Journal of Data Mining in Genomics & Proteomics

Alam et al., J Data Mining Genomics Proteomics
2016, 7:1
http://dx.doi.org/10.4172/2153-0602.1000187

Research Article Open Access

Imex Based Analysis of Repeat Sequences in *Flavivirus* Genomes, Including Dengue Virus

Chaudhary Mashhhood Alam[1,2], Asif Iqbal[1], Babita Thadari[2] and Safdar Ali[2*]

[1]*PIRO Technologies Private Limited, New Delhi-110025, India*

[2]*Department of Biomedical Sciences, SRCASW, University of Delhi, Vasundhara Enclave, New Delhi-110096, India*

[*]**Corresponding author:** Safdar Ali, Assistant Professor, Department of Biomedical Sciences, SRCASW, University of Delhi, New Delhi – 110096, India, Tel: 91-11-22623503; Fax: 91-11-22623504; E-mail: safdar_mgl@live.in; alisafd@gmail.com

Received date: January 21, 2016; **Accepted date:** January 28, 2016; **Published date:** January 31, 2016

Abstract

Simple sequence repeats (SSRs), also known as microsatellites, are 1-6 nucleotides repeat motif, present in varying number of iterations, across coding and non-coding regions of prokaryotes, eukaryotes and viruses. Present study focuses on simple sequence repeats (SSRs) in 27 *Flavivirus* genomes, which includes dengue virus. The comparative viral genomics in the light of SSRs would help us understand the diversity and adaptability to new hosts. A total of 1164 SSRs and 53 cSSRs were uncovered from the 27studied genomes. Mononucleotide A was the most prevalent repeat motif with an average distribution of around 6. This was followed by G (average distribution of 2). Amongst the dinucleotides AG/GA repeat motif was the most prevalent with an average distribution of 14 across studied genomes. The Flavivirus genomes lacked two essential features responsible for genome evolution, dinucleotide repeat motif AT/TA (least represented with average distribution of ~0.5) and cSSR in non-coding regions, suggesting a stable genome or evolution by hitherto unexplained mechanisms. The unveiling of conserved sequences in the isolates of Dengue virus suggests a basis for biomarker development for viral diagnostics.

Keywords: *Flavivirus*; Simple sequence repeats; Imperfect microsatellite extraction IMEx; dMAX

Abbreviations

SSR: Simple Sequence Repeat; cSSR: Compound Simple Sequence Repeat; IMEx: Imperfect Microsatellite Extraction; RD: Relative Density; RA: Relative Abundance

Introduction

Viruses utilize almost all spectra of the living world for their survival, as in host for infection and survival. The classification and evolution of viruses have been based either on the genome features (size/type) or on their host range [1,2]. A single viral genome encodes from 2 to about a thousand proteins [3,4]. Though a complete understanding of the evolutionary mechanisms driving evolution of viruses is underway, however, transposable elements and tandemly repetitive sequences are believed to play a crucial role [5,6].

Simple sequence repeats (SSRs), also known as microsatellites, are 1-6 nucleotides repeat motif, present in varying number of iterations, across coding and non-coding regions of prokaryotes, eukaryotes and viruses [7-9]. SSRs, being recombination hot spots aid in genome evolution, sometimes being the basis of diseases [10,11]. Functionally, these sequences are reported to be associated with gene regulation, transcription and protein function [12,13]. The incidence of SSRs may be influenced by the genome features like size and GC content [14-16]. However, this correlation is not universal, adding to the enigma of SSRs.

Present study focuses on simple sequence repeats (SSRs) in 27 *Flavivirus* genomes, which includes Dengue virus. Dengue is a mosquito borne viral infection found in tropical and sub-tropical regions of the world and is caused by one of the four serotypes of dengue viruses (DENV1-DENV4). An increase in infection has been seen in recent years due to many factors including urbanization and air travel. Over 2.5 billion people of the world's population are now at risk for dengue. They may be asymptomatic or may give rise to undifferentiated fever, dengue fever, dengue haemorrhagic fever (DHF), or dengue shock syndrome.

