Infection, Genetics and Evolution 21 (2014) 287–294

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Genome-wide scan for analysis of simple and imperfect microsatellites in diverse carlaviruses

Chaudhary Mashhood Alam [a], Avadhesh Kumar Singh [b], Choudhary Sharfuddin [a], Safdar Ali [b,*]

[a] Department of Botany, Patna University, Bihar 800005, India
[b] Department of Biomedical Sciences, Shaheed Rajguru College of Applied Sciences for Women (SRCASW), University of Delhi, Vasundhara Enclave, New Delhi 110096, India

ARTICLE INFO

Article history:
Received 19 August 2013
Received in revised form 15 November 2013
Accepted 21 November 2013
Available online 28 November 2013

Keywords:
IMEx
dMAX
cSSR%
Relative density
Relative abundance

ABSTRACT

An exhaustive compilation and analysis of incidence, distribution and variation of simple sequence repeats (SSRs) in viruses are required to understand the evolution and functional aspects of repetitive sequences. Present study focuses on the analysis of SSRs in 32 species of carlaviruses. The full length genome sequences were assessed from NCBI (http://www.ncbi.nlm.nih.-gov/) and analyzed using IMEx software. Variance in incidence of SSRs was observed, independent of genome size. Though the conversion of SSRs to imperfect microsatellite or compound SSR is low; compound microsatellites constituted by variant motifs accounted for up to 12.5% of the SSRs. Mononucleotide A/T is most prevalent followed by dinucleotide GT/TG and trinucleotide AAG/GAA in these genomes. The SSR and cSSR are predominantly localized to the coding region RDRP (RNA dependent RNA polymerase) and ORF-6 (open reading frame). The relative frequency of different classes of simple and compound microsatellites has been highlighted in accordance with the biology of carlavirus. Characterization of such variations would be pivotal for deciphering the enigma of these widely used, but incompletely understood sequences.

1. Introduction

According to the Ninth Report of the International Committee on the Taxonomy of Viruses (ICTV) the carlavirus genus belonging to the family *Betaflexiviridae* has 52 species accounting for ~63% of species present in the family (King et al., 2012). These species are crop damaging pathogens infecting leguminous and solanaceous crops as well as wild plant species and are mostly spread by whitefly and aphids. Each carlavirus species contains a single-stranded positive sense RNA molecule of approximately 9000 nucleotides, a terminal un-translated region and between them, a single open reading frame (ORF) that is translated into a poly-protein. The poly-protein is cleaved after translation into at least six proteins by virus-encoded proteinases that are part of the polyprotein. Interestingly, all the carlavirus species possess similar genome organization, despite of genetic diversity at the sequence level.

Simple sequence repeats (SSRs), also called as micro- or minisatellites are tandem repetitions of relatively short motifs of DNA. Their presence in viral genomes such as Human immunodeficiency virus (HIV) and Human cytomegalovirus (HCMV) extends their existence beyond prokaryotes and eukaryotes (Picone et al., 2005; Chen et al., 2012; Tóth et al., 2000; Mrazek et al., 2007), as believed otherwise. Their repeat number, length, and motif size influence microsatellite mutability (Pearson et al., 2005). Moreover, variations in copy number due to strand slippage and unequal recombination highlight the instability of the microsatellites (Tóth et al., 2000); which in turn makes it a predominant source of genetic diversity and a crucial player in genome evolution (Deback et al., 2009; Kashi and King, 2006). Their role in gene regulation, transcription and protein function is being elucidated (Kashi and King, 2006; Usdin, 2008) wherein variable length of microsatellites may affect local DNA structure or the encoded proteins and hence influence the expression profile of the corresponding genes (Mrazek et al., 2007). Though genome size and GC content might influence the incidence and polymorphic nature of microsatellites (Coenye and Vandamme, 2005; Dieringer and Schlotterer, 2003; Kelkar et al., 2008) the lack of a universal correlation ensures no single priority rule for predicting their occurrence and density.

On the basis of presence of interruptions in microsatellites; they are classified as interrupted, pure, compound, interrupted compound, complex and interrupted complex (Chambers and MacAvoy, 2000). The primary focus herein is on compound

Abbreviations: SSR, simple sequence repeat; cSSR, compound simple sequence repeat; IMEx, imperfect microsatellite extraction; RD, relative density; RA, relative abundance; RDRP, RNA dependent RNA polymerase; TGB1-2-3, triple gene block 1-2-3; ORF, open reading frame; CP, coat protein; MP, movement protein.

* Corresponding author. Tel.: +91 11 22623503; fax: +91 11 22623504.
E-mail addresses: safdar_mgl@live.in, alisafd@gmail.com (S. Ali).

1567-1348/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.meegid.2013.11.018

microsatellites; composed of two or more microsatellites adjacent to each other. Their presence has been reported in diverse taxa across viruses, prokaryotes and eukaryotes (Chen et al., 2012; Gur-Arie et al., 2000; Kofler et al., 2008). Interestingly, microsatellites are more abundant in coding regions than in non-coding regions in eukaryotes (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 (Karaoglu et al., 2005; Ellegren, 2004). Also, accumulation of microsatellites in the viral coding regions can be attributed to the high coding density of viral genome (Chen et al., 2009; George et al., 2012). The compound microsatellites constitute 4–25% of genomes of *Homo sapiens*, *Macaca mulatta*, *Mus musculus* and *Rattus norvegicus* which include some highly polymorphic compound repeats such as (dC-dA)n(dG-dT)n (Weber, 1990; Bull et al., 1999; Kofler et al., 2008). Further, 22 complete *Escherichia coli* (*E. coli*) genomes had a frequency of 1.75–2.85% while those from 81 HIV type-1 genomes had up to 24.24% cSSRs highlighting the variations across genomes (Chen et al., 2012). An in depth study of the diversifications in satellite sequences would provide insight into the imperfections and evolution of microsatellites.

