Open access • Posted Content • DOI:10.1101/683938

Unexpected features of GC-AG introns in long non-coding and protein-coding genes suggest a new role as regulatory elements — Source link

Monah Abou Alezz, Ludovica Celli, Giulia Belotti, Antonella Lisa ...+1 more authors

Institutions: National Research Council

Published on: 27 Jun 2019 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: RNA splicing, splice, Intron, Alternative splicing and Gene

Unexpected features of GC-AG introns in long non-coding and protein-coding genes suggest a new role as regulatory elements

Monah Abou Alezz[1], Ludovica Celli[1], Giulia Belotti[1], Antonella Lisa[1] and Silvia Bione[1*]

[1]Computational Biology Unit
Institute of Molecular Genetics Luigi Luca Cavalli-Sforza
National Research Council
Via Abbiategrasso 207
27100 PAVIA
Italy

[*]Corresponding author
bione@igm.cnr.it

ABSTRACT

Long non-coding (lnc) RNAs are today recognized as a new class of regulatory molecules despite very little is known about their functions in the cell. Due to their overall low level of expression and tissue-specificity, their identification and annotation in many genomes remains challenging. In this study, we exploited recent annotations provided by the GENCODE project to characterize the genomic and splicing features of lnc-genes in comparison to protein-coding (pc) ones, both in human and mouse. Our analysis highlighted slight differences between the two classes of genes in terms of genome organization and gene architecture. Significant differences in the splice sites usage were observed between lnc- and pc-genes. While the frequency of non-canonical GC-AG splice junctions represents about 0.8% of total splice sites in pc-genes, we identified a remarkable enrichment of the GC-AG splice sites in lnc-genes, both in human (3.0%) and mouse (1.9%). In addition, we found a positional bias of GC-AG splice sites being enriched in the first intron in both classes of genes. Moreover, a significant shorter length and weaker splice sites were found comparing GC-AG introns with the canonical GT-AG introns. The computational analysis of GC-AG splice sites strength revealed a strong reduction in both the donor and the acceptor splice sites scores especially in lnc first intron in both species. Genes containing at least one GC-AG intron were found conserved in many species and a functional enrichment analysis pointed toward their enrichment in specific biological processes. Furthermore, as previously suggested, GC-AG-containing genes were shown to be more prone to alternative splicing. Taken together, our study suggested that GC-AG introns could represent new regulatory elements mainly associated with lnc-genes.

INTRODUCTION

High-throughput sequencing technologies provided the opportunity to explore and to understand the complexity of the mammalian transcriptome whose extent and diversity remain to be fully assessed. It is now clear that a large portion of mammalian genomes is transcribed to produce non-coding RNAs which have been previously regarded as "transcriptional noise" (Johnson et al., 2005). The genomes of distantly related species house remarkably similar numbers of protein-coding (pc) genes prompting the notion that many aspects of complex organisms arise from non-coding regions (Liu et al., 2013). This stimulated the discovery of different classes of non-coding RNAs among which long non-coding RNAs (lncRNAs) are the most prevalent (Deveson et al., 2017).

LncRNAs are traditionally defined as transcripts longer than 200 nucleotides devoid of open reading frames. Like protein-coding genes, lncRNAs are transcribed by RNA Polymerase II, usually 5' capped, subjected to constitutive or alternative splicing and poly-adenylated (Harrow et al., 2012). However, their low levels of expression hindered their complete annotation and the assessment of their functional role that is today known for a very small fraction of them (Uszczynska-Ratajczak et al., 2018).

The lncRNA repertoire comprises thousands of transcripts which have distinct properties from protein-coding genes in terms of expression and subcellular localization. They are strikingly more tissue specific, expressed at much lower levels than mRNAs (Derrien et al., 2012), often co-expressed with their neighboring genes and many of them revealed a spatio-temporal dynamic expression (Cabili et al., 2011). Their sub-cellular localization range from strictly nuclear to almost exclusively cytoplasmic, albeit deeper analyses revealed a large variety of highly specific localizations in the cell (Carlevaro-Fita and Johnson, 2019).

In the past few years lncRNAs received growing attention as they emerged as an important regulatory layer of the transcriptome playing a role through distinct molecular mechanisms and in a variety of genomic contexts. LncRNAs, acting in cis or in trans, were described to be involved in many cellular processes like transcriptional and post-transcriptional regulation, splicing, mRNA translation and degradation, chromatin and DNA modifications, and spatial conformation of chromosomes (Jandura and Krause, 2017; Mattick, 2018). Moreover, it has also been reported that lncRNA transcription, rather than the RNA transcript itself, can have regulatory effects on neighboring genes (Engreitz et al., 2016; Long et al., 2017). LncRNAs have been implicated in several biological processes such as developmental patterning, dosage compensation, and genomic imprinting and in many pathological conditions (Ransohoff et al., 2018; Fernandes et al., 2019).

LncRNAs appeared less conserved than protein-coding genes due to the absence of constraints on coding sequences (Hezroni et al., 2015). Nevertheless, despite the lack of primary sequence conservation, lncRNAs are generally more conserved than neutrally evolving sequences, suggesting their conserved functions across species and highlighting focal areas of potential functional importance (Marques and Ponting, 2009). Indeed, lncRNAs exhibit conservation and selective constraints in their secondary structures and functional regulatory elements such as promoters and enhancers (Nitsche and Stadler, 2017). Furthermore, an evolutionary constraint on lncRNA sequences is also localized at splicing regulatory elements, suggesting that the recognition of the intron boundaries is a crucial step and the correct splicing of introns is required for their function (Ponjavic et al., 2007; Nitsche et al., 2015; Chernikova et al., 2016).