Dengue virus infection has been counted among emerging and re-emerging diseases because of (1) the increasing number of patients, (2) the expansion of epidemic areas, and (3) severe clinical manifestation of dengue hemorrhagic fever (DHF)/dengue shock syndrome (DSS), which is often fatal if not properly treated In the meantime, there are no effective dengue control measures: a dengue vaccine is still under development and vector control does not provide a long-lasting effect. Early recognition and prompt initiation of appropriate treatment are vital if disease related morbidity and mortality are to be limited. Our study proposes a biomarker based on repeat sequences, which can be used as an effective mode for diagnosing different strains of Dengue virus.

Materials and Methods

Genome sequences

Complete genome sequences of 27 *Flavivirus* were assessed and downloaded in both GenBank and FASTA formats from NCBI and subsequently analyzed for simple and compound microsatellites. The *Flaviviruses* included in the study and their genome features have been summarized in Table 1. *Flaviviruses* have monopartite linear genome of about 10-11kb length.

Microsatellite extraction

The search for microsatellites was performed using Imperfect microsatellite extractor (IMEx) software. The analysis was done using the 'Advance- Mode' of IMEx with parameters as reported for analysis of HIV genomes; as in Type of Repeat: perfect; Repeat Size: all; Minimum Repeat Number: 6, 3, 3, 3, 3, 3; Maximum distance allowed between any two SSRs (dMAX) is 10 [7]. Two SSRs separated by a distance of less than or equal to 10bp would be thus treated as compound SSR (cSSR).

Statistical analysis

Microsoft Office Excel 2007 was used to perform regression analysis to predict correlation of Genome size and GC content on different parameters of SSR and cSSR such as incidence, relative abundance and relative density. Our sample size was 27 genomes, which we used in our analysis.

MATLAB based tools for SSR analysis

IMEx has been widely used to obtain the SSRs in a genome [17-22]. However, for subsequent analysis we developed two MATLAB based tools namely Identification of Gene Location from NCBI Nucleotide File (IGLNNF) and In-corporation of Gene Location in SSR File (IGLSF). IGLNNF was used to obtain the gene locations from Genbank directly but some manual help was needed for incorporation of gene position, because only starting and end point of polyprotein were mentioned in NCBI file for species of Flavivirus genomes whereas individual member of polyprotein were mentioned separately as misc_feature. It was further saved further into (.xlsx) format. IGLSF was used to incorporate the gene location in the SSRs file.

Results and Discussion

SSR/cSSR incidence

A total of 1164 SSRs and 53 cSSRs were uncovered from the 27studied genomes. Though Flaviviruses are known to have comparable genome sizes (10-11kb), the SSR incidence per genome is varying from 27 (F12) to 67 (F21) (Figure 1 and Table 1, Supplementary file 1). These variations cannot be attributed to genome size owing to small range and further highlighted by lack of co-linearity between genome size and SSR incidence. For instance, F11 with genome length of 10871bp has 65 SSRs as compared to F19 (10892bp) has just 37 SSRs (Table 1).

