Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Biochemical and Biophysical Research Communications 514 (2019) 726–733

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

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

Conserved microsatellites may contribute to stem-loop structures in 5′, 3′ terminals of Ebolavirus genomes

Douyue Li [1], Hongxi Zhang [1], Shan Peng, Saichao Pan, Zhongyang Tan*

Bioinformatics Center, College of Biology, Hunan University, Changsha, China

ARTICLE INFO

Article history:
Received 9 April 2019
Received in revised form
25 April 2019
Accepted 28 April 2019
Available online 8 May 2019

Keywords:
Conserved microsatellite
Ebolavirus
Conserved stem-loop structure
Evolutionary selection

ABSTRACT

Microsatellites (SSRs) are ubiquitous in coding and non-coding regions of the Ebolavirus genomes. We synthetically analyzed the microsatellites in whole-genome and terminal regions of 219 Ebolavirus genomes from five species. The Ebolavirus sequences were observed with small intraspecies variations and large interspecific variations, especially in the terminal non-coding regions. Only five conserved microsatellites were detected in the complete genomes, and four of them which well base-paired to help forming conserved stem-loop structures mainly appeared in the terminal non-coding regions. These results suggest that the conserved microsatellites may be evolutionary selected to form conserved secondary structures in 5′, 3′ terminals of Ebolavirus genomes. It may help to understand the biological significance of microsatellites in Ebolavirus and also other virus genomes.

1. Introduction

Microsatellites, or simple sequence repeats (SSRs), are tandemly repeated tracts of 1–6 nucleotides that are repeated multiple times [1–3]. Microsatellites show wide distributions in eukaryotes, prokaryotes and viruses [4], and have length polymorphism, motif diversity and non-random distribution characteristics [5]. Microsatellites are highly mutable and vary widely among individuals possibly as a result of strand slippage during replication [6,7]. These mutations may be correlated with the genomic structures and functions [8,9]. Thus, the emergence and fixation of microsatellites could contribute to the genomic diversity and evolution [10]. In addition, microsatellites expansion has been proved to cause more than 30 neurological and neuromuscular diseases [11], for instance, epilepsy caused by abnormal expansions of TTTCA and TTTTA repeats [12], Huntington disease cansed by CAG repeats [13], etc. Also it is related to lung cancer [14], gastric cancer [15], etc. The occurrence of microsatellites in non-coding regions are also associated with diseases by regulating gene expression, such as Friedreich Ataxia [16].

Whole-genome-scale analysis of 257 viruses suggested that microsatellites were widely distributed in protein-coding and non-coding regions of viral genomes [4]. Microsatellites might potentially contribute to evolution of virus, through the comprehensive analysis of viral genomes in HIV-1 [17], Potyvirus [18] and HCV [19]. The distribution and composition of microsatellites in seven species of Filoviridae family, including Ebolavirus, also have salient features [20]. Ebolavirus comprises of five species: *Zaire ebolavirus*, *Bundibugyo virus*, *Reston ebolavirus*, *Sudan ebolavirus*, and *Taï Forest ebolavirus* [21]. The Ebolavirus genome is a negative and single-stranded RNA approximately 19 kb nucleotides. It is organized into a 3′ leader non-coding regions, followed by seven genes and a 5′ trailer non-coding regions [22]. Microsatellites in non-coding regions account for a large proportion of the Ebolavirus genomes [20].

The 5′ and 3′ terminal non-coding regions of virus likely regulate its replication, transcription, and assembly of new virions [23,24]. In Picornaviridae genomes, the terminal sequences can form stem-loop secondary structures to play a role in the viral life cycle [25]. Similarly, there are the conserved secondary structures in potato virus X RNA [26], hepatitis C viruses [27] and coronavirus [28]. Microsatellites mainly exist in non-coding regions [6,29]. The conservation of microsatellites in non-coding regions may be important to gene regulation in plants [30]. In plant viroids, the distribution of microsatellites probably contributes to the secondary structure [31]. The present study is a systematic analysis for the microsatellites in 219 genomic sequences from five Ebolavirus species. To our knowledge, this may be the first study to investigate the related biological significance of microsatellites in virus.

* Corresponding author.
E-mail address: zhongyang@hnu.edu.cn (Z. Tan).
[1] Co-First Author.

https://doi.org/10.1016/j.bbrc.2019.04.192
0006-291X/

2. Materials and methods

2.1. Collection of genomic sequences

All available 219 complete genomic sequences of Ebolavirus from five species were downloaded from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/genome/), and all were cDNA sequences. These genomes were considered as the references, which recorded more complete terminal sequence. In these sequences, 184 EBOV sequences were selected as the main reference sequence due to appearing with the most lethal rate and the most attention. The other BDBV, RESTV, SUDV and TAFV were respectively 8 (B1–B8), 9 (R1-R9), 15 (S1–S15) and 3 (T1-T3) (Table 1), and the detailed information about these sequences was also listed (Table in Supplementary Table S1).

2.2. Extraction of microsatellites

IMEx software was selected to extract microsatellites from 219 samples [1,32]. The lowest copy numbers for extracting each microsatellite motif, which was based on empirical criterion and referenced to previous microsatellite studies, were 6, 3, 3, 3, 3, 3 for the perfect mononucleotide to hexanucleotide microsatellites separately. Owing to extracting perfect microsatellites alone, the basic mode of IMEx software was chosen and the imperfect rate was 0 here [34]. Other parameters were used as default. In addition, the compound microsatellite and imperfect microsatellites were excluded in this analysis, thus there was no different result for the microsatellite extraction while we using other extracting tools or algorithms for checking, like MfSAT [35], RepeatMasker [36] and MISA [37]. The results of these methods were not significantly different (data not shown).

2.3. Alignment generation

In order to analyze the role of microsatellites in Ebola virus genome, sequence alignment was used to compare the distributional variations of microsatellites more intuitively. The sequence alignment was performed in Clustal W (version 2.1) and then detected the specifically conserved regions [38,39]. To weak the impact on gaps between microsatellites, the number of gaps was reduced and the length of gap was shorten, thus, the parameters were resetting: gap open penalty = 30.0, gap extension penalty = 10.0.