Though microsatellites have been the focus of studies for their origin, distribution, and evolution their presence and possible functional significance in plant viruses have been recognized only recently (Archak et al., 2007; Xiangyan et al., 2011; George et al., 2012). Concerted efforts are required to identify and confirm the presence, distribution and variations of SSRs in RNA viruses. Here, we systematically analyzed the occurrence, size, and density of different microsatellites in the highly divergent species of carlaviruses, which can serve as a model for understanding functional aspects, evolutionary relationships, and adaptation to divergent hosts.

2. Materials and methods

2.1. Assessing the genome sequences from public database

Complete genome sequence of 32 carlaviruses species was assessed form NCBI (http://www.ncbi.nlm.nih.-gov/) and analyzed for simple and compound microsatellite identification and analysis. Genome size of these species ranged from 7890nt (Acc. No.-EU433397) to 9104nt (Acc. No.-AY461421). The accession numbers and salient features of studied carlaviruses genomes have been summarized in Table 1.

2.2. Tools and techniques for microsatellite identification and investigation

The microsatellite search was performed using the IMEx software (Mudunuri and Nagarajaram, 2007). Earlier reports on eukaryotes and *E. coli* genomes focused on assessing microsatellites of 12 bp or more (Tóth et al., 2000) but these parameters did not yield any results in carlavirus. The onus for this observation may lie with relatively smaller size of carlavirus genomes. Subsequently, microsatellites from carlavirus genomes were extracted using the 'Advance-Mode' of IMEx using the parameters previously used for HIV (Chen et al., 2012) and potyvirus (Alam et al., 2013) which are 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) is 10. Other parameters were used as default. Compound microsatellites were not standardized in order to determine real composition.

Table 1
Overview of simple microsatellites in complete carlavirus genome sequences.

S. No	Name	Accession number	Genome size	GC content	SSR	RA	RD
C1	*Aconitum latent virus*	AB051848	8657	47	32	3.7	26.1
C2	*American hop latent virus*	JQ245696	8601	47.6	34	4.0	24.0
C3	*Blueberry scorch virus*	AY941199	8525	47.4	37	4.3	27.8
C4	*Butterbur mosaic virus*	AB517596	8662	44.5	30	3.5	21.2
C5	*Chrysanthemum virus B*	AM493895	8855	45.4	33	3.7	24.4
C6	*Coleus vein necrosis virus*	EF527260	8727	47	24	2.8	18.0
C7	*Cowpea mild mottle virus*	HQ184471	8127	41	32	3.9	26.0
C8	*Daphne virus S*	AJ620300	8739	45.1	42	4.8	31.4
C9	*Garlic common latent virus*	JQ899445	8614	44.9	30	3.5	25.1
C10	*Helleborus net necrosis virus*	AB623047	8541	45.1	30	3.5	22.5
C11	*Hippeastrum latent virus*	DQ098905	8500	47	26	3.1	19.8
C12	*Hop latent virus*	AB032469	8612	47.8	27	3.1	21.4
C13	*Hop mosaic virus*	EU527979	8550	48.5	36	4.2	28.1
C14	*Hydrangea chlorotic mottle virus*	EU754720	8433	46	32	3.8	25.3
C15	*Kalanchoë latent virus*	FJ531634	8517	46	28	3.3	20.7
C16	*Ligustrum necrotic ringspot virus*	EU074853	8412	46.5	30	3.6	22.2
C17	*Lily symptomless virus*	HM222522	8394	48.7	18	2.1	14.2
C18	*Mirabilis jalapa mottle virus*	JN039374	8315	49.1	27	3.2	21.6
C19	*Narcissus common latent virus*	AM158439	8539	48.4	31	3.6	23.7
C20	*Nerine latent virus*	JQ395044	8332	38.9	39	4.7	30.5
C21	*Passiflora latent virus*	DQ455582	8386	46.5	24	2.9	17.9
C22	*Phlox virus B*	EU162589	9058	45.2	24	2.6	16.7
C23	*Phlox virus S*	EF492068	8590	43.9	36	4.2	27.4
C24	*Poplar mosaic virus*	X65102	8737	45.5	36	4.1	26.6
C25	*Potato latent virus*	EU433397	7890	45.6	23	2.9	18.5
C26	*Potato virus M*	JN835299	8520	48.3	36	4.2	31.0
C27	*Potato virus P*	EU338239	8392	47.1	32	3.8	25.1
C28	*Potato virus S*	JX419379	8507	47.6	32	3.8	26.8
C29	*Red clover vein mosaic virus*	FJ685618	8604	43.6	32	3.7	24.5
C30	*Shallot latent virus*	HQ258896	8400	43.6	30	3.6	22.0
C31	*Sweet potato C6 virus*	JX212747	8857	39.9	20	2.3	14.8
C32	*Sweet potato chlorotic fleck virus*	AY461421	9104	42.3	37	4.1	25.5