Gene Cell Tissue. 2015 April; 2(2): e26204. DOI: 10.17795/gct-26204

Published online 2015 April 7. **Research Article**

Analysis of Simple and Imperfect Microsatellites in Ebolavirus Species and Other Genomes of *Filoviridae* Family

Chaudhary Mashhood Alam [1]; Choudhary Sharfuddin [1]; Safdar Ali [2,*]

[1]Department of Botany, Patna University, Bihar, India
[2]Department of Biomedical Sciences, Shaheed Rajguru College of Applied Sciences For Women (SRCASW), University of Delhi, New Delhi, India

*Corresponding author: Safdar Ali, Department of Biomedical Sciences, Shaheed Rajguru College of Applied Sciences For Women (SRCASW), University of Delhi, New Delhi, India. Tel: +91-1122623503, Fax: +91-1122623504, E-mail: safdar_mgl@live.inalisafd@gmail.com

Received: December 18, 2014**; Revised:** January 20, 2015**; Accepted:** January 26, 2015

Background: Microsatellites have evoked the interest of researchers owing to their applications in different fields such as DNA fingerprinting, genetic mapping, population genetics, forensics, paternity studies and evolution.

Objectives: The present study focused on the analysis of simple sequence repeats (SSRs) in genomes of seven species from three genera of the *Filoviridae* family.

Materials and Methods: Genome sequences of seven species from the *Filoviridae* family were assessed by the National Center for Biotechnology Information (NCBI), microsatellites were extracted using the IMEx software, and statistical analysis was performed Microsoft Office Excel 2007.

Results: A total of 516 microsatellites and 14 Compound Simple Sequence Repeats (cSSR) (also known as compound microsatellites) were extracted. Evidently, the conversion of SSRs to cSSR was low. Mononucleotide A/T was the most prevalent followed by dinucleotide AC/CA and trinucleotide AAC/CAA. Highest incidence of SSRs (mon-/di-nucleotide motif) was observed in RNA Dependent RNA Polymerase (RDRP) gene whereas tri-nucleotide motif was maximally localized in nucleoproteins (NP).

Conclusions: The salient features of simple and compound microsatellites in *Filoviridae* family have been highlighted herein. Microsatellite regions with higher mutation rates compared to the rest of the genome play a crucial role in genome evolution by acting as a source of quantitative genetic variation. The SSR mutation rate is known to be affected by motif length, motif sequence, and number of repeats and purity of repetition. The functional role of tandem repeats in viruses, remains to be fully elucidated. However, with the repetitive sequence allegedly acting as a hot spot for recombination, we postulate their involvement in genetic events such as recombination, replication, and repair mechanisms that drive sequence diversity leading to the formation of the genetic basis of adaptation.

Keywords: Max Protein, Drosophila; Microsatellite Repeats; Ebolavirus

1. Background

Microsatellites have evoked the interest of researchers owing to their applications in different fields such as DNA fingerprinting, genetic mapping, population genetics, forensics, paternity studies and evolution, to name a few (1, 2). However, the exploration of plausible functions and mutational mechanisms of microsatellites in genomes is at best in its nascent stage. Concerted efforts are underway to identify the presence, distribution and variations of Simple Sequence Repeats (SSRs) in RNA and DNA viruses.

Simple Sequence Repeats, also called microsatellites are tandem repetitions of one to six base-pair relatively short motifs of DNA, whereas minisatellites consist of a short series of nucleobases (10 - 60 base pairs); thus minisatellites are longer in length than microsatellites. The presence of SSRs in genomes of animal viruses such as Hepatitis C virus (HCV) (3) and Human Cytomegalovirus (HCMV) (4) has confirmed their existence beyond prokaryotes and eukaryotes (5, 6). Their repeat number, length, and motif size influence microsatellite mutability. For instance, the more the number of repeats, the higher the mutability (7). Moreover, variations in copy number due to strand slippage and unequal recombination highlight the instability of the microsatellites (5); which in turn makes them a predominant source of genetic diversity and a crucial player in viral genome evolution (8, 9). Variable length of microsatellites may affect local DNA structure or the encoded proteins (6) and hence influence the expression profile of the corresponding genes. Their role in gene regulation, transcription and protein function has been elucidated in a few cases (9, 10). Genome features such as size and GC content influence the incidence and polymorphic nature of microsatellites (11-13). However, owing to the absence of a universal correlation per se, a single priority rule cannot be forged for predicting their occurrence and density.

