Survey and analysis of simple sequence repeats (SSRs) present in the genomes of plant viroids

Lü Qin [a,b], Zhixiang Zhang [a], Xiangyan Zhao [b], Xiaolong Wu [b], Yubao Chen [c], Zhongyang Tan [a,b,*], Shifang Li [a,*]

[a] *State Key Laboratory of Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing 100193, China*
[b] *College of Biology, State Key Laboratory for Chemo/Biosensing and Chemometrics, Hunan University, Changsha 410082, China*
[c] *Department of Computational Biology, Beijing Computing Center, Yongfeng Industry Base, Beijing 100094, China*

ARTICLE INFO

ABSTRACT

Extensive simple sequence repeat (SSR) surveys have been performed for eukaryotic prokaryotic and viral genomes, but information regarding SSRs in viroids is limited. We undertook a survey to examine the presence of SSRs in viroid genomes. Our results show that the distribution of SSRs in viroids may influence secondary structure, and that SSRs could play a role in generating genetic diversity. We also discuss the potential evolutionary role of repeated sequences in the viroid genome. This is the first report of SSR loci in viroids, and our study could be helpful in understanding the structure and evolution of viroid genomes.
© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

1. Introduction

Viroids are single-stranded, circular, noncoding RNAs that cause a variety of infectious diseases in plants [1]. Viroids have unique structural, functional, and evolutionary characteristics [2]. Furthermore, they are the smallest self-replicating plant pathogens presently known, and possess very small genomes which range in size from 246 to 401 nucleotides [3]. As the lowest level of biological complexity, viroids do not code for any proteins, but the genomes contain a number of RNA structural elements which can interact with factors from the host. In the past decade, considerable attention has been paid to some new insights and novel ideas concerning the mechanisms of viroid pathogenesis, host defense and viroid survival, transcription initiation and regulation, processing, and RNA trafficking. Mutation and recombination are usually considered to be the driving forces behind viroid genome evolution [4–6].

Simple sequence repeats (SSRs), also known as microsatellites, are a class of DNA sequences consisting of simple motifs of 1–6 nt that are tandemly repeated a few to many times at a locus

in a genome [7,8]. SSRs have been extensively studied for their evolutionary and mutational properties in eukaryotes, prokaryotes, viruses, and also pre-miRNAs [9–13]. There is increasing evidence that SSRs can serve a functional role in the regulation of gene expression by affecting transcription, translational activity, DNA structure, and other metabolic activities [14,15]. Due to their high frequency of mutation [16], SSRs can increase the diversity of the pathogen population to counteract the host immune response by providing a source of genetic variation [17], which is important in the evolution history of pathogens [18,19]. In the present study, we can ask the following questions: why are SSRs ubiquitously present in viroid genomes? How do they arise? And what are the potential functions of these SSRs in the viroid genome? In this study we screened 36 reference genomes of different viroid species for the distribution of SSR tracts, and the results may offer novel insights into the occurrence, distribution, and evolution of sequence repeats in viroid genomes.

2. Materials and methods

2.1. RNA sequences

A total of 133 complete viroid genome sequences were downloaded from the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov/Taxonomy). These sequences included reference sequences for the 36 reported viroid species and the sequences of 97 *Potato spindle tuber viroid* (PSTVd)

---

isolates (Supplementary Tables 1 and 2). Microsatellites were extracted by using the IMEx (Imperfect Microsatellite Extractor) [20]. While searching a sequence for SSRs, definition of the minimum number of repeats is an important criterion. Since viroid RNAs are very small genomes, and this would preclude the presence of a significant number of repeat copies, we adopted a relatively new search criterion, and the parameters for minimum repeat numbers were set as 5, 3, 2, 2, 2, 2 for mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats respectively. Only perfect repeats were extracted in this study. Thus, mismatch limit and imperfection percentage in repeat tract were set as 0% and 0%. The relative SSR density is defined as the total length (in nucleotides) contributed by each SSR per kb of sequence analyzed.

