Available online at www.sciencedirect.com

SCIENCE DIRECT

Gene 346 (2005) 241–247

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Distribution and abundance of microsatellites in the genome of bivalves

Fernando Cruz, Montse Pérez, Pablo Presa*

University of Vigo, Faculty of Biology, Department of Biochemistry, Genetics and Immunology, 36310 Vigo, Spain

Received 2 April 2004; received in revised form 24 September 2004; accepted 17 November 2004
Available online 1 February 2005
Received by O. Clay

Abstract

Understanding how microsatellites are distributed in eukaryotic genomes is important to clarify the differential abundance of these repeats under an evolutionary scenario. We have concatenated data from 3165 DNA sequences of 326 Bivalvia species to search for taxonomic patterns of microsatellite distribution in genomic regions of markedly different functionality. Some microsatellite motifs in bivalves showed one of the lowest genomic densities observed among eukaryotes. Contrary to the expectation of a random distribution of microsatellites, they were overrepresented in introns (245 loci/Mb) compared to their frequency in exons (85 loci/Mb). Closely related species showed remarkable differences in microsatellite density suggesting species-specific properties as for mutation/repair efficiency on replication slippage. There was no evidence of a positive correlation between the density of microsatellites in intergenic DNA and the DNA-content. This research is relevant to better understand the forces shaping the distribution of microsatellites in the genome of bivalves.

Keywords: Genomic distribution of microsatellites; DNA-content; Bivalvia; Molluscs

1. Introduction

Microsatellites are ubiquitous DNA elements of eukaryotic genomes that consist of short combinations of simple nucleotide sequences repeated in tandem and usually flanked by non-repetitive sequences. Although microsatellites have been widely observed across genomes, their origin, evolution and genomic organization are only beginning to be understood. It is assumed that most microsatellites have evolved from frameshift mutations through slipped-strand mispairing during DNA replication or repair (e.g., Kornberg et al., 1964). Also interhelical junctions during chromosome alignment, base substitutions, and retrotransposition events can play a role in the generation of microsatellites (e.g., Wilder and Hollocher, 2001).

It has been shown that the overall frequency of microsatellites varies widely across genomes (e.g., Lagercrantz et al., 1993), and recent evidence points to their non-random genomic distribution. This offers an opportunity to test for possible selection on ubiquitous repeat motifs. The density of microsatellites is influenced by many features involved in shaping genomes, as the nucleotide composition, which can be addressed through large-scale genome sequencing (e.g., Bachtrog et al., 1999). Indeed, a differential abundance of repeats in exonic, intronic and intergenic regions has been observed in different eukaryotic taxa, suggesting that strand slippage theories alone are insufficient to explain microsatellite distributions (Tóth et al., 2000). Moreover microsatellite

Abbreviations: A, adenosine; $(AC)_n$, adenosine-cytosine tract repeated *n* times; $(AC)_{\geq n}$, adenosine-cytosine tract repeated at least *n* times; AC/TG, adenosine-cytidine and its complementary thymidine-guanosine; bp, base pair(s); C, cytidine; cDNA, DNA complementary to RNA; *C*-value, haploid DNA content of genomes; Da, dalton(s); DEAE, diethylaminoethyl; DIG, Digoxigenin; *e*, exons; *fe*, microsatellite frequency in exons; *fi*, microsatellite frequency in introns/*UTR*; G, guanosine; Gb, gigabase(s) or 1000,000,000 bp; *i*, introns plus *UTR*; kb, kilobase(s) or 1000 bp; Mb, megabase(s) or 1000,000 bp; ORF, open reading frame; p, plasmid; pg, picogramme(s); pmol, picomol(es); T, thymidine; *UTR*, untranslated region(s).

* Corresponding author. Tel./fax: +34 986 812567.
E-mail address: pressa@uvigo.es (P. Presa).

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.11.013

occurrence in exons seems to be limited by non-perturbation of the reading frame and tolerance of expanding amino acids repeat stretches in the encoded proteins (e.g., Katti et al., 2001).

Knowledge of the patterns of microsatellite distribution may also help to understand the evolutionary properties of these repeats. One intriguing question in this respect is why certain repeat motifs are more common than others, and why this varies among taxa. In humans, $(A)_n$ and $(AC)_n$ are by far the most common repeated motifs, the latter being the most abundant dinucleotide motif in eukaryotes. Microsatellite motifs, abundance, and mutation rates vary between species (e.g., Ross et al., 2003). Moreover, they are non-randomly distributed throughout eukaryotic genomes, and show different properties in genomic regions of different functionality (e.g., Katti et al., 2001).

