Conservation of Repeats at the Mammalian KCNQ1OT1-CDKN1C Region Suggests a Role in Genomic Imprinting

Evolutionary Bioinformatics
Volume 13: 1–14

DOI: 10.1177/1176934317715238

Marcos De Donato[1,2], Tanveer Hussain[1,3], Hectorina Rodulfo[2], Sunday O Peters[4], Ikhide G Imumorin[1,5,6] and Bolaji N Thomas[7]

[1]Animal Genetics and Genomics Laboratory, Office of International Programs, College of Agriculture and Life Sciences, Cornell University, Ithaca, NY, USA. [2]Escuela de Bioingenierias, Tecnologico de Monterrey, Campus Querétaro, Santiago de Querétaro, Mexico. [3]Department Molecular Biology, Virtual University of Pakistan, Lahore, Pakistan. [4]Department of Animal Science, Berry College, Mount Berry, GA, USA. [5]African Institute for Biosciences Research and Training, Ibadan, Nigeria. [6]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA. [7]Department of Biomedical Sciences, Rochester Institute of Technology, Rochester, NY, USA.

ABSTRACT: KCNQ1OT1 is located in the region with the highest number of genes showing genomic imprinting, but the mechanisms controlling the genes under its influence have not been fully elucidated. Therefore, we conducted a comparative analysis of the KCNQ1/KCNQ1OT1-CDKN1C region to study its conservation across the best assembled eutherian mammalian genomes sequenced to date and analyzed potential elements that may be implicated in the control of genomic imprinting in this region. The genomic features in these regions from human, mouse, cattle, and dog show a higher number of genes and CpG islands (detected using cpgplot from EMBOSS), but lower number of repetitive elements (including short interspersed nuclear elements and long interspersed nuclear elements), compared with their whole chromosomes (detected by RepeatMasker). The KCNQ1OT1-CDKN1C region contains the highest number of conserved noncoding sequences (CNS) among mammals, where we found 16 regions containing about 38 different highly conserved repetitive elements (using mVista), such as LINE1 elements: L1M4, L1MB7, HAL1, L1M4a, L1Med, and an LTR element: MLT1H. From these elements, we found 74 CNS showing high sequence identity (>70%) between human, cattle, and mouse, from which we identified 13 motifs (using Multiple Em for Motif Elicitation/Motif Alignment and Search Tool) with a significant probability of occurrence, 3 of which were the most frequent and were used to find transcription factor–binding sites. We detected several transcription factors (using JASPAR suite) from the families SOX, FOX, and GATA. A phylogenetic analysis of these CNS from human, marmoset, mouse, rat, cattle, dog, horse, and elephant shows branches with high levels of support and very similar phylogenetic relationships among these groups, confirming previous reports. Our results suggest that functional DNA elements identified by comparative genomics in a region densely populated with imprinted mammalian genes may be related to the regulation of imprinted gene expression.

KEYWORDS: genomic imprinting, conservation, mammalian, gene expression, transcription factors, repetitive elements

RECEIVED: March 8, 2017. **ACCEPTED:** May 23, 2017.

PEER REVIEW: Four peer reviewers contributed to the peer review report. Reviewers' reports totaled 1085 words, excluding any confidential comments to the academic editor.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by College of Agriculture and Life Sciences, Cornell University, Ithaca, NY and Pfizer Animal Health, Inc. (now Zoetis, Inc.). Additional support was provided by National Research Initiative Competitive Grant Program (grant no. 2006-35205-16864) from the USDA National Institute of Food and Agriculture, USDA-NIFA Research Agreements (nos. 2009-65205-05635, 2010-34444-20729), and USDA Federal formula Hatch funds appropriated to the Cornell University Agricultural Experiment Station. Visiting fellowships were awarded to T.H. by the Higher Education Commission of Pakistan.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Bolaji N Thomas, Department of Biomedical Sciences, Rochester Institute of Technology, Rochester, NY 14623, USA. Email: bolaji.thomas@rit.edu

Introduction

Genomic imprinting is the expression of only one allele from the chromosome of a specific parent, instead of the alleles of both parents, which is different from monoallelic expression where no bias exists toward a specific parental allele. This has been described in insects,[1–4] higher plants,[5–7] and mammals, but the latter is where it has mostly been studied.[8–12] This pattern of expression ranges from a small but significant bias toward one parental allele to a complete shutdown of one of the parental alleles.[13] To date, out of the about 25 000 genes described in the human genome, only 212 have been reported to show imprinted expression, 123 in mouse and 20 in cattle (Geneimprint, http://www.geneimprint.com). Even though many of these genes have not been studied in most mammals, they are predicted to be conserved due to their important roles in the development of the placenta, embryo, fetus, and neurons, among other functions.[14]

The p15.5 region of chromosome 11 in the human genome contains the highest density of imprinted genes, which is divided into 2 regions, 1 more telomeric with the H19-TH gene cluster and another toward the centromere, containing genes clustered around ASCL2-OSBPL5, where each block is independently regulated by imprinting control regions (ICRs) through differentially methylated regions (DMRs).[15] Mutations in any of these regions produce changes in the gene expression patterns leading to diseases such as the Beckwith-Wiedemann and Silver-Russell syndromes, producing either overgrowth or undergrowth of the fetus/newly born, respectively, as well as other complications.[14]

The control mechanism in the telomeric region containing the KCNQ1OT1 gene cluster has not been as well studied as the mechanisms in the H19-TH region, and it seems to be

more complex, implicating at least 2 types of control and regulating no less than 9 genes.[16] The KvDMR control element located at the promoter region of the KCNQ1OT1 gene is methylated in the maternal allele, thereby inhibiting the expression of this gene but allowing the expression of the maternal alleles of other genes. In the paternal chromosome, KvDMR is not methylated, allowing the binding of CTCF at this region and silencing the maternally expressed genes through its isolating property, and the formation of a DNA loop, mediated by the binding of the long, noncoding RNA from KCNQ1OT1.[17] This RNA also associates with the maternally expressed genes inducing its silencing via histone hypoacetylation and methylation of CpGs, which produces the condensation of the chromatin.[16] However, the mechanisms controlling the region affected by KvDMR and KCNQ1OT1 are complex and require the identification of all the regulatory sequences and their relative locations, their tissue specificity, and their 3-dimensional architecture to fully understand their developmental role in vivo.[16] Significantly, this region is an excellent model to understand mechanistically how enhancers, insulators, noncoding RNAs, and target genes are deployed to generate the appropriate expression outputs in the endogenous context. To this end, we conducted a comparative analysis of this region to study its conservation across several well sequenced eutherian mammalian genomes and to elucidate potential controlling elements participating in the control of genomic imprinting in the KCNQ1OT1-CDKN1C region.

Materials and Methods

We conducted a comparative analysis of the region containing the highest density of genes showing genomic imprinting, by examining the genomes of several mammalian species with high sequence depth. In *Homo sapiens*, this region, located in the p15.5 region of chromosome 11 is 1.2 Mbp in size, extending across H19 to OSBPL5 genes. For the gene nomenclature, we followed the guidelines for HUGO Gene Nomenclature Committee (HGNC[18]).