2.2. Software and analysis tools

In this study, random sequences with the same base composition and lengths were generated by the Perl programs. The Perl script for simulating random DNA sequences is based on the principle picking bases without replacement.

Statistical analysis and regression curves were generated with the software packages SPSS and R. Possible secondary structures were calculated using RNA structure 4.6 and RnaViz 2.0.

3. Results and discussion

In accordance with previous studies [21,22], we found that SSRs are distributed throughout viroid genomes, and there is no direct correlation between SSR content and genome size or GC content (for all two, $R^2 < 0.05$ and $P > 0.1$).

3.1. Viroid genomes contain a relatively high percentage of SSRs

In such small genomes, we detected the presence of a higher SSR density than in randomly-generated sequences similar in length and nucleotide percentage to viroid genomes, indicating that repeated sequences are probably an essential feature of the viroid genome, and are not nonfunctional "junk" or "selfish" sequences. Compared with other species with larger genomes, there are relatively few long repeats in viroids, showing that small viroid RNA genomes might preclude the presence of a significant number of repeat copies. Our results show that trinucleotide repeats occur with the highest frequency, followed by mono- and tetranucleotide repeats, with dinucleotide and hexanucleotide repeats being the least frequent.

Fig. 1. The occurrence of SSRs in 36 reference viroid genome sequences compared with randomly-generated, viroid-like sequences. ($P < 0.001$).

It is interesting to note that viroids are enriched with SSRs of all types. Evolution of SSRs in higher organisms has been owed to their occurrence in protein coding regions and the observed non-enrichment of SSRs other than tri- and hexa- has been linked to the selection against frameshift mutations caused by those repeats. There is no selection acting on enrichment bias of SSRs in viroid genomes as viroid genomes do not code for proteins.

We observe that the repeat content is higher in all of the reference sequences compared with the random sequences we generated (Fig. 1). The SSRs are not evenly distributed across the genomes. On average, 18% of the viroid genome is composed of repeats according to our analysis; the highest repeat content was found in *Iresine viroid* 1(IrVd-1) with 31.89%, and the lowest in *Coconut tinangaja viroid* (CTiVd) with 4.72%. The SSR analysis for the 97 PSTVd isolates revealed a similar distribution – the repeat compositions average 24.20% in the PSTVd genomes analyzed (Supplementary Table 2). The relatively high percentage of SSRs provides evidence that SSRs are significant in the formation of viroid genomes, given the important relationship between repeated sequences and positive selection [23,24].

Viroids have the highest mutation rate of any known biological entity. In our study, considering the high percentage of SSRs in viroid genomes compared with randomly-generated sequences, we predict a tendency that favors repeat generation (for example, replication slippage) in the evolutionary history of the viroid genome.

This tendency of producing repeats could be responsible for the observation that microsatellite polymorphisms are derived mainly from variability in length rather than in the primary sequence, and also for the observation that all three insertion mutations were adjacent repeats [25]. This repeat mechanism could also potentially lead to increases in genome size, and thus it is probably relevant to consider that lengths of viroid genomes tend to increase during evolution.

In the evolutionary history of viroid genomes, the occurrence of many repeated sequences may be harmful or even lethal, and they may be eliminated as a result of negative selection. This is one possible reason why genomes are not found to become longer very quickly; most very short SSRs may be selectively neutral, and become fixed by genetic drift. On the other hand, some longer repeat sequences may be beneficial to the genome, and can undergo positive selection to obtain functions that will enable the organism to survive changes to the environment.

Thus, we can ask – how important is this repeat tendency in the evolution of the viroid genome? A previous study has shown that a structural periodicity phenomenon has occured in most viroid genomes [26]. The formation of structural periodicity, which can also be considered to be an imperfect minisatellite, is likely due to the duplication of the consensus sequence during genome evolution. For example, the primary sequence AUCG (Fig. 3) can undergo segmental duplication during the process of evolution. In addition, the genome can be subjected to other dynamic process such as insertions/deletions, or point mutations, eventually resulting in an imperfect repeated sequence.

