Long Perfect Dinucleotide Repeats Are Typical of Vertebrates, Show Motif Preferences and Size Convergence

Paulo Almeida and Carlos Penha-Gonçalves
Instituto Gulbenkian de Ciência, Oeiras, Portugal

Microsatellites are simple sequence repeats (SSRs) showing complex patterns of length, motif sizes, motif sequences, and repeat perfection. We studied the structure of the dinucleotide SSR population at the genome level by analyzing assembled DNA sequence across species. Three dinucleotide populations were distinguished when SSR genome frequency was analyzed as a function of repeat length and repeat perfection. A population of low-perfection SSRs was identified, which is constituted by short repeats and represents the vast majority of genomic dinucleotide SSRs across eukaryotic genomes. In turn, the highly perfect repeats are 30 to 50 times less frequent and, in addition to short repeats, also contain a long repeat population that is uniquely represented in vertebrate species. Distinctive features of this population include the modal peak in the frequency distribution of repeat length and the strong preferential usage of the repeat motifs AC and AG. These results raise the hypothesis that the ability of carrying a distinct population of long, highly perfect dinucleotide repeats in the genome is a late acquisition in chordate evolution. Our analysis also suggests that different dinucleotide repeat populations have different dynamics and are likely to be underlined by different molecular mechanisms of generation and maintenance in the genome. Thus, these observations imply that caution should be taken in extrapolating results from studies on SSR mutability and on SSR phylogenetic comparisons that do not take into account the stratification of dinucleotide populations in the eukaryotic genome.

Introduction

Microsatellites, or simple sequence repeats (SSRs), are generally defined as genetic loci where one to six nucleotides are repeated in tandem. SSRs are present in eukaryotic genomes at higher frequencies than would be expected from random distribution, based on nucleotide composition (Tautz and Renz 1984; Metzgar, Bytof, and Wills 2000). The genomic abundance of the SSRs and their high degree of polymorphism have prompted many investigators to propose mechanisms of microsatellite mutability to explain observed patterns of SSR evolution across species and SSR variability within populations (Li et al. 2002). It is commonly assumed that SSR maintenance in the genome depends on a balance between the addition and deletion of repeat units, through replication slippage, and the correction of these errors by proofreading and mismatch repair (MMR) enzymes (Harfe and Jinks-Robertson 2000). In addition, point mutations could have an effect in limiting the expansion of microsatellites, by degrading perfect repeat sequences (Kruglyak et al. 2000). In turn, slippage is proposed to have the opposite effect, by purifying microsatellites of these interruptions (Harr and Schlötzer 2000). Recombination mechanisms are also suggested to play a role in SSR mutation, by unequal crossing over or gene conversion (Majewski and Ott 2000). However, a number of mutation biases related to the nature of the SSRs have been observed, including the increase of mutation rate with allele size and the higher rates of contraction for long alleles (Xu et al. 2000; Huang et al. 2002). These observations suggest that the mechanisms of SSR mutability are complex and may depend on several factors, including the structure of the SSRs.

The studies on microsatellite variability are frequently focused on very long, highly mutable loci, which are unrepresentative of the majority of simple repeats in the genome. On the other hand, several studies on microsatellite evolution compare closely related species representing very limited branching of the phylogenetic tree. In fact, the comparisons of microsatellite length and structure at orthologous loci in related species tend to analyze sets of loci selected on the basis of high polymorphism in the species from which they were isolated. As reasoned by H. Ellegren and colleagues (Webster, Smith, and Ellegren 2002), such interspecific comparisons are likely to be flawed due to ascertainment bias (Fitzsimmons, Moritz, and Moore 1995; Forbes et al. 1995; Ellegren et al. 1997). In addition, the criteria used for defining a microsatellite locus differ from one study to another and are not always clearly stated, making it difficult to compare results and to evaluate claimed findings. Most studies only considered completely perfect repeats, although it has been observed that several long repeats contain one or a few base substitutions, causing them to be counted as two or more shorter repeats. For instance, long alleles of CA repeats were found to have increased stability due to a TA interspersion (Bacon, Farrington, and Dunlop 2000) but there are few studies allowing for mismatches when scanning for SSRs (Katti, Ranjekar, and Gupta 2001). Here, we pursued an in silico analysis of the population structure of dinucleotide repeats across eukaryotic species, repeat motifs and repeat lengths to which we added an extra dimension by searching for differences in repeat perfection. To capture the bulk of the dinucleotide repeat population we aimed to work without a priori repeat selection criteria and used assembled DNA sequence available from current genome projects. This analysis revealed the existence of distinct groups of dinucleotide repeats that have specific structural properties and particular phylogenetic distributions.

