TÜBİTAK

Turkish Journal of Biology

http://journals.tubitak.gov.tr/biology/

Research Article

Turk J Biol
(2016) 40: 589-599

doi:10.3906/biy-1503-11

In silico genome-wide identification and analysis of microsatellite repeats in the largest RNA virus family (*Closteroviridae*)

Biju GEORGE[1], Binu GEORGE[1], Mayur AWASTHI[2], Ram Nageena SINGH[3,*]

[1]Blessy Software Solutions, Malviya Nagar, Jaipur, Rajasthan, India
[2]Mahatma Gandhi Chitrakoot Gramodaya Vishwavidhyalaya, Madhya Pradesh, India
[3]Lab No. 3, Division of Microbiology, Indian Agricultural Research Institute, Pusa, New Delhi, India

Received: 04.03.2015 • **Accepted/Published Online:** 29.07.2015 • **Final Version:** 18.05.2016

Abstract: Microsatellites are known to exhibit ubiquitous presence across all kingdoms of life, including viruses. Members of the family *Closteroviridae* are the largest RNA viruses and severely affect different agricultural crops worldwide, including citrus, grapevine, and vegetables. Here we identified and systematically analyzed the nature and distribution of both simple and complex microsatellites present in the complete genome of 36 species belonging to *Closteroviridae*. Our results showed, in all analyzed genomes, that neither genome size nor GC content had any influence on number, relative abundance, or relative density of microsatellites. For each genome, dinucleotide repeats were found to be highly predominant and AT/TA and AG/GA were the two most abundant dinucleotide repeat motifs. Repeats larger than trinucleotide were relatively rare in these viral genomes. Comparative study of occurrence, abundance, and density of microsatellites among available RNA and DNA viral genomes indicated that simple repeats were less abundant in genomes of *Closteroviridae*. To our knowledge, this is the first analysis of microsatellites occurring in the largest viral genome that infects plants. Characterization of such variations in repeat sequences would be important in deciphering the origin, mutational processes, and role of repeat sequences in viral genomes.

Key words: *Closteroviridae*, simple sequence repeats, relative density, relative abundance, compound microsatellite

1. Introduction

Closteroviridae is a family of plant viruses with filamentous particles varying in length from 650 nm to over 2000 nm. The genome is a single strand of positive-sense RNA whose size varies from 13 to 19 kb. The family *Closteroviridae* includes the largest and most complex viruses causing economic losses in different agricultural crops worldwide, including citrus, grapevine, and vegetables (Martelli et al., 2002). The family *Closteroviridae* includes 3 genera: *Ampelovirus*; *Closterovirus*, possessing a monopartite genome; and *Crinivirus*, which possess bi- or tripartite genomes. However, a few criniviruses such as Little cherry virus 1 (LChV-1), Grapevine leafroll-associated virus 7 (GLRaV-7), and Cordyline virus 1 (CoV-1) are monopartite closteroviruses and form a distinct clade within the family (Martelli et al., 2012). This has led to the proposal of the formation of a fourth genus, with the proposed name *Velarivirus*. The family *Closteroviridae* comprises 36 distinct species that are semipersistently transmitted by aphids (closteroviruses), whiteflies (criniviruses), or mealy bugs/scale insects (ampeloviruses) (Jarugula et al., 2010). Currently these viral diseases are controlled by either using methods to limit their dispersion or using resistant cultivars produced by genetic engineering or breeding methods. However, due to the rapid evolution potential of these viruses the resistance is often broken. Therefore, characterization of the genome variability in these viral populations is necessary because such information could provide useful information on the processes that regulate the virus' evolution.

Microsatellites are tandem repetitions of relatively short motifs of DNA and are found ubiquitously in all genomes analyzed. Strand slippage and unequal recombination leads to variation in the number of copies of microsatellites (Toth et al., 2000), thereby making them an important source of genetic diversity and critical players in genome evolution (Kashi and King, 2006; Deback et al., 2009). Indeed, polymorphic microsatellites have been used to identify relationships between virus isolates (Deback et al., 2009). Variable length of microsatellites may alter the structure of DNA or its encoded products (Mrazek et al., 2007). Though genome features such as genome

* Correspondence: singhcsjm@gmail.com

size and GC content have been reported to influence the occurrence and polymorphic nature of microsatellites (Dieringer and Schlotterer, 2003; Coenye and Vandamme, 2005; Kelkar et al., 2008), lack of a universal correlation makes the prediction of their occurrence and density a difficult task. Microsatellites can be compound (cSSR) where two or more simple sequence repeats (SSRs) lie adjacent to each other. Compound microsatellites have been reported in diverse taxa across viruses, prokaryotes, and eukaryotes (Gur-Arie et al., 2000; Kofler et al., 2008; Chen et al., 2012). Microsatellites are more abundant in coding regions than in noncoding regions of eukaryotic genomes (Metzgar et al., 2000; Tóth et al., 2000) and some prokaryotes (Gur-Arie et al., 2000; Li et al., 2004), most probably due to an enhanced selection in coding regions (Ellegren, 2004). In smaller viral genomes, accumulation of microsatellites in the coding regions is possibly due to high coding density of the viral genome (Chen et al., 2009; George et al., 2012).

Microsatellites are associated with genetic diseases (Usdin, 2008), bacterial pathogenesis, and virulence in eukaryotes and prokaryotes (Li et al., 2004; Mrazek et al., 2007). Several examples of functional microsatellite tracts having specific functions have been found among different classes of viruses. These microsatellite tracts function in different ways within each virus (Davis et al., 1999). Promoter microsatellites are known to modulate gene expression of organisms ranging from bacteria to humans (Sawaya et al., 2012). In the yeast genome, tandem repeats are frequently found in promoter regions and are directly responsible for divergence in transcription rates. Polymorphic repeats within the yeast promoter have been shown to alter promoter structure and the binding of transcription factor (Vinces et al., 2009). Identification and analysis of SSRs in diverse viral genomes would help in comparative analysis of these repeat sequences. Therefore, we systematically analyzed the occurrence, size, density, and distribution of different microsatellites in diverse species of *Closteroviridae*, which can help in understanding the origin and evolution of repeat sequences, genome evolution, and host adaptation.

