Molecular Phylogenetics and Evolution xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Insights on genome size evolution from a miniature inverted repeat transposon driving a satellite DNA

Thibault Scalvenzi, Nicolas Pollet *

Institute of Systems and Synthetic Biology, CNRS, Université d'Evry Val d'Essonne, Bâtiment 3, Genopole® campus 3, 1, rue Pierre Fontaine, F-91058 Evry, France

ARTICLE INFO

Article history:
Received 18 April 2014
Revised 20 July 2014
Accepted 12 August 2014
Available online xxxx

Keywords:
Transposon
Amphibian
Genome
Tc1/mariner
Miniature inverted repeat
Satellite DNA

ABSTRACT

The genome size in eukaryotes does not correlate well with the number of genes they contain. We can observe this so-called *C*-value paradox in amphibian species. By analyzing an amphibian genome we asked how repetitive DNA can impact genome size and architecture. We describe here our discovery of a Tc1/mariner miniature inverted-repeat transposon family present in *Xenopus* frogs. These transposons named miDNA4 are unique since they contain a satellite DNA motif. We found that miDNA4 measured 331 bp, contained 25 bp long inverted terminal repeat sequences and a sequence motif of 119 bp present as a unique copy or as an array of 2–47 copies. We characterized the structure, dynamics, impact and evolution of the miDNA4 family and its satellite DNA in *Xenopus* frog genomes. This led us to propose a model for the evolution of these two repeated sequences and how they can synergize to increase genome size.

1. Introduction

Many similar metazoan organisms diverge for their DNA content. Amphibian species exemplify this observation known as the *C*-value paradox. There is 100 times more nuclear DNA in the salamander *Necturus lewisi* in comparison to the frog *Limnodynastes ornatus*. We know that the presence of repetitive DNA sequences in genomes can explain the *C*-value paradox. Indeed, repetitive DNA can represent between 50% and 70% of metazoan genomes. But the identity and the contribution of the repetitive DNA sequences involved remain puzzling (Gregory, 2005).

Among repetitive DNA sequences, those derived from transposable elements (TEs) are widespread and can replicate themselves (Hua-Van et al., 2011). We know that the category of miniature inverted-repeat transposable elements (MITEs) can reach very high copy number in several eukaryote genomes (Feschotte et al., 2002; Lu et al., 2012). MITEs have been observed for the first time in the maize genome and they are the most abundant group of DNA TEs (Bureau and Wessler, 1992). MITEs are non-autonomous TEs and their transposition requires the activity of a class II DNA transposase acting via a cut-and-paste mechanism. MITEs present the structural hallmark of a typical DNA transposable element with conserved Inverted Terminal Repeats sequences (ITRs) flanked by target site duplications. They tend to be of small size (<500 bp), they lack protein-coding potential, they are interspersed and may reach high copy number with a high uniformity between copies. Yet these global characteristics suffer exceptions and some families of MITEs are not highly repeated and their sequences can be heterogeneous (Bergemann et al., 2008; Fleetwood et al., 2011).

Besides transposable elements, large genomes contain another category of repeated sequences named satellite DNAs. Satellite DNAs are made of non-coding tandemly repeated sequence motifs and can be divided in two classes. The first class, micro and minisatellites, is defined by the small size (2–20 bp) of its basic repeating units. In the second class, satDNA, the satellite monomers are larger and measure between one hundred to more than one kbp. SatDNA arrays can span from tens of kbp to several Mbp and they are commonly found in heterochromatic chromosomal compartments (Plohl et al., 2008). Satellite arrays can expand and contract, and the diversity of satDNA families varies rapidly within and between species. satDNA play functional roles in genome biology, for example a majority of plant and animal chromosomes contain centromeres made of a specific class of satellite DNA. Unequal crossing-over, replication slippage and rolling circle replication can explain the origin, expansion and evolution of satDNA (Krüger and Vogel, 1975; Levinson and Gutman, 1987; Okumura et al., 1987; Smith, 1976; Walsh, 1987). Plohl et al. proposed that

* Corresponding author. Address: Institute of Systems and Synthetic Biology, CNRS, Université d'Evry-Val-d'Essonne, Genavenir 3, Genopole® campus 3, 1, rue Pierre Fontaine, F-91058 Evry, France. Fax: +33 0169361119.
E-mail address: Nicolas.Pollet@issb.genopole.fr (N. Pollet).

http://dx.doi.org/10.1016/j.ympev.2014.08.014
1055-7903/

the major satDNA families found in a given species could be derived from an ancestral pool of sequences corresponding to satDNA monomers (Plohl et al., 2008). But how such a pool of sequence appears and how major satDNA families are preferentially amplified in each species remains unclear.

Transposable elements can give rise to satDNA and links between MITEs and satDNA have been discussed before (Plohl et al., 2012). Recent studies have clarified previous reports of MITEs containing satDNA motifs. The study of Yang and Barbash on DINE-1 and the report of Kapitonov and Jurka showed that TEs believed to be MITEs are non-autonomous *Helitron* elements (Jurka et al., 2005; Yang and Barbash, 2008). *Helitrons* are a particular class II DNA transposons that use a rolling-circle mechanism of transposition (Kapitonov and Jurka, 2007; Wicker et al., 2007). More recently, a composite repeat element from the clam was described as a MITE containing tandem repeats (Šatović and Plohl, 2013). Yet, the structure of the element named DTC84 resembles that of DINE-1 elements and the authors themselves conclude that DTC84 elements are likely derived from the rolling circle *Helitron* family of TE. Thus the links between *bona fide* MITEs derived from class II DNA transposase acting via a cut-and-paste mechanism and satDNA evolution remains unknown.