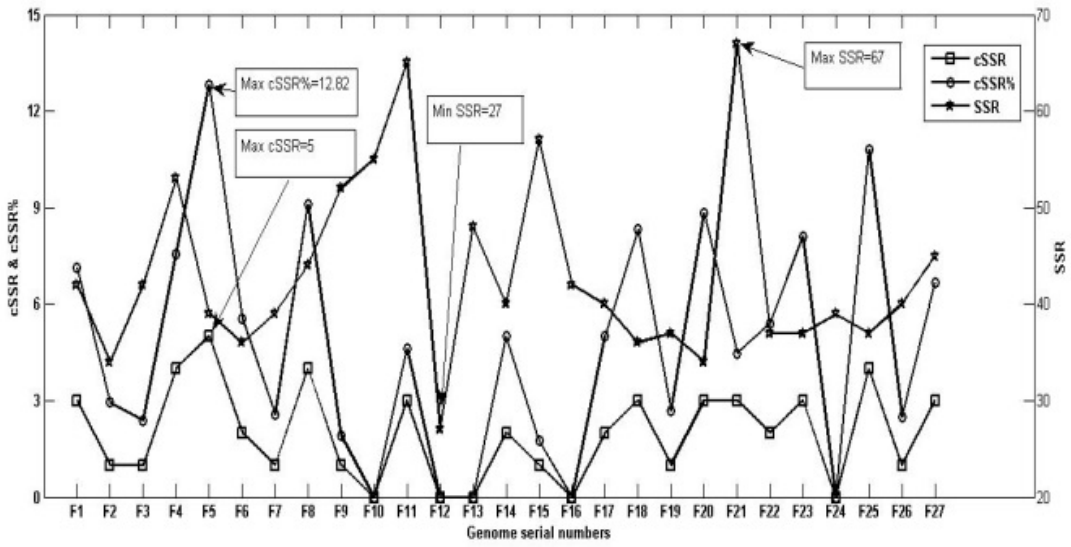


Figure 1: Incident frequency of SSRs, cSSR and cSSR%.

S.No	Genus: Flavivirus	Accession Number	GS*(bp)	GC**(%)	SSRa	cSSRa	RAb	RDc	cRAb	cRDc	cSSR%d
F1	*Apoi virus*	NC_003676.1	10116	48.3	42	3	4.15	23.03	0.30	4.05	7.14
F2	*Banzi virus*	DQ859056	10182	50.4	34	1	3.34	21.70	0.10	1.18	2.94
F3	*Bouboui virus*	DQ859057	10173	48.2	42	1	4.13	26.54	0.10	1.08	2.38
F4	*Dengue virus*	AF326827	10618	47	53	4	4.99	30.80	0.38	5.46	7.55
F5	*Edge Hill virus*	DQ859060	10206	47	39	5	3.82	24.99	0.49	8.72	12.82

F6	*Japanese encephalitis virus*	AF221499	10976	51.3	36	2	3.28	20.68	0.18	4.83	5.56
F7	*Jugra virus*	DQ859066	10173	48.2	39	1	3.83	24.38	0.10	1.57	2.56
F8	*Kedougou virus*	DQ859061	10227	53	44	4	4.30	29.53	0.39	5.87	9.09
F9	*Kyasanur Forest disease virus*	HM055369	10774	55.1	52	1	4.83	31.19	0.09	1.02	1.92
F10	*Langat virus*	AF253420	10943	54.3	55	0	5.03	32.99	0.00	0.00	0.00
F11	*Louping ill virus*	NC_001809	10871	54.8	65	3	5.98	37.72	0.28	4.14	4.62
F12	*Modoc virus*	AJ242984	10600	45.5	27	0	2.55	17.26	0.00	0.00	0.00
F13	*Montana myotis leukoencephalitis virus*	AJ299445	10690	44.1	48	0	4.49	27.60	0.00	0.00	0.00
F14	*Murray Valley encephalitis virus*	KF751871	10953	48.9	40	2	3.65	23.01	0.18	4.29	5.00
F15	*Powassan virus*	NC_003687	10839	53.3	57	1	5.26	34.41	0.09	1.57	1.75
F16	*Rio Bravo virus*	JQ582840	10742	43.4	42	0	3.91	24.30	0.00	0.00	0.00
F17	*Saboya virus*	DQ859062	10173	47.7	40	2	3.93	24.77	0.20	2.26	5.00
F18	*Sepik virus*	DQ859063	10218	47.2	36	3	3.52	24.08	0.29	4.60	8.33
F19	*St. Louis encephalitis virus*	KM267635.1	10892	49.78	37	1	3.40	23.60	0.09	1.01	2.70
F20	*Tembusu virus*	KR061333.1	10278	48.97	34	3	3.31	22.57	0.29	5.25	8.82
F21	*Tick-borne encephalitis virus*	NC_001672	11141	53.8	67	3	6.01	42.10	0.27	3.59	4.48
F22	*Uganda S virus*	DQ859065	10182	46.9	37	2	3.63	23.87	0.20	3.04	5.41
F23	*Usutu virus*	AY453411	11066	51	37	3	3.34	22.77	0.27	3.98	8.11
F24	*Wesselsbron virus*	JN226796	10814	47.7	39	0	3.61	23.58	0.00	0.00	0.00
F25	*West Nile virus*	NC_009942.1	11029	51.15	37	4	3.35	22.85	0.36	5.08	10.81
F26	*Yellow fever virus*	KM388815	10236	50.1	40	1	3.91	25.60	0.10	1.86	2.50
F27	*Zika virus*	DQ859059	10254	50.8	45	3	4.39	29.06	0.29	3.51	6.67