2. Materials and methods

2.1. Genome sequences

According to the International Committee on the Taxonomy of Viruses (2013), the family *Closteroviridae* comprises 3 genera with 36 distinct species (http://www.ictvonline.org/virustaxonomy). We selected all available complete viral genome sequences representing each of the three genera and sequences were downloaded in FASTA format from GenBank (http://www.ncbi.nlm.nih.gov). This included 10 species from *Ampelovirus*, 10 from *Closterovirus*, 12 from *Crinivirus*, 2 from *Velarivirus*, and 2 unclassified species. Of the 12 *Crinivirus* species, 11 viral genomes possess two genome components whereas one viral genome contains 3 genomic components. Accession numbers, genome sizes, and GC contents are summarized in Table 1. Existing annotation (the "CDS" features) was used for differentiating protein-coding and noncoding regions. In order to compare among genomic sequences of different lengths, we calculated the relative density and relative abundance values. Relative density is defined as the total length (bp) contributed by each microsatellite per kilobase of sequence analyzed, whereas relative abundance is the number of microsatellites present per kilobase of the genome.

2.2. Identification of microsatellites

Perfect di-, tri-, tetra-, penta-, and hexanucleotide repeats were detected using the Simple Sequence Repeat Identification Tool (Temnykh et al., 2001). Since viral microsatellites are known to be smaller in size, we have considered only those repeats wherein the motif was repeated continuously three or more times. Mononucleotide repeat motifs being repeated six or more times were surveyed manually as well as using IMEx software (Mudunuri and Nagarajaram, 2007). The parameters used were as follows: type of repeat: perfect; repeat size: all; minimum repeat number: 6, 90, 90, 90, 90, 90; maximum distance allowed between any two SSRs (dMAX): 10 nucleotides.

For identification of compound microsatellites, IMEx software (Mudunuri and Nagarajaram, 2007) was used. Microsatellites from genomes were extracted using the 'Advance-Mode' of IMEx using the parameters previously used for RNA viruses (Chen et al., 2012). The parameters used were as follows: type of repeat: perfect; repeat size: all; minimum repeat number: 6, 3, 3, 3, 3, 3; maximum distance allowed between any two SSRs (dMAX): 10 nucleotides. Few viral genomes contain multiple genomic components and in all such cases microsatellites found in various components were added and considered as a single genome.

2.2.1. Calculation of the expected number of microsatellites

In order to evaluate whether microsatellites were over- or underrepresented in genome sequences of members of *Closteroviridae*, we compared the observed number of microsatellites (O) with the expected number of microsatellites (E) in the form of a ratio of O/E (Mrazek, 2006). The expected number of microsatellites composed of M_t (M is the motif of the microsatellite with repeat number of t, and its length is L) in a genome of length G was calculated using the formula given by de Wachter (1981):

$$\mathrm{Exp}(M_t) = f(M)^t[1 - f(M)]\,[G'(1 - f(M)) + 2L] \quad (1)$$

$$G' = G - tL - 2L + 1 \quad (2)$$

where Exp(M_t) is the expected number of M_t, and f(M) is the probability of M.

2.3. Statistical analysis

Microsoft Office Excel 2007 was used to perform all statistical analysis. Linear regression was used to reveal the correlation between the genomic features and repeat sequences.

3. Results

3.1. Number, relative abundance, and density of various microsatellites in genomes of members of *Closteroviridae*

A genome-wide scan of 36 available genomes of various *Closteroviridae* genera revealed a total of 1852 SSR^{2-6} (dinucleotide to hexanucleotide SSR) distributed across all the species. On average 51 SSR^{2-6} were observed per genome (Table 1). The least incidence (29 SSR^{2-6}) was observed in Blackberry yellow vein-associated virus (NC_006962/NC_006963) while the maximum number (72 SSR^{2-6}) of SSR^{2-6} was observed in Lettuce chlorosis virus (NC_012909/NC_012910) (Table 1). The relative density of SSRs is highly variant, ranging from 12.59 bp/kb in the genome of Blackberry yellow vein-associated virus (NC_006962/NC_006963) to 29.17 bp/kb for Lettuce chlorosis virus (NC_012909/NC_012910) (Table 2). Similarly, relative abundance varied from a minimum of 1.84 in the Blackberry yellow vein-associated virus genome (NC_006962/NC_006963) to a maximum of 4.19 /kb in Lettuce chlorosis virus (NC_012909/NC_012910) (Table 2).

A genome-wide scan of *Closteroviridae* genomes revealed that cSSRs were present in all analyzed viral genomes except the Plum bark necrosis stem pitting-

Table 1. Overview of various microsatellites present in selected *Closteroviridae* genomes.