Interestingly, an experimental variant of citrus exocortis viroid (CEVd) was recently identified that has a 96-nucleotide duplicated sequence [27]. This could be considered evidence for the repeat tendency mechanism. Species with larger genomes, such as yeast [28], can also make use of repeat tendency to expand their genome size. In the human genome, the widespread occurrence of copy number variation also supports the significant role of repeat tendency in genome evolution. From these results we can see that repeats play a vital role in genome structure, and make a prominent contribution to genome expansion. Conceivably, the tendency to accumulate repeated sequences could be a driving force which can lead to the continual expansion of genome size in the continuing evolution of viroids.
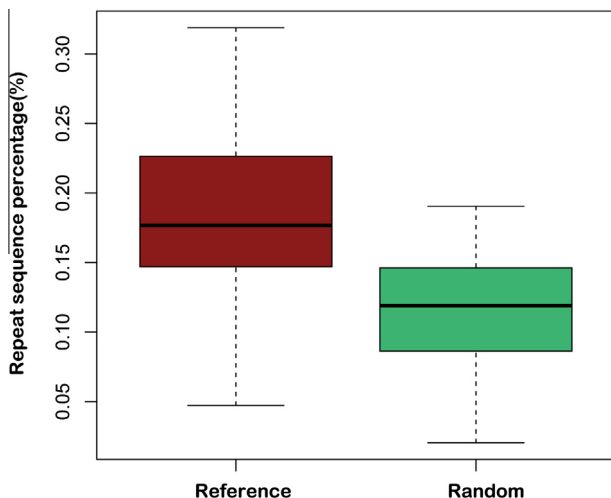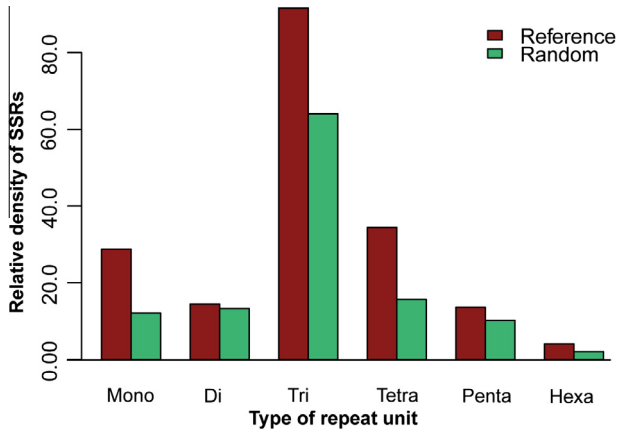
Fig. 2. Comparison of relative SSR densities in the 36 reference viroid genomes with SSR density in randomly-generated, viroid-like sequences. For the *t*-test of mononucleotide and tetranucleotide repeats between reference and random, the *p* values are less than 0.002.
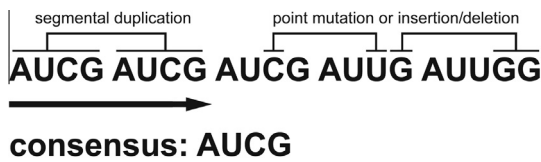

Fig. 3. A potential mechanism for imperfect repeat formation. Viroid genomes are subjected to segmental duplication during evolution, and at the same time the sequences can also experience point mutations and deletion/insertion, which can lead to imperfect repeats in the genome.

3.2. SSR polymorphism

Analysis of the occurrence of SSRs in 36 reference viroid genomes reveals a general trend of over-representation of every type of SSR repeat unit (Fig. 2), suggesting that the reference genome sequences tend to accumulate more repeats than the randomly-generated sequences. Mononucleotide and tetranucleotide repeats are significantly over-represented in viroid genomes.

Some repeat regions are highly polymorphic (Supplementary Table 5), such as the mononucleotide poly-A repeats of the PSTVd genome. In contrast, some repeated sequences are highly conserved in the genome; for instance, the tetranucleotide repeat UC-CUUCCU is present in all *Pospiviroid* reference genomes. We screened for sequence polymorphisms among strains of PSTVd;

RNA sequence alignments showed that some SSRs in the viroid genome are polymorphic, and mainly involved mononucleotide SSRs flanking region (Fig. 4). We can ask, what is the potential role of SSR variation in the viroid genome? We hypothesize that these SSRs may provide adaptive variation that is important to viroid evolution. The presence of polymorphic SSRs also reveals an important role for insertion–deletion mutations as a mechanism for producing genetic diversity in the viroid genome.