LncRNAs are encoded by multi-exonic genes and recognized by the same splicing machinery as well as mRNAs (Papasaikas and Valcàrcel, 2016). Like mRNAs, lncRNAs can undergo alternative splicing and exhibit alternative transcription start sites reflecting a complexity in their transcription regulation (Samudyata et al., 2018). Initial studies reported that lncRNAs harbor canonical splicing signals despite the fact they overall showed a decreased splicing efficiency (Derrien et al., 2012; Tilgner et al., 2012). A more recent study demonstrated that splicing inefficiency occurred in a number of lncRNAs as a consequence of inefficient U2AF65 binding to weak 3′ splice signals rather than a decrease in splicing enhancer density or U1 binding motif enrichment. Nevertheless, efficient splicing was observed among lncRNA with specific functions (Melè et al., 2017). In addition, a recent study reported an increased level of complexity in lncRNA splicing regulation: lncRNA loci were found highly prone to alternative splicing with their internal exons being almost universally alternatively spliced (Deveson et al., 2018).

In this study, we took advantage of data provided by the GENCODE project (Derrien et al., 2012) to characterize the genomic and splicing features of human and mouse lnc-genes in comparison to pc-genes. Interestingly, our study revealed a significant enrichment of GC-AG splice junctions in lncRNAs of both species together with their preferential localization in the first intron. Despite GC-AG introns were usually considered as canonical, our analyses demonstrated their different splicing behavior with respect to GT-AG introns, especially in lncRNAs. In addition, protein-coding genes owning GC-AG introns appeared highly conserved and significantly enriched in specific biological processes thus suggesting that GC-AG introns may represent specific regulatory elements involved in transcriptional regulation.

MATERIALS AND METHODS

Data collection and analysis

The lists of lncRNAs and protein-coding genes were downloaded from the GENCODE website (https://www.gencodegenes.org). Data from the release v27 were used for human genes annotated on the genome sequence GRCh38 (gencode.v27.long_noncoding_RNAs.gtf.gz; gencode.v27.basic.annotation.gtf.gz). Data from the release M16 were used for the mouse genes annotated on the genome sequence GRCm38 (gencode.vM16.long_noncoding_RNAs.gtf.gz; gencode.vM16.basic.annotation.gtf.gz). Protein-coding genes were selected based on the basic annotation when both gene and transcript were indicated as ‘‘protein_coding’’. The total number of genes, transcripts and exons considered in both species are reported in Table S1.

Data analyses and descriptive statistics were performed by custom scripts in RStudio version 1.1.456 (https://www.rstudio.com) and through the use of the "dplyr" R package version 0.8.1 (https://cran.r-project.org/web/packages/dplyr/dplyr.pdf). The Wilcoxon rank-sum test was applied to compare distributions and the Chi-square test was applied to compare groups. Correlation analysis was performed by estimating the Spearman correlation coefficient (*r*). For all statistical tests, a p-value < 0.05 was considered as significant.

To evaluate the conservation of genes containing GC-AG introns, we downloaded the list of orthologous genes in the human (GRCh38.p10) and mouse genomes (GRCm38.p5) from the Ensembl genome database (release 91) by using multi species comparisons in the BioMart data mining tool (Smedley et al., 2015). Multi-species conservation of 5' ss was assessed manually by aligning the sequences of corresponding introns in different organisms using the UCSC genome browser (http://genome.ucsc.edu) as data source. Species considered in this analysis were: human, chimp, macaque, mouse, rat, dog, pig, chicken, fugu and zebrafish.

Introns analyses

The sequences of introns were retrieved using the Table Browser tool from UCSC (http://genome.ucsc.edu/cgi-bin/hgTables; Karolchik et al., 2004) using human GRCh38 and mouse GRCm38 genome sequences. We excluded from the analysis all single-exon genes as they are not subjected to splicing: this resulted in a total of 56582 lnc- and 525149 pc-introns in human and 29612 lnc- and 393788 pc-introns in mouse.

The scores of splice junctions were calculated using the MaxEntScan web tool (Yeo and Burge, 2004), a program for predicting the strength of the splicing sequences based on the maximum entropy model. In particular, MaxEntScan::score5ss scores the donor splice site from a sequence motif of 9 nucleotides covering bases -3 to +6 and accounts for non-adjacent as well as adjacent dependencies between positions. MaxEntScan::score3ss scores the acceptor splice site from a sequence motif of 23 nucleotide covering bases -20 to +3. We evaluated the strength of 5' splice sites (ss) of human and mouse introns based on 4 probabilistic models (Maximum Entropy Model, Maximum Dependence Decomposition Model, First-order Markov Model and Weight Matrix Model) and the strength of 3'ss using 3 probabilistic models (Maximum Entropy Model, First-order Markov Model and Weight Matrix Model) as provided by the MaxEntScan tool. The evaluation of the polypyrimidine tract strength was performed using the "branchpointer" R package version 1.10.0 (Signal et al., 2018). The package predicted polypyrimidine tracts in query regions located at -18 to -44 nucleotides from the 3' splice sites.

Functional enrichment analysis

Gene list functional enrichment analyses were performed using the "The Database for Annotation, Visualization and Integrated Discovery" (DAVID: https://david.ncifcrf.gov) version 6.8 web tool (Huang et al., 2009) and with the "Protein Analysis Through Evolutionary Relationships" (PANTHER; Mi et al., 2019) overrepresentation test (release 20181113) implemented on the Gene Ontology (GO) website (http://www.geneontology.org). The lists of protein-coding genes containing a GC-AG intron from both human (n=1934) and mouse (n=1669) were analyzed for their enrichment on GO Biological Process terms and filtered based on their Benjamini adjusted p-value (DAVID) or FDR (PANTHER) applying a significance threshold of 0.05.