S. no.	Virus name	Genus	Genome size (nt)	GC%	Accession no.	No. of mono.	No. of SSR^{2-6}	No. of cSSRs
C1	Rose leaf rosette-associated virus	*Closterovirus*	17,656	46.7	NC_024906	21	62	4
C2	Mint-like virus	*Closterovirus*	15,362	43.8	NC_024448	17	50	3
C3	Carnation yellow fleck virus	Unclassified	15,602	45.3	NC_022978	19	37	1
C4	Blackberry vein banding-associated virus	*Ampelovirus*	18,643	47.6	NC_022072	5	52	3
C5	Blueberry virus A	Unclassified	17,798	45.6	NC_018519	5	55	1
C6	Cucurbit chlorotic yellows virus	*Crinivirus*	8607/8041	37.6/35.7	NC_018173/NC_018174	20	54	4
C7	Grapevine leafroll-associated virus 1	*Ampelovirus*	18,659	44.9	NC_016509	17	65	3
C8	Grapevine leafroll-associated virus 7	*Velarivirus*	16,404	44.6	NC_016436	22	56	4
C9	Grapevine leafroll-associated virus 6	*Ampelovirus*	13,807	44.6	NC_016417	5	45	1
C10	Grapevine leafroll-associated virus 4	*Ampelovirus*	13,830	43.2	NC_016416	9	44	1
C11	Grapevine leafroll-associated virus 5	*Ampelovirus*	13,384	44.3	NC_016081	4	48	1
C12	Lettuce chlorosis virus	*Crinivirus*	8591/8556	38.9/35.9	NC_012909/NC_012910	6	72	2
C13	Grapevine leafroll-associated virus 10	*Ampelovirus*	13,696	44.5	NC_011702	12	40	4
C14	Bean yellow disorder virus	*Crinivirus*	8965/8530	37.3/34.7	NC_010560/NC_010561	45	63	12
C15	Pineapple mealybug wilt-associated virus 1	*Ampelovirus*	13,071	42.6	NC_010178	2	48	3
C16	Plum bark necrosis stem pitting-associated virus	*Ampelovirus*	14,214	41.6	NC_009992	9	42	0
C17	Raspberry leaf mottle virus	*Closterovirus*	17,481	47.5	NC_008585	6	58	2
C18	Strawberry chlorotic fleck-associated virus	*Closterovirus*	17,039	41.1	NC_008366	10	45	1
C19	Grapevine leafroll-associated virus 2	*Closterovirus*	16,494	46	NC_007448	17	48	2
C20	Tomato chlorosis virus	*Crinivirus*	8595/8247	40.6/39.4	NC_007340/NC_007341	7	61	5
C21	Little cherry virus 1	*Velarivirus*	16,934	35.3	NC_001836	24	42	5
C22	Citrus tristeza virus	*Closterovirus*	19,296	45.3	NC_001661	8	53	1
C23	Beet yellows virus	*Closterovirus*	15,480	46	NC_001598	6	42	2
C24	Mint virus 1	*Closterovirus*	15,450	46.1	NC_006944	6	42	2
C25	Blackberry yellow vein-associated virus	*Crinivirus*	7800/7916	38.8/37.8	NC_006962/NC_006963	14	57	1
C26	Little cherry virus 2	*Ampelovirus*	15,045	40.8	NC_005065	5	56	4
C27	Strawberry pallidosis-associated virus	*Crinivirus*	8066/7978	38.7/36.4	NC_005895/NC_005896	15	47	4
C28	Potato yellow vein virus	*Crinivirus*	8035/5339/3892	37.9/35.8/35.1	NC_006062/AJ557129/AJ508757	19	63	5
C29	Grapevine leafroll-associated virus 3	*Ampelovirus*	17,919	46.1	NC_004667	16	44	2
C30	Beet pseudoyellows virus	*Crinivirus*	8006/7903	41.4/40.6	NC_005209/NC_005210	7	44	4
C31	Cucurbit yellow stunting disorder virus	*Crinivirus*	9123/7976	37.6/36.2	NC_004809/NC_004810	26	63	5
C32	Grapevine rootstock stem lesion associated virus	*Closterovirus*	16,527	46.3	NC_004724	20	55	3
C33	Sweet potato chlorotic stunt virus	*Crinivirus*	9407/8223	38.3/37.2	NC_004123/NC_004124	13	59	6
C34	Lettuce infectious yellows virus	*Crinivirus*	8118/7193	36.6/33.8	NC_003617/NC_003618	11	46	4
C35	Tomato infectious chlorosis virus	*Crinivirus*	8271/7913	38.3/35.8	FJ815440/FJ815441	17	67	6
C36	Carrot yellow leaf virus	*Closterovirus*	16,354	45.1	NC_013007	29	45	3
	Average*		16,008	42.06		13.7 ± 1.4	51.9 ± 1.4	3.1 ± 0.3

*Average is mean ± standard error for each.

Table 2. Relative abundance and density of various simple repeat sequences detected in genomes of members of *Closteroviridae*.

S. no.	Relative abundance (mononucleotide repeats)	Relative density (mononucleotide repeats)	Relative abundance (SSR^{2-6})	Relative density (SSR^{2-6})	Relative abundance (cSSRs)	Relative density (cSSRs)
C1	1.18	7.30	3.51	23.90	0.22	0.50
C2	1.10	6.63	3.25	21.15	0.19	0.31
C3	1.21	7.37	2.37	15.89	0.06	0.09
C4	0.26	1.87	2.78	18.02	0.16	0.33
C5	0.28	1.91	3.09	21.51	0.05	0.11
C6	1.20	7.44	3.24	22.10	0.24	0.09
C7	0.91	5.46	3.48	24.27	0.16	0.26
C8	1.34	8.16	3.41	23.89	0.24	0.38
C9	0.36	2.53	3.25	21.87	0.07	0.13
C10	0.65	3.97	3.18	20.53	0.07	0.13
C11	0.29	1.79	3.58	23.61	0.07	0.12
C12	0.34	2.09	4.19	29.27	0.11	0.21
C13	0.87	5.33	2.92	18.54	0.29	0.44
C14	2.57	11.88	3.60	24.34	0.68	0.07
C15	0.15	0.91	3.67	23.48	0.22	0.39
C16	0.63	3.79	2.95	20.75	0	0
C17	0.34	2.34	3.31	21.85	0.11	0.24
C18	0.58	3.52	2.64	17.37	0.05	0.09
C19	1.03	6.97	2.91	19.40	0.12	0.18
C20	0.41	2.49	3.62	24.16	0.29	0.57
C21	1.41	8.50	2.48	16.71	0.29	0.66
C22	0.41	2.48	2.74	19.17	0.05	0.08
C23	0.38	4.65	2.71	18.15	0.12	0.18
C24	0.38	2.33	2.71	17.47	0.12	0.31
C25	0.89	5.53	1.84	12.59	0.06	0.10
C26	0.33	2.39	3.72	25.92	0.26	0.53
C27	0.93	5.60	2.92	20.56	0.24	0.56
C28	1.10	6.77	3.64	25.48	0.28	0.50
C29	0.89	5.35	2.45	16.46	0.11	0.14
C30	0.44	2.64	2.76	18.29	0.25	0.53
C31	1.52	9.47	3.68	23.33	0.29	0.19
C32	1.21	7.80	3.32	21.90	0.18	0.38
C33	0.73	4.42	3.34	23.25	0.34	0.73
C34	0.71	4.37	3.00	20.37	0.26	0.42
C35	1.05	6.30	4.13	29.04	0.37	0.69
C36	1.77	10.88	2.75	19.07	0.18	0.28
Average*	0.82 ± 0.08	5.0 ± 0.46	3.1 ± 0.08	21.2 ± 6	0.18 ± 0.12	0.30 ± 0.19