aNumber of simple/compound microsatellites; bRelative abundance: number of simple/compound microsatellites present per kb of the genome (kb); cRelative density is defined as the total length (bp) contributed by each simple/compound microsatellite per kb of sequence analyzed; dcSSRs-% is the percentage of individual microsatellites being part of a compound microsatellite; *GS(bp); Genome Size ; **GC(%) Guanine and Cytosine in percentage.

Table 1: Overview of simple and compound microsatellites in genus *Flavivirus* genome including dengue Virus.

Two SSRs with a distance of <dMAX between them are considered as compound SSR (cSSR). Analysis of cSSR gives an insight into the uniformity in distribution of SSRs across genomes, wherein, a co-linearity between number of SSRs present and its conversion to cSSR would suggest existence of cSSR in an unbiased manner. However, the cSSR incident frequency ranged from zero to five (F5). A total of five species namely F10, F12, F13, F16 and F24 exhibited no cSSR in their genomes. These species had 55, 27, 48, 42 and 39 SSRs respectively. This variance in SSR to cSSR conversion across genomes is represented as cSSR% which reaches a maximum of 12.82% in F5 with 39 SSRs (Figure 1 and Table 1, Supplementary file 1).

These attributes highlight two aspects about repetitive sequences. First, the distribution of SSRs is non-uniform across genomes from which we can construe their emergence and maintenance in genomes to be based on functional and regulatory implications. Secondly, the variation in cSSR% across genomes is an outcome of differential clustering of SSRs in a genome which is suggestive of SSRs divergent roles in different genomes.

Relative abundance and relative density of SSR and cSSR

Relative abundance (RA) is number of SSRs present per Kb of genome while relative density (RD) is total SSR sequence per Kb of genome. The RA of SSR ranged from 2.55 (F12) to 6.01 (F21) and for cSSR it ranged from 0 (F10, F12, F13, F16, F24) to 0.49 (F5) (Table 1, Figure 2 and 3). The RD of SSR ranged from 17.26 (F12) to 42.10 (F21) and for cSSR it exhibited a maximum of 8.72 (F5) (Table 1, Figure 2

and 3). The range for RA and RD across *Flavivirus* genomes may be considered as a representative of potential for genome evolution.

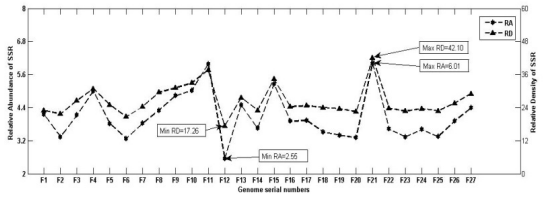


Figure 2: Relative abundance (Number of SSR per Kb of genome) and Relative density (Length occupied by SSR per Kb of genome) of SSRs.

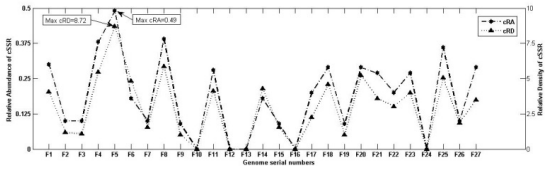


Figure 3: Relative abundance (Number of cSSR per Kb of genome) and Relative density (Length occupied by cSSR per Kb of genome) of cSSRs.