RESULTS

Long non-coding and protein-coding genes shared similar genomic features

In order to get insight into long non-coding RNAs features, a characterization of the genomic organization of human and mouse lncRNAs in comparison with protein-coding genes was performed. Our analysis was based on GENCODE human release 27 (containing 15778 lnc- and 19836 pc-genes) and mouse release M16 (containing 12374 lnc- and 21963 pc-genes). The total number of genes, transcripts and exons are reported in Table S1.

The genomic distribution of lnc- and pc-genes appeared highly similar in both species. Human and mouse lncRNAs appeared equally transcribed from the forward and the reverse strand as protein-coding genes (Table S2). In human, gene density at chromosomal level ranged from 1.2 to 12.0 genes/Mb for lnc- and from 1.1 to 25.0 genes/Mb for pc-genes (Figure 1a, Table S3). In mouse, gene density was lower, ranging from 1.1 to 7.4 genes/Mb for lnc- and from 2.0 to 13.9 genes/Mb for pc-genes (Figure 1b, Table S3). Interestingly, lnc-gene density on chromosome X resulted very low in both species (1.8 genes/Mb in human and 1.1 genes/Mb in mouse). Along each chromosome, lnc- and pc-genes were almost homogeneously interspersed although a relative abundance of genomic regions larger than 1 Mb containing only lnc-genes was observed and found conserved between human and mouse in a large proportion of cases (Figure S1).

Genome coverage of long non-coding genes was found considerably lower with respect to those of protein-coding ones. Indeed, long non-coding genes accounted for 12.5% of the human genome with respect to the 43.4% occupied by pc-genes. The reduced genome coverage was not entirely due to the smaller number of lnc-genes, as they account for about 80% of protein-coding ones; instead it appeared to be related to the lnc-gene length, that resulted significantly lower than that of protein-coding genes. Human lnc-genes resulted, on average, almost three times shorter than protein-coding ones (Table S4; Figure 1c). Similarly, the genome coverage of mouse lnc-genes was lower (6.8%) than that of pc-genes (39.2%). This lower genome coverage was in part due to the smaller number of lncRNA genes (56% of protein-coding genes) but also to the lnc-gene length that, as in human, resulted shorter than that of pc-genes (Figure 1d).

Lnc-genes short length appeared related to the lower number of exons composing them (Table S5). In human, more than 70% of lncRNA transcripts had 3 or less exons, compared to 16% of protein-coding transcripts bearing the same characteristics (Figure 1e). A large proportion of lncRNA transcripts was composed of 2 exons (34%) as previously reported (Derrien et al., 2012) and 14% are single-exon genes. In mouse, more than 75% of lncRNAs had 3 exons or less versus 23% in protein-coding transcripts and 24% of lncRNAs were single-exon genes versus 6.4% in protein-coding (Figure 1f). Also in the mouse genome, an enrichment of 2-exons transcripts (30%) was observed among all lncRNA transcripts.

Minor differences between lnc- and pc-genes were appreciated also in the length of exons and introns (Table S6). The last exon appeared significantly shorter in lncRNAs with respect to protein-coding transcripts both in human and mouse. Moreover, the first intron resulted shorter in lnc-genes with respect to pc-genes both in human and mouse, whereas inner introns resulted longer (Figure 1c,d). Interestingly, unlike what was described for pc-genes in which first introns are generally longer than inner introns (Bradnam and Korf, 2008), lnc first and inner introns appeared similar in length in both species.

The transcriptional complexity of long non-coding genes resulted lower with respect to protein-coding ones. Indeed, lnc-genes transcribed in a single isoform were more than double compared to pc-genes in human, and a similar trend was observed in mouse (Table S7). Nevertheless, genes with a large number of different isoforms (>10) were present in similar percentage in lnc and pc in both species.

In conclusion, lnc- and pc-genes shared similar features in terms of genomic distribution and organization although they showed differences in length, number of exons and transcriptional complexity both in human and mouse.

Unexpected assortment of splicing junctions consensus in long non-coding RNAs

We characterized the splicing features of lnc-introns (human n=56582; mouse n=29611) in comparison with those of protein-coding ones (human n=525149; mouse n=393788) (Table S6). Single-exon genes were excluded from this analysis as they are not subjected to splicing.

The sequence analysis of splicing junctions highlighted differences between lnc- and pc-genes consensus sequences (Table 1). The GC-AG splice junctions appeared strongly enriched in lnc-genes in which they represent 3.0% of the total, thus almost four times more than in pc genes (0.8%). The same enrichment was also found in mouse, in which GC-AG splicing junctions were more than double with respect to pc ones (lnc 1.9%, pc 0.8%).

Moreover, GC-AG introns showed a preferential location as first introns in both lncRNAs and protein-coding transcripts (Table 2). Indeed, in the human genome, their percentage resulted higher in the first intron (lnc: 4.2%; pc: 1.2%) with respect to inner introns (lnc: 2.1%; pc: 0.8%) and the same trend was observed in mouse (first: lnc 2.4%, pc: 1.2%; inner: lnc 0.4%, pc 0.8%). In lnc-genes, more than half of GC-AG introns were located in the first intron in both species (Table S8).

The enrichment of GC-AG introns in lnc-genes led us to further investigate their splicing features. In human, GC-AG introns resulted shorter both in lnc- and pc-genes and they showed the same trend whether they are first or internal introns (Table 3). For GC-AG first introns the average length resulted almost halved respect to GT-AG first introns in both human lnc- and pc-genes (lnc: GC 6700 +/-600, GT 12923 +/-201; pc: GC 8999 +/-648, GT 15335 +/-162). Human GC-AG inner introns showed the same decrease in length, albeit to a lesser extent (lnc: GC 8666 +/-827, GT 13995 +/-194; pc: GC 4165 +/-197, GT 5411 +/-25). In mouse, GC-AG introns appeared shorter but only when they are inner introns (lnc: GC 5190 +/-734, GT 7523 +/-148; pc: GC 3186 +/- 192, GT 4437 +/-27).

