Infection, Genetics and Evolution 27 (2014) 15–24

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Genome wide survey and analysis of small repetitive sequences in caulimoviruses

Biju George [a], Prabu Gnanasekaran [a], S.K. Jain [b], Supriya Chakraborty [a,*]

[a] Molecular Virology Laboratory, School of Life Sciences, Jawaharlal Nehru University, New Delhi 110067, India
[b] Department of Biotechnology, Jamia Hamdard University, New Delhi, Delhi 110062, India

ARTICLE INFO

Article history:
Received 6 February 2014
Received in revised form 1 June 2014
Accepted 22 June 2014
Available online 1 July 2014

Keywords:
Caulimoviridae
SSRs
Relative density
Relative abundance
Compound microsatellite

ABSTRACT

Microsatellites are known to exhibit ubiquitous presence across all kingdoms of life including viruses. Members of the *Caulimoviridae* family severely affect growth of vegetable and fruit plants and reduce economic yield in diverse cropping systems worldwide. Here, we analyzed the nature and distribution of both simple and complex microsatellites present in complete genome of 44 species of *Caulimoviridae*. Our results showed, in all analyzed genomes, genome size and GC content had a weak influence on number, relative abundance and relative density of microsatellites, respectively. For each genome, mono- and dinucleotide repeats were found to be highly predominant and are overrepresented in genome of majority of caulimoviruses. AT/TA and GAA/AAG/AGA was the most abundant di- and trinucleotide repeat motif, respectively. Repeats larger than trinucleotide were rarely found in these genomes. Comparative study of occurrence, abundance and density of microsatellite among available RNA and DNA viral genomes indicated that simple repeats were least abundant in genomes of caulimoviruses. Polymorphic repeats even though rare were observed in the large intergenic region of the genome, indicating strand slippage and/or unequal recombination processes do occur in caulimoviruses. To our knowledge, this is the first analysis of microsatellites occurring in any dsDNA viral genome. Characterization of such variations in repeat sequences would be important in deciphering the origin, mutational processes, and role of repeat sequences in viral genomes.

1. Introduction

Only three families of plant viruses have evolved to encode DNA genome: *Geminiviridae*, *Caulimoviridae*, and *Nanoviridae* (Hull, 2002). All plant viruses containing ds DNA genome having a reverse transcribing phase in its lifecycle are classified under *Caulimoviridae*. It includes 7 genera viz., *Badnavirus*, *Caulimovirus*, *Tungrovirus*, *Soymovirus*, *Cavemovirus*, *Petuvirus* and *Solendovirus*. Members of this family are unenveloped, nucleocapsid viruses ranging from 35 to 50 nm in diameter. The genomes of caulimoviruses contain monopartite, circular double-stranded DNA (6000–8000 base pairs) with one discontinuity in one strand and one or more discontinuity in the other. DNA can contain either one open reading frame (ORF) as observed in petuviruses, or up to eight ORFs such as in the soymoviruses. The viral DNA forms supercoiled mini-chromosome structures upon entering the host nucleus and is transcribed into terminally redundant polyadenylated RNA. Newly transcribed RNA moves into the cytoplasm and can be used as a template for either viral protein synthesis, or reverse transcription by viral encoded reverse transcriptase to make dsDNA (Hohn and Fütterer, 1992). This newly synthesized dsDNA can then reenter the nucleus for amplification. As replication involves use of RNA intermediate as seen in retroviruses, therefore, viruses from the *Caulimoviridae* family are termed as pararetroviruses (Hull, 2002). This property is shared by the *Hepadnaviridae* family which infect vertebrates (for review, Rothnie et al., 1994).

Simple sequence repeats (SSRs), also called as micro- or minisatellites are tandem repetitions of relatively short motifs of DNA. Their presence in viral genomes extends their existence beyond prokaryotes and eukaryotes. Strand slippage and unequal recombination leads to variation in number of copies of microsatellites (Tóth et al., 2000), thereby making it a predominant source of genetic diversity and a crucial player in genome evolution (Deback et al., 2009; Kashi and King, 2006). Indeed, polymorphic microsatellites have been used to identify relationship between virus isolates (Deback et al., 2009). Variable length of microsatellites may affect structure of DNA or its encoded products (Mrazek et al., 2007). Though genome size and GC content have been reported to influence the incidence and polymorphic nature

* Corresponding author. Tel.: +91 11 26704153.
E-mail address: supriyachakrasls@yahoo.com (S. Chakraborty).