*Average is mean ± standard error for each.

associated virus genome, which lacks cSSR sequences. A total of 114 cSSRs were observed in 36 viral genomes. The number of cSSRs ranged from 1 in 8 viral genomes to 12 in the Bean yellow disorder virus genome. An average of 3 cSSRs was observed in each genome (Table 1). The relative density of cSSRs changed drastically in selected *Closteroviridae* genomes, which ranged from 0.07 bp/kb for Bean yellow disorder virus (NC_010560/NC_010561) to 0.73 bp/kb in Sweet potato chlorotic stunt virus (NC_004123/NC_004124) (Table 2). Similarly, relative abundance varied from 0.05/kb in Blueberry virus A (NC_018519), Strawberry chlorotic fleck-associated virus (NC_008366), and Citrus tristeza virus (NC_001661) genomes to 0.68/kb in Bean yellow disorder virus (NC_010560/NC_010561) (Table 2).

Comparison of various microsatellite types indicated that mononucleotide repeats were the second most abundant microsatellite repeats and were present in all analyzed *Closteroviridae* genomes (Figure 1). In three viral genomic sequences mononucleotide repeats repeated 10 times or more were observed. A total of 495 mononucleotide repeats were observed in 36 selected

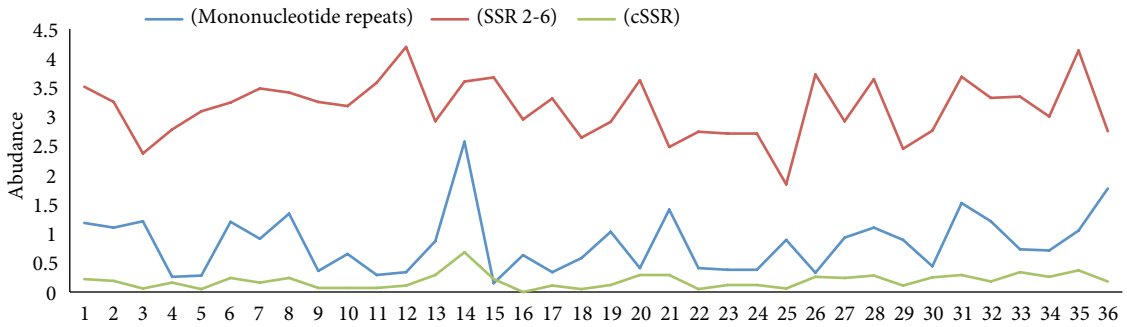

Figure 1. Comparison of relative abundance among various microsatellites observed in *Closteroviridae* genomes. Relative abundance = SSRs present per kilobase of genome.

genomes of the *Closteroviridae* members. Average mononucleotide repeats per genome was estimated as 13. A maximum number of 45 mononucleotide repeats was observed in the Bean yellow disorder virus (NC_010560/NC_010561) genome (Table 1), whereas a minimum of 2 mononucleotide repeats was observed in the Pineapple mealybug wilt-associated virus 1 (NC_010178) genome. Relative abundance of mononucleotide repeats ranged from 0.15/kb in Pineapple mealybug wilt-associated virus 1 (NC_010178) to 2.57/kb in Bean yellow disorder virus (NC_010560/NC_010561), whereas relative density was in the range of 0.91 to 11.8/kb in Pineapple mealybug wilt-associated virus 1 (NC_010178) and Bean yellow disorder virus (NC_010560/NC_010561), respectively (Table 2). The average values of number, relative abundance, and density of various microsatellite are shown in Tables 1 and 2. Microsatellite repeats were consistently underrepresented in the majority of the surveyed *Closteroviridae* members. O/E for mononucleotide repeats ranged from 0.16 to 2.16 (Table 3). The strongest underrepresentation of mononucleotide repeats was exhibited by the genome of Pineapple mealybug wilt-associated virus 1 (NC_010178).

3.2. Effect of dMAX on cSSR incidence

dMAX is the maximum distance between any two adjacent microsatellites and if the distance separating two microsatellites is less than or equivalent to dMAX, these microsatellites are classified as cSSRs (Kofler et al., 2008). To determine the impact of dMAX, 7 genome sequences representing all genera, namely Rose leaf rosette-associated virus (NC_024906), Blackberry vein banding-associated virus (NC_022072), Grapevine leafroll-associated virus 7 (NC_016436), Lettuce chlorosis virus (NC_012909/NC_012910), Blueberry virus A (NC_018519), Plum bark necrosis stem pitting-associated virus (NC_009992), and Tomato infectious chlorosis virus (FJ815440/FJ815441), were chosen to determine the variability of cSSRs with increasing dMAX. It is noteworthy that the dMAX value can only be set between 0 and 50 for IMEx (Mudunuri and Nagarajaram, 2007). The selected genomes show varied numbers of cSSRs at dMAX 10 (Table 2). Our analysis revealed an overall increase in number of cSSRs with higher dMAX in all selected viral genomes (Figure 2).

3.3. Genomic parameters influencing microsatellite distribution

We tested for the correlation between genome size/GC content and incidence, relative abundance/relative density of mononucleotide repeats, and SSR^{2-6} and cSSRs. Incidence of SSRs is not correlated ($R^2 = 0.24$, $P < 0.05$) with genome size and GC content ($R^2 = 0.01$, $P < 0.05$). Similarly, relative density ($R^2 = 0.002$, $P < 0.05$) and relative abundance ($R^2 = 0.0004$, $P < 0.05$) of SSRs were not correlated with genome size or with GC content ($R^2 = 0.01$, $P < 0.05$ and $R^2 = 0.01$, $P < 0.05$, respectively). The regression analysis of cSSRs for number of cSSRs ($R^2 = 0.05$, $P < 0.05$), relative density ($R^2 = 0.002$, $P < 0.5$), and relative abundance ($R^2 = 0.02$, $P < 0.05$) showed no correlation with genome size. Similarly, no correlation was observed between GC content and number of cSSRs ($R^2 = 0.04$, $P < 0.05$), relative density ($R^2 = 0.0008$, $P < 0.05$), or relative abundance ($R^2 = 0.04$, $P < 0.05$) of cSSRs.