To evaluate the functional behavior of GC-AG splicing junctions, we computed their strength using standard position weight-matrix models implemented in the MaxEntScan tool (Yeo and Burge, 2004), which assigns a computationally predicted score for 5′ and 3′ splice sites (ss). Overall, the strength of 5'and 3'ss resulted lower in lnc- than in pc-genes both in human and mouse (Table 4; Figure 2) and it was presumably one of the causes of the previously reported inefficiency of lnc-genes splicing (Melè et al., 2017; Tilgner et al., 2012). Despite lower weight-matrix scores for 5′ss-GC were expected, because of their imperfect pairing with the U1 snRNA, 5'ss-GC scores of lncRNAs resulted strongly reduced respect to 5'ss-GC of pc-genes in both species (human: lnc 5'ss-GC 0.50 WM, pc 5'ss-GC 2.76 WM; mouse: lnc 5'ss-GC 1.63 WM, pc 5'ss-GC 3.38 WM). The reduced strength of lncRNAs 5'ss-GC appeared to be due almost exclusively to the first intron junctions, whose scores resulted lower compared to those of inner introns, both in human and mouse (human: lnc first intron 5'ss-GC -0.93 WM, inner intron 5'ss-GC 2.60 WM; mouse: lnc first intron 5'ss-GC 0.78 WM, inner intron 5'ss-GC 2.65 WM). Despite owning the same consensus sequence, the 3'ss average weight-matrix scores for GC-AG introns appeared overall lower with respect to GT-AG acceptor sites, due to their shorter polypyrimidine tracts (Table S9). As it occurred for 5'ss, very weak 3'ss appeared preferentially located in the lnc first intron in both human and mouse.

To test whether the 5'ss and 3'ss weight-matrix scores and the introns length showed any correlation, the Spearman test was applied (Table S10). The strength of 5'ss and 3'ss was found positively correlated when located in the first intron of lnc-genes (human: $r = 0.58$, *p-value* <

2.2x10-16; mouse: $r = 0.51$, *p-value* < 2.2x10-16). The strengths of both 5'ss and 3'ss positively correlated to intron length and this correlation was found more pronounced in the first intron in both species (Table S10). Differently from what was reported for protein-coding genes, in which weak donor sites appeared flanked by stronger consensus at the acceptor sites (Kralovicova et al., 2011; Thanaraj and Clark, 2001), this analysis demonstrated that lnc-genes contained a class of very weak introns, preferentially located as first. Taken together, our data revealed that the first intron of lncRNAs with GC-AG splice junctions displayed peculiar features being shorter in length and owning particularly weak 5'ss and 3'ss.

GC-AG containing genes appeared conserved and not randomly assorted

In human, GC-AG introns were present in 1224 lnc-genes and in 1934 pc-genes, respectively representing the 7.8% and 9.7% of each type of genes. In mouse, GC-AG introns were present in 473 lnc-genes and in 1669 pc-genes, respectively representing the 3.8% and 7.6% of each type of genes. The great majority of transcripts included one single GC-AG intron, especially lncRNAs; few pc-transcripts owned more than two GC-AG introns per transcript.

Based on the human-mouse orthologue information provided by the Ensembl project (http://www.ensembl.org/index.html), a total of 908 protein-coding genes were conserved between the two species, respectively representing 47% and 54% of total GC-AG containing genes. Remarkably, in more than 75% of cases the GC-AG introns also shared the same ordinal position in homologous genes. Moreover, we found many examples in which the conservation of the GC-AG introns together with their relative position inside the gene was not limited to the mouse but extended across evolutionary distant species (Figure S2). For example, the GC-AG splice sites of the first intron of human ABI family member 3 binding protein (*ABI3BP*) and piccolo presynaptic cytomatrix protein (*PCLO*) genes were shown to be conserved in chimp, macaque, mouse, rat, dog, pig, chicken, fugu and zebrafish, always as first intron. The GC-AG splice sites of the human genes ceramide kinase like (*CERKL*) and 5-azacytidine induced 2 (*AZI2*) were shown to be conserved in inner introns of mammals, while the canonical GT was found in chicken, fugu and zebrafish. Despite the assessment of the conservation of lncRNAs was hindered by the lack of annotation in most species, a number of GC-AG splice junctions conserved between human and mouse was determined. Indeed, the TMEM51 antisense RNA 1 (*TMEM51-AS1*), the metastasis associated lung adenocarcinoma transcript 1 (*MALAT1*) and the nuclear paraspeckle assembly transcript 1 (*NEAT1*) genes contained a first GC-AG intron in both species whereas the JPX gene contained an inner GC-AG intron in both human and mouse. As the presence of a GC-AG intron was proposed to increase the level of alternative splicing (Churbanov et al., 2008), we compared the transcriptional complexity of both lnc- and pc-genes owning at least one GC-AG intron with respect to the ones containing only GT-AG introns (Table S11). In human, both long non-coding and protein-coding GC-AG-containing genes being transcribed in more than three isoforms exceeded the number of GT-AG-containing genes (lnc-GC 26.7% vs lnc-GT 14.2%; pc-GC 62.1%, vs pc-GT 43.0%). The same trend was confirmed in mouse, where long non-coding and protein-coding GC-AG-containing genes with more than three isoforms resulted more abundant than their GT-AG counterpart (lnc-GC 25.8% vs lnc-GT 11.9%; pc-GC 38.6%, vs pc-GT 25.6%).