Mollusca are expected to become an important model for evolutionary radiation since they have played a crucial role in morphological and molecular attempts to unravel the phylogeny of major animal groups (Schilthuizen, 2002). In this study we have investigated the frequencies of microsatellite repeats in exons, introns/*UTR* and intergenic DNA of 3165 published Bivalvia DNA sequences. This investigation has been reinforced by sequencing of 16 kb from 31 recombinant DNA clones of *Mytilus galloprovincialis* chosen as a representative species, which were largely composed of intergenic DNA.

Although there seems to exist a general tendency of length and density of microsatellites to increase with genome size (e.g., Primmer et al., 1997), this relationship is not universal. For instance, microsatellites in the pufferfish (*Fugu rubripes*) are denser and longer than in humans, even if the pufferfish genome is eight times smaller (e.g., Elgar et al., 1999). To address this question we have investigated the putative correlation between the genome size of several Bivalvia species and the microsatellite density they contain.

2. Materials and methods

2.1. DNA sequencing in M. galloprovincialis

A partial genome library was constructed with *M. galloprovincialis* DNA extracted from mantle tissue adding a mucopolysaccharides precipitation step (Sokolov, 2000). The DNA was digested with *Mbo*I and electrophoresed in preparative low-melting agarose gels. Fragments between 200 and 800 bp were size-fractionated by reverse electrophoresis on a DEAE cellulose membrane and then recovered using a standard salt method (Sambrook et al., 1989). Subsequent ligation into pSK(+) cloning vector and transformation into *E. coli MRF'Kan* Supercompetent Cells followed the instructions of the PCR-Script™ *Amp* SK(+) Cloning Kit (Stratagene). The transformation mixture was incubated at 37 °C overnight in agar plates, ink-labelled, and plate-filter replicated. About 6000 recombinant clones from this library were screened independently with the synthetic probes $(TG)_{10}$, $(TC)_{10}$, $(GC)_{10}$, $(AT)_{10}$, $(CCT)_6$, $(CTG)_6$, $(CAG)_6$, $(AAT)_6$, 3′-end labelled with the DIG-oligonucleotide Tailing kit (Innogenetics). Replica filters were two-fold hybridised with 12 pmol per probe both at 45 °C and 55 °C, following the recommendations of the DIG-DNA Labelling and Detection kit (Innogenetics). Thirty-one double positive recombinant clones accounting for 16 kb were sequenced on both DNA strands with the BigDye Terminator method in an ABIprism 377 automatic DNA sequencer (Applied Biosystems). Microsatellite tracts were systematically screened on those clones as described below for database sequences. Repeats with reverse complements of each other and equivalent motifs (see Table 1) were considered a single repeat type.

2.2. Screening of microsatellites in databases

We made a systematic survey of published Bivalvia DNA sequences using the GenBank release 132.0 at the site http://www.ncbi.nlm.nih.gov/ between 23th October and 26th November 2002. A total of 3739 genomic sequences were retrieved in FASTA form and compared with the BLAST package (BLASTN 2.2.5. [Nov-16-2002] to avoid sequence redundancy. When two sequences, not coming from gene families, shared an identity above 90%, only a single match was considered. After this filtering step the number of sequences considered for Class Bivalvia was 3165 (the accession numbers are given as Supplementary Material at the author's website http://webs.uvigo.es/c03/webc03/XENETICA/XB4/xb4.htm). The nucleotides corresponding to the poly-A tail next to the 3′-end of cDNA clones were excluded from the analyses because most of these tracts are believed to arise from the polyadenylation of RNA transcripts. Taxonomic groups were defined using the updated taxonomy of Mollusca from the NCBI Taxonomy Browser (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi).

2.3. Microsatellite frequency in intergenic DNA

Unbiased estimates of microsatellite frequency from intergenic DNA entries presented several caveats: (1) intergenic DNA is underrepresented in databases, (2) most of the published microsatellites were isolated from enriched libraries and consequently their frequencies may be overestimated, (3) most publications of partial genomic libraries only report sequences containing polymorphic microsatellites that are useful as genetic markers (e.g., paternity tests) and (4) the estimation of microsatellite frequencies across species is biased due to the use of different probes among studies. However, since the number of positive clones in non-enriched libraries (excluding false positives) is usually reported, this allowed