For mononucleotide repeats, the number of repeats decreases exponentially as the repeat iteration increases. This observation strongly suggests that it is very easy for short sequence repeats to form in the genome. In our opinion, such short repeats are the foundation for forming higher iteration repeats, and the repeated sequences might have arisen gradually from iterations of 5 to 6, and then to 7, and so on. This suggests that variation in the numbers of sequence repeats could proceed mainly through the process of strand-slippage during replication. Stand-slippage may be the main mechanism for producing insertion–deletion mutations during replication. For viroid genome, the slippage activities could be expected to occur frequently since the replication process depends on an DNA-dependent RNA polymerase, which is characterized by relatively lower replication fidelity.

3.3. SSR distribution and viroid secondary structure

To get a detailed picture of the distribution of repeats in the different genomes, we selected the reference sequences of *Potato spindle tuber viroid* (PSTVd), *Apple scar skin viroid* (ASSVd), *Hop stunt viroid* (HSVd), *Coconut cadang-cadang viroid* (CCCVd) and *Coleus blumei viroid* 1(CbVd-1) as representative genomes for analysis.

The secondary structure of viroid genomes consists of double-stranded stem–loop regions. The distribution of repeats in relation to viroid secondary structure is shown in Fig. 5. From Fig. 5, we can see that many more repeats tend to occur in the pathogenic and variable regions, which are related to viroid pathogenicity. For instance, the mononucleotide SSRs are predominantly distributed in the pathogenicity region (Supplementary Table 4). Thus, is there a correlation between repeat sequence and pathogenicity? A previous study showed that there are 27 nucleotide changes between isolate CEV-A and isolate CEV-DE26 of citrus exocortis viroid [29]. Based on the published sequences, we find that there are three deletions present in the long repeat region of the genome of isolate CEV-DE26 as compared to the sequence of CEV-A and there are also some substitutions in the short repeat region. These results could potentially provide clues for understanding the role of repeat sequences in the P and V domains.

Fig. 4. RNA sequence alignments of five regions containing mononucleotide repeat polymorphisms for eight isolates of PSTVd. Only the viroid genome sequences from nucleotides 1 to 160 are shown.
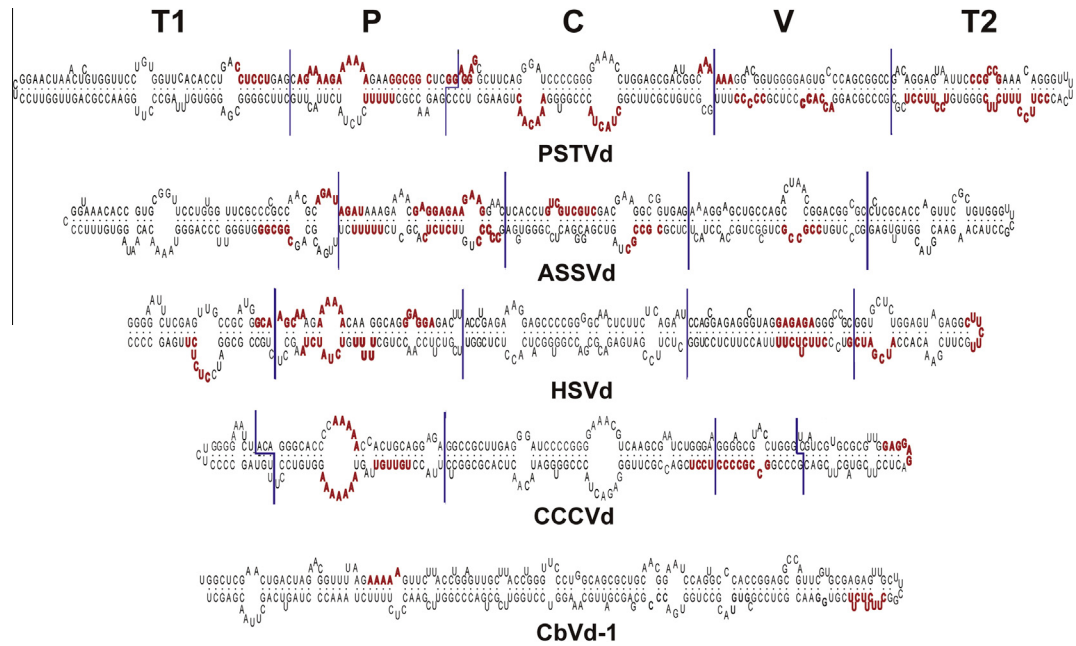
Fig. 5. The distribution of repeat sequences in different viroid genomes. The genome sequences can be divided into five domains according to Keese [30]. There is no detailed information at present about domain boundaries for CbVd-1.