Repetitive elements including transposable elements span more than thirty percent of the *Xenopus tropicalis* frog genome sequence (Hellsten et al., 2010). Yet our knowledge of transposons in *Xenopus* is fragmentary and is progressing slowly. The transposon landscape in the *Xenopus* genome is a complex mixture of many different families of DNA transposons and retrotransposons (Hikosaka and Kawahara, 2010; Hikosaka et al., 2011; Pollet and Mazabraud, 2006; Shen et al., 2013; Sinzelle et al., 2011, 2006, 2005). In this amphibian genome and unlike the mammalian genomes, DNA transposons predominate over retrotransposons. All these transposons are present in intergenic regions, in introns and even in untranslated exons, such as the last exons corresponding to the 3′ untranslated region of mRNAs. These transposons have an effect upon transcription and mRNA stability and they can fulfill specific roles when they are "domesticated" (Sinzelle et al., 2009).

While *X. tropicalis* genome sequencing enabled the identification of many DNA TEs, some elements identified by automatic bioinformatic pipelines remain relatively mysterious. A good example is given by the DNA4_Xt element found in RepBase (Jurka et al., 2005). The sequence of DNA4_Xt is annotated as a dimer of a Tc1/mariner element, and linked to a derived satellite. We became interested by DNA4_Xt when we found that it corresponds to a peculiar Tc1/mariner MITE family. This MITE named miDNA4 shares the characteristic features of a MITE but it is unique in that it contains a satellite DNA motif. We characterized the structure, dynamics and evolution of the miDNA4 MITE family and its satellite DNA in *Xenopus* frogs. We conclude by proposing a model for the evolution of these two repeated elements and how they can synergize to increase genome size.

2. Materials and methods

2.1. Bioinformatics and databases

We used the following softwares: Dotter (Sonnhammer and Durbin, 1995), NCBI BLAST (v2.2.16), EMBOSS package (www.emboss.org), Tandem Repeat Finder (TRF, V4.04) (Benson, 1999), Clustal Omega V1.2.0 (Sievers et al., 2011), DNAfold implemented in Genious 6.1.6, weblogo V3.3 (Crooks et al., 2004), UCLUST (Edgar, 2010), R (V2.15.1, http://cran.r-project.org/bin/), and PhyML V3.1 (Guindon et al., 2010). We downloaded the *X. tropicalis* genome assembly 7.1 and mRNA databases on Xenbase FTP server (Hellsten et al., 2010). We used Repbase V15.03 (Jurka et al., 2005).

2.2. MITE analysis

We extracted from RepBase all non-autonomous DNA transposable elements plus all autonomous DNA transposable element of less than 1000 bp for each superfamily (hAT, Kolobok, PIF/Harbinger, PiggyBac and Tc1/Mariner). We used each superfamily non-autonomous TE sequences as queries in megablast search of the *X. tropicalis* genome to count the corresponding non-autonomous TEs in the *X. tropicalis* genome. The parameters used were: megablast –e 1e-10 –b 100000 –v 100000. We used the values of similarity provided by megablast to compare the families and to generate the identity distribution using R scripts. A prototype miDNA4 sequence containing a single satellite motif is given in Supplementary Document 1 and another is available in GenBank under the accession number AAMC02025625.1 at positions 75473–75802.

We used Clustal Omega to align all 5′ and 3′ ITRs (31092 ITRs in total) from miDNA4 containing a single satellite motif and to estimate their similarity. This alignment was also used for Weblogo analysis. We used Needle to compare 5′ and 3′ ITR from the same miDNA4 for all miDNA4 containing a single satellite motif. We obtained the histogram of identity percent using R. We measured 6252 distances between 5′ ITR and 3′ ITR from the results of a BLASTN search of the *X. tropicalis* genome using ITR sequence queries to measure miDNA4 sizes. We parsed BLAST results and computed the distances using a homemade PERL script. We used BLASTN on a *X. tropicalis* mRNA database using miDNA4 as a query to keep only mRNA with one or more copy of miDNA4. We then crossed the data between miDNA4 coordinates in mRNA and mRNA regions (5′ UTR/CDS/3′ UTR).

2.3. miDNA4 in silico quantification

We quantified the number of miDNA4 elements using BLASTN on *X. tropicalis* genome with the miDNA4 consensus sequence from Repbase as a query (from 1 to 331 of RepBase DNA4_Xt entry). The BLASTN parameters used were "-e 1e-6 –v 100000 –b 100000". We parsed these BLASTN results using a PERL script to estimate the number of miDNA4 for several thresholds of percent identity and percent coverage (Table 1). We used TRF to find and count the number of miDNA4 satellite motif in the *X. tropicalis* genome. We searched all satellites in the *X. tropicalis* genome and we extracted flanking genomic regions at the 5′ and 3′ end of the tandem repeats. We then parsed these results to keep only those corresponding to the miDNA4 satellite motif and we searched ITRs sequences in the 5′ and 3′ end of the satellite motif. Thus we counted only miDNA4 satellite motifs with flanking miDNA4 ITR sequences.

2.4. miDNA4 satellite analysis

We used the satellite motif of the DNA4_Xt element sequence available in Repbase as a reference query for BLASTN sequence similarity searches against databases composed of miDNA4

Table 1
Count of miDNA4 as a function of sequence identity and coverage.

Identity	Coverage				
	>80%	>85%	>90%	>95%	>99%
>99%	0	0	0	0	0
>95%	174	160	151	137	100
>90%	9401	8721	8206	7505	6279
>85%	19311	17764	16646	14966	12440
>80%	20362	18652	17407	15546	12856

Please cite this article in press as: Scalvenzi, T., Pollet, N. Insights on genome size evolution from a miniature inverted repeat transposon driving a satellite DNA. Mol. Phylogenet. Evol. (2014), http://dx.doi.org/10.1016/j.ympev.2014.08.014