Available online at www.sciencedirect.com
ScienceDirect
Genomics 88 (2006) 323–332
GENOMICS
www.elsevier.com/locate/ygeno

Tandem repeats in the CpG islands of imprinted genes

Barbara Hutter [a], Volkhard Helms [a], Martina Paulsen [b,*]

[a] Bioinformatik, FR 8.3 Biowissenschaften, Universität des Saarlandes, Postfach 151150, D-66041 Saarbrücken, Germany
[b] Genetik/Epigenetik, FR 8.3 Biowissenschaften, Universität des Saarlandes, Postfach 151150, D-66041 Saarbrücken, Germany

Received 21 December 2005; accepted 30 March 2006
Available online 11 May 2006

Abstract

In contrast to most genes in mammalian genomes, imprinted genes are monoallelically expressed depending on the parental origin of the alleles. Imprinted gene expression is regulated by distinct DNA elements that exhibit allele-specific epigenetic modifications, such as DNA methylation. These so-called differentially methylated regions frequently overlap with CpG islands. Thus, CpG islands of imprinted genes may contain special DNA elements that distinguish them from CpG islands of biallelically expressed genes. Here, we present a detailed study of CpG islands of imprinted genes in mouse and in human. Our study shows that imprinted genes more frequently contain tandem repeat arrays in their CpG islands than randomly selected genes in both species. In addition, mouse imprinted genes more frequently possess intragenic CpG islands that may serve as promoters of allele-specific antisense transcripts. This feature is much less pronounced in human, indicating an interspecies variability in the evolution of imprinting control elements.

Keywords: Imprinting; CpG islands; Tandem repeats; DNA methylation; Repetitive elements; Epigenetics; Regulatory elements

To date, approximately 40 imprinted genes have been identified in the human and mouse genomes, respectively. Since their monoallelic expression depends on the parental origin, it is assumed that distinct DNA elements in imprinted genes are differentially recognized and modified in the parental germ lines and that the resulting epigenetic marks are maintained after fertilization [1]. Therefore, it has been questioned what sequence features distinguish imprinted from nonimprinted genes [2,3]. In several studies it was observed that many imprinted genes possess an unusual density of repetitive retroviral elements. Imprinted genes showed a depletion of short interspersed transposable elements (SINEs) [2] and an enrichment of long interspersed nuclear element 1 (LINE-1) repeats [4,5] compared to randomly selected genes. Although it was shown that LINE-1 and some SINEs are indeed differentially marked in the female and male germ lines [6–9], there is so far no evidence that these methylation marks can be stably maintained after fertilization. Thus, it is likely that additional elements are needed. Potential candidate elements for this purpose are so-called differentially methylated regions (DMRs) that exhibit distinct DNA methylation patterns on their parental alleles. Many DMRs have been shown to acquire a specific methylation pattern in one germ line but remain unmethylated in the other germ line [10]. Such DMRs represent key regulatory elements that are responsible for imprinted expression of neighboring genes, thereby contributing to the domain-like organization of imprinted genes.

Most DMRs overlap with CpG islands. These are characterized by a high density of CpG dinucleotides that can be targeted by DNA methylation. However, CpG islands are regulatory elements that are also common to biallelically expressed genes [11,12]. Usually they are located in the promoter region [13–15]. In contrast, DMRs of imprinted genes are not always related to transcriptional start sites of protein-coding genes. Some DMRs reside in introns where they serve as promoters for antisense transcripts [16–19]. Other DMRs are located upstream of imprinted genes, at some distance from the promoter region [20–22]. These DMRs may act as insulator or silencer elements as was shown for the DMR 5′ of the imprinted *H19* gene [23,24].

* Corresponding author. Fax: +49 0 681 302 2703.
E-mail address: m.paulsen@mx.uni-saarland.de (M. Paulsen).

0888-7543/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2006.03.019

Table 1
General sequence properties of human and mouse imprinted genes and control groups

Group	Total sequence length (bp)	Median gene length (bp)	Total G + C content (%)	G + C content in repetitive elements (%)	Total CpG content (%)	CpG content in repetitive elements (%)
Mouse imprinted	2,452,775	20,697	45.43 ± 4.65	43.90 ± 3.18	1.28 ± 0.46	0.87 ± 0.42
Mouse ctrl1	5,864,555	21,603	46.37 ± 4.22	45.46 ± 2.43	1.35 ± 0.47	0.93 ± 0.29
Mouse ctrl2	5,615,657	15,103	45.87 ± 4.11	45.59 ± 2.30	1.35 ± 0.52	0.93 ± 0.26
Mouse all	13,932,987	20,912	45.98 ± 4.26	45.21 ± 2.62	1.34 ± 0.49	0.95 ± 0.35
Human imprinted	3,259,027	26,583	46.93 ± 7.68	44.96 ± 5.65	1.82 ± 0.97	1.54 ± 0.79
Human ctrl1	6,181,623	27,992	46.27 ± 6.37	45.82 ± 4.18	1.66 ± 0.83	1.68 ± 0.88
Human ctrl2	6,783,964	24,270	46.02 ± 6.14	46.29 ± 4.37	1.70 ± 0.83	1.72 ± 0.87
Human all	16,224,614	27,247	46.30 ± 6.53	45.84 ± 4.57	1.71 ± 0.86	1.67 ± 0.86