Correlation studies

We tested for correlation between genome size/GC content and number/relative abundance/relative density of SSRs and cSSRs. Incidence of SSRs is non-significantly correlated with genome size (R^2=0.15, P>0.05) and GC content (R^2=0.12, P>0.05). Similarly relative density (R^2 =0.09, P>0.05) and relative abundance (R^2=0.06, P>0.05) were non- significantly correlated with genome size and GC content respectively R^2=0.36, P>0.05; and R^2=0.27, P>0.05. The regression analysis of cSSR (R^2=0.02, P>0.05), relative density (R^2=0.02, P>0.05) and relative abundance (R^2=0.04, P>0.05) shows non-significant correlation with genome size. Similarly GC content is also not significantly correlated for cSSR (R^2=0.04, P>0.05), relative density (R^2=0.02, P>0.05) and relative abundance (R^2=0.03, P>0.05).

cSSR and dMAX

The uncovered cSSRs in present study were with a dMAX value of 10 as mentioned in section 2.2. However, IMEx has an option of varying the dMAX value between 0 and 50 [23]. So, in order to determine the impact of varying dMAX on cSSR incidence, five genomes F1, F7, F14, F21 and F27 were chosen at random and cSSR were extracted with increasing dMAX from 10 to 50. Expectedly, an increase in cSSRs% with higher dMAX values were observed as represented in Figure 4. However non-linearity in the increase further corroborates our initial suggestion of unequal distribution of SSRs, as in the distances between iterations differs across genomes. The ability of motifs to induce variations is often dependent on the proximity with other motifs and hence the differences therein would lead to different genome evolution potential. The repeat sequences induce variations by strand slippage and unequal recombination, chances of which are enhanced when different SSRs are in close proximity to one another [24].

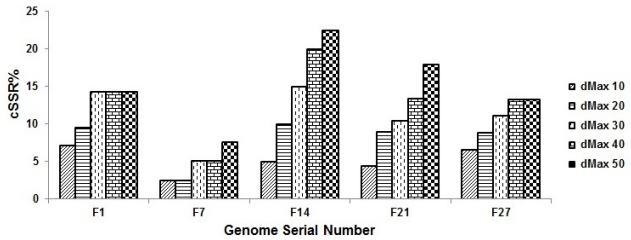
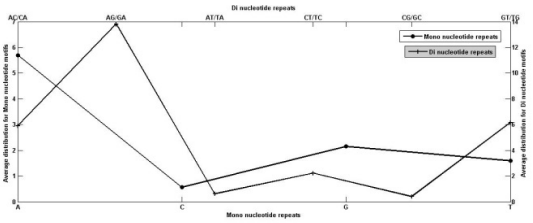


Figure 4: Frequency of cSSR-% (Percentage of individual microsatellites being part of a compound microsatellite) in relation to varying dMAX (10 to 50) across 5 randomly selected genomes.

Motif types in iterations

We further looked into the divergence of repeat motifs extracted from *Flavivirus* genomes. The SSRs repeat motif ranged from mono- to penta-nucleotides. With the GC content in the genomes lying close to 50%, a bias in the iterations was not expected. However, in mononucleotides, A was the most prevalent repeat motif with an average distribution of around 6. This was followed by G (average distribution of 2). The least represented mononucleotide repeat motif was C as represented in Figure 5A. Amongst the dinucleotides AG/GA repeat motif was the most prevalent with an average distribution of 14 across studied genomes (Figure 5A). AAG/GAA was the most represented trinucleotide repeat as illustrated in Figures 5B.



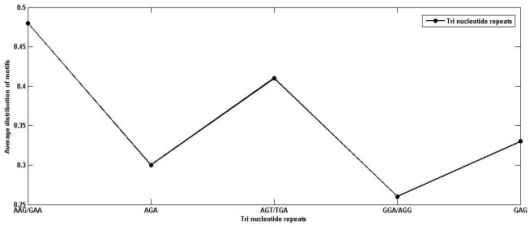


Figure 5: Average distribution of repeat motifs. 5A) Mono- or di-nucleotide repeat motifs. 5B) Tri-nucleotide repeat motifs.