Genome sequences and analytical tools

The genomic sequences from human (NC_000011.10, assembly GRCh38.p7, TaxID: 9606), marmoset (NC_013906.1, assembly 3.2, TaxID: 9483), horse (NC_009155.2, assembly EquCab2.0, TaxID: 9796), dog (NC_006600.3, assembly CanFam3.1, TaxID: 9615), cattle (AC_000186.1, assembly UMD_3.1.1, TaxID: 9913), mouse (NC_000073.6, assembly GRCm38.p4 C57BL/6J, TaxID: 10090), rat (NC_005100.4, assembly Rnor_6.0, TaxID: 10116), elephant (NW_003573565.1, assembly Loxafr3.0, TaxID: 9785), and chicken (NC_006092.3, assembly 5.0, TaxID: 9031) were obtained from the GenBank (www.ncbi.nlm.nih.gov/genbank) for the purpose of comparative analysis. Gene annotation were performed on GenBank, using the latest version of Map Viewer (January 26, 2004) for each species. CpG islands were identified using the program cpgplot from EMBOSS (V: 2.0; www.ebi.ac.uk/Tools/seqstats/emboss_cpgplot), which defines an island as a region where the proportion of CGs and/or GCs observed over the expected was above 0.6 and with the %G + %C content above 50%, calculated as a sliding average over 10 windows with a minimum size of 100 nucleotides.[19] The number and type of repetitive sequences were detected with RepeatMasker, version 4.0.5 (www.repeatmasker.org/cgi-bin/WEBRepeatMasker).[20] The genomic sequences were compared using the program LAGAN from the suite mVISTA (V: 2.0; genome.lbl.gov/vista/index.shtml),[21] which performs progressive pairwise alignments, guided by a phylogenetic tree, aligned to other alignments using the sum-of-pairs metric.[22]

Comparative genome analysis between mammals

Matched sequences with 70% or higher identity with human were used to find sequence motifs, a sequence pattern repeated in a group of DNA sequences, with the Multiple Em for Motif Elicitation (MEME) suite (Motif-based sequence analysis tools, version 4.10.0; meme.nbcr.net/meme/tools/meme). We calculated the log-likelihood ratio of the occurrences producing a probability that the motif is found by chance. A motif was taken as statistically significant when the probability of occurrence was below 0.05, with motifs between 6 and 100 nucleotides, located in 1 or more positions for each sequence. The program Motif Alignment & Search Tool (MAST) was used to find the motifs, identified by MEME in the conserved sequences. We calculated the match scores for each conserved sequence converted into various types of *P* values that were used to determine the overall match of the sequence to the motifs and the probable order and spacing of occurrences of the motifs in the sequences.[23]

The most frequent motifs found in the conserved sequences were used to search for transcription factor binding sites through the JASPAR suite (version 5.0_ALPHA; jaspar.genereg.net).[24] This is a collection of transcription factor DNA-binding preferences, modeled as matrices, converted into position weight matrices or position-specific scoring matrix and used for scanning those sequences. The JASPAR CORE database contains a curated, nonredundant set of profiles from published articles, where the transcription factor–binding sites were experimentally defined for multicellular eukaryotes.[25] The prime differences to similar resources (TRANSFAC, etc) consist of open data access, nonredundancy, and quality of the binding sites.

Data analysis and statistics

In addition, the conserved noncoding sequences (CNS) found in the introns of the region KCNQ1/KCNQ1OT1-CDKN1C with sequence identity >70%, (which was the region with the highest conservation) were divided into repetitive conserved noncoding sequences (RCNS) and

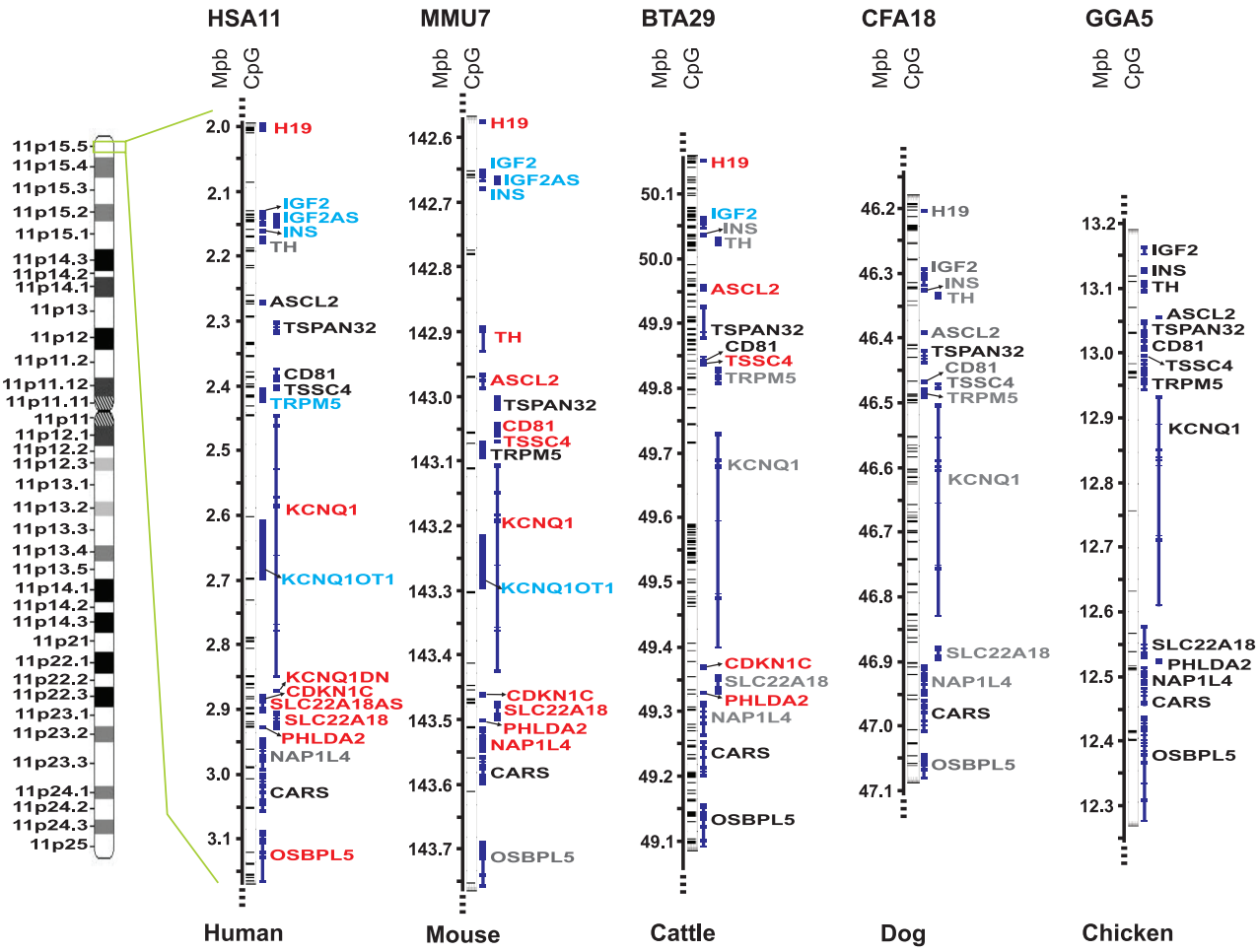

Figure 1. Region in the genomes of human (HSA11), mouse (MMU7) cattle (BTA29), dog (CFA18), and chicken (GGA5) containing the highest density of genes with genomic imprinting. The genes in red have shown preferential maternal expression, those in blue preferential paternal expression, the ones in gray have no evidence of either pattern of expression and those in black have shown biallelic expression. The measures of the maps are in millions of base pairs (Mbp). The genes are represented with their exons (boxes) and introns (thin lines). The black horizontal lines represent the CpG islands present in each genome.

unique conserved noncoding sequences (UCNS). Along with the coding sequences (CDS) for the genes KCNQ1 and CDKN1C, they were used to determine evolutionary pattern by determining the phylogenetic relationships among human, marmoset, horse, dog, cattle, mouse, rat, and elephant. First, the Neighbor-Joining Method was used,[26] computing the evolutionary distances by Maximum Composite Likelihood method.[27] Second, the Maximum Likelihood method[28] with a discrete Gamma distribution was used to model evolutionary rate differences among sites. Both analyses were conducted in MEGA7.[29] Finally, a Bayesian phylogenetic analysis was conducted using MrBayes, v 3.2.1,[30] implementing the general time-reversible model with the rate at each site as random variable with a gamma distribution (G) and a proportion of invariable sites. A bootstrap test[31] of 1000 replicates was used to determine the statistical support of the branches in the most likely tree. The evolutionary distances used to infer the phylogenetic tree were computed using the Maximum Composite Likelihood method[27] and are in the units of the number of base substitutions per site. In the analysis, all positions containing gaps and missing data were eliminated.

Results

Comparing the region containing the imprinted genes in the block p15.5 of human chromosome 11 with those of mouse, cattle, dog, and chicken, we found an extensive conservation in the gene order and number of intron/exons (Figure 1), although the size of the introns and intergenic spacing was highly variable. In addition, we found CpG islands in almost all of the promoter regions, which should be involved in the control of gene expression.