2.4. Prediction of RNA secondary structure

Three common programs and a freely available web server platform were chosen to predict the RNA secondary structure of conserved regions in Ebola virus terminal microsatellite, including IPKNOT 1.3.1 (McCaskill), RNAfold, MFOLD v2.3 [40]. In order to obtain succinct images of subsequent processing, the predicted results of IPKNOT 1.3.1 were selected as the final result graphs, and the outputs from the other two prediction platforms were used as reference to adjustment and improvement. In addition, we also used the R-scape to perform pairwise covariation analysis to support our predictions (data not shown) [41].

2.5. Statistical analysis

Means, standard deviations, and graphs were generated with Microsoft excel. SPSS 22.0 was utilized to analyze statistical differences, and the independent sample *t*-test was used for comparison between the two groups. $P < 0.0001$ indicated that the difference was statistically significant.

3. Results

3.1. High sequence diversity of 5′, 3′ terminal regions and low diversity in complete genomes of ebolavirus

To measure sequence variances of five species, we aligned the complete genomes and 5′, 3′ terminal non-coding regions of 219 sequences. Size of all analyzed regions were ranged from (18874.73 ± 0.44) nt (SUDV) to (18990.04 ± 158.91) nt (EBOV) of complete genomes, (457.00 ± 0.00) nt (BDBV and SUDV) to (468.89 ± 0.53) nt (EBOV) of 5′ terminals, (703.22 ± 0.44) nt (RESTV) to (741.25 ± 0.46) nt (BDBV) of 3′ terminals, respectively (Table 1). Intraspecific variations were compared by using the first sequence of each species (Z1, B1, T1, R1, S1) as reference sequences (Supplementary Table S1). Compared with the aligned scores of complete genomes, there were no statistical significance ($P > 0.05$) in these scores of 5′ and 3′ terminals, which were above 90 (Fig. 1A–C, Supplementary Table S2). However, in a comparison of same-region genomes, tremendous variances could be found among five species, although they have similar genome sizes. The aligned scores of the complete genomes were only in the range of 60–65, and those of 5′ terminals ranged from 17 to 25. This trait was even more evident in 3′ terminal sequences, in which the lowest aligned score was 5 (EBOV *vs.* BDBV). And significance analysis showed that interspecies variations of two terminal regions were significantly higher than those of complete genomes ($P < 0.0001$) (Fig. 1D–F). Results here indicated that the terminal non-coding regions of Ebolavirus may have greater sequence variances than other parts of the genomes.

3.2. High conserved microsatellites loci mainly occurred in 5′, 3′ terminal non-coding regions

Microsatellites on all of the analyzed sequences were extracted to explore the relationship between microsatellites and Ebolavirus genomes. All microsatellites were counted for each of these genomes and the average summated microsatellites were ranged

Table 1
List of 5 species Ebolavirus in complete genomes and 5′, 3′ terminal sequences.

Species	Abbreviation	Number of sequences	Genome size (nt)	5′ terminal		3′ terminal	
				Size (nt)	Length[b]	Size (nt)	Length
Zaire ebolavirus	EBOV	184	18990.04 ± 158.91[a]	468.89 ± 0.53	47.90 ± 0.53	739.66 ± 0.63	50.62 ± 0.51
Bundibugyo virus	BDBV	8	18939.63 ± 0.74	457.00 ± 0.00	48.00 ± 0.00	741.25 ± 0.46	50.13 ± 0.35
Taï Forest ebolavirus	TAFV	3	18935.00 ± 0.00	463.00 ± 0.00	48.00 ± 0.00	737.00 ± 0.00	51.00 ± 0.00
Reston ebolavirus	RESTV	9	18890.44 ± 2.40	462.78 ± 0.67	48.00 ± 0.00	703.22 ± 0.44	50.25 ± 0.46
Sudan ebolavirus	SUDV	15	18874.73 ± 0.44	457.00 ± 0.00	48.00 ± 0.00	707.73 ± 0.44	50.64 ± 0.50

[a] The sizes of 184 sequences are averaged and the standard deviation is calculated to indicate the degree of dispersion.
[b] Length for Prediction: regions including conservative microsatellites is intercepted for secondary structure prediction.