Methods
Genomic DNA Sequence

All genome sequences were downloaded in FASTA format from the National Center for Biotechnology Information database.
behave independently of one another. For simplicity, counted as two repeats (see table 2) and are assumed to repeated stretches separated by a large interruption are sequence data that do not represent SSRs. Particularly, a low threshold of identity would risk the inclusion of repeats) and also because accepting short sequences with time and processing power it would take to include smaller sequences (table 1). We show the results of analyzing two mammals (human and mouse), one fish (Fugu rubripes), one invertebrate chordate (Cliona intestinalis), one insect (Drosophila melanogaster), and one nematode (Caenorhabditis elegans). The analysis incorporated three different parameters simultaneously: (1) the frequency of SSRs per megabase of DNA; (2) the number of repeated dinucleotide units; and (3) the identity of the repeated sequence. Identity is a measure of repeat perfection and was defined as the quotient between the length of repetitive DNA and the total size of the SSR. The length of repetitive DNA is found by counting, within the SSR, all the nucleotides belonging to a given repeat motif. This means that interspersed nucleotides will decrease identity (table 2).

Our analysis clearly shows that in all species analyzed, the vast majority of dinucleotide SSRs are shorter than 15 units long, and under 97% identity (fig. 1) and their frequency in the eukaryotic genome is considerably constant varying within the range 500-1,000/Mb, depending on the species considered (table 3). These SSRs were here classified as low-identity repeats and represent a population in which the frequency distribution exponentially decays with both repeat length and repeat identity. To assure that the low-identity repeats were not a set of random sequences, we performed several analyses with randomly generated sequences. Such simulations led to the conclusion that low-identity microsatellites occurred in the genome in much higher frequencies than would be expected by chance (data not shown).

The dinucleotide SSRs showing identity above 97% have a relatively low frequency in the genome, as they are 15–80 times less frequent than the low-identity repeats (fig. 2). The high-identity SSRs were detected in all the eukaryotic genomes analyzed, from yeast to human. Among the non-vertebrate species, the frequency of high-identity SSRs decays exponentially with the repeat length in a similar curve as compared to the low-identity

<table>
<thead>
<tr>
<th>Species</th>
<th>Source</th>
<th>Total Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>Chr 1</td>
<td>220 Mb</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>Chr 1</td>
<td>100 Mb</td>
</tr>
<tr>
<td>Fugu rubripes</td>
<td>Unknown</td>
<td>330 Mb</td>
</tr>
<tr>
<td>Cliona intestinalis</td>
<td>Whole genome</td>
<td>117 Mb</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>Whole genome</td>
<td>123 Mb</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>Whole genome</td>
<td>100 Mb</td>
</tr>
</tbody>
</table>


### Results

#### Low-Identity Repeats

To investigate the heterogeneity of dinucleotide repeats across the eukaryotic genomes, we performed an analysis of significant amounts of assembled genomic sequence (table 1). We show the results of analyzing two mammals (human and mouse), one fish (Fugu rubripes), one invertebrate chordate (Cliona intestinalis), one insect (Drosophila melanogaster), and one nematode (Caenorhabditis elegans). The analysis incorporated three different parameters simultaneously: (1) the frequency of SSRs per megabase of DNA; (2) the number of repeated dinucleotide units; and (3) the identity of the repeated sequence. Identity is a measure of repeat perfection and was defined as the quotient between the length of repetitive DNA and the total size of the SSR. The length of repetitive DNA is found by counting, within the SSR, all the nucleotides belonging to a given repeat motif. This means that interspersed nucleotides will decrease identity (table 2).

Our analysis clearly shows that in all species analyzed, the vast majority of dinucleotide SSRs are shorter than 15 units long, and under 97% identity (fig. 1) and their frequency in the eukaryotic genome is considerably constant varying within the range 500-1,000/Mb, depending on the species considered (table 3). These SSRs were here classified as low-identity repeats and represent a population in which the frequency distribution exponentially decays with both repeat length and repeat identity. To assure that the low-identity repeats were not a set of random sequences, we performed several analyses with randomly generated sequences. Such simulations led to the conclusion that low-identity microsatellites occurred in the genome in much higher frequencies than would be expected by chance (data not shown).