There were many CNS between the human genome and those of marmoset, horse, dog, cattle, mouse, rat, and elephant in the introns and the intergenic spacing in the whole region (Figure 2). Even though KCNQ1OT1 is only annotated in the human and mouse genomes, its high conservation in the other mammals suggests the possibility that it is also present in these genomes. In the rat genome, there is no sign of conservation because the sequences are undetermined for this region. The

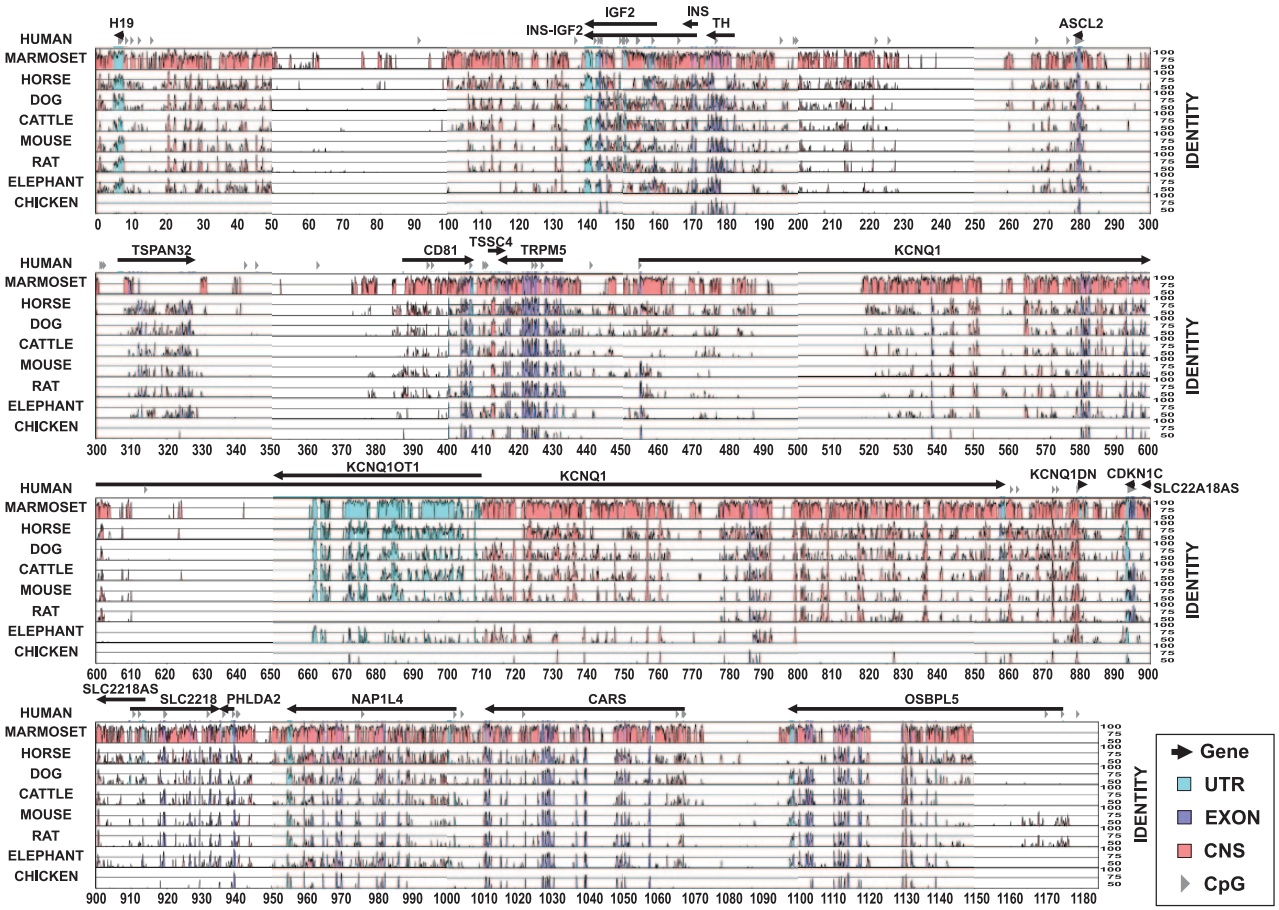

Figure 2. Comparison of the human DNA sequences with those of other mammals. The conserved noncoding sequences (CNS) of intergenic and intronic regions are shown in red, as well as the untranslated transcribed regions (UTR) and exons of the genes are shown in light and dark blue, respectively, whereas CpG islands are shown as gray arrowheads. The numbers in the *X* axis correspond to thousands of base pairs (kbp). The right *Y* axis to the left shows the percentage of identity to the human sequences.

KvDMR element located at the promoter and first part of the transcription sequence of KCNQ1OT1 shows very high conservation. The proportion of CNS between the human and the horse genomes is the highest, followed by dog and cattle, and much less than with the mouse and rat, even though rodents have been shown to be the closest group to primates. In addition, all the CDS show high conservation between human and all other mammalian genomes as well as with chicken.

By analyzing the region containing imprinted genes compared with the rest of the chromosomes, we found a higher gene density per million base pairs (Mbp) in the imprinted region in the genomes of human, mouse, cattle, and dog but not for chicken (Figure 3). A similar pattern was shown for the density of CpG islands, although, there are differences on the density among species, with mouse showing the lowest density of CpG islands and cattle showing about 6 times higher density than mouse. In addition, all mammalian regions show a lower density of repetitive elements than the rest of the chromosome. In chicken, however, there is no difference between this region and the rest of the chromosome.

To delineate the different types of repetitive sequences, we found a significant reduction in short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs) in the imprinting region among all the species, with some exceptions as shown in Figure 4. The Alu SINEs or Alu-like (shown as other SINEs) were significantly reduced in all the species. Long interspersed nuclear elements show the highest reduction in LINE2 sequences for human and mouse; the highest reduction in LINE1 was found in dog, whereas cattle showed a reduction in both LINE1 and LINE2. In human and mouse, we observed a significant reduction in the mammalian-wide interspersed repeats (MIRs) sequences, which are otherwise increased in dog.

In the intergenic space between KCNQ1/KCNQ1OT1 and CDKN1C, we also found a very high conservation, including the human gene KCNQ1DN, which is not annotated in the other species. In the chicken genome, there were 11 short CNS in the introns of KCNQ1 and KCNQ1DN and in the intergenic spaces (Figure 5), both in the comparison with the unmasked and masked human sequences (related to UCNS), even though genomic imprinting has not been reported in chicken. To identify the CNS associated with repetitive elements, in the region with the highest number of CNS, located between the genes KCNQ1/KCNQ1OT1 and CDKN1C, we compared the