From the alignment analysis in Fig. 4, we also can see that most of the poly-A-type repeats are located in the loop regions, and it appears that mononucleotide A repeats tend to form loops in the secondary structure. One possible explanation is that the higher number of repeats, the more difficult it is to form stable base pairs on a complementary strand. Therefore, there is a higher probability of forming loop regions for longer poly-A repeats. In the stem region, there are more CG base pairs than AT base pairs (Supplementary Table 3), which is not only due to the CG content bias of the genome. It is more likely that stem structures in the genome contain more stable CG base pairs that AT base pairs. In contrast, A/T nucleotides are more prone to be located in the loop regions. It is easier for genomes to generate variability with poly (A/U) repeats than with poly (C/G); thus, the length of poly-A sequence repeats show a high degree of variability. In addition, all tetranucleotide repeats are related to loop structures, suggesting that they play an active role in the formation of these loops. Thus, we can expect a correlation between SSR distribution and secondary structure. If those loop structures are less constrained for stability, we suggest that they are the only place where an otherwise harmful SSR could lie on and thus a relaxed purifying selection in those regions could be expected. It appears that repeated sequences are an important component of the loop regions of the viroid genome. The functional role of tandem repeats in determining genome structure needs to be further tested by biological experimentation.

4. Conclusion

In this study, we investigated the occurrence and distribution of SSRs in viroid genomes, and compared this with the occurrence of SSRs in randomly-generated viroid-like sequences. Taking our computational and statistical data together, we conclude that the non-random distribution of SSRs can influence viroid secondary structure. We suggest that the presence of SSRs could lead to genome sequence diversity that drives adaptation, and also that these SSRs might play a significant role in the evolutionary history of the viroid genome. Our study of SSR distribution is a small step toward a better understanding of genome structure and evolution in viroids.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fob.2014.02.001.

References

[1] Sänger, H.L., Klotz, G., Riesner, D., Gross, H.J. and Kleinschmidt (1976) Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. Proc. Natl. Acad. Sci. USA 73, 3852–3856.
[2] Flores, R., Hernández, C., Alba, A.E.M., Daròs, J. and Serio, F.D. (2005) Viroids and viroid–host interactions. Annu. Rev. Phytopathol. 43, 117–139.
[3] Gómez, G., Martinez, G. and Pallás, V. (2009) Interplay between viroid-induced pathogenesis and RNA silencing pathways. Trends Plant Sci. 14, 264–269.
[4] Tabler, M. and Tsagris, M. (2004) Viroids: petite RNA pathogens with distinguished talents. Trends Plant Sci. 9, 339–348.
[5] Ding, B. and Itaya, A. (2007) Viroid: a useful model for studying the basic principles of infection and RNA biology. Mol. Plant Microbe Interact. 20, 7–20.
[6] Kawaguchi-Ito, Y. et al. (2009) Cultivated grapevines represent a symptomless reservoir for the transmission of hop stunt viroid to hop crops: 15 years of evolutionary analysis. PLoS One 4, e8386.
[7] Vogt, P. (1990) Potential genetic functions of tandem repeated DNA sequence blocks in the human genome are based on a highly conserved "chromatin folding code". Hum. Genet. 84, 301–336.
[8] Schlötterer, C. (2000) Evolutionary dynamics of microsatellite DNA. Chromosoma 109, 365–371.
[9] Ellegren, H. (2004) Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. 5, 435–445.
[10] Chen, M. et al. (2009) Similar distribution of simple sequence repeats in diverse completed Human Immunodeficiency Virus Type 1 genomes. FEBS Lett. 583, 2959–2963.