#### High-Identity Repeats

The dinucleotide SSRs showing identity above 97% have a relatively low frequency in the genome, as they are 15–80 times less frequent than the low-identity repeats (fig. 2). The high-identity SSRs were detected in all the eukaryotic genomes analyzed, from yeast to human. Among the non-vertebrate species, the frequency of high-identity SSRs decays exponentially with the repeat length in a similar curve as compared to the low-identity

### Table 1

Length of DNA Sequences Scanned for Each Species

<table>
<thead>
<tr>
<th>Species</th>
<th>Source</th>
<th>Total Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>Chr 1</td>
<td>220 Mb</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>Chr 1</td>
<td>100 Mb</td>
</tr>
<tr>
<td>Fugu rubripes</td>
<td>Unknown</td>
<td>330 Mb</td>
</tr>
<tr>
<td>Cliona intestinalis</td>
<td>Whole genome</td>
<td>117 Mb</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>Whole genome</td>
<td>123 Mb</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>Whole genome</td>
<td>100 Mb</td>
</tr>
</tbody>
</table>

### Table 2

Length of Repeats and Identity Scoring for Illustrative SSRs

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Repetitive DNA (nt)</th>
<th>Total Length (nt)</th>
<th>Identity (%)</th>
</tr>
</thead>
</table>
| GGACACACACACACACACACACACACACACACAGG | 30 | 30 | 100
| GGACACACACACACACACACACACACACACACAGG | 29 | 30 | 96.7
| GGACACACACACACACACACACACACACACACAGG | 27 | 30 | 90
| GGACACACACACACACACACACACACACACACAGG | 24 | 30 | 80
| GGACACACACACACACACACACACACACACACAGG | 21 | 30 | 70
| GGACACACACACACACACACACACACACACACAGG | 19 | 30 | 63.3
| GGACACACACACACACACACACACACACACACAGG | 22 | 30 | 73.3
| GGACACACACACACACACACACACACACACACAGG | 5/5 | 100/100
| GGACACACACACACACACACACACACACACACAGG | 16/14 | 16/14 | 100/100

Note: SSR scored as high identity, low identity; under minimum identity (not scored), under minimum length (not scored), and two repeats scored (compound repeat).
repeats (fig. 2). This suggests that the mechanisms that limit the elongation of low- and high-identity repeats could be similar in nonvertebrate species. We noted that the long high-identity SSRs population was present in the genome of vertebrates like human, mouse, and fish (Fugu rubripes). Strikingly, this population was virtually absent in the chordate C. intestinalis, in the fly D. melanogaster and in the nematode C. elegans (fig. 2). This long high-identity population was also evident in the rat and in the fish Danio rerio, but was not present in the mosquito Anopheles gambiae, in the plant Arabidopsis thaliana, or in the yeast Saccharomyces cerevisiae (data not shown). When identity is included as a variable in this analysis, the long high-identity repeats show a population characterized by a repeat count higher than 15 and an identity higher than 91% (fig. 3). As the long high-identity dinucleotide repeats were found in the genome of vertebrates, but not in the genome of lower species, we suggest that the ability to generate and carry long, perfect SSRs was a late acquisition in the eukaryotic genome evolution within the chordate phylum.

Regardless of repeat perfection, the short repeat populations (under 15 repeat units) follow a power distribution

---

**Fig. 1.**—Distribution of dinucleotide repeats drawn from genomic DNA of the indicated species. Frequency represents the number of repeats per megabase of analyzed genomic DNA and is presented as a function of both the number of repeat units (repeat count) and the percent of repeated sequence (identity) found across the sequence of each SSR.
in the species analyzed and the frequency of both low-
identity and high-identity SSRs exponentially decays as 
repeat length increases (fig. 2). In contrast, the long high-
identity SSRs display a modal peak in the frequency 
distribution (fig. 3). In fact, the average size of the long 
high-identity repeats varies from 21 to 25 repeat units,
in vertebrate species, suggesting that the convergence mecha-
nisms constraining this SSR population are conserved. 
These observations suggest that the long high-identity SSRs 
are under a distinct population dynamics and that the 
mechanisms underlying the generation and maintenance of 
these repeats are virtually absent in nonvertebrates.