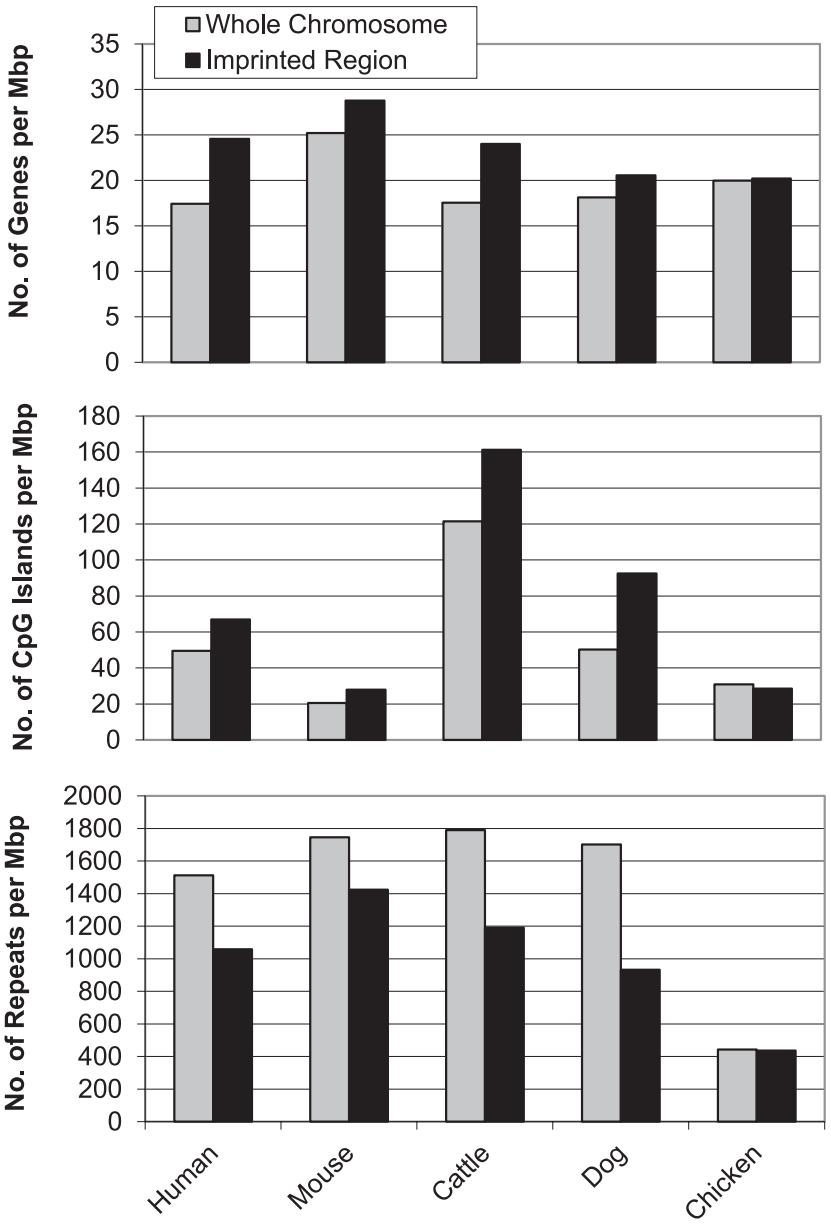

Figure 3. Genomic features of the regions containing the genes showing genomic imprinting and the whole chromosome where those regions are located for the human, mouse, cattle, dog, and chicken genomes.

sequences of human, cattle, and mouse, using chicken as an outgroup, with the human repetitive sequences unmasked and then masked. We identified the presence of CNS containing highly conserved repetitive elements (>70% identity), which disappeared when the human repeats were masked (Figure 5), evidencing the presence of RCNS in addition to the UCNS.

We identified 16 regions of sizes between 94 and 1519 nucleotides long, containing a total of 74 CNS with identities >70%, showing different repetitive elements. In these CNS, 38 different types of repetitive elements were detected, most of them of type LINE1, such as L1M4, L1MB7, HAL1, L1M4a, and L1Med, as well as 1 LTR element: MLT1H (Table 1). The frequencies of these elements identified in the CNS were a lot higher in the imprinted region than in the whole chromosome, implying they were distributed differently, with elements mostly or only found in these CNS. Interestingly, 19 CpG islands were detected in the region, all of which were located in segments with CNS.

The 74 CNS identified were used to detect conserved motifs with the program MEME, being present in more than 1 CNS. Our analysis identified 13 motifs, showing probabilities of random occurrence less than 0.05 (Table 2), of which motifs 2, 9, and 13 were the most frequently found and in higher number in the CNS. There was no significant association between any pair of motifs. However, of the 74 CNS, only 40 showed the presence of these motifs (Table 3). The repetitive elements where the motifs were most frequently found were the LINE1 elements: L1M4, L1MB7, L1Med,

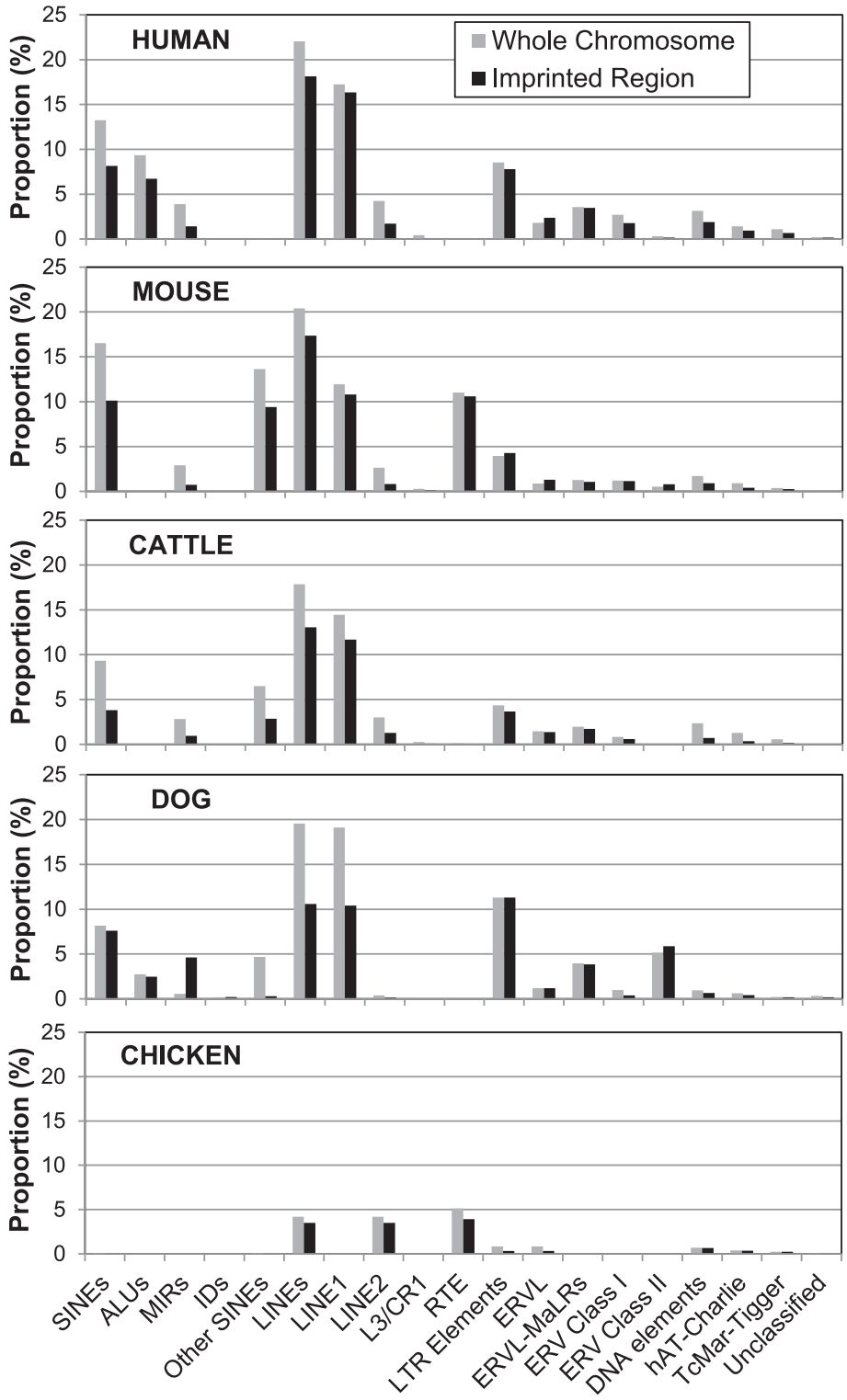

Figure 4. Comparison of the proportion of each type of repetitive element found in the regions containing the genes showing genomic imprinting and the whole chromosome where those regions are located for the human, mouse, cattle, dog, and chicken genomes.

and L1PA5. The consensus sequences of motifs 2, 9, and 13 that were used to find transcription factors binding sites using JASPAR indicated 40 different transcription factors that could recognize these motifs, revealing most of them to be associated with transcription factor families SOX, FOX, and GATA (Table 4).