3.3.1. Types of repeat motifs

Mononucleotide repeats were observed in all members of the family *Closteroviridae* analyzed and poly (A/T) repeats were found to be more prevalent than poly (G/C) repeats (Table S1). Mononucleotide repeats with six or more repetitions were considered. The longest mononucleotide repeat $(A)_{18}$ was found in genomes of Grapevine leafroll-associated virus 2 (NC_007448) followed by $(A)_{15}$ in Grapevine rootstock stem lesion associated virus (NC_004724). The longest mononucleotide with C nucleotide $(C)_{10}$ was found in Cucurbit chlorotic yellows virus (NC_018174) (Table S1). Dinucleotide repeats were the most abundant microsatellite types in all *Closteroviridae* genomes analyzed. Among the six possible types of dinucleotide repeat motifs (AG/GA, AT/TA, AC/CA, GC/CG, TG/GT), AT/TA followed by AG/GA was the most abundant motif in most of the viral genomes analyzed here, whereas the GC/CG motif was very rare and even not

Table 3. Microsatellite representation of mononucleotide repeats in analyzed *Closteroviridae* genomes.

S. no.	Total number of mononucleotide repeats with 6–10 repeats	Total number of expected mononucleotide repeats with 6–10 repeats	Observed number of mononucleotide repeats with 6–10 repeats / expected no. of mononucleotide repeats with 6–10 repeats
C1	21	13.59	1.54
C2	17	13.45	1.26
C3	19	12.68	1.49
C4	5	13.98	0.35
C5	5	14.28	0.35
C6	20	23.25	0.86
C7	17	15.45	1.09
C8	22	13.78	1.59
C9	5	11.59	0.43
C10	9	12.5	0.71
C11	4	11.41	0.35
C12	6	23.94	0.25
C13	12	11.56	1.03
C14	45	26.38	1.70
C15	2	12.27	0.16
C16	9	14.22	0.63
C17	6	13.14	0.45
C18	10	17.64	0.56
C19	16	13.02	1.22
C20	7	18.93	0.36
C21	24	26.95	0.89
C22	8	15.69	0.50
C23	6	12.22	0.49
C24	6	12.15	0.49
C25	14	20.34	0.68
C26	5	15.89	0.31
C27	15	20.76	0.72
C28	19	26.03	0.72
C29	16	14.09	1.13
C30	7	16.58	0.42
C31	26	23.88	1.08
C32	19	12.90	1.47
C33	13	22.82	0.56
C34	11	24.94	0.44
C35	17	22.60	0.75
C36	29	13.41	2.16
Average*	13.6	16.8	0.81

*Average is mean for each.

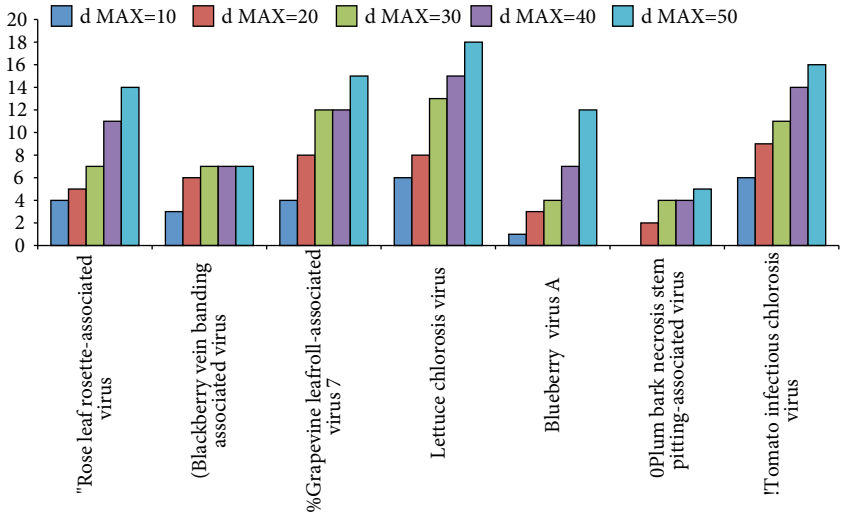

Figure 2. Number of cSSRs in relation to varying dMAX (10–50) among selected *Closteroviridae* genomes.

found in many genomes (Figure 3). Trinucleotide repeats were the third most abundant SSRs present within the viral genomes tested here (Table 4, Table S2). Repeats above trinucleotide motifs were less frequent; for example, tetra-, penta-, and hexanucleotide repeats were found in 16, 9, and 3 viral genomes, respectively. Although the majority of cSSRs were composed of two motifs, cSSRs with 3 SSRs were observed in 8 viral genomes (Table S3).

4. Discussion

It has been shown that smaller genomes such as those of viruses possess short microsatellite repeat tracts. However, such small microsatellite repeats have been previously shown to be useful as polymorphic markers (Deback et al., 2009). In this report we analyzed simple and complex SSRs from 36 completely sequenced genomes belonging to the family *Closteroviridae*, representing at least one member from each of the three genera. A minimum of 29 SSR^{2-6} and maximum of 72 such repeats were found in the 36 analyzed genomes of *Closteroviridae*. In comparison, members of ssDNA virus families such as *Geminiviridae* possess at least 4 such repeats per genome and up to 19 SSRs were observed in certain genomes (George et al., 2012). However, it should be noted that geminiviruses possess a genome size that is comparatively smaller (~2.9 kb). Larger RNA viruses such as *Caulimoviridae* possess a minimum of 12 SSR^{2-6} and maximum of 47 such repeats (George et al., 2014). The carlaviruses possess a minimum of 12 and maximum of 34 SSR^{2-6} (Alam et al., 2014). In tobamoviruses SSR^{2-6} repeats range from 7 to 30 (Alam et al., 2013). Human immunodeficiency virus (HIV) isolates possess 22 to 48 SSRs (Chen et al., 2009), whereas hepatitis C virus (HCV) possesses 16 to 30 SSRs (Chen et al., 2011). The number of SSRs may vary according to the size of the genome, with larger genomes possessing higher numbers of SSRs. Therefore, we analyzed and