In order to assess if the presence of a GC-AG intron may represent a regulatory motif involved in specific biological processes, we performed an enrichment analysis of Gene Ontology (GO) terms of human and mouse protein-coding genes. By means of the DAVID Functional Annotation Tool and the PANTHER Overrepresentation Test, we selected those terms that resulted significantly enriched in both species and by both tools (Figure 3; Table S12). This resulted in the identification of three groups of linked terms in the biological process ontology. The first group comprised the GO term "microtubule-based movement" (GO:0007018) and its ancestors "movement of cell or subcellular component" (GO:0006928) and "microtubule-based process" (GO:0007017) and included 221 human and 176 mouse genes. Despite very little is known about the biological processes in which lncRNAs are involved, at least three of the GC-AG-containing lncRNAs were described to have a role in the regulation of the movement of cells or subcellular components: the maternally expressed 3 (*MEG3)* gene (Xu et al., 2018; Wang et al., 2018), the SOX2 overlapping transcript (*SOX2-OT*) gene (Wang et al., 2017) and the spermatogenesis associated 13 (*SPATA13*) gene (Jean et al., 2013). The second group contained the GO term "DNA Repair" (GO:0006218) and its ancestors "cellular response to DNA damage stimulus" (GO:0006974) and "cellular response to stress" (GO:0033554) and accounted for 257 human and 179 mouse genes. Interestingly, two of the GC-AG-containing lncRNAs were described to be involved in DNA repair: the *MALAT1* gene (Hu et al., 2018) and the *NEAT1* gene (Adriaens et al., 2016). In the third group, the GO term "neuron projection development" (GO:0031175) with its ancestors "neuron development" (GO:0048666), "generation of neurons" (GO:0048699), "neurogenesis" (GO:0022008) and "nervous system development" (GO:0007399) were included and contained 273 and 220 human and mouse genes. Several lncRNAs containing a GC-AG intron were described to play a role in neuron development and growth like the *MEG3* gene (You and You, 2019), the *NEAT1* gene (Barry et al., 2017), the *SOX2-OT* gene, the GDNF antisense RNA 1 (*GDNF-AS1*) gene and the myocardial infarction associated transcript (*MIAT*) (Clark and Blackshaw, 2014).

In conclusion, genes containing GC-AG introns appeared highly conserved, more subjected to alternative splicing and enriched in specific biological process thus underlining a putative regulatory role of these introns.

DISCUSSION

In this study, we report a genome-wide comparison of genomic and splicing features of long non-coding and protein-coding genes in human and mouse. Being based on GENCODE releases 27 and M16, our analysis considered a conspicuously higher number of genes with respect to previous studies (Cabili et al., 2011; Derrien et al., 2012) and it was strengthened by the comparison between the two species.

The characterization of genomic features revealed slight differences between long non-coding and protein-coding genes in both human and mouse. In agreement with previous studies (Ravasi et al., 2006; Cabili et al., 2011; Derrien et al., 2012), lnc-genes resulted shorter than protein-coding ones in both species, with an average length of about 24 kb versus 68 kb in human and about 15 kb versus 49 kb in mouse. The difference in gene length was shown to be mainly due to the lower number of exons composing lncRNAs: genes composed of less than 3 exons were definitely more abundant among lnc-genes compared to pc-genes in both species (in human: 48% of lnc and 8% of pc; in mouse: 54% of lnc and 14% of pc). The shorter gene length and the limited number of exons could be attributable to the incomplete annotation of long non-coding genes whose low expression level and high tissue specificity hampers the complete characterization, as suggested by the studies of Lagarde and colleagues (Lagarde et al., 2016; Lagarde et al., 2017). Nevertheless, our results did not appear to be driven by this bias as we used a recent and more complete GENCODE release, whose annotation was based on stronger experimental and computational evidence (Frankish et al., 2019). Moreover, our results were confirmed in the FANTOM5 collection of human lncRNAs, accurately annotated at their 5'ends (Hon et al., 2017) in which we demonstrated the same trend for lnc-genes length (mean=28.2 kb) and the reduced number of exons (lncRNAs with less than 3 exons: 56%). Taken together, our data suggested that genomic organization and gene architecture did not significantly differ between lnc- and pc-genes, thus implying that they could be subjected to the same mechanisms of genomic control and gene regulation.

The characterization of splicing features revealed a significant enrichment of introns harboring GC-AG splice sites in lncRNAs of both species. GC-AG splice sites were generally considered as a non-canonical variant of the major U2-type GT-AG splice junctions, accounting for 0.865% and 0.817%, respectively, in human and mouse genomes (Sheth et al., 2006; Parada et al., 2014). In agreement with what previously reported, we assessed the same frequency of GC-AG introns in both species when considering only protein-coding genes (0.83% in human and

0.81% in mouse). When lncRNAs were taken into account, the frequency of GC-AG splice sites resulted more than three time higher in human and more than two times higher in mouse, accounting for 3.0% and 1.9% of their total splice junctions. Notably, the enrichment of GC-AG splice sites did not appear to be evenly distributed, as it emerged more prominent in the first intron of both types of genes. In human, lncRNA GC-AG first introns corresponded to 4.2% of total first introns, whereas they account for 2.1% of total inner introns; in protein-coding genes they corresponded to 1.2% and 0.8% of total first and inner introns respectively. The same trend was observed in mouse in which a higher ratio of GC-AG splice junctions were found in the first intron with respect to inner ones in both lncRNA (2.4% vs 0.4%) and protein-coding genes (1.2% vs 0.8%). The significant increase of GC-AG introns in lnc-genes, together with their non-random distribution along the gene, led us to hypothesize that they may represent unique regulatory elements. The preferential localization of GC-AG splice sites in the first intron provided a clear indication of their role in gene regulation. Indeed, first introns were described to possess particular regulatory features, as they were shown to be more conserved with respect to inner introns and to be enriched in epigenetics marks associated with active transcription, such as H3K4me3 and H3K9ac (Park et al., 2014; Bieberstein et al., 2012), thus being likely involved in gene expression and splicing regulation. In many cases, first introns were demonstrated to be responsible for transcription initiation and increase of mRNA transcriptional rates (Rose, 2019). Moreover, the binding of the U1-complex to 5'ss was demonstrated to be involved not only in splicing regulation but also in polyadenylation control and in regulation of gene expression through its interaction with promoter elements (Berg et al., 2012; Almada et al., 2013; Singh and Singh, 2019) suggesting that the non-canonical GC 5'ss could in some way perturb this mechanism of action.