Based on interruptions of microsatellites, they are classified as interrupted, pure, compound, interrupted compound, complex and interrupted complex (14). The present study also focused on compound microsatellites, which are composed of two or more microsatellites adjacent to each other. Their presence has been reported in diverse taxa across viruses, prokaryotes and eukaryotes (15-17). Interestingly, microsatellites are more abundant in non-coding regions than in coding regions in eukaryotes (5, 18), and some prokaryotes (16, 19). The presence of minisatellites in coding regions of different species has formed the basis of excavation and exploration of genes from novel genomes; therein acting as probes for cDNA preparations, thus ruling out the screening of cDNA library, giving fast insight into hitherto unexplored genomes and transcriptomes. It has been reported that ~3% of the human genome is composed of microsatellites (20), whereas the conversion of the SSR to compound microsatellites in eukaryotic genomes like *Homo sapiens*, *Macaca mulatta*, *Mus musculus* and *Rattus norvegicus* has been 4 - 25% (17).

In the present study we analyzed repeat sequences in the *Filoviridae* family. The *Filoviridae* is one of the families of order *Mononegavirales*. It comprises of three Genera, which constitute seven species. Genera Ebolavirus comprises of five species whereas the other two genera have one species each according to the Ninth Report of the International Committee on the Taxonomy of Viruses (ICTV) (21). The negative-strand RNA viruses of *Filoviridae* family are feared for their ability to cause hemorrhage and death in infected individuals. The fatality rate can range from 30% to 90% (22). Ebola Virus (EBOV), Sudan Virus (SUDV) and Bundibugyo Virus (BDBV) are the three clinically significant species that cause Ebola Virus Disease (EVD). These viruses are typically found in sub-Saharan Africa, and cause sporadic outbreaks that can be brought under control with appropriate rapid public health responses. However, the containment and management of the recent emergence of EBOV in West Africa in early 2014 has been challenging (23).

The EBOV genome is a single-stranded RNA approximately 19000 nucleotides long that encodes for seven structural proteins; nucleoproteins (NP), polymerase cofactors VP35 and VP40, glycoproteins (GP), transcription activators VP30 and VP24, and RNA Dependent RNA Polymerase (RDRP). The EBOV surface glycoprotein (GP1, 2) plays important roles in virus infection and pathogenesis, and its expression is tightly regulated by an RNA editing mechanism during virus replication. An understanding of how GP1, 2 expression affects virus production and infectivity may enable us to identify targets in the virus life cycle for vaccines and antivirals.

2. Objectives

Here, we systematically analyzed the *Filoviridae* family genomes for microsatellites, with an attempt to put forth a virus genome model for understanding functional aspects, evolutionary relationships, and adaptation to divergent hosts.

3. Materials and Methods

3.1. Genome Sequences

Complete genome sequences of seven species from the *Filoviridae* family were assessed by the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) and used for identification and analysis of simple and compound microsatellites. Genome size of the studied species ranged from 18796 nt (Acc No-FJ621585) to 19111nt (Acc No-DQ217792). The accession numbers and salient features of *Filoviridae* family genomes have been summarized in Table 1.

3.2. Microsatellite Identification and Investigation

The microsatellite search was performed using the IMEx software (24). Earlier studies on eukaryotes and *E. coli* genomes have focused on microsatellites with lengths of 12 bp or more (5), yet due to the smaller size of the *Filoviridae* family genome, simple and compound microsatellite search using these parameters did not yield any results. Therefore, the search for simple and compound microsatellites was accomplished with the 'Advance-Mode' of IMEx using the parameters used for HIV (15), potexvirus, carlavirus, tobamovirus and potyvirus (25-28). The parameters were set as follows: type of repeat: perfect; repeat size: all; minimum repeat number: 6, 3, 3, 3, 3, 3; and maximum distance allowed between any two SSRs (dMAX) was 10. Other parameters were used as default. Compound microsatellites were not standardized in order to determine real composition.

Table 1. Salient Features of the Studied Genomes and the Observed Microsatellites [a]

S. No	Name	Acc No	Genome Size (bp)	GC%	SSR	RA	RD	cSSR	cRA	cRD	cSSR%
E1	*Bundibugyo ebolavirus*	KC545395	18939	41.84	74	3.91	25.77	2	0.11	1.53	2.70
E2	*Reston ebolavirus*	FJ621585	18796	40.67	75	3.99	26.71	1	0.05	0.90	1.33
E3	*Sudan ebolavirus*	FJ968794	18875	41.27	64	3.39	22.41	2	0.11	1.64	3.13
E4	*Tai Forest ebolavirus*	FJ217162	18935	42.25	79	4.17	27.25	2	0.11	2.17	2.53
E5	*Zaire ebolavirus*	AF499101	18960	41.11	78	4.11	26.69	4	0.21	3.43	5.13
E6	*Lloviu cuevavirus*	JF828358	18927	45.99	80	4.23	28.69	1	0.05	1.32	1.25
E7	*Marburg marburgvirus*	DQ217792	19111	38.29	66	3.45	21.45	2	0.10	1.94	3.03

[a] Abbreviations: Acc No, accession number; bp, base pair; cRA, compound simple sequence repeats relative abundance; cRD, compound simple sequence repeats relative density; cSSR, compound simple sequence repeats; GC%, guanine cytosine %; RA, relative abundance; RD, relative density; SSR, simple sequence repeats.