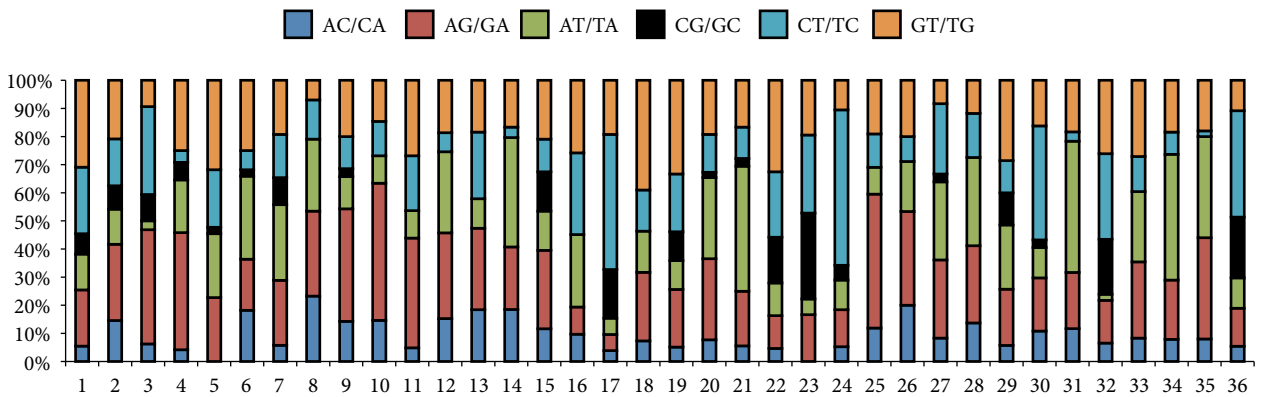

Figure 3. Relative abundance of dinucleotide motifs in *Closteroviridae* genomes.

Table 4. Occurrence of various types of SSR[2–6] in *Closteroviridae* genomes.

S. no.	No. of dinucleotid repeats	No. of trinucleotide repeats	No. of tetranucleotide repeats	No. of pentanucleotide repeats	No. of hexanucleotide repeats
C1	55	5	0	1	1
C2	48	1	0	1	0
C3	32	3	1	0	0
C4	48	3	0	1	0
C5	44	9	2	0	0
C6	44	10	0	0	0
C7	52	11	1	1	0
C8	43	11	1	1	0
C9	35	10	0	0	0
C10	41	2	0	0	1
C11	41	6	1	0	0
C12	91	12	0	1	0
C13	38	2	0	0	0
C14	90	8	0	1	0
C15	43	5	0	0	0
C16	31	10	1	0	0
C17	52	6	0	0	0
C18	41	4	0	0	0
C19	39	8	1	0	0
C20	84	7	1	1	0
C21	36	6	0	0	0
C22	43	9	1	0	0
C23	36	6	0	0	0
C24	38	4	0	0	0
C25	42	15	0	0	0
C26	45	10	1	0	0
C27	59	10	1	0	0
C28	135	11	0	0	1
C29	35	9	0	0	0
C30	59	7	0	0	0
C31	93	3	0	0	0
C32	46	8	1	0	0
C33	81	10	1	0	0
C34	62	7	1	0	0
C35	87	15	2	0	0
C36	37	6	1	1	0

compared the abundance and density of SSRs in terms of their relative values. Among all RNA and DNA viruses analyzed, the average relative abundance of SSR^{2-6} in members of *Closteroviridae* was found to be lower than in the majority of RNA viruses such as *Tobamovirus*, *Carlavirus*, *Potyvirus*, and HIV. However, the relative abundance of SSR^{2-6} in *Closteroviridae* was found to be higher as compared to HCV genomes and members of *Caulimoviridae* (Table 5).

Thus, it can be concluded that smaller RNA virus genomes possess comparably more SSRs as compared to viruses with larger genomes. Similarly, average relative density of SSR^{2-6} in *Closteroviridae* was lower as compared to other RNA viruses, except HCV and members of *Caulimoviridae* (Table 5). We previously reported that relative abundance and relative density values of SSR^{2-6} in geminiviruses (~2.8 kb ssDNA virus) are comparable to those of RNA viruses. A similar conclusion can be drawn from the analysis presented here. The highly recombinogenic nature (Padidam et al., 1999) and high mutation rate (Duffy and Holmes, 2008) could be responsible for the high abundance of microsatellite repeats in the small genomes of geminiviruses.

Genome size and GC content have been shown to have a certain influence on the occurrence of microsatellites in several species (Dieringer and Schlotterer, 2003; Coenye and Vandamme, 2005). For example, SSR density tends to be positively correlated with the genome size in some fungal (Karaoglu et al., 2005) and plant genomes (Hancock, 2000; Morgante et al., 2002). In general, a larger genome contributes to more microsatellites than smaller ones, which does not hold true if we compare the relative abundance and density of SSRs in members of *Closteroviridae* with members of *Geminiviridae* and *Tombusviridae*.

In addition genome features such as genome size and GC content were not correlated with the number, relative abundance, and relative density of microsatellites among members of *Closteroviridae*. A genome-wide scan among members of *Closteroviridae* revealed 0–6 cSSRs (except Bean yellow disorder virus, which possesses 12 cSSRs). Other RNA viruses such as *Potyvirus*, HIV, *Calaravirus*, and *Tobamovirus* genomes possess cSSRs in similar ranges of 0–5, 0–8, 0–4, and 0–4 cSSRs, respectively. Incidence of cSSRs in caulimoviruses and geminiviruses was in the range of 0 to 11 and 0 to 4, respectively. Average relative abundance and relative density values of cSSRs in *Potyvirus*, HIV, *Calaravirus*, and *Tobamovirus* genomes were 0.16 and 2.2 nt/kb, 0.32 and 4.5 nt/kb, 0.13 and 2 nt/kb, and 0.13 and 2 nt/kb, respectively. In members of *Closteroviridae* the average relative abundance (0.18/kb) of cSSR repeats was higher than that of most RNA viruses. However, the relative density of cSSRs was smaller, indicating that SSRs of smaller length constitute the cSSRs in *Closteroviridae* genomes. In contrast, in DNA viruses like caulimoviruses and geminiviruses, the average RA and RD values for cSSRs were 0.20 and 4 nt/kb and 0.4 and 7.2 nt/kb, respectively, indicating that abundance and density of cSSRs are higher in DNA viruses (including viruses possessing DNA in any stage of their life cycle) as compared to viruses with RNA genomes (George et al., 2014).

