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

Table 1
Observed number and rough abundance of microsatellite motifs in 3 Mb of *Mytilus galloprovincialis* genomic DNA

Motif [a]	No.	Spacing (kb)[b]	No. of loci ($\times 10^3$)[c]	GenBank accession no.
$(A/T)_n$, $(G/C)_n$	131	0.09	14000	AF445370–AF445375, AY102075–AY102095
$(AC/TG)_n$	5	429	2890	AF445370, AF445372, AF445373, AF445375
$(AT/TA)_n$	2	1072	1157	AF445370, AF445375
$(GC/CG)_n$	1	–	–	AF445374
$(CT/GA)_n$	1	–	–	AF445371
$(AAT/TTA)_n$	1	–	–	AF445374
$(AAAC/TTTG)_n$	1	–	–	AF445374
Total dinucleotides	9	238	5202	–
Total tri-tetranucleotides	2	1072	1156	–
Total (excluding polymononucleotides)	11	195	6358	–

[a] Being $n \geq 5$ for all motifs.
[b] Estimated spacing between consecutive microsatellite repeats in the intergenic DNA screened from libraries (2143 kb). Calculation of average spacing was only possible when at least two microsatellites were observed in the DNA screened. Spacing between polymononucleotides was directly estimated on the 16 kb DNA sequenced in this species, by considering a rough 70% of intergenic DNA (11200 bp).
[c] Calculated by extrapolating the spacing value to the genome size of *M. galloprovincialis* (1.77×10^9 bp; Rodríguez-Juíz et al., 1996), and considering a rough 70% of intergenic DNA (1.24×10^9 bp) in this genome.

us to roughly estimate the abundance of microsatellites per species in intergenic regions. In order to compare the rough frequency of microsatellites in intergenic DNA of *M. galloprovincialis* with other bivalves, we selected several species in which partial estimates of microsatellites were feasible, e.g., *Crassostrea gigas*, *Dreissena bugensis*, *D. polymorpha*, *Ostrea edulis*, and *Placopecten magellanicus*. Using the observed number of positive clones and the total DNA length screened in genomic libraries, we calculated the average spacing between two consecutive tracts in the intergenic DNA of those Bivalvia species. Since the genome fraction occupied by genes is inversely correlated to the genome size (e.g., 54% in *C. elegans* (*C*-value ≈ 0.1 Gb) and 10–15% in mammals (*C*-value ≈ 1 Gb) (Cavalier-Smith, 1978) due to the accumulation of repetitive DNA across genomes, and provided that the average Molluscan genome size (2.02 Gb in this study) overlaps with that of insects (0.8–7 Gb), we have considered as plausible that a 30% of Mollusca DNA corresponds to genes. Estimates of microsatellite spacing in intergenic DNA were then calculated as the number of loci (positive clones) related to the amount of intergenic DNA of bivalves (≈ 70%). For instance the estimated DNA length of *M. galloprovincialis* is 1.77×10^9 bp (conversion from the *C*-value given by Rodríguez-Juíz et al., 1996) and therefore should contain approximately 1.24×10^9 bp of intergenic DNA. This DNA length is useful for estimating the theoretical number of intergenic microsatellites in this species, given the spacing figures calculated from randomly-built genomic libraries described in Section 2.1 (see also Table 1 for calculation details).

The program GENEID (http://genome.imim.es/software/geneid/index.html) was applied to search for ORF in the 16 kb DNA sequenced from those libraries. Since no apparent coding frames were observed, a rough 70% of this DNA length (11200 bp) was considered as intergenic DNA, and used to calculate the microsatellite density in this window. Following the above estimation of gene abundance in Mollusca, the remaining 30% DNA from libraries would correspond to putative partial coding frames and/or to 5′and 3′ *UTR* that could be inadvertently present in the sequenced clones.

