Genomics Proteomics Bioinformatics **12** (2014) 164–171

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

REVIEW

Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function

Shweta Mehrotra *, **Vinod Goyal**

Department of Botany, University of Delhi, Delhi 110007, India

Received 12 April 2014; revised 29 June 2014; accepted 3 July 2014
Available online 15 August 2014

Handled by Xiangfeng Wang

KEYWORDS

Repetitive sequences;
Satellites;
Tandem;
Dispersed;
Concerted evolution;
Next-generation sequencing

Abstract Repetitive DNA sequences are a major component of eukaryotic genomes and may account for up to 90% of the genome size. They can be divided into minisatellite, microsatellite and satellite sequences. Satellite DNA sequences are considered to be a fast-evolving component of eukaryotic genomes, comprising tandemly-arrayed, highly-repetitive and highly-conserved monomer sequences. The monomer unit of satellite DNA is 150–400 base pairs (bp) in length. Repetitive sequences may be species- or genus-specific, and may be centromeric or subtelomeric in nature. They exhibit cohesive and concerted evolution caused by molecular drive, leading to high sequence homogeneity. Repetitive sequences accumulate variations in sequence and copy number during evolution, hence they are important tools for taxonomic and phylogenetic studies, and are known as "tuning knobs" in the evolution. Therefore, knowledge of repetitive sequences assists our understanding of the organization, evolution and behavior of eukaryotic genomes. Repetitive sequences have cytoplasmic, cellular and developmental effects and play a role in chromosomal recombination. In the post-genomics era, with the introduction of next-generation sequencing technology, it is possible to evaluate complex genomes for analyzing repetitive sequences and deciphering the yet unknown functional potential of repetitive sequences.

Introduction

Genomes of higher eukaryotes contain more DNA than expected when estimates are based on the length and number of coding genes in the genomes. The amount of DNA in the unreplicated genome, or the haploid genome, of a species is known as C-value or Constant-value [1,2]. The lack of correlation between size and complexity of eukaryotic genomes, largely due to the presence of noncoding highly repetitive DNA, is termed as the C-value paradox, which is a common phenomenon observed in higher plants. It is believed that the proportion of protein-coding sequences is generally similar for different plant species, with variation in genome size mainly due to the presence of repetitive DNA [3,4] that has accumulated in the genomes during evolution, since ancestral angiosperms had been indicated to possess small genomes [5]. The term "repetitive sequences" refers to homologous DNA fragments that are present in multiple copies in the genome. Repetitive DNA sequences are present in all higher plants and can account for

---

* Corresponding author.
E-mail: shwetamehrotra@rediffmail.com (Mehrotra S).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2014.07.003

up to 90% of the genome size in some species. These repetitive DNA sequences are considered to generate major differences between genomes, which may reflect evolutionary distances between species. The repetitive sequences were once thought to be “selfish elements” or “junk DNA” [6,7], since they do not harbor genes.

Knowledge of the distribution, genomic organization, chromosomal location and evolutionary origin of repetitive DNA sequences is necessary for insight into the organization, evolution, behavior and functional potential of repetitive sequences in eukaryotic genomes [8]. In the last few decades, several repetitive sequences have been analyzed (**Table 1**) to gain more information on primary structure, molecular organization, evolution and function of repetitive sequences, whereas more recently genome-sequencing technology has produced an unprecedented wealth of information about origin, diversity and genomic impact of repetitive sequences [9].

Types and distribution of repetitive DNA sequences

The genomes of most eukaryotes contain a variety of repetitive DNA sequences [10,11]. Repetitive DNA may be dispersed throughout the genome or may be restricted at specific locations in a tandem configuration. According to the length of the repeated unit and array size, tandem repeated DNA sequences can be classified into three groups: (i) microsatellites with 2–5 bp repeats and an array size of the order of 10–100 units, (ii) minisatellites with 6–100 bp (usually around 15 bp) repeats and an array size of 0.5–30 kb and (iii) satellite DNA (satDNA) with a variable AT-rich repeat unit that often forms arrays up to 100 Mb. The monomer length of satDNA sequences ranges from 150–400 bp in majority of plants and animals. satDNA sequences are located at heterochromatic regions, which are found mostly in centromeric and subtelomeric regions in the chromosomes but also at intercalary positions. **Figure 1** shows a diagrammatic representation of different types of repetitive sequences on a plant chromosome. Satellite DNAs are portions of the genome that can be separated as satellite peaks from primary DNA peaks [12] using the methods including restriction endonuclease digestion, colony filter hybridization with relic DNA, amplification with specific primers, colony filter hybridization and genomic self-amplification (also known as self-priming) [13].