3.3. Statistical Analysis

We used the Microsoft Office Excel 2007 to perform all statistical analysis. Linear regression was used to reveal the correlation between the relative abundance, and relative density of microsatellites with genome size.

4. Results

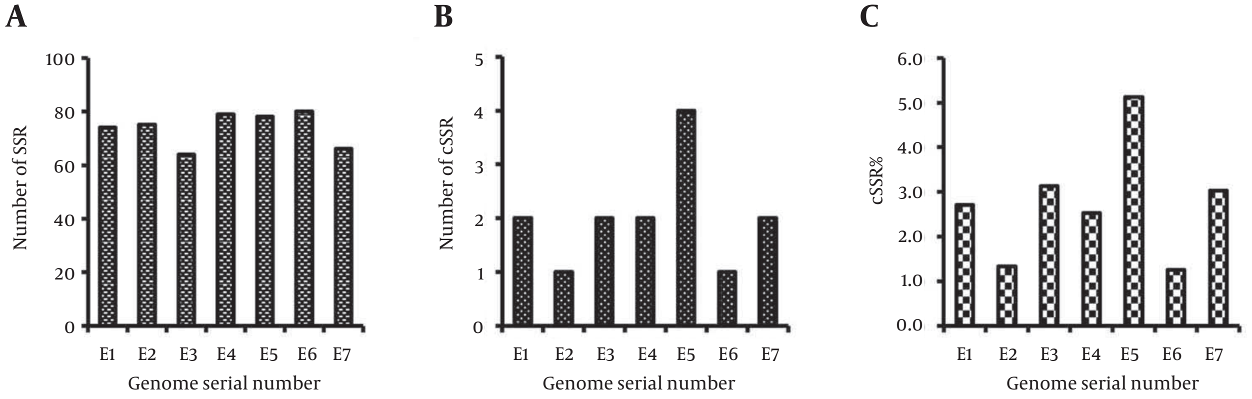
4.1. Presence of Simple Sequence Repeats, Compound Simple Sequence Repeats and Compound Simple Sequence Repeats Percentage in Analyzed Genomes

Genome-wide scan of *Filoviridae* genomes revealed that SSRs are present across all the species with the least being 64 in E3 (Acc No-FJ968794) and the most being 80 in E6 (Acc No-JF828358) (Table 1, Suppl 1 and Figure 1 A). A total of 516 SSRs were observed in the seven *Filoviridae* genomes (Table 1) whereas 14 Compound Simple Sequence Repeats (cSSR) (also known as compound microsatellites) were also present across all the analyzed genomes with the least being one in E2 (Acc No-FJ621585) and E6 (Acc No-JF828358) and the most being four in E5 (Acc No-AF499101) (Table 1, Suppl 1 and Figure 1 B). The percentage of individual microsatellites being part of a compound microsatellite (cSSRs%) (Figure 1 C) was 1.25% in E6 (Acc No-JF828358) with 80 microsatellites, and 5.13% in E5 (Acc No-AF499101) with 78 microsatellites.