Number of mononucleotide SSRs declined sharply beyond the length of 6 bp, possibly because of an active selection against long SSRs. A similar trend was also seen in prokaryotes, where such long tracts promoted reversible mutations affecting specific genes, typically those encoding

Table 5. Comparative analysis of average relative abundance and density of SSR^{2-6} in various RNA and DNA viruses.

Virus genome	Genome size and (GC%)	Average SSR^{2-6} RA†	Average SSR^{2-6} RD‡	Reference
Caulimoviridae	7808 (40)	2.03/kb	19.7 nt/kb	George et al., 2014
Geminivirividae	2677 (43.2)	3.04/kb	21.43 nt/kb	George et al., 2012
HIV	8989.4 (41.7)	3.84/kb	26.14 nt/kb	Chen et al., 2009
HCV	9496 (57.9)	2.53/kb	17.2 nt/kb	Chen et al., 2011
Potyviruses	9703 (42)	3.58/kb	24.0 nt/kb	Alam et al., 2013*
Carlavirus	8553(45.6)	3.6/kb	23.5 nt/kb	Alam et al., 2014*
Tobamovirus	6358 (43.1)	4.0/kb	27.0 nt/kb	Alam et al., 2013*
Closteroviridae	16008 (42.06)	3.14/kb	21.2 nt/kb	In this report

*Data for mononucleotide repeats were omitted and average RA and RD of SSR^{2-6} were calculated.
†Relative abundance.
‡Relative density.

surface antigens in pathogens (Groisman and Casadesus, 2005). In prokaryotic genomes, a mononucleotide repeat of eight nucleotides in length was found to be polymorphic (Gur-Arie et al., 2000). Polymorphic microsatellites have been observed in virus genomes such as those of HIV, HCV (Chen et al., 2011), human cytomegalovirus (Davis et al., 1999), and herpes simplex virus type 1 strains (Deback et al., 2009). In CaMV the CT-rich motif located downstream of the transcription start site of the CaMV 35S promoter is involved in enhancing gene expression and in interaction with plant nuclear proteins (Pauli et al., 2004).

The sequence composition of the repeat type is also an important factor in determining the abundance of microsatellites. Variability was observed in abundance of repeat types among members of *Closteroviridae*. GC/CG repeats were found to be very rare in these viruses (Figure 3). Similarly, dinucleotide repeats such as GC/GC and GT/TG were rarely found in the genomes of other viruses (George et al., 2014). The lower frequencies of CG/GC repeats can be explained on the basis of A/T richness and the relative difficulty of strand separation for CG as compared to AT and other tracts, thus increasing slipped strand mispairing.

Compatible RNA virus infection is known to destabilize the plant genome in multiple ways, resulting in large rearrangements, point mutations, double strand breaks, mutation frequency, and microsatellite instability (Kovalchuk et al., 2003; Boyko et al., 2007; Kathiria et al., 2010). It has been hypothesized that this mechanism is used by eukaryotes including plants for faster adaptation to environmental stresses (Kashi et al., 1997; Boyko et al., 2007; Boyko and Kovalchuk, 2008). In the SV40 genome $d(GA{\cdot}TC)_n$ microsatellite DNA sequences enhance homologous DNA recombination (Benet et al., 2000). In addition, a repetitive sequence has been proposed to be correlated with recombination hot spots (Murphy and Stringer, 1986; Napierala et al., 2002, 2004). However, the functional significance of small microsatellites is not clearly understood. Therefore, we postulate that such repeats in the *Closteroviridae* genomes could also be involved in generating sequence diversity.

In conclusion, our study showed that mononucleotide repeats were underrepresented in most of the *Closteroviridae* genomes analyzed. Dinucleotide repeats were the most abundant microsatellite types, suggesting that they might play important roles in genome organization and generation of genome diversity. Functional roles of tandem repeats are still poorly understood, especially in viruses; therefore, the biological relevance of our findings remains to be elucidated. In addition, the presence of microsatellites in genomes of the *Closteroviridae* members may be useful for better understanding of the diversity and evolutionary biology of RNA viruses that infect plants.

Acknowledgment
We acknowledge Blessy Software Solutions for support.

References

Alam CM, George B, Sharfuddin C, Jain SK, Chakraborty S (2013). Occurrence and analysis of imperfect microsatellites in diverse potyvirus genomes. Gene 521: 238-244.

Alam CM, Singh AK, Sharfuddin C, Ali S (2014). Genome-wide scan for extraction and analysis of simple and imperfect microsatellites in diverse carlaviruses. Infect Genet Evol 21: 287-294.

Benet A, Mollà G, Azorín F (2000). Microsatellite DNA sequences enhance homologous DNA recombination in SV40 minichromosomes. Nucleic Acids Res 28: 4617-4622.

Boyko A, Kathiria P, Zemp FJ, Yao Y, Pogribny I, Kovalchuk I (2007). Transgenerational changes in the genome stability and methylation in pathogen-infected plants: (virus-induced plant genome instability). Nucleic Acids Res 35: 1714-1725.

Boyko A, Kovalchuk I (2008). Epigenetic control of plant stress response. Environ Mol Mutagen 49: 61-72.

Chen M, Tan Z, Jiang J, Li M, Chen H, Shen G, Yu R (2009). Similar distribution of simple sequence repeats in diverse completed human immunodeficiency virus type 1 genomes. FEBS Lett 583: 2959-2963.

Chen M, Tan Z, Zeng G, Zhuotong Z (2012). Differential distribution of compound microsatellites in various human immunodeficiency virus type 1 complete genomes. Infect Genet Evol 12: 1452-1457.