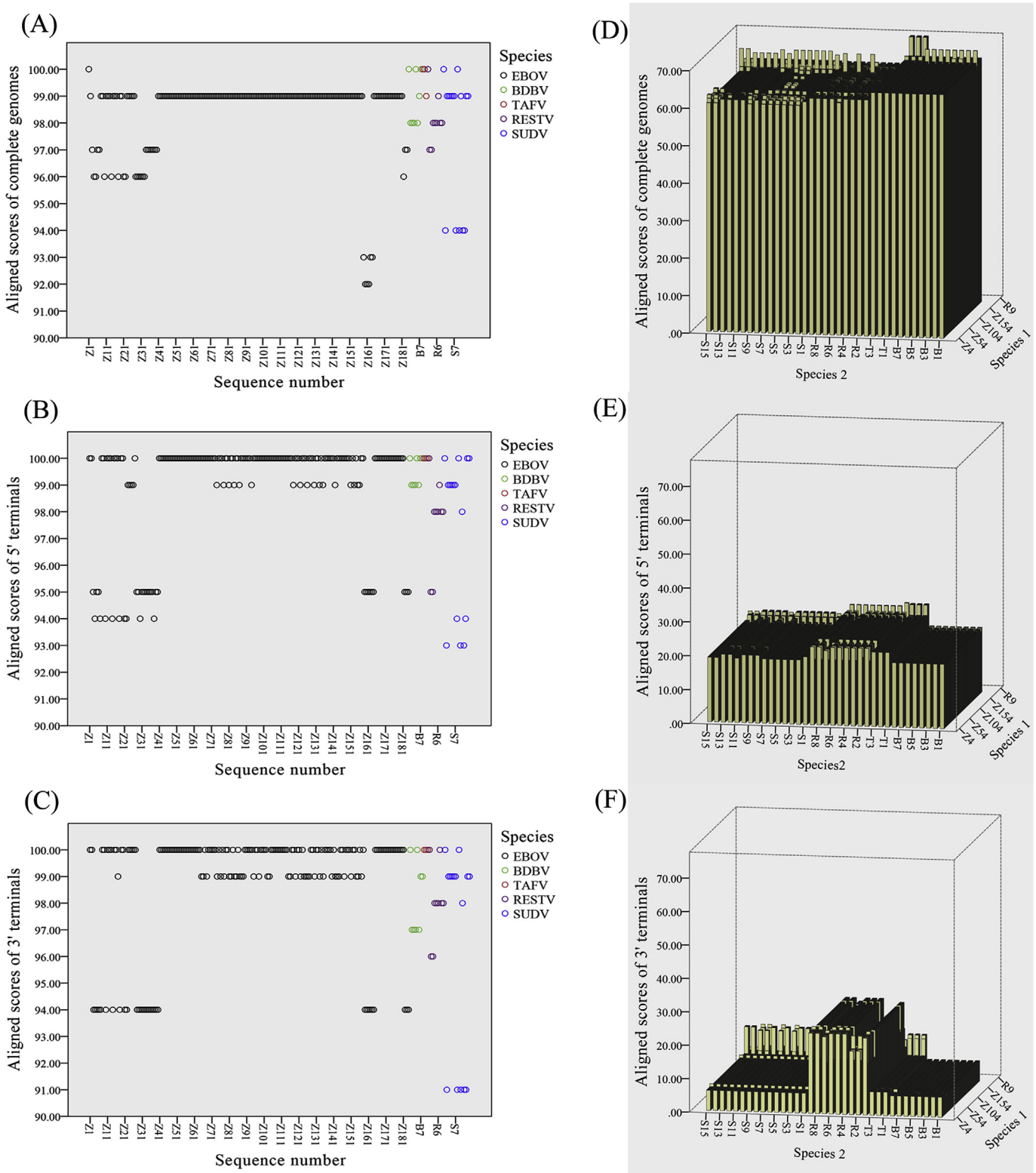

Fig. 1. Sequence diversity in 5′, 3′ terminal regions and complete genomes of Ebolavirus. (A)Intraspecies variations of complete genomes. (B) Intraspecies variations of 5′ terminal sequences. (C) Intraspecies variations of 3′ terminal sequences. (D) Interspecific variations of complete genomes. (E) Interspecific variations of 5′ terminal sequences. (F) Interspecific variations of 3′ terminal sequences. Species 1 and Species 2 are the distinctions between different sequences for pairwise alignment, respectively. Species 1 includes the sequences of Z1-Z184, B1−B8, R1-R9, and Species 2 includes the sequences of B1−B8, R1-R9, S1−S15.

from 61.07 ± 2.31 (SUDV) to 81.78 ± 0.83 (RESTV) (Fig. 2A, Table 2, Supplementary Table S2). Then the extracted microsatellites were labeled on the sequences to observe different locations had the identical microsatellite in five species, which was called conserved loci in this research. These included positions of 5–10 nt, 39–44 nt, 6918–6924 nt, 18951–18956 nt (Z1 as reference), and the microsatellites were respectively CACACA, UGUGUG, AAAAAAA, UGUGUG that were named as completely conserved microsatellites. In addition, there was a relatively conserved locus at 18912 nt, similar to CACACA of the 3′ terminals, however, it was imperfect microsatellite UACACA in BDBV and TAFV except for the other species. It was also classified as a conserved locus, considering its particularity and there was no such a phenomenon elsewhere in the genome. And the imperfect microsatellites on these conserved loci were defined as highly conserved microsatellites. These five conserved loci were named for Conserved 1–5 (Fig. 2B, Table 2, Supplementary Fig. S1). Moreover, nearly all (100%) sequences had these conserved microsatellites (Table 2).

Although there is no direct correlation of microsatellites with sequence variances, it is generally assumed that the more similar genomes are expected to contain more conserved microsatellites than do the more different ones [4]. On the contrary, conserved microsatellite loci were mainly located on the terminal non-coding regions with the exception of Conserved 3 that were located on GP gene. In view of this particular phenomenon, we mainly analyzed the terminal sequences further, and continue to study the Conserved 3 in the future. In order to verify the extremely high conservatism of conserved microsatellites at the terminal sequences, the conserved rate was analyzed and the counts subsequently compared across complete genomes. The highest conserved rate of the complete genomes was 8.20% ± 0.30, while the highest one in the terminal regions was 100% ± 0.00 (EBOV), which meant that there were only those two conserved microsatellites in all sequences. Conserved rate of the terminal sequences were considerably higher than those of complete genomes ($P < 0.0001$) (Fig. 2C, Supplementary Table S2). To observe the proportion of conserved microsatellites in the sequences, the relative abundance of the microsatellites was calculated to eliminate the influence of different lengths. Additionally, the perfect microsatellites were mainly analyzed in this step, so only completely conserved microsatellites were considered for further analysis. Relative abundances of conserved microsatellites in 5′, 3′ terminal sequences were significantly higher than those of complete genomes ($P < 0.0001$) (Fig. 2D, Supplementary Table S2).