4.2. Relative Abundance and Relative Density of Simple Sequence Repeats and Compound Simple Sequence Repeats

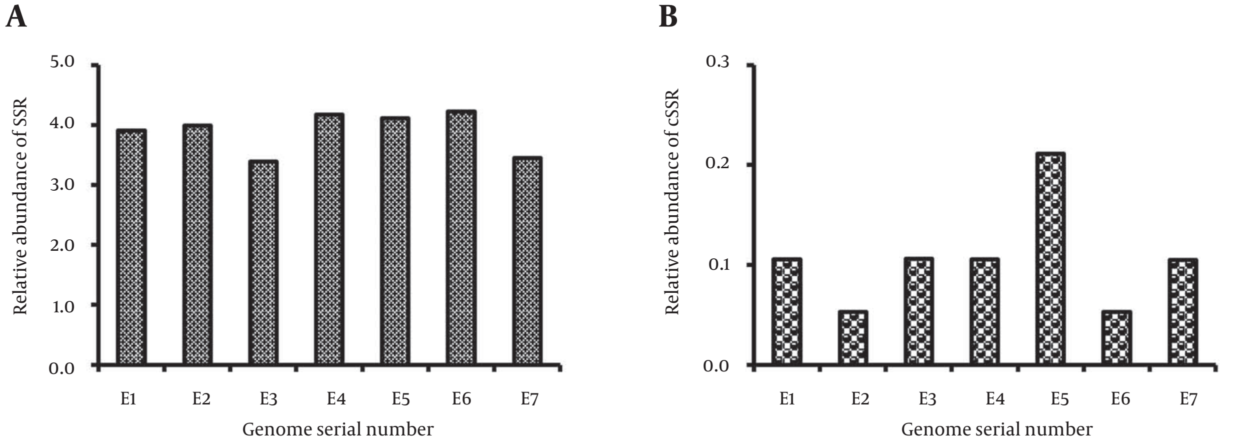
The relative abundance of SSR was highly variant ranging from 3.39 bp/kb in E3 to 4.23bp/kb for E6 (Table 1, Figure 2 A). Similarly, relative abundance for cSSR varied from a minimum of 0.05, for E2 and E6, to a maximum of 0.21 bp/kb for E5. (Table 1, Figure 2 B). However, relative density of SSR varied from 21.45 in E7 to 28.69 in E6 (Table 1, Figure 3 A), and similarly relative density of cSSR ranged from 0.9 in E2 to 3.43 in E5 (Table 1, Figure 3 B).

Figure 1. Analysis of Simple Sequence Repeats



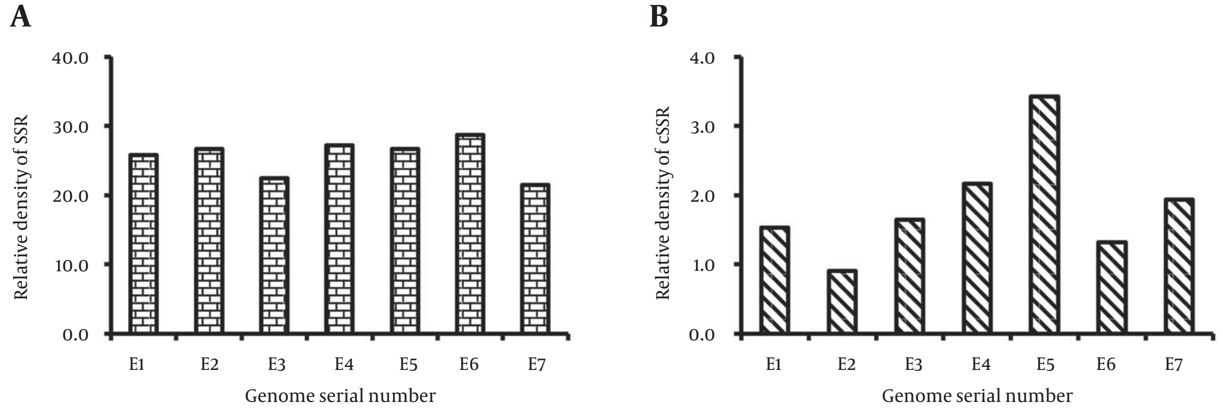
A, Distribution of SSRs; B, distribution of cSSRs; C, cSSR% across the *Filoviridae* family.

Figure 2. Relative Abundance: Simple Sequence Repeats / Compound Simple Sequence Repeats Present per Kilo Base of Genome



A, SSR; B, cSSR across *Filoviridae* family.

Figure 3. Relative density: Total length covered by Simple Sequence Repeats Per Kilo Base of Genome



A, SSR; B, cSSR across *Filoviridae* family.

4.3. The Effect of Maximum Distance Allowed Between Any Two SSRs on the Occurrence of Compound Microsatellites

To determine the impact of dMAX, incidence of compound microsatellites of the seven genomes with increasing dMAX was studied by analyzing cSSRs percentage (13). It is important to mention that the dMAX value can only be set between 0 and 50 for IMEx (24). Our analysis revealed an overall increase in cSSRs percentage with higher dMAX in all the seven *Filoviridae* genomes (Figure 4).

4.4. Sequence Repeats and Compound Microsatellite Distribution

We tested for the correlation between genome size/GC content and number/relative abundance/relative density of SSRs and cSSRs. Incidence of SSRs was non-significantly correlated (R2 = 0.06, P > 0.05) with genome size and GC content (R2 = 0.44, P > 0.05). Similarly, relative abundance (R2 = 0.08, P > 0.05) and relative density (R2 = 0.2, P > 0.05) were non-significantly correlated with genome size and GC content, R2 = 0.45, P > 0.05; and R2 = 0.62, P > 0.05 respectively. The regression analysis of cSSR for cSSR percentage (R2 = 0.15, P > 0.05), relative abundance (R2 = 0.11, P > 0.05) and relative density (R2 = 0.18, P > 0.05) showed a non-significant correlation with genome size. Similarly, GC content was also not significantly correlated with cSSR percentage (R2 = 0.21, P > 0.05), relative abundance (R2 = 0.12, P > 0.05) and relative density (R2 = 0.06, P > 0.05).