GC-AG introns displayed peculiar splicing features in comparison with GT-AG introns, in particular when located in the first intron of lnc-genes. Introns harboring GC-AG splice sites appeared significantly shorter than GT-AG introns, in both lncRNA and protein coding genes. This trend was more prominent in human GC-AG first introns, having an average length of ~6.7 kb in lnc-genes and ~9 kb in pc-genes, and significantly shorter than GT-AG first introns (~13 kb and ~15 kb in lnc- and pc-genes respectively). In addition to their shorter length, GC-AG splice sites appeared significantly weaker than GT-AG ones. A reduction in the 5'ss strength of GC-AG introns was expected because of the mismatch at position +2 with the U1 snRNA consensus. Nevertheless, the reduction of 5' ss strength was more evident in GC splice sites of lnc rather than in protein-coding genes and it was more prominent in the first intron rather than in internal ones (Table 4). Similar results were obtained for 3’ss, whose average weight-matrix scores for GC-AG introns appeared lower compared to GT-AG junction, especially when located in lnc first introns. Interestingly, the Spearman correlation test demonstrated a positive correlation among intron length and 5’/3’ ss strength for the first intron of lnc-genes, thus implying the enrichment of short and very weak first introns in this class of molecules. It was suggested that the base pairing between 5’ss and U1 regulates alternative versus constitutive splicing, hence suggesting that weak splice sites are more prone to undergo alternative splicing (Stamm et al., 1994; Sorek et al., 2004). In agreement with a previous report (Thanaraj and Clark, 2001), our analysis at gene level confirmed that GC-AG containing genes were more prone to alternative splicing than genes harboring GT-AG introns. Moreover, Churbanov and colleagues (Churbanov et al., 2008) demonstrated that an excess of GT to GC 5’ss conversions occurred both in primates and rodents, hypothesizing that the accumulation of GC sites in mammals might arise from positive selection in favor of alternative splicing. Taken together, these results further supported the role of GC-AG introns as regulatory elements putatively involved in the control of alternative splicing events. How GC-AG introns could contribute to increase alternative splicing levels and which type of alternative splicing they could favor will require further investigations.

Despite the percentage of GC 5’ss is relatively small, the number of genes containing at least one GC-AG intron is not irrelevant, as they account for about 10% of pc-genes and 8% of lnc-genes in human (in mouse: about 8% of pc-genes and 4% of lnc-genes). The relevance of GC-AG-containing genes emerged also from the analysis of their conservation: about 50% of GC-AG containing pc-genes resulted conserved between human and mouse and in the majority of cases (75%) also the intron position resulted conserved. Moreover, in many instances the GC-AG splice sites appeared to be conserved not only in the mouse genome but also in other species and across large evolutionary distance. The evaluation of the conservation of GC-AG splice sites in lncRNA genes was hindered by their current incomplete annotation in many species. However, among the well-studied and annotated lncRNAs, we could still identify examples of the conservation of GC-AG splice sites between human and mouse. Indeed, the two well characterized nuclear lncRNAs *NEAT1* and *MALAT1* juxtaposed on human chromosome 11 (on chromosome 19 in mouse) share similar gene features: both are transcribed in long unspliced isoforms as well as in shorter and spliced transcripts starting from the same promoter. Moreover, both *NEAT1* and *MALAT1* shorter transcripts contain a GC-AG first intron in human and mouse thus suggesting similar regulatory functions.

The functional enrichment analysis of human and mouse protein-coding genes provided further evidence that GC-AG introns could represent a specific regulatory motif as it revealed a significant enrichment of GO terms related to DNA repair, neurogenesis, and microtubule-based movements. Despite the enrichment analysis for lncRNA genes was obstructed by the lack of their functional annotation, we reported several examples of the involvement of lncRNA genes harbouring a GC-AG introns in these biological processes. This analysis suggested that GC-AG introns may be involved in the expression control of genes involved in specific cellular functions, presumably needing a concerted regulation.

In few cases, the functional relevance of GC-AG introns was already demonstrated. In the study of Farrer and colleagues (Farrer et al., 2002) it was demonstrated that the weak GC 5'ss located in intron 10 of the Collagen alpha-2(IV) chain (*let-2)* gene in *C.elegans* was essential for developmentally regulated alternative splicing, and that its replacement with a stronger GT splice site suppressed the alternative splicing regulation occurring during embryos development. In the inhibitor of growth family member 4 (*ING4*) gene, the selection between a weak GC 5'ss or a near-located canonical GT was shown to result into alternative transcript isoforms which diverged for the presence of a nuclear localization signal thus affecting the subcellular localization of the encoded protein (Tsai et al., 2008). In the work of Palaniswamy and colleagues (Palaniswamy et al., 2010), a single nucleotide polymorphism converting a 5'ss GT to GC, present with varying frequencies in different mouse strains, was shown to be responsible for an alternative splicing event affecting the length and the translational efficiency of the GLI-Kruppel family member GLI1 (*Gli1*) gene in mouse. Moreover, for the PR/SET domain (*PRDM*) gene family in human (Fumasoni et al., 2007) and for the starch synthase (*SS*) gene family in rice (Chen et al., 2017) the activation of a GC 5'ss was shown to contribute to the diversification and the evolution of both gene families.