http://dx.doi.org/10.1016/j.meegid.2014.06.018
1567-1348/

of microsatellites (Coenye and Vandamme, 2005; Dieringer and Schlotterer, 2003; Kelkar et al., 2008), lack of a universal correlation makes prediction of their occurrence and density a difficult task. Microsatellites can be complex where two or more simple sequence repeats lie adjacent to each other. Compound microsatellites have been reported in diverse taxa across viruses, prokaryotes and eukaryotes (Chen et al., 2012; Gur-Arie et al., 2000; Kofler et al., 2008). Microsatellites are more abundant in coding regions than in non-coding regions of eukaryotic genomes (Tóth et al., 2000; Metzgar et al., 2000) and some prokaryotes (Gur-Arie et al., 2000; Li et al., 2004) possibly due to an enhanced selection in coding regions (Ellegren, 2004). In smaller viral genomes, accumulation of microsatellites in the coding regions is probably due to high coding density of 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 and references therein). Promoter microsatellites are known to modulate gene expression of organisms ranging from bacteria to human (Sawaya et al., 2012). In 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 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, and density of different microsatellites in diverse species of the *Caulimoviridae*, which can help in understanding 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 (2012), the family *Caulimoviridae* comprises of 44 distinct species (http://www.ictvonline.org/virusTaxonomy). We selected fifty-four available complete viral genome sequences representing each of the seven genera and sequences were downloaded in FASTA format from the GenBank (http://www.ncbi.nlm.nih.gov). This includes 31 viral sequences from the genus *Badnavirus* (23 species), 11 species of the *Caulimovirus*, 4 of the *Soymovirus*, 3 of the *Cavemovirus*, 1 from each *Tungrovirus*, *Petuvirus*, *Solendovirus* and 2 unclassified sequences assigned to the family *Caulimoviridae*. Accession numbers, genomes size, and GC content are summarized in Table 1. Existing annotation (the "CDS" features) were used for differentiating protein-coding and non-coding 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 kb of sequence analyzed whereas, relative abundance is number of microsatellites present per kb of the genome (kb).

2.2. Identification of microsatellites

Perfect di-, tri-, tetra-, penta-, and hexanucleotide repeats were detected using simple sequence repeat identification tool (SSRIT) (Temnykh et al., 2001). Since viral microsatellites are known to be smaller in size therefore, we have considered only those repeats, wherein the motif was repeated continuously for three or more times. Mononucleotide repeats motifs, being repeated for five or more times were surveyed manually. Similar threshold values have been previously used for analyzing the microsatellite distribution in viral genomes (George et al., 2012).

For identification of compound microsatellite, 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; Alam et al., 2013). The parameters used are as follows: type of repeat: perfect; repeat size: all; minimum repeat number: 6, 3, 3, 3, 3, 3 for mono, di, tri, tetra, penta and hexanucleotide repeats, respectively. Maximum distance allowed between any two SSRs (dMAX) is 10 nucleotide.

2.3. Calculation of the expected number of microsatellites

In order to evaluate whether microsatellites were over- or underrepresented in genome sequences of members of the *Caulimoviridae*, we compared the observed number of microsatellites (O) with the expected number of microsatellites (E) in the form of a ratio of O/E. Statistical significance of the microsatellite representation (O/E), was assessed using Z scores defined as $(O-E)/\sqrt{E}$ (Mrazek, 2006). The expected number of microsatellite composed of *Mt* (M is 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 as 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 $\mathrm{Exp}(M_t)$ is the expected number of M_t, and f(M) is the probability of *M*.

2.4. 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 caulimovirus genomes

Genome-wide scan of 54 available genomes of caulimoviruses revealed a total of 1204 SSR^{2-6} (dinucleotide to hexanucleotide SSR) distributed across all the species. On an average 22 SSR^{2-6} were observed per genome (Table 1). The least incidence (12) was observed in Banana streak IM virus (HQ593112) and maximum number (47) of SSR^{2-6} was observed in Tobacco vein clearing virus (AF190123) (Table 1, Supplementary Table S1). The relative density of SSRs is highly variant ranging from 10.6 bp/kb in the genome of Banana streak IM virus to 45.1 bp/kb for Sweet potato caulimo-like virus (HQ694978) (Table 2, Fig. 1A). Similarly, relative abundance varied from a minimum of 1.5 in Banana streak IM virus genome to a maximum of 6.3 bp/kb in Sweet potato caulimo-like virus genome (Table 2, Fig. 1B).

Genome-wide scan of caulimovirus genomes revealed 0–11 compound microsatellites (cSSRs) in each of the analysed sequences (Table 1 and Supplementary Table S2). Interestingly, 21 genomes of the *Caulimoviridae* family lacked cSSR. A total of 87 cSSRs was observed in 54 genomes (Table 1). The relative density of cSSRs changed drastically in selected caulimovirus genomes, which ranged from 1.71 bp/kb (*Bougainvillea spectabilis* chlorotic vein-banding virus, EU034539) to 29.3 bp/kb in Tobacco vein clearing virus