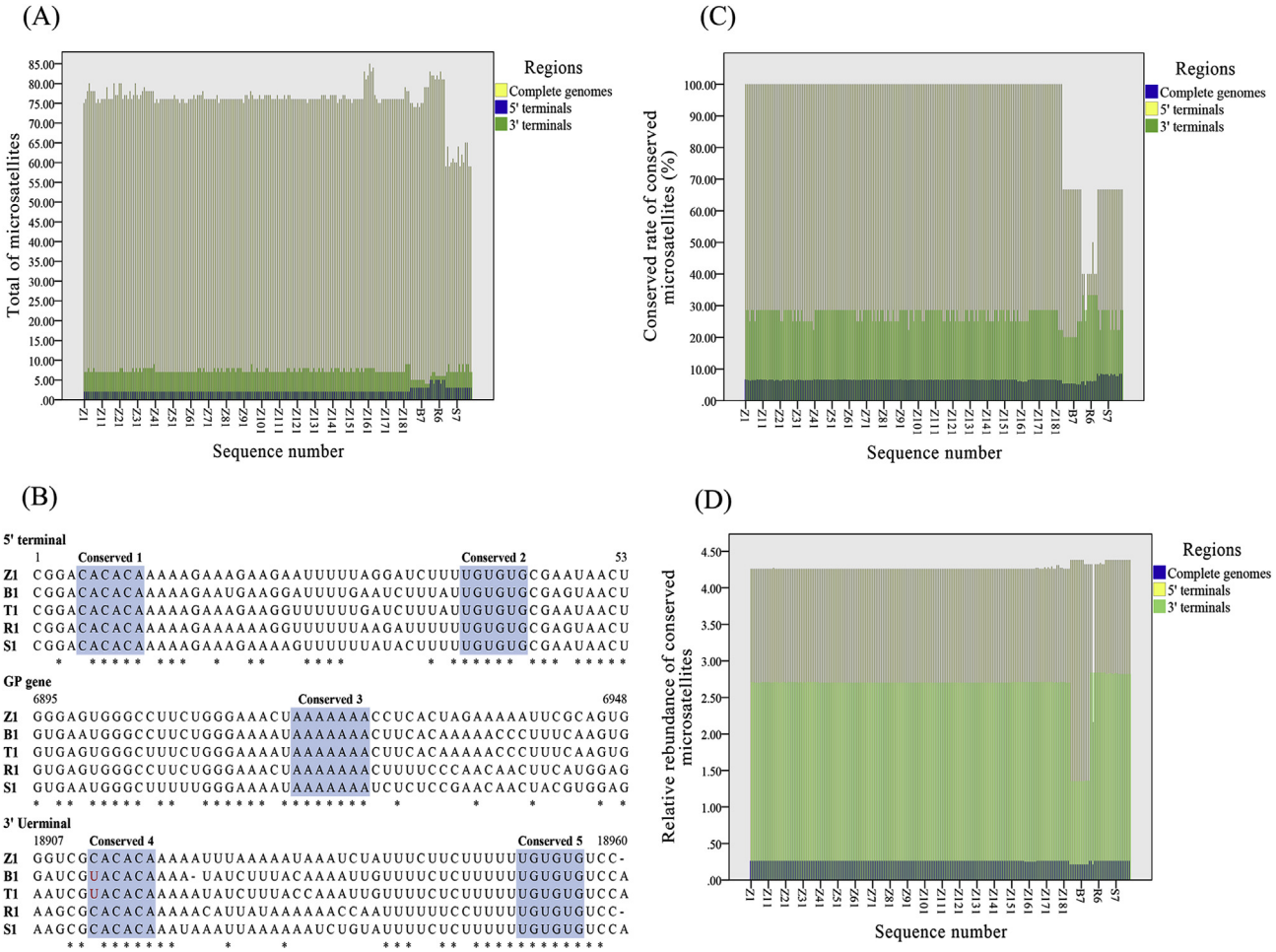

Fig. 2. High conserved microsatellites loci mainly occurred in 5′, 3′ terminal sequences. (A) Total of microsatellites in the complete genomes and 5′, 3′ terminal sequences. (B) Sequences of five conserved microsatellite loci. (C) Conserved rate of complete genomes and terminal sequences. (D) Relative abundance of complete genomes and terminal sequences. * indicates that they have the same base. Blue is used to represent conserved microsatellites, and different bases on highly conserved microsatellites are represented in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Statistics of conserved microsatellites loci in Ebolavirus genomes.

Species	Regions	Total of SSRs	Conserved SSRs (%)[a]	Relative abundance[b]	Motif[c]	Rate of sequences (%)[d]
EBOV	Complete genome	76.66 ± 1.55[e]	6.53 ± 0.13	0.26 ± 0.00	–[f]	–
	5′ terminal	2.00 ± 0.00	**100 ± 0.00**	4.27 ± 0.01	CACACA	100.00
					TGTGTG	100.00
	3′ terminal	7.35 ± 0.53	27.33 ± 1.84	2.70 ± 0.00	CACACA	100.00
					TGTGTG	100.00
BDBV	Complete genome	74.63 ± 0.52	5.36 ± 0.04	0.21 ± 0.00	–	–
	5′ terminal	3.00 ± 0.00	66.67 ± 0.00	4.38 ± 0.00	CACACA	100.00
					TGTGTG	100.00
	3′ terminal	5.00 ± 0.00	20.00 ± 0.00	1.35 ± 0.00	**TACACA**	100.00
					TGTGTG	100.00
TAFV	Complete genome	79.00 ± 0.00	5.06 ± 0.00	0.21 ± 0.00	–	–
	5′ terminal	3.00 ± 0.00	66.67 ± 0.00	4.32 ± 0.00	CACACA	100.00
					TGTGTG	100.00
	3′ terminal	4.00 ± 0.00	25.00 ± 0.00	1.36 ± 0.00	**TACACA**	100.00
					TGTGTG	100.00
RESTV	Complete genome	81.78 ± 0.83	5.98 ± 0.42	0.26 ± 0.02	–	–
	5′ terminal	4.78 ± 0.44	39.44 ± 6.35	4.08 ± 0.72	CACACA	100.00
					TGTGTG	88.89 g
	3′ terminal	6.22 ± 0.44	32.28 ± 2.10	2.84 ± 0.00	CACACA	100.00
					TGTGTG	100.00
SUDV	Complete genome	61.07 ± 2.31	8.20 ± 0.30	0.26 ± 0.00	–	–
	5′ terminal	3.00 ± 0.00	66.67 ± 0.00	4.38 ± 0.00	CACACA	100.00
					TGTGTG	100.00
	3′ terminal	7.67 ± 0.98	26.46 ± 3.10	2.83 ± 0.00	CACACA	100.00
					TGTGTG	100.00

[a] Conserved of conserved microsatellites, number of conserved microsatellites/total of microsatellites × 100%.
[b] Relative abundance of conserved microsatellites, number of conserved microsatellites/kb.
[c] Conserved microsatellites of Conserved 1–5.
[d] Conserved rate of conserved microsatellites in the sequences, such as 184 sequences all have this motif, the conservative rate is 100%.
[e] Number of microsatellites in all sequences is averaged and the degree of dispersion is expressed by standard deviation.
[f] The motifs of complete genome contains those in the terminal sequences, and Conserved 3 was not analyzed in this study and is therefore not listed.