Further, in terms of number of iterations present at a stretch, a maximum of 49 A repeat motif were observed in F21 followed by G repeat motif of 11(F15) and 10(F21) respectively. Whereas in di-nucleotide (AG/GA) repeat motifs maximum iteration were 5 in F10, F16 and F20. Tri-nucleotide maximum iteration repeats were found to be 4 in F1, F10, F11, F15, F21, F22 and F27 respectively.

Furthermore, the AT/TA dinucleotide motif were the least represented with average distribution of ~0.5. This motif is an established platform for SSR mutability and their low incidence is possibly suggestive of genome stability. Though repeats are known to be associated with copy number variations, strand slippage and polymorphisms accounting for genome evolution and adaptation; [6,25,26] their absence can lead to converse outcomes as well.

SSRs/cSSRs in coding regions

The distribution of SSRs across coding and non-coding regions of the genomes was accomplished by first extracting the locations of genes/proposed genes in the genomes in excel format using IGLNNF. A total of ~50 proteins were thus obtained. Subsequently, this data was simulated with the SSR data through IGLSF to get the distribution across coding and non-coding regions. For our analysis, we used 11 proteins present in most number of species (Figure 6). Coding regions accounting for over 80% of the total SSRs has been observed in earlier studies [17-22] across a diverse set of viruses suggestive of their role in gene expression, regulation and evolution. In the present study, interestingly, there was no cSSR present in the non-coding region (Figure 6). This further corroborates the idea that these repeat sequences have a role to play in gene regulation and expression when present in the coding regions. And when in non-coding regions they have a role to play in introducing variations leading to genome evolution. However, the present set of genomes have a low frequency of AT/TA repeat motifs as well as cSSRs are absent in non-coding regions suggestive of relatively stable genome of *Flaviviruses*.

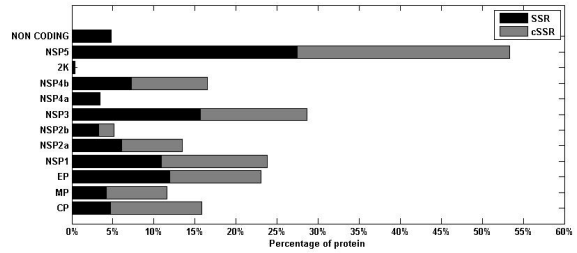


Figure 6: Differential distribution of SSRs (%) in coding/non-coding regions of *Flavivirus* genomes.

Conserved motifs in dengue virus

The weak platform for genome evolution across *Flavivirus* genomes and the clinical significance of Dengue virus edged us to explore the possibility of conserved regions across the different isolates of Dengue virus*es* which can be used as biomarkers for diagnostics. The details of sequences of Dengue virus*es* used in the study have been listed in Table 2. These sequences were retrieved from NCBI (http://www.ncbi.nlm.nih.-gov/) and analyzed for conserved cSSR motifs. A total of 7 motifs were subsequently analyzed and the results have been summarized in Table 3. There are 2 such motifs which were present in all the 32 studied sequences whereas another 2 motifs were represented in all but one isolate sequences. We postulate the possibility of these sequences as candidate biomarkers for a common diagnostics of different isolates of Dengue viruses.

S. No	Dengue virus isolate	Accession No
D1	Dengue virus vector p4(Delta30)	AY376438.1|
D2	Dengue virus type 4 recombinant clone 2Adel30	AF326826.1|
D3	Dengue virus type 4 vector p4	AY648301.1|
D4	Dengue virus type 4 recombinant clone rDEN4	AF326825.1|
D5	DENSTRA Dengue virus type 4	M14931.2|
D6	Dengue virus type 4 recombinant clone 2A	AF375822.1|
D7	Dengue virus type 4 strain 814669	AF326573.1|
D8	Dengue virus strain Dakar HD 34460	KF907503.1|
D9	Dengue virus 4 strain 341750	GU289913.1|