Clustered DNA repeats can be detected in centromeric and telomeric heterochromatin, and are transcriptionally inert. Centromeric DNA (such as *CENH3*) is the most abundant tandem repeats found in both plants and animals [14]. In contrast, subtelomeric repetitive sequence families are usually genus-specific, such as TrsA and Os48 in *Oryza* species [15,16]; TrsB in *Oryza brachyantha* [17]; *Sac*I family in *Silene latifolia* [18]; and pAv34, pAc34, pRp34, pRn34 and pRs34 in *Beta* species [19], or chromosome (Chr)-specific, such as WE 35 on Chr 5B in *Triticum aestivum* [20], JNK family on Chr 2R in *Secale cereale* [21], *Afa*I family in Poaceae members [22], TRI on Chr Y in *Silene latifolia* [23] and RUSI on Chr 1 in *Rumex* species [24].

A particular sequence may be either species-specific or present in many species within a taxonomic family or various families, indicating that some repetitive sequences evolve rapidly, whereas others may be conserved [25]. For instance, the CL600 satellite repeat isolated from *Citrus limon* was detected in other Rutaceae members (*C. aurantium*, *C. paradisi*, *Poncirus trifoliata* and *Fortunella margarita*) [26]. Similarly, the PCvKB repetitive DNA isolated from *Crocus vernus* was present in 16 species of *Crocus* analyzed and 2 species of Iridaceae, 3 species of Liliaceae and 1 species of Amaryllidaceae [27]. The *Hin*dIII repeat of *Brassica campestris* shows homology not only to that of other *Brassica* species but also to that of *Raphanus sativus*, *Sinapis alba*, *Diplotaxis muralis* and *Erucastrum* sp. [28]. Repetitive elements in cauliflower, mustard and radish, all belonging to Brassicaceae family, show 75%–80% homologies [29,30]. Mehrotra et al. [31] recently reported that the pCt*Kpn*I-I satellite repeat, initially isolated from *Carthamus tinctorius* and other species of *Carthamus* [31], is present in widely divergent families of angiosperms [32].

Various repetitive sequences from diverse taxa have been integrated into a database for easy accessibility. PlantSat, a database specialized for plant satDNA, has integrated sequence data from several resources such as NCBI and DNA Data Bank of Japan (DDBJ) [33], to provide a list of satDNA sequences for members of many plant families including Poaceae, Brassicaceae, Chenopodiaceae, Cucurbitaceae and Solanaceae [34–36], as well as many other plants. Plant satDNA sequences commonly have monomer unit lengths of 135–195 bp or 315–375 bp, which are consistent with reports that the basic monomer units of plant repetitive DNA sequences correspond to the length of DNA wrapped around a nucleosomal particle [10,33,37].

Different families of repetitive sequences show consistent presence of motifs like AA/TT dinucleotides, pentanucleotide CAAAA, *etc*. The presence of conserved motifs in unrelated repetitive sequence families suggests their significance for molecular mechanisms underlying the amplification and maintenance of tandem repeats in a genome, and the determination of specific chromatin properties of loci containing the repetitive DNA [33]. The occurrence of short, direct and inverted repeats and short palindromes is a characteristic feature of various plant satellite families. These may act as preferential sites for changes and as potential substrates for homologous recombination allowing rearrangements [38–42]. The repetitive sequences have the nearest-neighbor characteristics with high frequencies of GG, AG and GA nearest neighbors [43], which originate during the repair of heteroduplex intermediates of the exchange [44,45]. Frequent occurrence of GGT and GTT trinucleotides in the monomers of repeat sequences makes the sequence suitable substrate for the *de novo* telomere synthesis in the repairing process of broken chromosome ends [46].

Another characteristic feature of repetitive sequences is methylation. Methylation of DNA sequences is considered to trigger structural changes in DNA strands [21,22]. Methylation has been observed in satellite repeats following restriction analysis of genomic DNA with *Msp*I and *Hpa*I. Methylation has been reported to occur in a 500 bp satellite repeat family in *Arabidopsis* [47], in a JNK repeat family in Japanese rye [21], in a pCt*Kpn*I repeat family in *Carthamus* [31], and many other repeat families [26,48].

Functions of repetitive sequences

Repetitive DNA sequences are present in the heterochromatin region. Heterochromatin has been associated with several functions ranging from regulating gene expression to protecting chromosomal integrity. Heterochromatin can incur different