3.3. Conserved microsatellites were located on conserved stem-loop structures

To identify potential functional consequences of conserved microsatellites, RNA secondary structure prediction was performed for the further analysis on the above conserved microsatellites. All terminal sequences of five species were identified and they all form different secondary structures. However, the intuitive analysis of conserved microsatellite regions was affected by the multiple possibilities resulted from the algorithms in the process of prediction. While base pairing of the two conserved regions could lead to the formation of the base of the putative stem-loop structures, the intervening sequence variations altered the predictions for the rest of the structures [42]. Therefore, a segment of the conserved microsatellite region was intercepted for prediction, where the 5′ terminal was truncated to about 48 nt, and the 3′ terminal was truncated to about 50 nt (Table 1). Then these sequences of conserved microsatellite regions were put into an online secondary structure website for prediction, and accomplished with the comparison of multiple websites to ensure its accuracy. Thermodynamic modeling of Ebola viral RNA predicted the formation of RNA stem-loop structures at the 3′ and 5′ terminal. All 5′ terminal sequences formed similar stem-loop structures. Unlike the other three species, RESTV and SUDV did not appear the corner, however, they generally exhibited similar structures (Fig. 3A). Consistent with the 5′ terminal, the secondary structures of 3′ terminal were predicted (Fig. 3B). Sequence analysis showed a highly identity among five species in their 3′ and 5′ terminal.

Furthermore, two conserved microsatellites in 5′ terminal sequences were complementary pairing and formed stems on the stem-loop structure (Fig. 3A). Similarly, two conserved microsatellites on 3′ terminal sequence also complemented each other to form stems on the stem-loop structures. Although the 3′ terminal of BDBV and TAFV were high conserved microsatellites (UACACA) and complete conserved microsatellite (UGUGUG) pairs, they were also highly homologous to the other three species (Fig. 3B). The prediction of RNA secondary structure of 438 terminal sequences revealed all Ebolavirus sequences follow above rules. Summarily, only one sequence of each species was selected to display, and the others were in the supplementary figure (Supplementary Figs. S2–S19).

4. Discussion

In this study, five conserved microsatellites were found in complete genomes of Ebolavirus and mainly distributed in the terminal regions. The four conserved microsatellites were revealed to help forming conserved stem-loop structures by the RNA secondary structure prediction method, and it is tempting to speculate that the conserved microsatellites may contribute to the formation of conserved RNA secondary structures. The former comprehensive analysis [4,17,19] is little known about the relationship between microsatellites and functions and structures of viral genomes. And this study may help to understand the biological significance of microsatellites in Ebolavirus and also other virus genomes.

4.1. Conserved microsatellites may contribute to conserved stem-loop structures in the terminal regions

The selections to the microsatellite appearances in genomes might have some association with the important biological significance [43]. The 5′, 3′ terminal non-coding regions were thought to regulate viral transcription, replication, and assembly of new virions [23]. The 3′ trailer of Ebolavirus were essential for replication by formed a small stem-loop structure, and participated in RNA-Protein interaction [44]. The related results suggested that