D10	Dengue virus 4 strain H402276	JN559740.2|
D11	Dengue virus 4 isolate DENV-4/US/BID-V2429/1994	GQ199878.1|
D12	Dengue virus 4 isolate DENV-4/US/BID-V2437/1996	GQ199883.1|
D13	Dengue virus 4 isolate DENV-4/US/BID-V2440/1996	FJ850058.1|
D14	Dengue virus 4 isolate DENV-4/US/BID-V860/1994	FJ226067.1|
D15	Dengue virus 4 isolate DENV-4/US/BID-V2438/1996	GQ199884.1|
D16	Dengue virus 4 isolate DENV-4/US/BID-V2435/1996	GQ199881.1|
D17	Dengue virus 4 isolate Haiti73	JF262782.1|
D18	Dengue virus 4 isolate DENV-4/US/BID-V2439/1996	GQ199885.1|
D19	Dengue virus 4 isolate DENV-4/US/BID-V1094/1998	EU854297.1|
D20	Dengue virus 4 isolate DENV-4/US/BID-V2436/1996	GQ199882.1|
D21	Dengue virus 4 isolate INH6412	JF262781.1|
D22	Dengue virus 4 isolate DENV-4/VE/BID-V2163/1998	FJ639736.1|
D23	Dengue virus 4 isolate DENV-4/US/BID-V2446/1999	FJ882599.1|
D24	Dengue virus 4 isolate DENV-4/US/BID-V2448/1999	FJ882601.1|
D25	Dengue virus 4 isolate DENV-4/VE/BID-V2172/1999	FJ639744.1|
D26	Dengue virus 4 isolate DENV-4/CO/BID-V1600/1997	FJ024476.1|
D27	Dengue virus 4 isolate DENV-4/VE/BID-V2607/2006	JN819406.1|
D28	Dengue virus 4 strain H780120	JQ513341.1|
D29	Dengue virus 4 strain H772854	JN559741.2|
D30	Dengue virus 4 isolate Br246RR/10	JN983813.1|
D31	Dengue virus 4 strain H779228	JQ513338.1|
D32	Dengue virus 4 strain H772846	JQ513330.1|

Table 2: Details of *Dengue virus* sequences used in the study.

S. No	Motif	Motif present/total sequences analyzed	Candidate Biomarker
1	$(AG)_3$-x_1-$(AG)_3$	31/32 (Absent in D32)	Yes
2	$(AG)_3$-x_3-$(A)_6$	32/32	Yes
3	$(AC)_3$-x_4-$(AG)_3$	32/32	Yes
4	$(GA)_3$-x_6-$(GA)_3$	31/32(Absent in D26)	Yes
5	$(TG)_3$-x_6-$(C)_6$	2/32	No
6	$(AG)_4$-x_2-$(TC)_3$	3/32	No
7	$(AG)_3$-x_9-$(A)_6$	20/32	Maybe

Table 3: Conserved iterations in *dengue virus*.

Recent studies have demonstrated that discrete steps in the replication cycles of these viruses can be inhibited by pharmacological agents that target host factors mediating lipid synthesis, metabolism, trafficking, and signal transduction. Lipids are necessary for every step in the replication cycle of *Hepatitis C virus* (HCV) and Dengue virus (DENV), members of the family *Flaviviridae*. Despite this, targeting host lipid metabolism and trafficking as an antiviral strategy by blockade of entire biosynthetic pathways may be limited due to host

toxicity therein highlighting the need for better diagnostics to counter the challenge of these viruses.

Conclusion

The comparative viral genomics in the light of SSRs would help us understand the diversity and adaptability to new hosts. The *Flavivirus* genomes lacked two essential features responsible for genome evolution, dinucleotide repeat motif AT/TA (least represented with average distribution of ~0.5) and cSSR in non-coding regions, suggesting a stable genome or evolution by hitherto unexplained mechanisms. The unveiling of conserved sequences in the isolates of Dengue virus suggests a basis for biomarker development for viral diagnostics.