Sequence lengths were accumulated per group. For mouse, undetermined nucleotides were subtracted. Columns 4 to 7 list means and standard deviations. The control groups (79 genes each) have approximately twice the number of sequences as the imprinted groups (39 murine and 38 human genes). Median gene lengths and G + C content are similar in imprinted and control groups.

Frequent features of DMRs are arrays of tandem repeat motifs, and it has been suggested that such repeats are involved in the regulation of imprinting [25]. In addition to a potential role in imprinting, tandem repeat arrangements of DNA sequences are likely to be involved in various other epigenetic silencing and heterochromatin formation processes [25–27]. Many multicopy transgene arrays in mammalian genomes show increased methylation levels and reduced gene expression compared to corresponding single-copy transgenes [28]. Similarly, tandem repeat arrangements can attract DNA methylation in meiotic processes in filamentous fungi [29]. Tandem repeats are found within or adjacent to both paternally and maternally methylated DMRs. The so far described tandem repeat motifs vary from 5 to 400 bp in length (Supplementary Material Table S8, [5]). Although repeat arrays have been identified in mouse as well as in human, only a few orthologous DMRs possess highly conserved tandem repeat motifs. Likewise, the number of repeated motifs and their arrangement within the DMR are highly variable in different species [19,21, 30,31]. Even though the frequent appearance of tandem repeat arrays associated with imprinted genes has been noted in many publications (see references in [5], Supplementary Material Table S8), it has so far not been systematically investigated whether such sequence elements are indeed enriched in imprinted genes compared to biallelically expressed genes.

Therefore, we compared in this work the CpG islands of imprinted genes against CpG islands of randomly selected genes in mouse and human and investigated the positions of CpG islands within the genes and their overlaps with repetitive elements. Searching specifically for tandem repeat arrays, we observed a significant enrichment of these motifs in CpG islands of imprinted genes.

Results

General sequence properties

For the analyses of CpG islands, genomic sequences of 39 murine and 38 human genes that showed pronounced imprinting effects in at least one of the two species were selected from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) and Ensembl (http://www.ensembl.org) databases. The genomic DNA segment used for each gene comprised the entire intragenic region as annotated in the database plus 10 kb upstream of the transcriptional start site and 10 kb downstream of the last exon. Likewise, two control groups, each including 79 randomly selected genes, were collected for mouse and for human (see Material and methods). In most cases, genes used in the human control groups were also included into the murine groups. Thus, similar to the selected imprinted genes, the murine and human control groups contained a large number of homologous genes (Supplementary Material Table S1). Since it cannot be excluded that the random selection of genes is biased, for example toward strongly expressed genes or toward sequences with a high G + C content, we collected the two control groups using different approaches. Actual differences between imprinted and nonimprinted genes should differ significantly for both control groups. As the assembly/annotation of the mouse genome is still not finished, some mouse sequences contain stretches of undefined nucleotides (termed N), with an average N content of 0.71% (standard deviation 1.22%). The transcribed regions of most genes are less than 25 kb long (Table 1). In general, mouse genes are shorter than human ones [32]. The gene length distribution is similar in imprinted and randomly selected genes (Wilcoxon–Mann–Whitney test, $p > 0.4$).

Comparing the G + C content did not reveal any notable differences, neither between human and mouse nor between imprinted and control sequences. The mean is 46.30% for human and 45.99% for mouse. The standard deviation of the G + C content is larger in human (6.53%) than in mouse (4.26%), indicating that the G + C content of some human sequences is more extreme, as was reported before [32]. Human sequences have a mean CpG content of 1.71% (std dev 0.86%) and are significantly enriched in CpG dinucleotides compared to mouse with 1.34% (std dev 0.49%) (t test, $p < 0.001$). This difference is predominantly caused by a higher CpG content of interspersed repeat elements in human sequences. SINE repeats are significantly reduced in imprinted genes ([2], Supplementary Material Table S2). In all analyzed groups the CpG and G + C contents are linearly