It is today clear that organisms complexity does not correlate with genome size or gene content, but it is instead more consistently related to the level of gene expression regulation. Higher level of gene regulation is thought to ensure the development of more sophisticated capabilities of higher organisms, despite the fact that the number of protein-coding genes is similar in evolutionary distant species. Furthermore, the amount of alternative splicing, which allows the production of a wide variety of proteins starting from a smaller number of genes, is known to be positively correlated with eukaryotic complexity (Bush et al., 2017; Schaefke et al., 2018). Moreover, the amount of transcribed ncDNA resulting in the production of a large collection of ncRNAs mainly involved in the regulation of gene expression, is known to increase together with organisms' complexity (Liu et al., 2013; Jandura et al., 2017). As it occurs for alternative splicing and for non-coding transcripts, also the frequency of GC-AG splice sites was reported to correlate with metazoan complexity (Sheth et al., 2006), hence indicating that this class of introns may represent a new layer of gene regulation. Interestingly, the conversion of donor splice sites from GT to GC was demonstrated to be an evolutionary driven mechanism, putatively due to the increased amount of alternative splicing occurring at weak GC-AG introns (Churbanov et al., 2008; Abril et al., 2005).

Taken together, our data suggested that GC-AG introns represent new regulatory elements mainly associated with lncRNAs and preferentially located in their first intron. Their increased frequency in higher organisms suggested that they could contribute to the evolution of complexity, adding a new layer in gene expression regulation. How they exerted their regulatory role remains to be elucidated despite preliminary evidence suggested that they could favor alternative splicing. The elucidation of the mechanisms of action of GC-AG introns would contribute to a better understanding of gene expression regulation and could address the comprehension of the pathological effects of mutations affecting GC donor sites contained in several disease-causing genes.

REFERENCES

Abril JF, Castelo R and Guigò R. "Comparison of splice sites in mammals and chicken". Genome Res. 2005. 15:111-119. doi: 10.1101/gr.3108805.

Adriaens C, Standaert L, Barra J, Latil M, Verfaillie A, Kalev P et al. "p53 induces formation of NEAT1 lncRNA-containing paraspeckles that modulate replication stress response and chemosensitivity". Nat Med. 2016. 22:861-868. doi: 10.1038/nm.4135.

Almada AE, Wu X, Kriz AJ, Burge CB and Sharp PA. "Promoter directionality is controlled by U1 snRNP and polyadenylation signals". Nature. 2013. 499: 360-363. doi: 10.1038/nature12349.

Barry G, Briggs JA, Hwang DW, Nayler SP, Fortuna PR, Jonkhout N et al. "The long non-coding RNA NEAT1 is responsive to neuronal activity and is associated with hyperexcitability states". Sci Rep. 2017. 7:40127. doi: 10.1038/srep40127.

Berg MG, Singh LN, Younis I, Liu Q, Pinto AM, Kaida D et al. "U1 snRNP determines mRNA length and regulates isoform expression". Cell. 2012. 150: 53-64. doi: 10.1016/j.cell.2012.05.029.

Bieberstein NI, Carrillo Oesterreich F, Straube K, Neugebauer KM. "First exon length controls active chromatin signatures and transcription". Cell Rep. 2012. 2: 62-68. doi: 10.1016/j.celrep.2012.05.019.

Bradnam KR and Korf I. "Longer first introns are a general property of eukaryotic gene structure". PLoS One 2008; e3093. doi: 10.1371/journal.pone.0003093.

Bush SJ, Chen L, Tovar-Corona JM and Urrutia AO. "Alternative splicing and the evolution of phenotypic novelty". Philos Trans R Soc Lond B Biol Sci. 2017. 372:1713. doi: 10.1098/rstb.2015.0474.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A et al. “Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.” Genes Dev. 2011. 25: 1915–27. doi: 10.1101/gad.17446611.

Carlevaro-Fita J and Johnson R. “Global positioning system: understanding long noncoding RNAs through subcellular localization”. Mol Cell. 2019. 73:869-883. doi: 10.1016/j.molcel.2019.02.008.

Chen C, Gao S, Sun Q, Tang Y, Han Y, Zhang J et al. “Induced splice site mutation generates alternative intron splicing in starch synthase II (SSII) gene in rice”. Biotechnology & Biotechnological Equipment. 2017. 31: 1093–1099. Doi:10.1080/13102818.2017.1370984.

Chernikova D, Managadze D, Glazko GV, Makalowski W and Rogozin IB. “Conservation of the Exon-Intron Structure of Long Intergenic Non-Coding RNA Genes in Eutherian Mammals”. Life. 2016. 6: E27. doi: 10.3390/life6030027.

Churbanov A, Winters-Hilt S, Koonin EV and Rogozin IB. “Accumulation of GC donor splice signals in mammals”. Biol Direct. 2008. 3:30. doi: 10.1186/1745-6150-3-30.

Clark BS and Blackshaw S. “Long non-coding RNA-dependent transcriptional regulation in neuronal development and disease”. Front Genet. 2014. 5:164. doi: 10.3389/fgene.2014.00164.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. “The GENCODE v7 catalogue of human long non-coding RNAs : analysis of their structure, evolution and expression”. Genome Res. 2012; 22:1775–89. doi: 10.1101/gr.132159.111.

Deveson IW, Hardwick SA, Mercer TR and Mattick JS. “The Dimensions, Dynamics, and Relevance of the Mammalian Noncoding Transcriptome”. Trends in Genet. 2017. 33: 464–478. doi: 10.1016/j.tig.2017.04.004.

Deveson IW, Brunck ME, Blackburn J, Tseng E, Hon T, Clark TA et al. “Universal alternative splicing of noncoding exons”. Cell Syst. 2018. 6:245-255.e5. doi:10.1016/j.cels.2017.12.005.

Engreitz JM, Haines JE, Perez EM, Munson G, Chen J, Kane M et al., "Local regulation of gene expression by lncRNA promoters, transcription and splicing". Nature. 2016. 539:452-455. doi: 10.1038/nature20149.