Chen M, Zeng G, Tan Z, Jiang M, Zhang J, Zhang C, Lu L, Lin Y, Peng J (2011). Compound microsatellites in complete Escherichia coli genomes. FEBS Lett 585: 1072-1076.

Coenye T, Vandamme P (2005). Characterization of mononucleotide repeats in sequenced prokaryotic genomes. DNA Res 12: 221-233.

Davis CL, Field D, Metzgar D, Saiz R, Morin PA, Smith IL, Spector SA, Wills C (1999). Numerous length polymorphisms at short tandem repeats in human cytomegalovirus. J Virol 73: 6265-6270.

Deback C, Boutolleau D, Depienne C, Luyt CE, Bonnafous P, Gautheret-Dejean A, Garrigue I, Agut H (2009). Utilization of microsatellite polymorphism for differentiating herpes simplex virus type 1 strains. J Clin Microbiol 47: 533-540.

de Wachter R (1981). The number of repeats expected in random nucleic acid sequences and found in genes. J Theor Biol 91: 71-98.

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.

Duffy S, Holmes EC (2008). Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. J Virol 82: 957-965.

Ellegren H (2004). Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5: 435-445.

George B, Gnanasekaran P, Jain SK, Chakraborty S (2014). Genome wide survey and analysis of small repetitive sequences in caulimoviruses. Infect Genet Evol 27: 15-24.

George B, Mashhood AC, Jain SK, Sharfuddin C, Chakraborty S (2012). Differential distribution and occurrence of simple sequence repeats in diverse geminivirus genomes. Virus Genes 45: 556-566.

Groisman EA, Casadesus J (2005). The origin and evolution of human pathogens. Mol Microbiol 56: 1-7.

Gur-Arie R, Cohen CJ, Eitan Y (2000). Simple sequence repeats in Escherichia coli: abundance, distribution, composition, and polymorphism. Genome Res 10: 62-71.

Hancock JM (2002). Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. Genetica 115: 93-103.

Jarugula S, Gowda S, Dawson WO, Naidu RA (2010). 3'-Coterminal subgenomic RNAs and putative cis-acting elements of *Grapevine leafroll-associated virus 3* reveals 'unique' features of gene expression strategy in the genus *Ampelovirus*. Virol J 7: 180.

Karaoglu H, Lee CM, Meyer W (2005). Survey of simple sequence repeats in completed fungal genomes. Mol Biol Evol 22: 639-649.

Kashi Y, King D, Soller M (1997). Simple sequence repeats as a source of quantitative genetic variation. Trends Genet 13: 74-78.

Kashi Y, King DG (2006). Simple sequence repeats as advantageous mutators in evolution. Trends Genet 22: 253-259.

Kathiria P, Golubov A, Sidler C, Kalischuk M, Kawchuk LM, Kovalchuk I (2010). Tobacco mosaic virus infection results in an increase in recombination frequency and resistance to viral, bacterial and fungal pathogens in the progeny of infected tobacco plants. Plant Physiol 153: 1859-1870.

Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD (2008). The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res 18: 30-38.

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.

Kovalchuk I, Kovalchuk O, Kalck V, Boyko V, Filkowski J, Heinlein M, Hohn B (2003). Pathogen-induced systemic plant signal triggers DNA rearrangements. Nature 423: 760-762.

Li YC, Korol AB, Fahima T, Nevo E (2004). Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 21: 991-1007.

Martelli GP, Abou Ghanem-Sabanadzovic N, Agranovsky AA, Al Rwahnih M, Dolja VV, Dovas CI, Fuchs M, Gugerli P, Hu JS, Jelkmann W (2012). Taxonomic revision of the family *Closteroviridae* with special reference to the grapevine leafroll-associated members of the genus *Ampelovirus* and the putative species unassigned to the family. J Plant Pathol 94: 7-19.

Martelli GP, Agranovsky AA, Bar-Joseph M, Boscia D, Candresse T, Coutts RHA, Dolja VV, Falk BW, Gonsalves D, Jelkmann W et al. (2002). The family *Closteroviridae* revised. Arch Virol 147: 2039-2044.

Metzgar D, Bytof J, Wills C (2000). Selection against frame shift mutations limits microsatellite expansion in coding DNA. Genome Res 10: 72-80.

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

Mrazek J (2006). Analysis of distribution indicates diverse functions of simple sequence repeats in *Mycoplasma* genomes. Mol Biol Evol 23: 1370-1385.

Mrazek J, Guo X, Shah A (2007). Simple sequence repeats in prokaryotic genomes. P Natl Acad Sci USA 104: 8472-8477.

Mudunuri SB, Nagarajaram HA (2007). IMEx: Imperfect Microsatellite Extractor. Bioinformatics 23: 1181-1187.

Murphy KE, Stringer JR (1986). RecA independent recombination of poly[d(GT)-d(CA)] in pBR322. Nucleic Acids Res 14: 7325-7340.

Napierala M, Dere R, Vetcher A, Wells RD (2004). Structure-dependent recombination hot spot activity of GAA.TTC sequences from intron 1 of the Friedreich's ataxia gene. J Biol Chem 279: 6444-6454.

Napierala M, Parniewski P, Pluciennik A, Wells RD (2002). Long CTG.CAG repeat sequences markedly stimulate intramolecular recombination. J Biol Chem 277: 34087-34100.

Padidam M, Sawyer S, Fauquet CM (1999). Possible emergence of new geminiviruses by frequent recombination. Virology 265: 218-225.

Pauli S, Rothnie HM, Chen G, He X, Hohn T (2004). The cauliflower mosaic virus 35S promoter extends into the transcribed region. J Virol 78: 12120-12128.

Sawaya SM, Bagshaw AT, Buschiazzo E, Gemmell NJ (2012). Promoter microsatellites as modulators of human gene expression. Adv Exp Med Biol 769: 41-54.

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001). Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11: 1441-1152.

Toth G, Gáspári Z, Jurka J (2000). Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10: 967-981.

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

Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ (2009). Unstable tandem repeats in promoters confer transcriptional resolvability. Science 324: 1213-1216.