Using CNS from the region between KCNQ1OT1 and CDKN1C genes, as shown in Figure 5, we could construct phylogenetic trees showing the relationships among mammals, whose genomes have been totally sequenced. The branches in the tree reveals a very high level of support, with rodents shown to be the most related group to primates and carnivores, as well

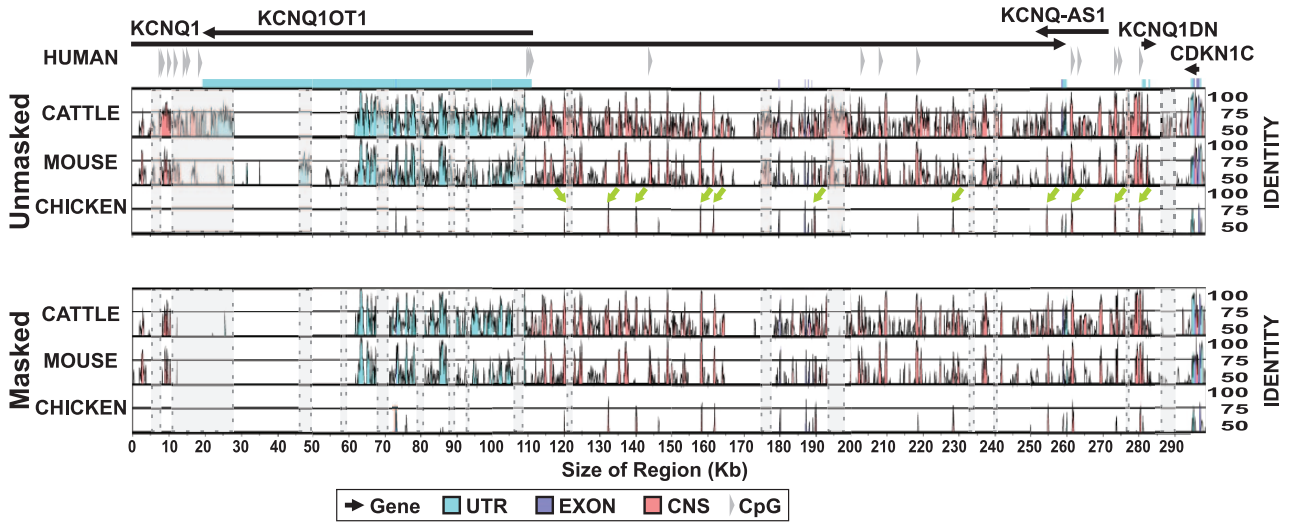

Figure 5. Comparison of the human DNA sequences with the sequence of cattle, mouse, and chicken, as outgroup, for the region between the genes KCNQ1OT1 and CDKN1C where the human repetitive elements have been unmasked or masked. The conserved noncoding sequences (CNS) of intergenic and intronic regions are shown in red, as well as the untranslated transcribed regions (UTR) and exons of the genes are shown in light and dark blue, respectively, whereas CpG islands are shown as gray arrowheads. The numbers in the *X* axis correspond to thousands of base pairs (kbp). The right *Y* axis to the left shows the percentage of identity to the human sequences.

Table 1. Repetitive elements associated to conserved noncoding sequences (CNS) from Figure 5, ordered according to the number of times present in the 16 conserved regions analyzed, showing also how many times per Mbp they are found in the imprinted region and in the whole chromosome.

TYPE OF REPEAT	CLASS OR FAMILY	ORIENTATION	CNS	NUMBER (MBP)	
				REGION	CHROMOSOME
L1M4	LINE/L1	C/+	9	63.5	2.9
L1MB7	LINE/L1	C	6	35.3	1.6
HAL1	LINE/L1	+	6	31.7	1.4
L1M4a	LINE/L1	C	5	<<1	<<1
MLT1H	LTR/ERVL-MaLR	C/+	5	10.6	0.5
L1MEd	LINE/L1	C	4	14.1	0.6
Charlie1	DNA/hAT-Charlie	+	2	3.5	0.2
Charlie25	DNA/hAT-Charlie	+	2	17.6	0.8
L1M2	LINE/L1	C	2	14.1	0.6
L1MC5a	LINE/L1	C/+	2	7.1	0.3
L1ME4b	LINE/L1	C	2	10.6	0.5
L1MEg	LINE/L1	C	2	17.6	0.8
MIRb	SINE/MIR	C/+	2	42.3	1.9
MLT1H1	LTR/ERVL-MaLR	+	2	7.1	0.3
AluJr4	SINE/Alu	C	1	3.5	0.2
AluSz	SINE/Alu	C	1	10.6	0.5
L1M5	LINE/L1	C	1	14.1	0.6
L1MB5	LINE/L1	C	1	3.5	0.2
L1MC	LINE/L1	C	1	10.6	0.5

Table 1. (Continued)

TYPE OF REPEAT	CLASS OR FAMILY	ORIENTATION	CNS	NUMBER (MBP)	
				REGION	CHROMOSOME
L1MC3	LINE/L1	C	1	7.1	0.3
L1MC4	LINE/L1	C	1	3.5	0.2
L1MD1	LINE/L1	C	1	14.1	0.6
L1ME1	LINE/L1	+	1	7.1	3.5
L1ME3Cz	LINE/L1	+	1	3.5	0.2
L1ME4a	LINE/L1	+	1	10.6	0.5
L1ME4c	LINE/L1	+	1	3.5	0.2
L1ME5	LINE/L1	C	1	3.5	0.2
L1PA5	LINE/L1	C	1	3.5	0.2
L2b	LINE/L2	+	1	45.9	2.1
LTR84b	LTR/ERVL	+	1	3.5	0.2
MER1B	DNA/hAT-Charlie	+	1	7.1	0.3
MIR1_Amn	SINE/MIR	C	1	24.7	1.1
MIR3	SINE/MIR	C	1	38.8	1.8
MIRc	SINE/MIR	C	1	45.9	2.1
MSTB	LTR/ERVL-MaLR	+	1	3.5	0.2
Tigger18a	DNA/TcMar-Tigger	+	1	7.1	0.3
UCON55	DNA/TcMar-Tigger	C	1	3.5	0.2
X6B_LINE	LINE/CR1	+	1	3.5	0.2

The orientation refers to the sense (+) or antisense (C) direction of the repeat.

as Perissodactyla shown to be very close together and related to Cetartiodactyla (Figure 6). Elephants were shown to be the most basal group among the mammals studied and as such was used as outgroup.

Discussion

So far, genomic imprinting has been confirmed in a little over 200 human genes (Geneimprint, www.geneimprint.com), but the total number of genes with imprinting has been estimated to be between several hundred to 2000, using whole genome search methods.[32-35] Many of these genes show imprinting in a tissue-specific or stage-specific manner,[36] which makes it more difficult to detect this type of expression because the analysis should use the right tissue and stage to be able to detect this pattern. Because of this, computational methods have been used to predict genomic imprinting, based on the features of the DNA sequences, CpG islands, repetitive elements, blocks of micro RNA, and epigenetic modifications,[33,35,37,38] because these features can be detected at any point in time in the DNA. However, the success rate of these methods is still low due to the lack of knowledge of all the signals and features associated with genomic imprinting, especially those defining the temporal and spatial control of imprinting expression.[39] Thus, this study attempts to further identify DNA elements associated with imprinting control in the selected region.

Here, we identify repetitive sequences located in regions highly conserved in several mammalian species. Because of its repetitive nature and the millions of years of divergence among them, we assumed that the conservation of these elements can only be associated with an important function in the control of gene expression, as previously suggested for CNS that are highly conserved.[40] Differentially methylated regions are one of the elements implicated in the control of blocks containing imprinted genes because they allow the marking of alleles to be turned on or shut down when transcription factors/enhancers do not recognize or start recognizing these sites, and these markers can be maintained for many rounds of cell division.[14] These DMRs are usually located in CpG islands that are highly conserved in human and mouse genomes,[40,41] which is consistent with our findings here in those highly conserved CpG islands in the KCNQ1OT1-CDKN1C region in all the mammals analyzed.

Name	Sequence Logo	Probability	N	CNS
Motif 1		>0.0001	10	6
Motif 2		>0.0001	47	27
Motif 3		>0.0001	4	4
Motif 4		>0.0001	5	5
Motif 5		>0.0001	6	6
Motif 6		>0.0001	5	5
Motif 7		>0.0001	9	9
Motif 8		>0.0001	6	5
Motif 9		>0.0001	32	19
Motif 10		0.0071	4	4
Motif 11		0.0013	6	6
Motif 12		0.0480	5	4
Motif 13		0.0013	43	29

Table 2. Sequence logo of the nucleotides in the motifs identified by the MEME program in the CNS from the regions in Figure 5. In each motif, the *Y* axis shows the relative frequency of each nucleotide. The probability refers to the likelihood of identifying a motif at random in the sequences analyzed. The number of times the motif is found in all the regions and in the CNS is shown.