4.5. Related Parameters Influencing Single Motif Types in Analyzed Genomes

Mononucleotide repeats were observed in all the analyzed *Filoviridae* genomes. Poly A/T repeats were significantly more prevalent (89%) than poly G/C repeats in every *Filoviridae* genome (Suppl 1), which can be attributed to the high A/T content of the genomes. However, the A/T content being only slightly higher than G/C content in each of the analyzed sequences suggests a weak influence on the occurrence of poly G/C repeats. Similar to earlier studies of eukaryotic and prokaryotic genomes with more abundant poly A/T tracts; in the analyzed sequences poly A or poly T mononucleotide repeats prevailed over poly G or poly C repeats (Figure 5).

The present study revealed di-nucleotide repeats of five types: AG/GA, GT/TG, AC/CA, CT/TC, and AT/TA, which exhibited variations in occurrence across different *Filoviridae* genomes. The AC/CA repeats were the most common motifs followed by AT/TA, whereas, CG/GC was the least represented. The distribution results of di-nucleotide repeats in the studied genomes have been summarized in Suppl 1 (Figure 5).

Tri-nucleotide repeats were the third most abundant SSRs within the *Filoviridae* genomes. Of the 64 triplet repeat types, the density of AAC/CAA coding for asparagine/glutamine was the most abundant (Suppl 1) followed by ACC/CCA coding for threonine/proline, respectively (Figure 5). Differences exist in abundance of trinucleotide repeats among different *Filoviridae* genomes species yet AAC/CAA showed the highest prevalence in most species. Tetra-nucleotide repeat motifs CTTC (E1), GTTT (E2), TAAC (E2), ACCA (E3), CCAA (E4), TCGA (E5) and TTCT (E6) were present in *Filoviridae* genomes whereas pentanucleotide and hexanucleotide motifs were absent.

4.6. Single Sequence Repeats in Coding Regions

The distribution of SSRs in coding/non-coding regions revealed that 31.4% of SSRs were located in the RDRP protein followed by 11.24% in NP. Furthermore, mono and dinucleotide repeats were most prevalent in RDRP proteins whereas tri-nucleotide repeats had the highest incidence in nucleoproteins (NP) (Figure 6). Similarly distribution of cSSRs in coding/non-coding regions revealed that 21% were located in RDRP followed by 14.3% in GP, VP30 and VP40 while a maximum of 37% of cSSR were present in non-coding regions (Figure 7).

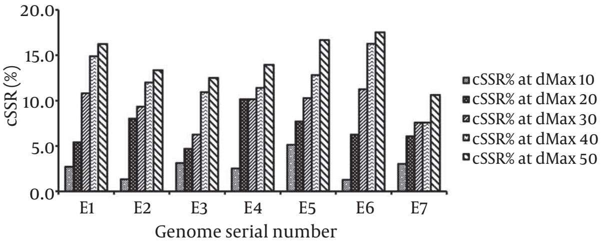


Figure 4. Analysis of Percentage of Individual Microsatellites Being Part of a Compound Microsatellite Across the *Filoviridae* Family With Varying dMAX From 10 to 50

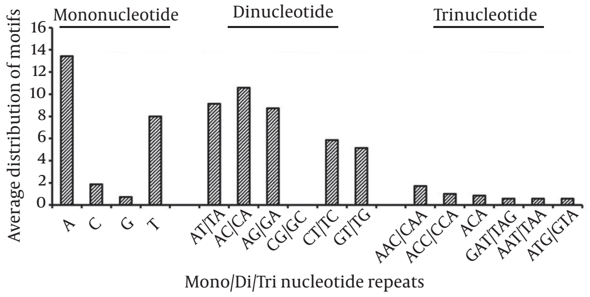


Figure 5. Average Distribution of Mono-di-Tri-Nucleotide Repeat Motifs Across the *Filoviridae* Family