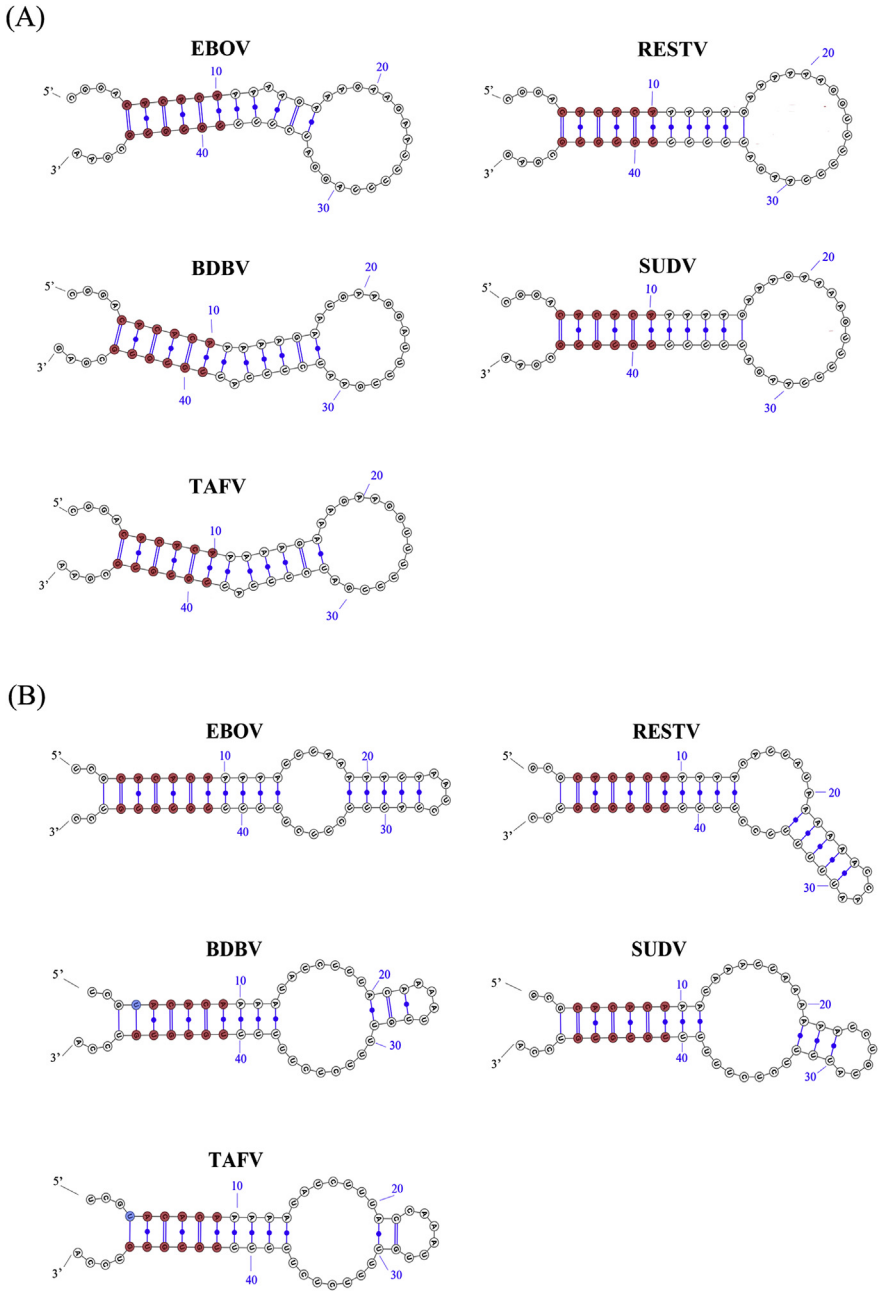

Fig. 3. Conserved microsatellites were located on conserved stem-loop structures. (A) Stem-loop structures of 5 species of Ebolavirus in the 5′ terminal sequences. (B) Stem-loop structures of 5 species of Ebolavirus in the 3′ terminal sequences. Red is used to represent conserved microsatellites, and different bases on highly conserved microsatellites are represented in blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

conserved microsatellites in the terminal regions of Ebolavirus were well base-paired to help forming conserved stem-loop structures (Fig. 3, Supplementary Figs. S2–S19). Sequence analysis showed that the nucleic acid of Ebolavirus (EBOV, SUDV, RESTV, TAFV) between their 3′ terminal (18 nt) and 5′ terminal (20 nt) were highly identical [42], and conserved stem-loop structures were found [42,44,45], which proved the accuracy of the prediction results in this study. Here, we displayed the relationship between such structures and microsatellites to speculate on the role of microsatellites in Ebolavirus genome. Dinucleotide repeats have been demonstrated that can form secondary structure in eukaryotic genome [8]. And Qin et al. found that microsatellite distribution affects their secondary structure in plant viroids [31]. Therefore, the conserved microsatellites analyzed here indicated that they may contribute to the formation of the stem-loop structures in the terminal regions of Ebolavirus.

4.2. Conserved microsatellites under greater evolutionary pressure

The distribution of microsatellites on Ebolavirus genomes had regional preference between and among species [20]. Due to the isolation of many viruses still require the differentiation of thousands of years [46], the viral genome often produces more mutations. Li et al. calculated that the common ancestor for five species of Ebolavirus was about 1257 years old [43]. According to the novel filamentous virus classification process, nucleotide level difference

of less than 30% can be divided into the same species, and less than 50% can be divided into the same genera [21]. Therefore, Ebolavirus genomes had high similarities in same species, and there are large sequence variances among five species (Fig. 1, Supplementary Table S2). However, conserved microsatellites loci were mainly distributed in the terminal non-coding regions, and the terminal regions had larger sequence variances that might face with more selective pressures in contrast to complete genomes. Microsatellites were previously reported to be more abundant in non-coding regions [29,47]. According to statistics, except for the longest L gene, the microsatellites in the non-coding regions of Ebolavirus are the most [20]. However, in this research, the variances of the terminal sequences were clearly more than other sequences in Ebolavirus genomes. In addition, the microsatellites with CA and UG as motifs are not high in viral genome [4]. These conserved microsatellites can be preserved through the long evolutionary time, which mean that they perhaps occurred in the importantly locus.

4.3. Conserved microsatellites might be preserved depending on functional selection

Selective pressure is widely deemed to the causes for the unequal distributions of microsatellites [48,49]. Some microsatellites might be selected by their functions in the process of mutation [50] and could be retained for countless years. For example, some non-coding microsatellites were conserved in plants between species across hundreds of millions of years [30], and human genome have many highly conserved promoter microsatellites [51]. The original Ebolavirus genomes probably experienced powerful pressures and different fates through the series of selection, consequently, the adaptive ones were preserved [50,52].

It is noteworthy that in Conserved 4, the imperfect microsatellite UACACA of BDBV and TAFV was most likely due to single point mutation. According to the stepwise mutation model, the microsatellite will gradually disappear into the process of evolution by means of point mutation [2]. BDBV, which was discovered and named only in 2007, has the highest similarity with TAFV [53]. This conserved locus may serve as another proof that these two species are more homologous than other Ebolavirus species. So the role of Conserved 4 in the Ebola virus deserves further study. In addition, there was a sequence (R3) of RESTV in Conserved 2, and UGUGUG is also mutated to UGUGUA, which might be the same case (Fig. 2B, Supplementary Fig. S1). However, these mutations did not affect the formation of secondary structures (Supplementary Fig. S8). In summary, conserved microsatellites might be a characteristic product of natural variation and deserves more research.

Acknowledgments

This work was jointly supported by funding from the “National Key Plan for Scientific Research and Development of China” (grant No. 2016YFC1200200 and 2016YFD0500300).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbrc.2019.04.192.

Transparency document

Transparency document related to this article can be found online at https://doi.org/10.1016/j.bbrc.2019.04.192.