Acknowledgement

We thank Department of Biomedical Sciences, Shaheed Rajguru College of Applied Sciences for Women, University of Delhi, Delhi-96, India and PIRO Technologies Private Limited, New Delhi-25, India for the financial and infrastructural support provided.

References

1. Gao L, Qi J (2007) Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. BMC Evol Biol 7: 41.
2. Iyer LM, Balaji S, Koonin EV, Aravind L (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res 117: 156–184.
3. Mrázek J, Karlin S (2007) Distinctive features of large complex virus genomes and proteomes. Proc Natl Acad Sci U S A 104: 5127–5132.
4. Van Etten JL, Lane LC, Dunigan DD (2010) DNA Viruses: The Really Big Ones (Giruses). Annu Rev Microbiol 64: 83–99.
5. Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42: 251–269.
6. Hancock JM (2002) Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. Genetica 115: 93–103.
7. Chen M, Tan Z, Zeng G, Zhuotong Z (2012) Differential distribution of compound microsatellites in various Human Immunodeficiency Virus Type 1 complete genomes. Infection Genetics and Evolution 12: 1452–1457.
8. Gur-Arie R, Cohen CJ, Eitan Y (2000) Simple sequence repeats in Escherichia coli: abundance, distribution, composition, and polymorphism. Genome Res 10: 62–71.
9. 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.
10. Jeffreys AJ, Holloway JK, Kauppi L, May CA, Neumann R et al (2004) Meiotic recombination hot spots and human DNA diversity. Philos Trans R Soc Lond B Biol Sci 359: 141–152.
11. Kovtun IV, McMurray CT (2008) Features of trinucleotide repeat instability in vivo. Cell Res 18: 198–213.
12. Kashi Y, King DG (2006) Simple sequence repeats as advantageous mutators in evolution. Trends Genet 22: 253–259.
13. Usdin K (2008) The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res 18: 1011–1019.
14. Coenye T, Vandamme P (2005) Characterization of mononucleotide repeats in sequenced prokaryotic genomes. DNA Res 12: 221–233.
15. Dieringer D, Schlotterer C (2003) Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res 13: 2242–2251.
16. Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res 18: 30–38.
17. Alam CM, Singh AK, Sharfuddin C, Ali S (2013) In silico analysis of simple and imperfect microsatellites in diverse tobamovirus genomes. Gene 530: 193-200.
18. Alam CM, Singh AK, Sharfuddin C, Ali S (2014a) Genome-wide scan for extraction and analysis of simple and imperfect microsatellites in diverse carlaviruses. Infection, Genetics and Evolution 21: 287-294.
19. Alam CM, Singh AK, Sharfuddin C, Ali S (2014b) Incidence, complexity and diversity of simple sequence repeats across potexvirus genomes. Gene 537: 189-196.
20. Alam CM, Singh AK, Sharfuddin C, Ali S (2014c) In- silico exploration of thirty alphavirus genomes for analysis of the simple sequence repeats. Meta Gene 2: 694-705.
21. Alam CM, Sharfuddin C, Ali S (2015) Analysis of simple and imperfect microsatellites in Ebolavirus species and other genomes of Filoviridae family. Gene Cell and Tissue 2: e26204.
22. Singh AK, Alam CM, Sharfuddin C, Ali S (2014) Frequency and distribution of simple and compound microsatellites in forty-eight Human Papillomavirus (HPV) genomes. Infection, Genetics and Evolution 24: 92-98.
23. Mudunuri SB, Nagarajaram HA (2007) IMEx: imperfect microsatellite extractor. Bioinformatics 23: 1181–1187.
24. Li Y, Korol AB, Fahima T, Nevo E (2004) Microsatellites Within Genes: Structure, Function, and Evolution. Mol Biol Evol 21: 991-1007.
25. Tóth G, Gáspári Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10: 967–981.
26. Deback C, Boutolleau D, Depienne C, Luyt CE, Bonnafous P et al. (2009) Utilization of microsatellite polymorphism for differentiating herpes simplex virus type 1 strains. J Clin Microbiol 47: 33–540.