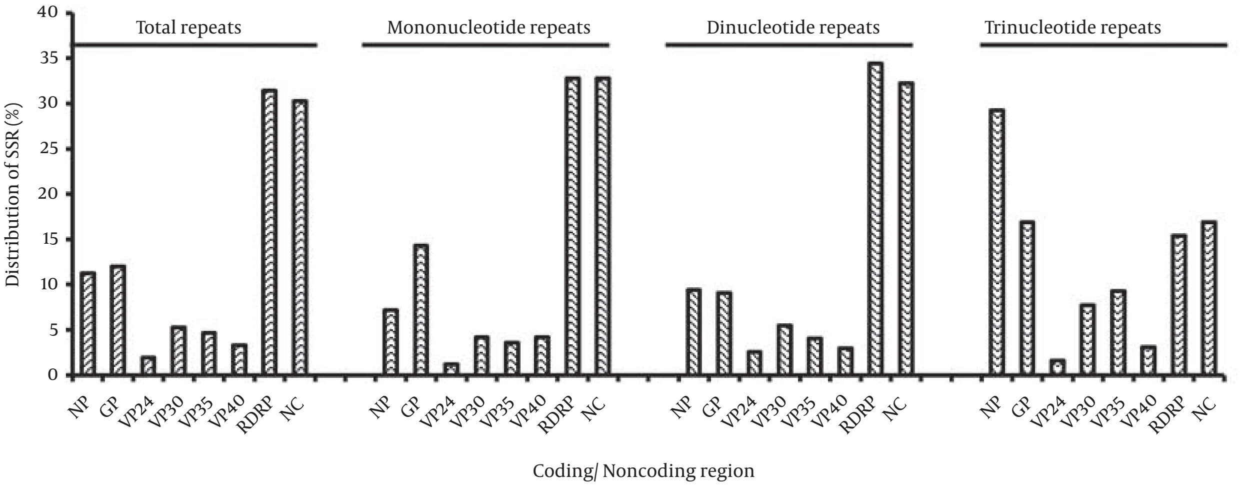


Figure 6. Differential Distribution of Total Single Sequence Repeat Motifs and Individual Mono-, di- and tri-Nucleotide Single Sequence Repeat Motifs (%) Across Coding/Non-Coding Regions of *Filoviridae* Family

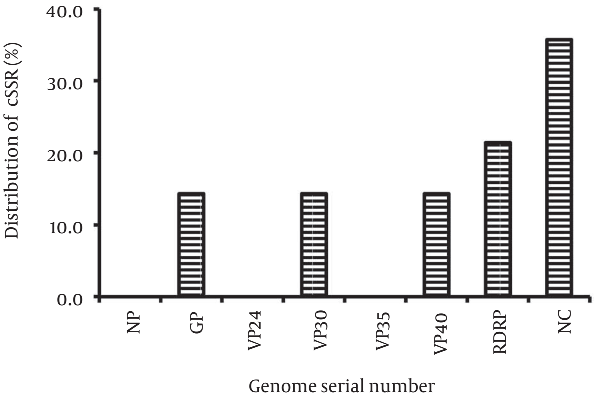


Figure 7. Differential Distribution of Compound Single Sequence Repeat Motifs (%) Across Coding/Non-Coding Regions of the *Filoviridae* Family

5. Discussion

In this study, we screened seven *Filoviridae* family genomes for the presence, abundance and composition of SSR tracts. The incidence of SSRs (mononucleotide to hexanucleotide repeats) was proportional to the genome size of the *Filoviridae* family with 64 - 80 SSRs per genome as compared to potyviruses (23 - 45 SSRs) (29) or Human immunodeficiency virus isolates (22 - 48 SSRs) (30) yet higher than geminivirus (4-19) with a smaller genome. Though relative density tends to be positively correlated with genome size in some fungal and other genomes (31-33) yet for *Filoviridae* family species both relative density and relative abundance were non-significantly correlated with genome size and GC content.

The sequence composition of repeats determines the abundance of microsatellites. In the *Filoviridae* family, AC/CA repeats predominated whereas GC/CG repeats were rare. Furthermore, CG/GC repeats were also rare in geminivirus, human, Drosophila, Arabidopsis thaliana, *Caenorhabditis elegans*, yeast (3), fungi (31, 34) and some eukaryotes (35). Di-nucleotide repeats are more prevalent than trinucleotide repeats due to instability of dinucleotide repeats because of higher slippage rate (35). The repeat sequences may provide a molecular device for faster adaptation to environmental stresses (9, 19, 36); thus may accelerate the evolution of the *Filoviridae* family.