2.4. Analysis of microsatellites in exons and introns

From a non-redundant data set of about 100 Mb of Bivalvia DNA, we extracted all the nucleotide tracts with a repeat unit size ranging 2–10 bp, using the online computer program *Alex Dong Li's RepeatFinder* v0.4 (http://www.genet.sickkids.on.ca/~ali/repeatfinder.html), where the number of tandem repeats was allowed to be ≥5. We considered two loci as independent when two repeated tracts were separated by more than 5 bp. Database subsets were broken down into two genomic windows, e.g., *e*: exons and *i*: introns plus transcribed but not translated DNA (5′-*UTR* and 3′-*UTR*). Flanking regions 5′ and 3′ were defined as the adjacent parts of the DNA outwards from the 5′ and 3′ ends of a gene entry. The sequence length of each genome window as well as the microsatellite motifs found therein were added successively to the species account. In order to avoid that microsatellite frequencies would be overestimated they were calculated only for those species in which at least 3 kb were available.

We applied Robust Statistical Methods (e.g., Hampel, 1974) to study microsatellite frequencies in exons and introns. These methods are appropriate in case of contaminated distributions, overlapped or even highly deviated from normality. Their application in our study is justified by the scarcity of microsatellite frequencies recovered from available Bivalvia sequences and because these microsatellite

frequencies come from unknown distributions. All calculations were performed with the software of the postgraduate course "Métodos Avanzados de Estadística Aplicada" at UNED University (http://www.uned.es/experto-metodos-avanzados/). This program is a modification of the *R* package (http://www.r-project.org/index.html) that uses the bootstrap method of Efron and Tibshirani (1993) and includes calculations for Robust Methods (Hampel, 1974). Knowing the *location* and *dispersion* parameters could provide some interesting information about how microsatellite frequencies are distributed in exons and introns. Therefore, we calculated two statistical descriptors, e.g., the *Sample α-Winsorized Mean* for location and the *Sample α-Winsorized Deviation* for dispersion. We applied bootstrapping methods to calculate the *bias-corrected and accelerated* (BC_a) *bootstrap* confidence interval that includes both corrections for bias due to unequal variance and deviations of the statistical descriptor from the confidence interval. For further comparisons between mean microsatellite frequencies in exons and introns, we compared the results of two parametric and two non-parametric robust methods. The parametric methods were the Welch ANOVA and the Box ANOVA. The non-parametric methods were the Rust and Fligner Test, a robust analogous of Kruskal-Wallis, and the bootstrap with Box ANOVA.

2.5. Rough correlation between C-value and microsatellite spacing in intergenic DNA

The haploid DNA-length of species was calculated from the haploid DNA-content assuming the equivalences of 1 Da=1.67×10^{-24} g and 1 bp=649 Da. The DNA-content of Bivalvia species were taken from Hinegardner (1974), Rodríguez-Juíz et al. (1996), González-Tizón et al. (2000) and from the Animal Genome Size Database at site http://www.genomesize.com. Since the *C*-value of genus *Dreissena* was not available, we used the average value of its Order Veneroida (1.64, S.E.=0.43) as an approximate estimate. The *population percentage bend midcorrelation* (r_{bp}) is a robust analogue to the Spearman's *Rho* coefficient, that was used together with its contrast statistic (t_{bp}), to unveil any evidence of relationship between genome size and microsatellite abundance as independent variables. We used several breakpoint values (β), e.g., the recommended 0.1 and higher values (0.2 and 0.5) which increase the robustness of this test.

3. Results

The abundance of microsatellite motifs in the intergenic DNA window of *M. galloprovincialis* was roughly estimated from partial genomic libraries constructed in our laboratory. These figures have been compared with those recovered from the literature for other Bivalvia species. The frequency of microsatellites in exons and introns-*UTR* was also compared in Bivalvia through the screening of 3165 genomic sequences from 326 species, accounting for 101 Mb.