Table 3. Human CNS containing repeats with high sequence identity where the motifs were detected.

START	END	SIZE	PROBABILITY	MOTIF 2	MOTIF 9	MOTIF 13	REPEATS
2596080	2596936	856	9.50E−27	3	0	0	L1ME1
2602330	2602812	482	5.50E−38	1	1	0	L1MD1
2605116	2605661	545	4.80E−51	1	0	1	L1M4
2605687	2606310	623	5.10E−39	1	0	2	L1M4
2606724	2606899	175	3.00E−08	1	0	0	L1M4
2607199	2607376	177	0.00016	1	0	0	L1M4
2608093	2608438	345	7.30E−26	0	0	2	L1MB7
2608918	2609139	221	2.10E−06	0	0	1	L1MB7
2610566	2610811	245	2.50E−14	1	0	2	L1MB7
2610827	2611252	425	3.10E−07	1	1	1	L1M4
2611513	2612011	498	0.00013	1	0	0	L1M4a1
2612422	2613608	1186	1.20E−20	3	2	2	L1M4a2
2613609	2614154	545	0.036	0	2	1	L1MC5a
2614771	2615448	677	2.00E−46	1	0	2	L1M4
2635703	2637219	1516	1.30E−217	5	1	4	L1PA5
2636053	2636861	808	6.00E−162	1	0	1	L1PA5
2636933	2637219	286	8.50E−29	3	1	1	L1PA5
2647199	2647379	180	7.00E−15	0	0	1	L1M2
2657493	2658770	1277	1.60E−13	1	5	2	L1MC3
2659913	2660309	396	2.70E−20	1	0	1	L1MC3
2676843	2677521	678	3.60E−05	0	2	0	Tigger18a
2677665	2678144	479	1.10E−08	0	2	1	L1ME4b
2678177	2678848	671	0.00078	0	0	1	L1ME4b
2682169	2682552	383	0.016	0	2	0	L2b
2695163	2695882	719	2.10E−15	0	3	0	L1ME5, L1M4
2696123	2696925	802	8.30E−26	1	2	1	L1M4
2697056	2697892	836	5.70E−20	2	0	1	L1M4
2709442	2710316	874	2.00E−67	1	2	1	L1MB7
2710401	2711083	682	1.10E−23	1	0	0	L1MB7
2710657	2711083	426	1.10E−69	1	0	0	L1MB7
2764437	2764818	381	2.60E−18	0	0	1	L1Meg
2765545	2766293	748	4.10E−08	2	1	2	L1Meg
2783496	2784821	1325	2.70E−125	3	0	3	L1Med
2784964	2785317	353	7.80E−34	2	0	1	L1Med
2785987	2787505	1518	1.30E−25	3	1	3	L1Med
2822845	2823092	247	1.90E−06	2	1	1	HAL1
2828998	2829814	816	5.80E−11	3	1	1	HAL1
2865770	2866004	234	9.90E−11	0	0	1	L1ME4a
2866067	2866454	387	1.20E−11	0	1	0	MIRc
2875331	2875482	151	5.20E−05	0	1	0	Charlie25

Abbreviation: CNS, noncoding sequences.
The probability refers to the likelihood of finding at random 1 of the 13 identified motifs in the sequences analyzed and the frequency of the motifs 2, 9, and 13 which were the most frequent. The repeats found in each CNS are shown. The position of the sequences is according to the GenBank (accession number: NC_000011.10).

Table 4. Transcription factor–binding sites detected by JASPAR in the main motifs identified in the noncoding sequences.

TRANSCRIPTION FACTORS	MOTIFS			TOTAL
	2	9	13	
SOX3	7	0	8	15
FOXP1	2	0	5	7
SOX10	4	0	3	7
SOX6	4	0	2	6
GATA3	1	0	5	6
MZF1_5-13	0	5	0	5
MZF1_1-4	0	5	0	5
HLTF	1	0	4	5
GATA4	0	0	5	5
FOXP2	2	0	2	4
GATA2	0	0	4	4
GATA1	0	0	4	4
FOXD3	2	0	2	4
FOXO1	2	0	2	4
SP1	0	3	0	3
KLF5	0	3	0	3
ZNF263	1	0	1	2
MECOM	0	0	2	2
FOXI1	2	0	0	2
ZNF354C	0	2	0	2
MEF2C	1	0	1	2
FOXL1	0	0	1	1
CRX	0	0	1	1
EGR1	0	1	0	1
E2F1	0	1	0	1
SPIB	0	0	1	1
SOX5	1	0	0	1
E2F4	0	1	0	1
IRF1	0	0	1	1
EGR2	0	1	0	1
SOX9	1	0	0	1
EN1	0	0	1	1
SPI1	0	0	1	1
SP2	0	1	0	1
NKX2-5	0	0	1	1
KLF4	0	1	0	1
CDX2	0	0	1	1
E2F6	0	1	0	1
SOX17	1	0	0	1
SOX2	1	0	0	1
Total	33	25	58	116