Notably, no significant correlation was observed between genome size and two of the microsatellite features (relative density and relative abundance), concurrent with *E. Coli*/HIV-1. The analysis of cSSRs revealed some interesting results. These compound microsatellites are reportedly involved in regulation of gene expression and at functional level of proteins in several species (3). Though their significance in the *Filoviridae* family is not clear, our results suggest the presence of a possibly complex regulation at the functional level. Further, the analysis of dMAX (10 to 50) showed that cSSR percentage in the five analyzed species of *Filoviridae* family increased with increase in dMAX, though not in a linear fashion. Approximately, 97% of the extracted cSSRs constituted of two motifs only. The largest compound microsatellite in the *Filoviridae* family was composed of three SSRs whereas, in prokaryotes the largest microsatellite has four and in eukaryotes three SSRs. In general, the cSSR incidence decreases with increase in complexity. Interestingly, cSSRs percentage varied between 1.25 - 5.13% in the *Filoviridae* family genome; 0 - 15.15% in potyvirus genome, (28) 0 - 24.24% in HIV-1 genomes, 4 - 25% in eight eukaryotic genomes (17), and 1.75 - 2.85% in *E. coli* genomes (37). The distribution of microsatellite in the viral genome is organism specific rather than host specific. This is supported by the fact that the taxonomy of *Filoviridae* family shows no comparable congruence with host taxonomy, and species from the same lineage may have quite unrelated hosts (38). Interestingly, each *Filoviridae* family species possesses at least one cSSR, which might be causing their variation and evolution.

Microsatellite regions with higher mutation rates as compared to the rest of the genome (16, 39) play a crucial role in genome evolution by acting as a source of quantitative genetic variation (40). The SSR mutation rate is known to be affected by motif length, motif sequence, number of repeats and purity of repetition (41). Single base substitution can stabilize pure microsatellites by reducing the purity of repetition. The functional role of tandem repeats in viruses, remains to be fully elucidated. However, with the repetitive sequence allegedly acting as a hot spot for recombination (42), we postulate their involvement in genetic events such as recombination, replication, and repair mechanisms that drive sequence diversity leading to formation of the genetic basis of adaptation. The microsatellites in *Filoviridae* family genomes may serve as one of the tools for better understanding of viral genetic diversity and its implications.

Acknowledgements

We thank the Department of Biomedical Sciences, Shaheed Rajguru College of Applied Sciences for Women, University of Delhi, Delhi 110096 and Department of Botany, Patna University, Bihar 800005, for all their financial and infrastructural support.

Authors' Contributions

Chaudhary Mashhood Alam: carried out the work. Chaudhary Sharfuddin: planned the work and analyzed the data. Safdar Ali: planned the work and wrote the manuscript.

References

1. Goldstein D, Schlotterer C. *Microsatellites:Evolution and applications.*Oxford, UK: Oxford University Press; 1999.
2. Queller DC, Strassmann JE, Hughes CR. Microsatellites and kinship. *Trends Ecol Evol.* 1993;**8**(8):285–8.
3. Chen M, Zeng G, Tan Z, Jiang M, Zhang J, Zhang C, et al. Compound microsatellites in complete Escherichia coli genomes. *FEBS Lett.* 2011;**585**(7):1072–6.
4. Picone O, Ville Y, Costa JM, Rouzioux C, Leruez-Ville M. Human cytomegalovirus (HCMV) short tandem repeats analysis in congenital infection. *J Clin Virol.* 2005;**32**(3):254–6.
5. Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* 2000;**10**(7):967–81.
6. Mrazek J, Guo X, Shah A. Simple sequence repeats in prokaryotic genomes. *Proc Natl Acad Sci U S A.* 2007;**104**(20):8472–7.
7. Pearson CE, Nichol Edamura K, Cleary JD. Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet.* 2005;**6**(10):729–42.
8. Deback C, Boutolleau D, Depienne C, Luyt CE, Bonnafous P, Gautheret-Dejean A, et al. Utilization of microsatellite polymorphism for differentiating herpes simplex virus type 1 strains. *J Clin Microbiol.* 2009;**47**(3):533–40.
9. Kashi Y, King DG. Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.* 2006;**22**(5):253–9.
10. Usdin K. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. *Genome Res.* 2008;**18**(7):1011–9.
11. Coenye T, Vandamme P. Characterization of mononucleotide repeats in sequenced prokaryotic genomes. *DNA Res.* 2005;**12**(4):221–33.
12. Dieringer D, Schlotterer C. Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Res.* 2003;**13**(10):2242–51.
13. Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res.* 2008;**18**(1):30–8.
14. Chambers GK, MacAvoy ES. Microsatellites: consensus and controversy. *Comp Biochem Physiol B Biochem Mol Biol.* 2000;**126**(4):455–76.
15. Chen M, Tan Z, Zeng G, Zeng Z. Differential distribution of compound microsatellites in various Human Immunodeficiency Virus Type 1 complete genomes. *Infect Genet Evol.* 2012;**12**(7):1452–7.
16. Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM, Kashi Y. Simple sequence repeats in Escherichia coli: abundance, distribution, composition, and polymorphism. *Genome Res.* 2000;**10**(1):62–71.
17. Kofler R, Schlotterer C, Luschutzky E, Lelley T. Survey of microsatellite clustering in eight fully sequenced species sheds light on the origin of compound microsatellites. *BMC Genomics.* 2008;**9**:612.
18. Metzgar D, Bytof J, Wills C. Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res.* 2000;**10**(1):72–80.
19. Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol.* 2004;**21**(6):991–1007.
20. Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grutzner F, et al. Genome analysis of the platypus reveals unique signatures of evolution. *Nature.* 2008;**453**(7192):175–83.
21. King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ. *Virus Taxonomy: Classification and Nomenclature of Viruses : Ninth Report of the International Committee on Taxonomy of Viruses.*: Elsevier; 2011.
22. Sanchez A, Geisbert T, Feldmann H. *Filoviridae: Marburg and Ebola Viruses.* 5th edPhiladelphia,: Wolters kluwer/ Lippincott Williams & Wilkins; 2007.
23. Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, et al. Emergence of Zaire Ebola Virus Disease in Guinea. *N Engl J Med.* 2014;**371**(15):1418–25.
24. Mudunuri SB, Nagarajaram HA. IMEx: Imperfect Microsatellite Extractor. *Bioinformatics.* 2007;**23**(10):1181–7.
25. Alam CM, Singh AK, Sharfuddin C, Ali S. Incidence, complexity