3.1. Microsatellite distribution in intergenic DNA

Excluding polymononucleotide repeats, the sequence $(AC/TG)_n$ was the most abundant microsatellite motif in *M. galloprovincialis* (Table 1), and the ratio $(AC)_n/(AT)_n$ was 3:1. An inverse relationship was observed between abundance and motif size. Dinucleotide repeats clearly outnumbered tri- and tetranucleotides, and the global abundance of microsatellites was approximately one repeat every 195 kb in this species. This figure provides a rough estimate of 6360 microsatellite loci in the genome of *M. galloprovincialis*. The large dispersion of microsatellite spacing figures in intergenic DNA of several species (Table 2) did not support any taxonomic patterning of microsatellite densities within Class Bivalvia. Closely related species showed marked differences in microsatellite spacing, e.g., the two species of the Family Ostreidae, *O. edulis* and *C. gigas*, differed by a factor of three, and also congeneric species showed conspicuous differences. The *population percentage bend midcorrelations* between *C*-value and microsatellite spacing in intergenic DNA showed nonsignificant correlations between both estimates (Table 3). In any case we could reject the null hypothesis of no correlation (r_{bp}=0), even increasing the breakpoint value to β=0.5 (data not shown).

3.2. Microsatellite distribution in exons and introns

The BC_a bootstrap confidence interval of introns was larger than that of exons, though their distributions were slightly overlapped (Table 4). The *Sample α-Winsorized mean* (x_α^W) of microsatellite frequencies was significantly lower in exons than in introns (Table 4), as was shown using parametric robust methods, i.e., the Welch ANOVA and the Box ANOVA (*F*-ratio=15.19, *p*-value=0.018) and the Rust and Fligner Test (p_{ij}=4.86, *p*-value=0.027). Also the Box ANOVA applying bootstrap revealed significant differences between the *Sample α-Winsorized* means for microsatellite frequencies at exons and introns (Table 5). Although

Table 2
C-value and spacing figures of microsatellite loci in intergenic DNA of several bivalvia species

Species	Spacing (kb)[a]	*C*-value (Gb)
Crassostrea gigas	172	0.84
Dreissena bugensis	438	1.51
Dreissena polymorpha	367	1.51
Mytilus galloprovincialis	71	1.77
Ostrea edulis	59	1.08
Placopecten magellanicus	90	1.94

[a] These calculations were performed considering the number of positive clones detected in genomic libraries and assuming 70% of intergenic DNA in the genome size screened.

Table 3
Robust correlation coefficient (r_{bp}) and its contrast statistic (t_{bp}) between *C*-value and microsatellite spacing in intergenic DNA

Breakdown point	r_{bp}	t_{bp}	*p*-value
β=0.1	−0.012	−0.025	0.981
β=0.2	−0.169	−0.343	0.749

bootstrap sampling distribution converges to the true sampling distribution rather quickly, the asymptotic values could be theoretically reached above 1000 replicates. In our study, the bootstrapping asymptotic value was reached above 7000 replicates. Excluding the marginal exception of the 5000 replicates case, the microsatellite frequencies differed significantly between exons and introns for any replicate number.

4. Discussion

4.1. Microsatellite density in M. galloprovincialis

Unbiased estimates of microsatellite abundance using all possible dinucleotide probes, showed that $(AC/TG)_n$ is the most abundant motif in *M. galloprovincialis*, excluding polymononucleotide motifs (Pérez et al., 2005), and that the ratio between the $(AC/TG)_n$ motif and the second most abundant motif (in our case $(AT/TA)_n$) was nearly 3:1. These data were also observed in many other eukaryotes (e.g., Ross et al., 2003), excluding plants where $(AT)_n$ is the most abundant motif (Lagercrantz et al., 1993). Overall, most eukaryotes show a high frequence of dinucleotide repeats compared to other motif classes, except yeast and fungi where di- and tetranucleotide repeats are the least abundant motifs (Tóth et al., 2000).