<table>
<thead>
<tr>
<th>Species</th>
<th>Low Identity</th>
<th>Short High Identity</th>
<th>Long High Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Freq</td>
<td>AC %</td>
<td>AG %</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>961</td>
<td>27</td>
<td>40</td>
</tr>
<tr>
<td>M. musculus</td>
<td>1089</td>
<td>37</td>
<td>37</td>
</tr>
<tr>
<td>F. rubripes</td>
<td>1046</td>
<td>49</td>
<td>32</td>
</tr>
<tr>
<td>C. intestinalis</td>
<td>908</td>
<td>27</td>
<td>10</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>1061</td>
<td>35</td>
<td>18</td>
</tr>
<tr>
<td>C. elegans</td>
<td>574</td>
<td>22</td>
<td>35</td>
</tr>
</tbody>
</table>

* The low-identity population contains repeats up to 91% identity, the short high-identity population is defined by identity higher than 97% and repeat count shorter 
than 15, and the long high-identity population is defined by identity higher than 91% and longer than 15 repeat counts. Frequency is expressed as number of SSRs per megabase.

![Graphs](image-url)
Repeat Motif Usage

The low-identity repeats show no clear preferential use of repeat motifs, although it was noted that nonvertebrate species use AT repeats more frequently (table 3). We did not find any common trend in motif usage among nonvertebrate species when comparing low and high-identity repeats. However, within each of the vertebrate species, the relative usage of repeat motifs in high-identity repeats differs markedly from that observed for low-identity repeats. Specifically, the use of non-AT repeats is increased both in short and long high-identity repeats, when comparing them to the low-identity SSRs (table 3). Our results clearly show that AC dinucleotide repeats are the major contributors for the long high-repeats population, in human, mouse, and fish (table 3 and fig. 4). AG and AT repeats were also present in the mouse but are almost absent in human and fish. This indicates that the composition of high-identity repeats in vertebrate species is influenced by strong repeat preference of the long perfect repeats and supports the notion that the mechanisms of generation of long high-repeats may have specific

Fig. 3.—Distribution of long dinucleotide repeats drawn from genomic DNA of the indicated species. Frequency represents the number of repeats per megabase of analyzed genomic DNA and is presented as a function of both the number of repeat units (repeat count) and the percent of repeated sequence (identity) found across the sequence of each SSR.
Discussion

In contrast with previous analyses of dinucleotide repeats, which focused on perfect or near perfect repeats, we used the repeat identity as a variable in our analysis, together with repeat frequency and repeat length. This allowed us to discriminate three populations of dinucleotide repeats and to reveal a complex stratification of these SSRs in the genome. These dinucleotide populations display markedly different frequency in the genome, length distribution, and repeat motif usage, suggesting that distinct mechanisms are in play to their generation, expansion, and maintenance. The identification of a population of long high-identity repeats, in mammals and fish, but not in nonvertebrate species, suggests that the ability to carry, generate, and/or maintain these SSRs is a late acquisition in the chordate genome evolution. When
scanning the genome for SSRs (especially in human studies), this population could be easily contaminated by other dinucleotide SSRs if locus selection does not use a stringent identity criterion, or if short repeats are included (see figs. 2 and 3). The identification of this population is in accordance with the findings by Xu et al. (2000), who reported an increase in contraction rates with allele length and suggested this mechanism might explain the stationary allele distribution of microsatellites from different species. Such a mechanism would also explain the modal peak in the population of long, perfect dinucleotides observed here.

The findings reported here raise the issue that many studies, which do not take into account the stratification of the dinucleotide SSR population, may be flawed due to ascertainment bias or confounding and should be analyzed carefully. In fact, many discrepancies have been reported on SSR mutation rates within and between species, and between types of repeat motifs (Weber and Wong 1993; Chakraborty et al. 1997; Kruglyak et al. 1998; Schug et al. 1998). For instance, studies focused on Drosophila dinucleotide repeats to calculate mutation rates (Bachtrog 2000) and to study repeat frequency (Katti, Ranjekar, and Gupta 2001) may lead to illegitimate interpretations when establishing comparisons with mammalian genomes that carry high-identity repeats.