[11] Chen, M., Tan, Z., Zeng, G. and Peng, J. (2010) Comprehensive analysis of simple sequence repeats in pre-miRNAs. Mol. Biol. Evol. 27, 2227–2232.

[12] Gemayel, R., Vinces, M.D., Legendre, M. and Verstrepen, K.J. (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu. Rev. Genet. 44, 445–477.

[13] Zhao, X.Y. et al. (2012) Coevolution between simple sequence repeats (SSRs) and virus genome size. BMC Genomics 13, 435.

[14] Li, Y.C., Korol, A.B., Fahima, T., Beiles, A. and Nevo, E. (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol. Ecol. 11, 2453–2465.

[15] Hefferon, T.W. (2004) A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. Proc. Natl. Acad. Sci. USA 101, 3504–3509.

[16] Pearson, C.E., Edamura, K.N. and Cleary, J.D. (2005) Repeat instability: mechanisms of dynamic mutations. Nat. Rev. Genet. 6, 729–742.

[17] Kashi, Y. and King, D.G. (2006) Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22, 253–259.

[18] Schlötterer, C. (1998) Genome evolution: are microsatellites really simple sequences? Curr. Biol. 8, 132–134.

[19] Moxon, E.R. and Wills, C. (1999) DNA microsatellites: agents of evolution. Sci. Am. 280, 94–99.

[20] Mudunuri, S.B. and Nagarajaram, H.A. (2007) IMEx: imperfect microsatellite extractor. Bioinformatics 23, 1181–1187.

[21] Lawson, M.J. and Zhang, L.Q. (2006) Distinct patterns of SSR distribution in the Arabidopsis thaliana and rice genomes. Genome Biol. 7, R14.

[22] Toth, G., Gaspari, Z. and Jurka, J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10, 967–981.

[23] Loire, E., Praz, F., Higuet, D., Netter, P. and Achaz, G. (2009) Hypermutability of genes in Homo sapiens due to the hosting of long mono-SSR. Mol. Biol. Evol. 26, 111–121.

[24] Field, D. and Wills, C. (1998) Abundant microsatellite polymorphism in Saccharomyces cerevisiae, and the different distributions of microsatellites in eight prokaryotes and S. cerevisiae, result from strong mutation pressures and a variety of selective forces. Proc. Natl. Acad. Sci. USA 95, 1647–1652.

[25] Gago, S., Elena, S.F., Flores, R. and Sanjuán, R. (2009) Extremely high mutation rate of a hammerhead viroid. Science 323, 1308.

[26] Juhász, A., Hegyi, H. and Solymosy, F. (1988) A novel aspect of the information content of viroids. Biochim. Biophys. Acta 950, 455–458.

[27] Fadda, Z., Daròs, J.A., Flores, R. and Duran-Vila, N. (2003) Identification in eggplant of a variant of citrus exocortis viroid (CEVd) with a 96 nucleotide duplication in the right terminal region of the rod-like secondary structure. Virus Res. 97, 145–149.

[28] Liti, G. and Louis, E.J. (2005) Yeast evolution and comparative genomics. Annu. Rev. Microbiol. 59, 135–153.

[29] Visvader, J.E. and Symons, R.H. (1983) Comparative sequence and structure of different isolates of citrus exocortis viroid. Virology 130, 232–237.

[30] Keese, P. and Symons, R.H. (1985) Domains in viroids: Evidence of intermolecular RNA rearrangements and their contribution to viroid evolution. Proc. Natl. Acad. Sci. USA 82, 4582–4586.