References

[1] L. Zhou, L. Deng, Y. Fu, et al., Comparative analysis of microsatellites and compound microsatellites in T4-like viruses, Gene 575 (2016) 695–701. https://doi.org/10.1016/j.gene.2015.09.053.
[2] H. Ellegren, Microsatellite mutations in the germline: implications for evolutionary inference, Trends Genet. 16 (2000) 551–558. https://doi.org/10.1016/S0168-9525(00)02139-9.
[3] H. Karaoglu, C.M. Lee, W. Meyer, Survey of simple sequence repeats in completed fungal genomes, Mol. Biol. Evol. 22 (2005) 639–649. https://doi.org/10.1093/molbev/msi057.
[4] X. Zhao, Y. Tian, R. Yang, et al., Coevolution between simple sequence repeats (SSRs) and virus genome size, BMC Genomics 13 (2012) 435. https://doi.org/10.1186/1471-2164-13-435.
[5] R. Gur-Arie, C.J. Cohen, Y. Eitan, et al., Simple sequence repeats in Escherichia coli: abundance, distribution, composition, and polymorphism, Genome Res. 10 (2000) 62–71. https://doi.org/10.1101/gr.10.1.62.
[6] H. Ellegren, Microsatellites: simple sequences with complex evolution, Nat. Rev. Genet. 5 (2004) 435–445. https://doi.org/10.1038/nrg1348.
[7] C.L. Galindo, L.J. McIver, J.F. McCormick, et al., Global microsatellite content distinguishes humans, primates, animals, and plants, Mol. Biol. Evol. 26 (2009) 2809–2819. https://doi.org/10.1093/molbev/msp192.
[8] T.W. Hefferon, J.D. Groman, C.E. Yurk, et al., A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 3504–3509. https://doi.org/10.1073/pnas.0400182101.
[9] A. Bacolla, J.E. Larson, J.R. Collins, et al., Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties, Genome Res. 18 (2008) 1545–1553. https://doi.org/10.1101/gr.078303.108.
[10] R.J. Haasl, B.A. Payseur, Microsatellites as targets of natural selection, Mol. Biol. Evol. 30 (2013) 285–298. https://doi.org/10.1093/molbev/mss247.
[11] A.J. Hannan, Tandem repeats mediating genetic plasticity in health and disease, Nat. Rev. Genet. 19 (2018) 286–298. https://doi.org/10.1038/nrg.2017.115.
[12] H. Ishiura, K. Doi, J. Mitsui, et al., Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy, Nat. Genet. 50 (2018) 581–590. https://doi.org/10.1038/s41588-018-0067-2.
[13] R. Batra, D.A. Nelles, E. Pirie, et al., Elimination of toxic microsatellite repeat expansion RNA by RNA-targeting Cas9, Cell 170 (2017) 899–912, e10, https://doi.org/10.1016/j.cell.2017.07.010.
[14] K.R. Velmurugan, R.T. Varghese, N.C. Fonville, et al., High-depth, high-accuracy microsatellite genotyping enables precision lung cancer risk classification, Oncogene 36 (2017) 6383–6390. https://doi.org/10.1038/onc.2017.256.
[15] M. Ratti, A. Lampis, J.C. Hahne, et al., Microsatellite instability in gastric cancer: molecular bases, clinical perspectives, and new treatment approaches, Cell. Mol. Life Sci. 75 (2018) 4151–4162. https://doi.org/10.1007/s00018-018-2906-9.
[16] H. Tang, E.F. Kirkness, C. Lippert, et al., Profiling of short-tandem-repeat disease alleles in 12,632 human whole genomes, Am. J. Hum. Genet. 101 (2017) 700–715. https://doi.org/10.1016/j.ajhg.2017.09.013.
[17] M. Chen, Z. Tan, J. Jiang, et al., Similar distribution of simple sequence repeats in diverse completed Human Immunodeficiency Virus Type 1 genomes, FEBS Lett. 583 (2009) 2959–2963. https://doi.org/10.1016/j.febslet.2009.08.004.
[18] X. Zhao, Z. Tan, H. Feng, et al., Microsatellites in different Potyvirus genomes: survey and analysis, Gene 488 (2011) 52–56. https://doi.org/10.1016/j.gene.2011.08.016.
[19] M. Chen, Z. Tan, G. Zeng, Microsatellite is an important component of complete hepatitis C virus genomes, Infect. Genet. Evol. 11 (2011) 1646–1654. https://doi.org/10.1016/j.meegid.2011.06.012.
[20] C. Alam, Analysis of simple and imperfect microsatellites in Ebola virus species and members of Filoviridae family, Gene Cell Tissue 2 (2015), e26204. https://doi.org/10.17795/gct-26204.
[21] J.H. Kuhn, K.G. Andersen, Y. Bao, et al., Filovirus RefSeq entries: evaluation and selection of filovirus type variants, type sequences, and names, Viruses 6 (2014) 3663–3682. https://doi.org/10.3390/v6093663.
[22] W. Wan, L. Kolesnikova, M. Clarke, et al., Structure and assembly of the Ebola virus nucleocapsid, Nature 551 (2017) 394–397. https://doi.org/10.1038/nature24490.
[23] L. Markoff, 5'- and 3'-noncoding regions in flavivirus RNA, Adv. Virus Res. 59 (2003) 177–228. https://doi.org/10.1016/S0065-3527(03)59006-6.
[24] A. Sugai, H. Sato, M. Yoneda, et al., Gene end-like sequences within the 3' non-coding region of the Nipah virus genome attenuate viral gene transcription, Virology 508 (2017) 36–44. https://doi.org/10.1016/j.virol.2017.05.004.
[25] C. Witwer, S. Rauscher, I.L. Hofacker, et al., Conserved RNA secondary structures in Picornaviridae genomes, Nucleic Acids Res. 29 (2001) 5079–5089. https://doi.org/10.1128/AEM.69.1.715-718.2003.
[26] K.H. Kim, S.J. Kwon, C. Hemenway, Cellular protein binds to sequences near the 5' terminus of potato virus X RNA that are important for virus replication, Virology 301 (2002) 305–312. https://doi.org/10.1006/viro.2002.1559.
[27] M. Fricke, N. Dunnes, M. Zayas, et al., Corrigendum: conserved RNA secondary structures and long-range interactions in hepatitis C viruses, RNA 22 (2016) 1640–1641. https://doi.org/10.1261/rna.058123.116.
[28] D. Yang, J.L. Leibowitz, The structure and functions of coronavirus genomic 3' and 5' ends, Virus Res. 206 (2015) 120–133. https://doi.org/10.1016/j.

