a length >10 nt were found. The absence of very long SSRs in mitochondrial, chloroplast [3] and fungal genomes [9] may be due to their smaller genome size, a relatively stable nature, downward mutation bias and short existence time [31]. However, HIV-1 genomes are extremely instable, and being found to be hyper mutative [21]. Thus, the absence of very long SSRs in HIV-1 genomes suggests the involvement of additional mechanisms.

Acknowledgement

This work was made possible by funding from the Chinese Key National Technology R&D Program: 2006BAD08A13.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2009.08.004.

References

[1] Kim, T.S., Booth, J.G., Gauch Jr., H.G., Sun, Q., Park, J., Lee, Y.H. and Lee, K. (2008) Simple sequence repeats in *Neurospora crassa*: distribution, polymorphism and evolutionary inference. BMC Genomics 9, 31.
[2] Li, Y.C., Korol, A.B., Fahima, T. and Nevo, E. (2004) Microsatellites within genes: structure, function, and evolution. Mol. Biol. Evol. 21, 991–1007.
[3] Rajendrakumar, P., Biswal, A.K., Balachandran, S.M., Srinivasarao, K. and Sundaram, R.M. (2007) Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intercoding regions. Bioinformatics 23, 1–4.
[4] Hong, C.P., Piao, Z.Y., Kang, T.W., Batley, J., Yang, T.J., Hur, Y.K., Bhak, J., Park, B.S., Edwards, D. and Lim, Y.P. (2007) Genomic distribution of simple sequence repeats in *Brassica rapa*. Mol. Cells 23, 349–356.
[5] Cardle, L., Ramsay, L., Milbourne, D., Macaulay, M., Marshall, D. and Waugh, R. (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156, 847–854.
[6] Toth, G., Gaspari, Z. and Jurka, J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10, 967–981.
[7] Gur-Arie, R., Cohen, C.J., Eitan, Y., Shelef, L., Hallerman, E.M. and Kashi, Y. (2000) Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism. Genome Res. 10, 62–71.
[8] Mrazek, J., Guo, X. and Shah, A. (2007) Simple sequence repeats in prokaryotic genomes. Proc. Natl. Acad. Sci. USA 104, 8472–8477.
[9] Karaoglu, H., Lee, C.M. and Meyer, W. (2005) Survey of simple sequence repeats in completed fungal genomes. Mol. Biol. Evol. 22, 639–649.
[10] Katti, M.V., Ranjekar, P.K. and Gupta, V.S. (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol. Biol. Evol. 18, 1161–1167.
[11] Riley, D.E. and Krieger, J.N. (2009) Embryonic nervous system genes predominate in searches for dinucleotide simple sequence repeats flanked by conserved sequences. Gene 429, 74–79.
[12] Riley, D.E. and Krieger, J.N. (2009) UTR dinucleotide simple sequence repeat evolution exhibits recurring patterns including regulatory sequence motif replacements. Gene 429, 80–86.
[13] Madsen, B.E., Villesen, P. and Wiuf, C. (2008) Short tandem repeats in human exons: a target for disease mutations. BMC Genomics 9, 410.
[14] Heesacker, A., Kishore, V.K., Gao, W., Tang, S., Kolkman, J.M., Gingle, A., Matvienko, M., Kozik, A., Michelmore, R.M., Lai, Z., Rieseberg, L.H. and Knapp, S.J. (2008) SSRs and INDELs mined from the sunflower EST database: abundance, polymorphisms, and cross-taxa utility. Theor. Appl. Genet. 117, 1021–1029.
[15] Dereeper, A., Argout, X., Billot, C., Rami, J.F. and Ruiz, M. (2007) SAT, a flexible and optimized Web application for SSR marker development. BMC Bioinformatics 8, 465.
[16] Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S. and 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.
[17] Kashi, Y., King, D. and Soller, M. (1997) Simple sequence repeats as a source of quantitative genetic variation. Trends Genet. 13, 74–78.
[18] Huang, T.S., Lee, C.C., Chang, A.C., Lin, S., Chao, C.C., Jou, Y.S., Chu, Y.W., Wu, C.W. and Whang-Peng, J. (2003) Shortening of microsatellite deoxy(CA) repeats involved in GL331-induced down-regulation of matrix metalloproteinase-9 gene expression. Biochem. Biophys. Res. Commun. 300, 901–907.
[19] Ellegren, H. (2004) Microsatellites: simple sequence sequences with complex evolution. Nat. Rev. Genet. 5, 435–445.
[20] Kuiken, C., Leitner, T., Foley, B., Hahn, B., Marx, P., McCutchan, F. et al. (2008) HIV Sequence Compendium 2008. Published by Theoretical Biology and Biophysics. URL: http://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/2008/frontmatter.pdf.
[21] Reeves, J.D. and Doms, R.W. (2002) Human immunodeficiency virus type 2. J. Gen. Virol. 83, 1253–1265.
[22] Morgante, M., Hanafey, M. and Powell, W. (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat. Genet. 30, 194–200.
[23] 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.

[24] Field, D. and Wills, C. (1998) Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. Proc. Natl. Acad. Sci. USA 95, 1647–1652.
[25] Usdin, K. (2008) The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res. 18, 1011–1019.
[26] Kashi, Y. and King, D.G. (2006) Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22, 253–259.
[27] Stallings, R.L., Ford, A.F., Nelson, D., Torney, D.C., Hildebrand, C.E. and Moyzis, R.K. (1991) Evolution and distribution of (GT)*n* repetitive sequences in mammalian genomes. Genomics 10, 807–815.
[28] Lagercrantz, U., Ellegren, H. and Andersson, L. (1993) The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. Nucleic. Acids Res. 21, 1111–1115.
[29] Paxton, R.J., Thoren, M.P.A., Tengo, J., Estoup, A. and Pamilo, P. (1996) Mating structure and nestmate relatedness in a communal bee, *Andrena jacobi* (Hymenoptera, Andrenidae), using microsatellites. Mol. Ecol. 5, 511–519.
[30] Wierdl, M., Dominska, M. and Petes, T.D. (1997) Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics 146, 769–779.
[31] Harr, B. and Schlötterer, 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.