In *M. galloprovincialis* the average spacing between consecutive dinucleotide microsatellites (238 kb) and between higher size motifs (1072 kb), reflects an inverse relationship between motif size and microsatellite abundance. A possible explanation for this fact seems to be a higher slippage rate of short motifs (mono- and dinucleotide repeats) compared to larger ones (Kruglyak et al., 1998). When we extrapolate the density of microsatellites observed in libraries of *M. galloprovincialis* to its whole genome (3.84 pg, Rodríguez-Juíz et al., 1996), the density of $(AC)_{\geq 5}$ averages 1 every 429 kb, e.g., considerably less than in other bivalves such as the European flat oyster (1 every 139 kb, Naciri et al., 1995). Exceptions to the general $(AC)_n$ richness of eukaryotes have also been observed in other taxa such as birds, e.g., one $(AC)_{\geq 10}$ every 141–182 kb, assuming an even genomic distribution of repeats throughout the genome (Primmer et al., 1997). In that study it is argued that the $(AC)_n$ scarcity could be due to the short length of the avian genome and hence to the low amount of non-coding DNA, which could limit the expansion of microsatellites. Indeed, it has been shown that Simple Sequence Repeats are less abundant in exons than in non-coding regions (Hancock, 1995). Therefore, a low genomic proportion of intergenic DNA could be a factor influencing a low abundance of microsatellites. However, genome size alone cannot explain the $(AC)_n$ scarcity observed in the *M. galloprovincialis* genome (1.77×10^9 bp, Rodríguez-Juíz et al., 1996) that it is larger than the average avian genome (1.2×10^9 bp, see Primmer et al., 1997) but it is less enriched in $(AC)_n$ microsatellites. Moreover, *O. edulis* shows spacing values closer to avian estimates even if its genome (1.08×10^9 bp, Rodríguez-Juíz et al., 1996) is shorter than the average avian genome. The genomic abundance of microsatellites is probably not causally related to the genome size (e.g., *F. rubripes*; Elgar et al., 1999; Comeron, 2001) but to other species-specific molecular and evolutionary mechanisms such as recombination rates, and/or genomic mutation/repair rates (e.g., Deka et al., 1999).

Table 4
Statistical descriptors for data distributions of microsatellite frequency (expressed as number of loci per Mb) in exons (f_e) and in introns/*UTR* (f_i) of some bivalvia species: *Sample α-Winsorized mean* (x_α^W), *Sample α-Winsorized deviation* (S_W), Bias-corrected and accelerated bootstrap (BC_a) confidence intervals (95%) calculated using 1000 replicates

Species	f_e	f_i
Dreissena polymorpha	74	500
Spisula solidissima	119	303
Anadara trapezia	–	89
Mytilus edulis	47	198
Mytilus galloprovincialis	54	–
Crassostrea gigas	97	187
Crassostrea virginica	74	–
Mizuhopecten yessoensis	–	291
Pinctada fucata	403	–
x_α^W	84.52	245.03
S_W	27.80	59.56
BC_a Confidence Interval	(59.87, 221.35)	(141.72, 364.69)

Table 5
Box ANOVA applying the bootstrap method to the average microsatellite frequencies of exons and introns

Bootstrap replicates	Critical value[a]
1000	5941.95*
3000	6181.42*
5000	6493.55
7000	6166.73*
10000	6287.03*

Test value=6440.44, for each case.
[a] Critical values smaller than the test value indicate significant differences between *Sample α-Winsorized* means of microsatellite frequencies in exons and introns.
* Significant differences at α=0.05.