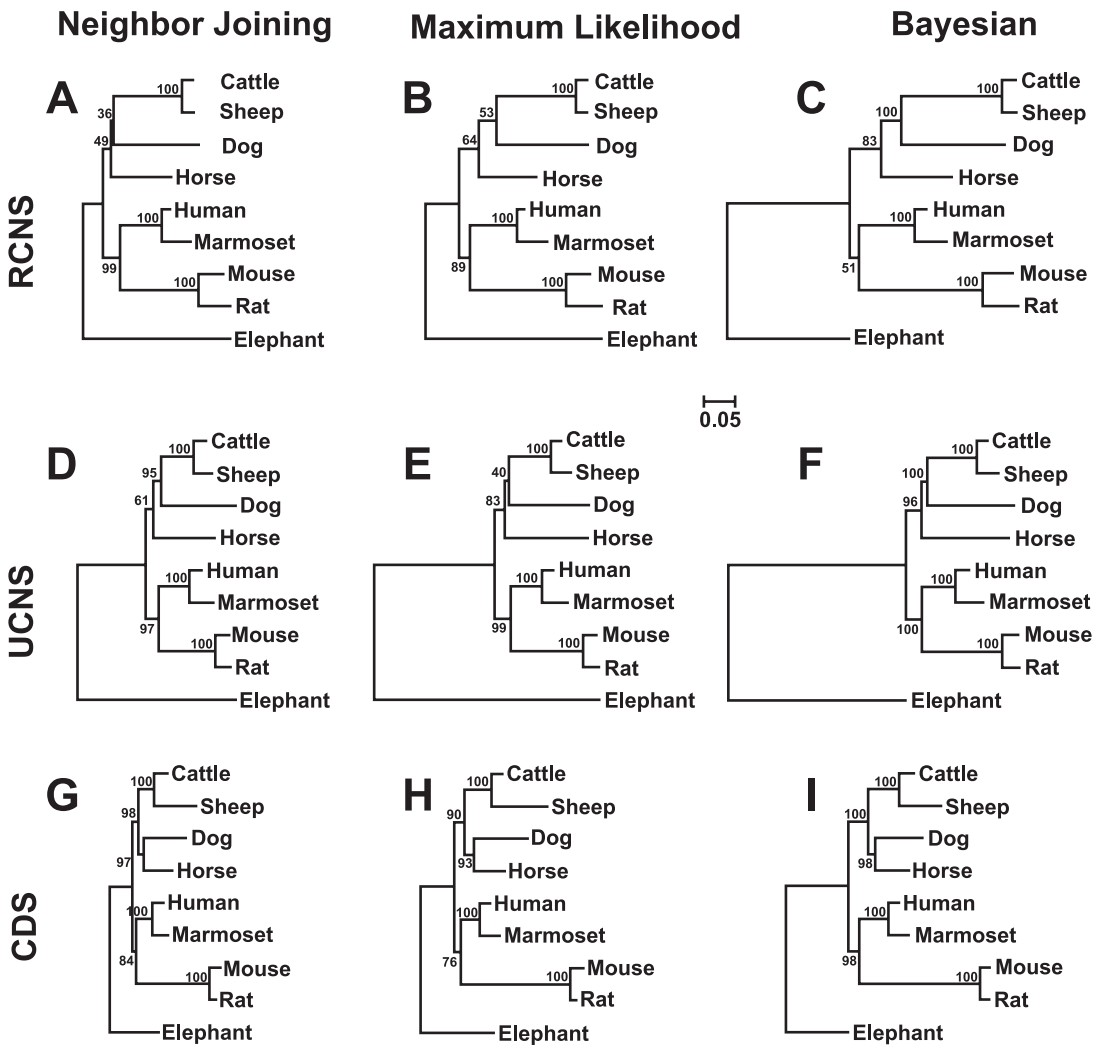

Figure 6. Phylogenetic relationship among human, marmoset, mouse, rat, dog, horse, cattle, sheep, and elephant, using the (A, D, and G) Neighbor-Joining, (B, E, and G) Maximum Likelihood, (C, F, and I) and Bayesian methods, based on the (A, B, and C) repetitive conserved noncoding sequences (RCNS), (D, E, and F) unique conserved noncoding sequences (UCNS), and (G, H, and I) coding sequences (CDS) from the region KCNQ1OT1-CDKN1C with sequence identity >70%. The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. All trees are drawn to scale with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree which are the number of base substitutions per site.

The regions showing genomic imprinting are highly conserved in monotremata and other vertebrates, but the distribution of repetitive elements are not, and they have expanded significantly in the regions with DMRs, only in the eutherians, so it has been suggested that retrotransposition is associated with the acquisition of new DMRs that regulate genomic imprinting,[42,43] but the mechanisms by which they affect it are unknown.

Repetitive elements have also been associated with regions rich in genes showing genomic imprinting, where several reports show a reduction in the number of SINEs but not a reduction in the number of LINEs,[44–46] suggesting that they can attract epigenetic modifications to the DMRs nearby or protect them from being epigenetically marked in the germinal lines.[47] However, we instead found a reduction in the number of LINEs in the imprinted region. Other studies have shown that the reduction in SINEs and the increase in LINEs are not widespread features of regions with genomic imprinting.[48]

One proof of the relation between repetitive elements and genomic imprinting is that some imprinted genes are of retroelement origin, such as PEG10, an essential gene for placental development which has a high similarity with the retrotransposon sushi-ichi.[49] Tandem repeats are also associated with DMRs, suggesting a role as isolators or silencers, or controlling the methylation status of nearby regions,[47,50] but its low conservation suggests more of a structural role than the presence of specific motifs that can participate in expression control.[51]

Even though primates and rodents are the most evolutionarily closest among the mammals analyzed, a higher proportion of shared CNS was found between human and horse, dog, and cattle. This could be related to the fact that previous studies have found that mouse repetitive elements have been amplifying at a relatively constant rate through evolution,

whereas primate elements underwent a sharp peak of activity about 40 million years ago and are currently amplifying relatively slowly.[52] This could mean a higher diversification of repetitive sequences in the rodent genomes, leading to less conservation of CNS.

Hutter et al[40] compared the conservation of sequences of autosomal regions with genes showing biallelic expression, with those with imprinted expression containing more conserved repetitive elements located in both intergenic and intronic regions. However, they did not find differences in the number of LINE1 elements being conserved, differing with our results where we saw a reduction in the LINE1 and LINE2 elements in the imprinted region. Khatib and Kim[53] also found a reduction in LINE1 elements in the genes with genomic imprinting they analyzed, although Paço et al[54] did find an increase in these elements in the genes with genomic imprinting. It appears that the relationship between repetitive elements and imprinting regions is more complex and region specific and could be more related to the presence of specific repeats for specific regions acquiring functions in these regions just by chance.

We detected binding sites for 3 main transcription factor families in the motifs identified: SOX, FOX, and GATA. The 20 members of the SOX family (sex-determining region Y boxes) are transcriptional regulators, containing the high-mobility group domain that binds to DNA and have been divided into 8 groups.[55] The proteins from the same group share biochemical properties and have redundant and synergistic functions, but those from different groups have different functions as development regulators, going from sex determination, hematopoiesis, neural crest development, and neurogenesis.[56] SOX and OCT motifs have been shown to coincide with the ICR in the H19/IGF2 region, allowing the active demethylation of this element in the maternal allele in mouse.[57–59] However, these motifs were not associated with demethylation of DNA during the erasure of epigenetic marks in primordial germinal cells in paternal alleles. Their role is likely to be associated more with the maintenance of the demethylated status of the maternal alleles in the embryo postimplantation.[59] It is likely that the SOX sites identified in our study are involved in the maintenance of the demethylated status of the DMRs to regulate the expression of the maternal alleles in the genes with this type of imprinting pattern.

The forkhead box (FOX) proteins constitute a family of evolutionarily conserved transcription factors with functions not only during development but also during the adult functions.[60] There are about 100 proteins in the family divided in groups designated from FOXA through FOXS, with high affinity to a very similar core sequences, but with differences in the surrounding sequences for each group, providing them different functions in development, cell proliferation, and differentiation; stress resistance; apoptosis; metabolism; and reproduction.[61] The GATA transcription factors are also highly conserved, along with the dedicated cofactors named friends of GATA controlling differentiation and cell fate of multiple cell types from *Drosophila* to human.[62] Due to the different functions of these transcription factors in the cell, it is possible that some are associated with mechanism of control of allele-specific expression or to the temporal or tissue-specific pattern of expression in this region. However, it is necessary to conduct experimental essays to prove whether these DNA motifs are recognized and bound by these transcription factors.

In conclusion, we show that comparative genomics can be used to identify functional DNA elements related to the regulation of gene expression in the specific regions. The approach of comparing the sequence alignment between species with and without masking the repetitive elements can be used to identify conserved repetitive elements in a large region. The identified elements may be playing important roles in the control of imprinting, as the search continues to better understand the whole process.

Author Contributions

MDD, SOP, IGI and BNT conceived the project; MDD, TH, HR, SOP and IGI designed and carried out the experiments; MDD, SOP, IGI and BNT analyzed the data; TH, HR and SOP contributed to the scientific content; MDD, IGI and BNT wrote the manuscript. All authors read and approved the final manuscript.

REFERENCES

1. Lemos B, Branco AT, Jiang PP, Hartl DL, Meiklejohn CD. Genome-wide gene expression effects of sex chromosome imprinting in Drosophila. *G3 (Bethesda)*. 2014;4:1–10.
2. Sanchez L. Sex-determining mechanisms in insects based on imprinting and elimination of chromosomes. *Sex Dev*. 2014;8:83–103.
3. Eggert H, Kurtz J, Diddens-de Buhr MF. Different effects of paternal transgenerational immune priming on survival and immunity in step and genetic offspring. *Proc Biol Sci*. 2014;281:20142089.
4. McEachern LA, Bartlett NJ, Lloyd VK. Endogenously imprinted genes in *Drosophila melanogaster*. *Mol Genet Genomics*. 2014;289:653–673.
5. Florez-Rueda AM, Paris M, Schmidt A, Widmer A, Grossniklaus U, Städler T. Genomic imprinting in the endosperm is systematically perturbed in abortive hybrid tomato seeds. *Mol Biol Evol*. 2016;33:2935–2946.
6. Furihata HY, Suenaga K, Kawanabe T, Yoshida T, Kawabe A. Gene duplication, silencing and expression alteration govern the molecular evolution of PRC2 genes in plants. *Genes Genet Syst*. 2016;91:85–95.
7. Hatorangan MR, Laenen B, Steige KA, Slotte T, Köhler C. Rapid evolution of genomic imprinting in two species of the Brassicaceae. *Plant Cell*. 2016;28:1815–1827.
8. Barlow DP, Bartolomei MS. Genomic imprinting in mammals. *Cold Spring Harb Perspect Biol*. 2014;6:a018382.
9. Chess A. Monoallelic gene expression in mammals. *Annu Rev Genet*. 2016;50:317–327.
10. Saito T, Hara S, Tamano M, Asahara H, Takada S. Deletion of conserved sequences in IG-DMR at Dlk1-Gtl2 locus suggests their involvement in expression of paternally expressed genes in mice. *J Reprod Dev*. 2017;63:101–109.
11. Kaut O, Sharma A, Schmitt I, Wüllner U. DNA methylation of imprinted loci of autosomal chromosomes and IGF2 is not affected in Parkinson's disease patients' peripheral blood mononuclear cells. *Neurol Res*. 2017;39:281–284.
12. John RM. Imprinted genes and the regulation of placental endocrine function: pregnancy and beyond [published online ahead of print January 10, 2017]. *Placenta*. doi:10.1016/j.placenta.2017.01.099
13. Imumorin IG, Peters SO, De Donato M. Genomic imprinting and imprinted gene clusters in the bovine genome. In: Khatib H, ed. *Livestock Epigenetics*. Edinburgh: Wiley-Blackwell; 2012:89–111.
14. Bartolomei MS, Ferguson-Smith AC. Mammalian genomic imprinting. *Cold Spring Harb Perspect Biol*. 2011;3:a002592.
15. Nativio R, Sparago A, Ito Y, Weksberg R, Riccio A, Murrell A. Disruption of genomic neighbourhood at the imprinted IGF2-H19 locus in Beckwith-Wiedemann syndrome and Silver-Russell syndrome. *Hum Mol Genet*. 2011;20:1363–1374.