virusres.2015.02.025.
[29] Y.C. Li, A.B. Korol, T. Fahima, et al., Microsatellites within genes: structure, function, and evolution, Mol. Biol. Evol. (2004) 991–1007. https://doi.org/10.1093/molbev/msh073.
[30] L. Zhang, K. Zuo, F. Zhang, et al., Conservation of noncoding microsatellites in plants: implication for gene regulation, BMC Genomics 7 (2006) 323. https://doi.org/10.1186/1471-2164-7-323.
[31] L. Qin, Z. Zhang, X. Zhao, et al., Survey and analysis of simple sequence repeats (SSRs) present in the genomes of plant viroids, Febs Open Bio 4 (2014) 185–189. https://doi.org/10.1016/j.fob.2014.02.001.
[32] S.B. Mudunuri, S. Patnana, H.A. Nagarajaram, MICdb3.0: a Comprehensive Resource of Microsatellite Repeats from Prokaryotic Genomes, Database (Oxford) 2014, 2014 bau005, https://doi.org/10.1093/database/bau005.
[34] M. Chen, G. Zeng, Z. Tan, et al., Compound microsatellites in complete Escherichia coli genomes, FEBS Lett. 585 (2011) 1072–1076. https://doi.org/10.1016/j.febslet.2011.03.005.
[35] M. Chen, Z. Tan, G. Zeng, MfSAT: detect simple sequence repeats in viral genomes, Bioinformation 6 (2011) 171–172. https://doi.org/10.6026/97320630006171.
[36] S. Tempel, Using and understanding RepeatMasker, Methods Mol. Biol. 859 (2012) 29–51. https://doi.org/10.1007/978-1-61779-603-6_2.
[37] S. Beier, T. Thiel, T. Munch, et al., MISA-web: a web server for microsatellite prediction, Bioinformatics 33 (2017) 2583–2585. https://doi.org/10.1093/bioinformatics/btx198.
[38] L. Tian, S. Liu, S. Wang, et al., Ligand-binding specificity and promiscuity of the main lignocellulolytic enzyme families as revealed by active-site architecture analysis, Sci Rep-Uk 6 (2016). https://doi.org/10.1038/srep23605.
[39] R. Chenna, H. Sugawara, T. Koike, et al., Multiple sequence alignment with the Clustal series of programs, Nucleic Acids Res. 31 (2003) 3497–3500. https://doi.org/10.1093/nar/gkg500.
[40] K. Sato, Y. Kato, M. Hamada, et al., IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming, Bioinformatics 27 (2011) i85–i93. https://doi.org/10.1093/bioinformatics/btr215.
[41] E. Rivas, J. Clements, S.R. Eddy, A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs, Nat. Methods 14 (2017) 45–48. https://doi.org/10.1038/nmeth.4066.
[42] S.M. Crary, J.S. Towner, J.E. Honig, et al., Analysis of the role of predicted RNA secondary structures in Ebola virus replication, Virology 306 (2003) 210–218. https://doi.org/10.1016/S0042-6822(02)00014-4.
[43] Y.H. Li, S.P. Chen, Evolutionary history of Ebola virus, Epidemiol. Infect. 142 (2014) 1138–1145. https://doi.org/10.1017/S0950268813002215.
[44] J. Sztuba-Solinska, L. Diaz, M.R. Kumar, et al., A small stem-loop structure of the Ebola virus trailer is essential for replication and interacts with heat-shock protein A8, Nucleic Acids Res. 44 (2016) 9831–9846. https://doi.org/10.1093/nar/gkw825.
[45] V.E. Volchkov, V.A. Volchkova, A.A. Chepurnov, et al., Characterization of the L gene and 5' trailer region of Ebola virus, J. Gen. Virol. 80 (Pt 2) (1999) 355–362. https://doi.org/10.1099/0022-1317-80-2-355.
[46] H.D. Nguyen, Y. Tomitaka, S.Y. Ho, et al., Turnip mosaic potyvirus probably first spread to Eurasian brassica crops from wild orchids about 1000 years ago, PLoS One 8 (2013), e55336. https://doi.org/10.1371/journal.pone.0055336.
[47] J. Mrazek, X. Guo, A. Shah, Simple sequence repeats in prokaryotic genomes, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 8472–8477. https://doi.org/10.1073/pnas.0702412104.
[48] D. Metzgar, J. Bytof, C. Wills, Selection against frameshift mutations limits microsatellite expansion in coding DNA, Genome Res. 10 (2000) 72–80. https://doi.org/10.1101/gr.10.1.72.
[49] R. Gemayel, M.D. Vinces, M. Legendre, et al., Variable tandem repeats accelerate evolution of coding and regulatory sequences, Annu. Rev. Genet. 44 (2010) 445–477. https://doi.org/10.1146/annurev-genet-072610-155046.
[50] R.J. Haasl, R.C. Johnson, B.A. Payseur, The effects of microsatellite selection on linked sequence diversity, Genome Biol Evol 6 (2014) 1843–1861. https://doi.org/10.1093/gbe/evu134.
[51] S.M. Sawaya, A.T. Bagshaw, E. Buschiazzo, et al., Promoter microsatellites as modulators of human gene expression, Adv. Exp. Med. Biol. 769 (2012) 41–54. https://doi.org/10.1007/978-1-4614-5434-2_4.
[52] D. Li, W. Jiao, S. Zhou, et al., Comparative analysis on precise distribution-patterns of microsatellites in HIV-1 with differential statistical method, Gene Reports 12 (2018) 141–148. https://doi.org/10.1016/j.genrep.2018.06.007.
[53] J.S. Towner, T.K. Sealy, M.L. Khristova, et al., Newly discovered ebola virus associated with hemorrhagic fever outbreak in Uganda, PLoS Pathog. 4 (2008), e1000212. https://doi.org/10.1371/journal.ppat.1000212.