4.2. Interspecific comparisons

The number of positive recombinant clones recovered from genomic libraries is an indirect measure of the genome-specific microsatellite richness. A recent study showed a large difference in the number of microsatellites recovered from five species of *Drosophila* using the same

experimental protocols (Ross et al., 2003). Likewise the abundance of microsatellites was very dissimilar between evolutionarily close Bivalvia species. For instance, the two species of Family *Ostreidae*, *O. edulis* and *C. gigas*, showed microsatellite spacing values of 59 kb and 172 kb, respectively. Even the two species of genus *Dreissena*, *D. bugensis* and *D. polymorpha*, showed microsatellite spacing differences of about 80 kb. Those differences could be explained by different evolutionary properties of microsatellites between species, and presumably relate to species-specific mutation/repair mechanisms of replication slippage underlying those properties (Schug et al., 1998). For example, it has been shown in *E. coli* that the combination of a proofreading defect with a mismatch repair deficiency results in extreme microsatellite instability across the whole genome (Morel et al., 1998). This means that the mutation rate results from a balance between the mutability of genomes and the counteracting species-specific efficiency of the mismatch repair systems (e.g., Harr et al., 2002). This molecular balance could affect the dynamics of birth and death of microsatellites and influence their differential abundance between genomes much more than differences in genome size (e.g., Comeron, 2001).

4.3. Microsatellite frequency and C-value

Although gene numbers do not vary too much even between distantly related species, there are huge differences in their genome sizes. Known as the *C*-value paradox (Cavalier-Smith, 1978), this suggests that the unknown mutational or selective forces operating on genome size should be quite different between species. Similarly to the *C*-paradox, the number of microsatellites and the proportion of the genome they occupy, vary significantly between species as we show here for bivalves. The *C*-paradox and the microsatellite-paradox have led to believe in their causal relationship (e.g., Primmer et al., 1997). Microsatellites are believed to be neutral markers because most of them map at intergenic DNA and show extensive polymorphism. However, microsatellite mutations (expansion/contraction) are non-neutral either in exons (except for moderately trinucleotide expansions with low functional impact), in 5′ or 3′ *UTR* (except at non-regulatory sites) or in intronic DNA (except at those internal regions not essential for correct transcript maturation). This means that microsatellite distribution at the genome level should be shaped, at least, by the selective constraints acting differentially between genome windows of different functionality (e.g., Rose and Falush, 1998). Therefore, we would expect intergenic DNA to contain a larger microsatellite density than functionally relevant DNA regions such as exons or introns-*UTR*. If this is so, genomes with larger *C*-values should contain a corresponding larger density of microsatellites (e.g., Primmer et al., 1997). Our results do not show any statistical evidence that reinforces the existence of a relationship between genome size and microsatellite abundance, even after increasing the recommended breakpoint value and consequently the test robustness. This result is congruent with a recent study in vertebrates where no relationship was observed between the number of microsatellite loci and the genome size of several taxonomic groups (Neff and Gross, 2001). The above results suggest that the amount of intergenic DNA, though necessary for microsatellites to appear and expand, it is not sufficient to explain the microsatellite density of genomes.

4.4. Distribution of microsatellites in exons and introns

In this study, we have estimated separately the microsatellite densities of exons and introns. In the absence of any other knowledge about the population of microsatellite frequencies in databases, microsatellite frequency distributions in a random sample of size n from that population is the best guide to their distributions in the population. This is the main reason for using bootstrapping techniques in this study as recommended (Efron and Tibshirani, 1993), i.e., the use of 100 replicates to estimate variance or standard error and 1000 or more replicates to estimate a confidence interval. The power of parametric methods is higher than that of non-parametric methods, the former tending to be more sensitive when the data set meet test requirements. Nevertheless, the results obtained by applying bootstrap methods were concordant with those recovered without bootstrapping (e.g., robust parametric methods). These methods showed that in bivalves, the microsatellite frequency at exons (x_{α}^{W}=84.52) is lower than at introns (x_{α}^{W}=245.03) as indicated by all the tests performed, e.g., Robust ANOVAs, Rust and Fligner Test and Robust ANOVAs with bootstrap (Tables 4 and 5). In fact, the estimated number of microsatellites per Mb in Bivalvia ranged from 60 to 221 in exons and from 142 to 365 in introns/*UTR*, so the frequency of microsatellites partially overlapped between these windows. However, the distribution of microsatellites was significantly skewed toward upper frequencies in introns/*UTR*. This larger abundance of microsatellites in introns/*UTR* could be explained in terms of stronger selective constraints on exons (e.g., Liu et al., 2001; see Section 4.3). Letting alone the role played by nucleotide composition in the genesis of repeats at coding DNA (larger slippage propensity of AT-rich tracts respect to GC tracts), the effect of expansions/contractions on the structure and function of the encoded proteins would be a major selective force (Katti et al., 2001). Although introns/*UTR* microsatellite mutations can affect transcription and transcript maturation (e.g., Gebhardt et al., 2000), the evidence for selection against frameshift mutations limiting the expansion of repeats in exons (Metzgar et al., 2000), could explain their larger abundance in introns/*UTR*.