16. Schultz BM, Gallicio GA, Cesaroni M, Lupey LN, Engel N. Enhancers compete with a long non-coding RNA for regulation of the Kcnq1 domain. *Nucleic Acids Res*. 2015;43:745–759.
17. Lewis A, Reik W. How imprinting centres work. *Cytogenet Genome Res*. 2006;113:81–89.
18. Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW, Povey S. Guidelines for human gene nomenclature. *Genomics*. 2002;79:464–470.
19. Goujon M, McWilliam H, Li W, et al. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res*. 2010;38:W695–W699.
20. Smit AFA, Hubley R, Green P. RepeatMasker. Open-4.0.5. 2014. http://www.repeatmasker.org.
21. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res*. 2004;32:W273–W279.
22. Brudno M, Do CB, Cooper GM, et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*. 2003;13:721–731.
23. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*. 1998;14:48–54.
24. Mathelier A, Zhao X, Zhang AW, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2014;42:D142–D147.
25. Sandelin A, Wasserman WW. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol*. 2004;338:207–215.
26. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4:406–425.
27. Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci U S A*. 2004;101:11030–11035.
28. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*. 1993;10:512–526.
29. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol*. 2016;33:1870–1874.
30. Ronquist F, Teslenko M, van der Mark P, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 2012;61:539–542.
31. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*. 1985;39:783–791.
32. Nikaido I, Saito C, Mizuno Y, et al; RIKEN GER Group; GSL Members. Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling. *Genome Res*. 2003;13:1402–1409.
33. Luedi PP, Hartemink AJ, Jirtle RL. Genome-wide prediction of imprinted murine genes. *Genome Res*. 2005;15:875–884.
34. Luedi PP, Dietrich FS, Weidman JR, Bosko JM, Jirtle RL, Hartemink AJ. Computational and experimental identification of novel human imprinted genes. *Genome Res*. 2007;17:1723–1730.
35. Wei Y, Su J, Liu H, et al. MetaImprint: an information repository of mammalian imprinted genes. *Development*. 2014;141:2516–2523.
36. Wang X, Sun Q, McGrath SD, Mardis ER, Soloway PD, Clark AG. Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS ONE*. 2008;3:e3839.
37. Yang HH, Hu Y, Edmonson M, Buetow K, Lee MP. Computation method to identify differential allelic gene expression and novel imprinted genes. *Bioinformatics*. 2003;19:952–955.
38. Brideau CM, Eilertson KE, Hagarman JA, Bustamante CD, Soloway PD. Successful computational prediction of novel imprinted genes from epigenomic features. *Mol Cell Biol*. 2010;30:3357–3370.
39. Wang X, Soloway PD, Clark AG. A survey for novel imprinted genes in the mouse placenta by mRNA-seq. *Genetics*. 2011;189:109–122.
40. Hutter B, Bieg M, Helms V, Paulsen M. Imprinted genes show unique patterns of sequence conservation. *BMC Genomics*. 2010;11:649.
41. Dindot SV, Person R, Strivens M, Garcia R, Beaudet AL. Epigenetic profiling at mouse imprinted gene clusters reveals novel epigenetic and genetic features at differentially methylated regions. *Genome Res*. 2009;19:1374–1383.
42. Pask AJ, Papenfuss AT, Ager EI, McColl KA, Speed TP, Renfree MB. Analysis of the platypus genome suggests a transposon origin for mammalian imprinting. *Genome Biol*. 2009;10:R1.
43. Renfree MB, Suzuki S, Kaneko-Ishino T. The origin and evolution of genomic imprinting and viviparity in mammals. *Philos Trans R Soc Lond B Biol Sci*. 2013;368:20120151.
44. Greally JM. Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. *Proc Natl Acad Sci U S A*. 2002;99:327–332.
45. Ke X, Thomas NS, Robinson DO, Collins A. The distinguishing sequence characteristics of mouse imprinted genes. *Mamm Genome*. 2002;13:639–645.
46. Allen E, Horvath S, Tong F, et al. High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes. *Proc Natl Acad Sci U S A*. 2003;100:9940–9945.
47. Walter J, Hutter B, Khare T, Paulsen M. Repetitive elements in imprinted genes. *Cytogenet Genome Res*. 2006;113:109–115.
48. Cowley M, de Burca A, McCole RB, et al. Short interspersed element (SINE) depletion and long interspersed element (LINE) abundance are not features universally required for imprinting. *PLoS ONE*. 2011;6:e18953.
49. Suzuki S, Ono R, Narita T, et al. Retrotransposon silencing by DNA methylation can drive mammalian genomic imprinting. *PLoS Genet*. 2007;3:e55.
50. Hutter B, Helms V, Paulsen M. Tandem repeats in the CpG islands of imprinted genes. *Genomics*. 2006;88:323–332.
51. Paulsen M. Unique patterns of evolutionary conservation of imprinted genes. *Clin Epigenetics*. 2011;2:405–410.
52. Deininger PL, Moran JV, Batzer MA, Kazazian HH Jr. Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev*. 2003;13:651–658.
53. Khatib H, Kim ES. The distribution and conservation of retrotransposable elements in cattle. *Epigenetics*. 2008;3:81–88.
54. Paço A, Adega F, Chaves R. LINE-1 retrotransposons: from "parasite" sequences to functional elements. *J Appl Genet*. 2015;56:133–145.
55. de la Rocha AM, Sampron N, Alonso MM, Matheu A. Role of SOX family of transcription factors in central nervous system tumors. *Am J Cancer Res*. 2014;4:312–324.
56. Sakaguchi R, Okamura E, Matsuzaki H, Fukamizu A, Tanimoto K. Sox-Oct motifs contribute to maintenance of the unmethylated H19 ICR in YAC transgenic mice. *Hum Mol Genet*. 2013;22:4627–4637.
57. Sarkar A, Hochedlinger K. The sox family of transcription factors: versatile regulators of stem and progenitor cell fate. *Cell Stem Cell*. 2013;12:15–30.
58. Hori N, Yamane M, Kouno K, Sato K. Induction of DNA demethylation depending on two sets of Sox2 and adjacent Oct3/4 binding sites (Sox-Oct motifs) within the mouse H19/insulin-like growth factor 2 (Igf2) imprinted control region. *J Biol Chem*. 2012;287:44006–44016.
59. Zimmerman DL, Boddy CS, Schoenherr CS. Oct4/Sox2 binding sites contribute to maintaining hypomethylation of the maternal igf2/h19 imprinting control region. *PLoS ONE*. 2013;8:e81962.
60. Benayoun BA, Caburet S, Veitia RA. Forkhead transcription factors: key players in health and disease. *Trends Genet*. 2011;27:224–232.
61. Thackray VG. Fox tales: regulation of gonadotropin gene expression by forkhead transcription factors. *Mol Cell Endocrinol*. 2014;385:62–70.
62. Chlon TM, Crispino JD. Combinatorial regulation of tissue specification by GATA and FOG factors. *Development*. 2012;139:3905–3916.