A similar argument can be made about theoretical models of SSR mutability that are applied to calculate time of divergence between species, among vertebrates and nonvertebrates. If the SSRs used in these calculations are drawn from SSR populations with different dynamics, this could partly explain the inconsistency of the results obtained (Calabrese, Durrett, and Aquadro 2001; Zhivotovsky, Rosenberg, and Feldman 2003). If there are mechanisms stabilizing the long high-identity SSRs, and causing their length to converge, this effect might counter the slippage mutations, on which is based the rationale to determine the distance between species, or strains. Calabrese, Durrett, and Aquadro (2001) found that two different methods for determining genetic distances, based on microsatellites lengths, systematically underestimate divergence time. We noted that the markers used by these authors, to compare human/chimp populations, mostly fall in the 15–30 repeat units range, where the distribution is normal and repeat lengths are likely to be under convergence constraints.

Another example of SSR ascertainment biases comes from studies based on microsatellite markers that are identified by laboratory methods that strongly favor the collection of perfect repeats. For instance, we verified that the average identity of mouse AC dinucleotides, in the Whitehead Institute/MIT database, is 96% and the length frequency distribution is similar to that observed for the long high-identity repeats. Several studies consistently showed that microsatellite markers, when isolated in one organism, tend to be longer than their orthologs in other organisms. This bias is likely caused by the laboratory methods that favor the selection of polymorphic markers and may be reflected in studies on SSR evolution using such genetic markers (Ellegren et al. 1997).

In addition, a significant number of studies on microsatellites just consider perfect repeats although it has been observed that several long repeats contain one or a few base substitutions, causing them to be counted as two or more shorter repeats (Katti, Ranjekar, and Gupta 2001). Such ascertainment methods artificially reduce the size of the repeats under analysis. In fact, we show that the perfection among the long high-identity SSR population varies from 91% to 100%. We treated SSR of different motifs as independent repeats although sometimes they may occur close together (see table 2). Such compound repeats do not fall into the identity criteria used in our analysis. However, the analysis of compound repeats has been seldom addressed, and it is a field for further investigation as it may indicate specific mechanisms underlying the generation of these repeats.

Our findings raise the hypothesis that long high-identity SSRs represent a late acquisition of the eukaryotic genome. Several mechanisms have been proposed to regulate the evolution of microsatellites including a DNA repair mechanism involving strand discrimination that would be responsible for the equilibrium of SSR distributions (Xu et al. 2000). In fact, slippage rates alone cannot explain the overall mutation rate of these loci, which can be the opposite of the primary slippage mutation rate (Harr, Todorova, and Schlötterer 2002). A mechanism of strand discrimination could account for the presence of high-identity SSRs in human and mouse and it could also explain the absence of this SSR population in organisms like D.melanogaster, whose DNA is not methylated, except in embryonic stages (Lyko, Ramsa-hoye, and Jaenisch 2000). It is possible that the lack of a strand discrimination mechanism might lead to simple deletion of any loops formed during replication, thus causing Drosophila microsatellites to be generally shorter and to lack the high-identity population. Recent studies suggested the need for supplementary mutation mechanisms, other than the slippage-point mutation model, to explain differences in the distribution of dinucleotide populations with different length and with low identity (Dieringer and Schlötterer 2003; Sibly et al. 2003).

Another possibility is that the size of the SSRs could be regulated by mismatch repair (MMR) enzymes that have not been isolated in organisms where long high-identity SSRs are absent. We noted that as opposed to human and mouse, the MLH3 and MSH3 orthologs of MutL and MutS (part of the Escherichia coli MMR system) have not been identified so far in the fly and C. elegans (Marti, Kunz, and Fleck 2002). These MMR enzymes are apparently involved in repairing frameshift intermediates and insertion/deletion loops and tend to cause insertion mutations (Harfe and Jinks-Robertson 2000; Kearney et al. 2001).

In this context, the mismatch repair system is a good candidate to modulate the outcome of the mutation process and regulate the size of SSRs. Therefore, we speculate that components of the MMR machinery, like the MSH3 protein, may lead to the production of the observed long high-identity repeats.

An alternative explanation for the existence of the normally distributed SSR population might be the selection of some function that could be assigned to long high-identity SSRs sized within a certain range. Candidate
functions include influence on recombination, gene expression and transcription, perhaps by enabling DNA to assume specific conformational changes (Li et al. 2002).

Acknowledgments

We acknowledge Instituto Gulbenkian de Ciência for financing our work and the Fundação para a Ciência e Tecnologia for supporting Paulo Almeida and Carlos Penha-Gonçalves.

Literature Cited


Pierre Capy, Associate Editor

Accepted January 21, 2004