Unfortunately, the densities of microsatellites in genes could not be compared with those of intergenic DNA because of the different methodological approaches used to estimate them. However the present results suggest that

microsatellite frequencies in intergenic DNA would be higher than in exons or in introns/*UTR*, because of the relaxed selection on the former (e.g., Comeron, 2001). The study of microsatellite distribution on functionally differentiated genome windows is limited by the scarcity of genomic sequences available from bivalves. Genetic maps and comparative genomics of important cultured bivalves would provide deeper information about microsatellite distribution patterns and their species-specific evolutionary properties.

Acknowledgements

This research was supported by the Spanish Ministerio de Ciencia y Tecnología Grant BIO2001-3659 to PP and by PhD grants of Xunta de Galicia and Ministerio de Ciencia y Tecnología to FC and MP, respectively. Authors are grateful to Dr. Guillermo Thode from Universidad de Málaga and to Alfonso García Pérez (UNED) for their help in the statistical treatment. Special thanks to our fellows A.P. Diz and A. Seoane for their help in the laboratory. We also thank the invaluable contribution of the referees and the Corresponding Editor at significantly improving the final presentation of this manuscript.

References

Bachtrog, D., Weiss, S., Zangerl, B., Brem, G., Schlötterer, C., 1999. Distribution of dinucleotide microsatellites in the *Drosophila melanogaster* genome. Mol. Biol. Evol. 16, 602–610.

Cavalier-Smith, T., 1978. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J. Cell. Sci. 34, 247–278.

Comeron, J.M., 2001. What controls the length of noncoding DNA? Curr. Opin. Genet. Dev. 11, 652–659.

Deka, R., Guangyun, S., Smelser, D., Ahong, Y., Kimmel, M., Chakraborty, R., 1999. Rate and directionality of mutations and effects of allele size constraints at anonymous, gene-associated, and disease-causing trinucleotide loci. Mol. Biol. Evol. 16, 1166–1177.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Elgar, G., Clark, M.S., Meek, S., Smith, S., Warner, S., Edwards, Y.J.K., Bouchireb, N., Cottage, A., Yeo, G.S.H., Umrania, Y., Williams, G., Brenner, S., 1999. Generation and analysis of 25 Mb of Genomic DNA from the Pufferfish *Fugu rubripes* by sequence scanning. Genome Res. 9, 960–971.

Gebhardt, F., Burger, H., Brandt, B., 2000. Modulation of EGFR gene transcription by a polymorphic repetitive sequence—a link between genetics and epigenetics. Int. J. Biol. Markers 15, 105–110.

González-Tizón, A.M., Martínez-Lage, A., Rego, I., Ausió, J., Méndez, J., 2000. DNA content, karyotypes, and chromosomal location of 18S-5.8S-28S ribosomal loci in some species of bivalve molluscs from the Pacific Canadian coast. Genome 43, 1065–1072.

Hampel, F.R., 1974. The influence curve and its role in robust estimation. J. Am. Stat. Assoc. 69, 330–336.

Hancock, J.M., 1995. The contribution of slippage-like processes to genome evolution. J. Mol. Evol. 41, 1038–1047.