and diversity of simple sequence repeats across potexvirus genomes. *Gene.* 2014;**537**(2):189–96.

26. Alam CM, Singh AK, Sharfuddin C, Ali S. Genome-wide scan for analysis of simple and imperfect microsatellites in diverse carlaviruses. *Infect Genet Evol.* 2014;**21**:287–94.
27. Alam CM, Singh AK, Sharfuddin C, Ali S. In-silico analysis of simple and imperfect microsatellites in diverse tobamovirus genomes. *Gene.* 2013;**530**(2):193–200.
28. Alam Ch M, George B, Sharfuddin C, Jain SK, Chakraborty S. Occurrence and analysis of imperfect microsatellites in diverse potyvirus genomes. *Gene.* 2013;**521**(2):238–44.
29. Zhao X, Tan Z, Feng H, Yang R, Li M, Jiang J, et al. Microsatellites in different Potyvirus genomes: survey and analysis. *Gene.* 2011;**488**(1-2):52–6.
30. Chen M, Tan Z, Jiang J, Li M, Chen H, Shen G, et al. Similar distribution of simple sequence repeats in diverse completed Human Immunodeficiency Virus Type 1 genomes. *FEBS Lett.* 2009;**583**(17):2959–63.
31. Karaoglu H, Lee CM, Meyer W. Survey of simple sequence repeats in completed fungal genomes. *Mol Biol Evol.* 2005;**22**(3):639–49.
32. Morgante M, Hanafey M, Powell W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet.* 2002;**30**(2):194–200.
33. Hancock JM. Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. *Genetica.* 2002;**115**(1):93–103.
34. Kim TS, Booth JG, Gauch HJ, Sun Q, Park J, Lee YH, et al. Simple sequence repeats in Neurospora crassa: distribution, polymorphism and evolutionary inference. *BMC Genomics.* 2008;**9**:31.
35. Hong CP, Piao ZY, Kang TW, Batley J, Yang TJ, Hur YK, et al. Genomic distribution of simple sequence repeats in Brassica rapa. *Mol Cells.* 2007;**23**(3):349–56.
36. Kashi Y, King D, Soller M. Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet.* 1997;**13**(2):74–8.
37. Kruglyak S, Durrett R, Schug MD, Aquadro CF. Distribution and abundance of microsatellites in the yeast genome can Be explained by a balance between slippage events and point mutations. *Mol Biol Evol.* 2000;**17**(8):1210–9.
38. Gibbs AJ, Ohshima K, Phillips MJ, Gibbs MJ. The prehistory of potyviruses: their initial radiation was during the dawn of agriculture. *PLoS One.* 2008;**3**(6).
39. Xu X, Peng M, Fang Z. The direction of microsatellite mutations is dependent upon allele length. *Nat Genet.* 2000;**24**(4):396–9.
40. Tautz D, Trick M, Dover GA. Cryptic simplicity in DNA is a major source of genetic variation. *Nature.* 1986;**322**(6080):652–6.
41. Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet.* 2004;**5**(6):435–45.
42. Jeffreys AJ, Murray J, Neumann R. High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Mol Cell.* 1998;**2**(2):267–73.