Harr, B., Todorova, J., Schlötterer, C., 2002. Mismatch repair driven mutational bias in *D. melanogaster*. Mol. Cell 10, 199–205.

Hinegardner, R., 1974. Cellular DNA-content of the Mollusca. Comp. Biochem. Physiol. 47, 447–460.

Katti, M.V., Ranjekar, P.K., Gupta, V.S., 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol. Biol. Evol. 18, 1161–1167.

Kornberg, A., Bertsch, L.L., Jackson, J.F., Khorana, H.G., 1964. Enzymatic synthesis of deoxyribonucleic acid: XVI. Oligonucleotides as templates and the mechanisms of their replication. Proc. Natl. Acad. Sci. U. S. A. 51, 315–323.

Kruglyak, S., Durrett, R.T., Schug, M.D., Aquadro, C.F., 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. U. S. A. 95, 10774–10778.

Lagercrantz, U., Ellegren, H., Andersson, L., 1993. The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. Nucleic Acids Res. 21, 1111–1115.

Liu, Z., Li, P., Kocabas, A., Karsi, A., Ju, Z., 2001. Microsatellite-containing genes from the Channel Catfish brain: evidence of trinucleotide repeat expansion in the coding region of nucleotide excision repair gene RAD23B. Biochem. Biophys. Res. Commun. 289, 317–324.

Metzgar, D., Bytof, J., Wills, C., 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res. 10, 72–80.

Morel, P., Reverdy, C., Michael, B., Ehrlich, D., Cassuto, E., 1998. The role of SOS and flap processing in microsatellite instability in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. 95, 10003–10008.

Naciri, Y., Vigoroux, Y., Dallas, J., Desmarais, E., Delsert, C., Bonhomme, F., 1995. Identification and inheritance of (GA/TC)n and (AC/GT)n repeats in European flat oyster *Ostrea edulis* (L.). Mol. Mar. Biol. Biotechnol. 4, 83–89.

Neff, B.D., Gross, M.R., 2001. Microsatellite evolution in vertebrates: inference from AC dinucleotide repeats. Evolution 55, 1717–1733.

Pérez, M., Cruz, F., Presa, P., 2005. Distribution properties of polymononucleotide repeats in molluscan genomes. J. Heredity 96, 1–12.

Primmer, C.R., Raudsepp, T., Chowdhary, B.P., Møller, A.P., Ellegren, H., 1997. Low frequency of microsatellites in the avian genome. Genome Res. 7, 471–782.

Rodríguez-Juíz, A.M., Torrado, M., Méndez, J., 1996. Genome-size variation in bivalve molluscs determined by flow cytometry. Mar. Biol. 126, 489–497.

Rose, O., Falush, D., 1998. A threshold size for microsatellite expansion. Mol. Biol. Evol. 15, 613–615.

Ross, C.L., Dyer, K.A., Erez, T., Miller, S.J., Jaenike, J., Markow, T.A., 2003. Rapid divergence of microsatellite abundance among species of Drosophila. Mol. Biol. Evol. 20, 1143–1157.

Sambrook, J., Fritsch, E.F., Maniatis, T., 1989. Molecular cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, New York.

Schilthuizen, M., 2002. Mollusca: an evolutionary cornucopia. Trends Ecol. Evol. 17, 8–9.

Schug, M.D., Wetterstrand, K.A., Gaudette, M.S., Lim, R.H., Hutter, C.M., Aquadro, C.F., 1998. The distribution and frequency of microsatellite loci in Drosophila melanogaster. Mol. Ecol. 7, 57–70.

Sokolov, E.P., 2000. An improved method for DNA isolation from mucopolysaccharide-rich Molluscan tissues. J. Molluscan Stud. 66, 573–575.

Tóth, G., Gáspári, Z., Jurka, J., 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10, 967–981.

Wilder, J., Hollocher, H., 2001. Mobile elements and the genesis of microsatellites in dipterans. Mol. Biol. Evol. 18, 384–392.