Farrer T, Roller AB, Kent WJ and Zahler AM. "Analysis of the Role of Caenorhabditis Elegans GC-AG Introns in Regulated Splicing". Nucleic Acids Res. 2002. 30: 3360–3367. doi: 10.1093/nar/gkf465.

Fernandes J, Acuña S, Aoki J, Floeter-Winter L and Muxel S. "Long Non-Coding RNAs in the Regulation of Gene Expression: Physiology and Disease". Noncoding RNA. 2019. 5:E17. doi:10.3390/ncrna5010017.

Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J et al. "GENCODE reference annotation for the human and mouse genomes". Nucleic Acids Res. 2019. 47: D766–773. doi: 10.1093/nar/gky955.

Fumasoni I, Meani N, Rambaldi D, Scafetta G, Alcalay M and Ciccarelli FD. "Family Expansion and Gene Rearrangements Contributed to the Functional Specialization of PRDM Genes in Vertebrates". BMC Evol Biol. 2007. 7: 187. doi: 10.1186/1471-2148-7-187.

Harrow J. Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokacinski F et al. "GENCODE: The Reference Human Genome Annotation for The ENCODE Project". Genome Res. 2012. 22: 1760–1774. doi: 10.1101/gr.135350.111.

Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP and Ulitsky I. "Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species". Cell Rep. 2015. 11:1110-1122. doi: 10.1016/j.celrep.2015.04.023.

Hon CC, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJ, Gough J et al. "An atlas of human long non-coding RNAs with accurate 5′ ends." Nature. 2017. 543: 199–204. doi: 10.1038/nature21374.

Hu Y, Lin J, Fang H, Fang J, Li C, Chen W et al. “Targeting the MALAT1/PARP1/LIG3 complex induces DNA damage and apoptosis in multiple myeloma”. Leukemia. 2018. 32:2250-2262. doi: 10.1038/s41375-018-0104-2.

Huang da W, Sherman BT and Lempicki RA. “Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources”. Nat Protoc. 2009. 4:44–57. doi: 10.1038/nprot.2008.211.

Jandura A and Krause HM. “The new RNA world: growing evidence for long noncoding RNA functionality”. Trends Genet. 2017. 33:665-676. doi: 10.1016/j.tig.2017.08.002.

Jean L, Majumdar D, Shi M, Hinkle LE, Diggins NL, Ao M et al. “Activation of Rac by Asef2 promotes myosin II-dependent contractility to inhibit cell migration on type I collagen”. J Cell Sci. 2013. 126:5585-5597. doi: 10.1242/jcs.131060.

Johnson JM, Edwards S, Shoemaker D and Schadt EE. “Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments”. Trends in Genet. 2005. 21:93-102. doi:10.1016/j.tig.2004.12.009.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D et al. “The UCSC Table Browser Data Retrieval Tool”. Nucleic Acids Res. 2004. 32:493D-496. doi: 10.1093/nar/gkh103.

Kralovicova J, Hwang G, Asplund AC, Churbanov A, Smith CI and Vorechovsky I. “Compensatory signals associated with the activation of human GC 5′ splice sites”. Nucleic Acids Res. 2011. 39: 7077–7091. doi: 10.1093/nar/gkr306.

Lagarde J, Uszczynska-Ratajczak B, Santoyo-Lopez J, Gonzalez JM, Tapanari E, Mudge JM et al. “Extension of Human LncRNA Transcripts by RACE Coupled with Long-Read High-Throughput Sequencing (RACE-Seq)”. Nat Commun. 2016. 7:12339. doi: 10.1038/ncomms12339.

Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C et al. “High-Throughput Annotation of Full-Length Long Noncoding RNAs with Capture Long-Read Sequencing”. Nat Genet. 2017. 49: 1731–1740. doi: 10.1038/ng.3988.

Liu G, Mattick JS and Taft RJ. “A meta-analysis of the genomic and transcriptomic composition of complex life.” Cell Cycle. 2013. 12: 2061–2072. doi: 10.4161/cc.25134.

Long Y, Wang X, Youmans DT and Cech TR. "How do lncRNAs regulate transcription?" Sci Adv. 2017. 3:eaao2110. doi:10.1126/sciadv.aao2110.

Marques AC and Ponting CP. “Catalogues of Mammalian Long Noncoding RNAs: Modest Conservation and Incompleteness”. Genome Biol. 2009. 10: R124. doi: 10.1186/gb-2009-10-11-r124.

Mattick JS. “The central role of RNA in human development and cognition”. FEBS Lett. 2011. 585: 1600-1616. doi: 10.1016/j.febslet.2011.05.001.

Mattick JS. “The state of long non-coding RNA biology”. Noncoding RNA. 2018. 4:E17. doi: 10.3390/ncrna4030017.

Melé M, Mattioli K, Mallard W, Shechner DM, Gerhardinger C and Rinn JL. “Chromatin envirinment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs”. Genome Res. 2017. 1: 27-37. doi: 10.1101/gr.214205.116.

Mi H, Muruganujan A, Ebert D, Huang X and Thomas PD. “PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools”. Nucleic Acids Res. 2019. 47:419-426. doi: 10.1093/nar/gky1038.

Nitsche A, Rose D, Fasold M, Reiche K and Stadler PF. “Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved”. RNA. 2015. 21:801-812. doi: 10.1261/rna.046342.114.

Nitsche A and Stadler PF. “Evolutionary clues in lncRNAs”. Wiley Interdiscip Rev RNA. 2017. doi: 10.1002/wrna.1376.

Palaniswamy R, Teglund S, Lauth M, Zaphiropoulos PG and Shimokawa T. “Genetic Variations Regulate Alternative Splicing in the 5’ Untranslated Regions of the Mouse Glioma-Associated Oncogene 1, Gli1”. BMC Mol Biol. 2010. 11: 32. doi: 10.1186/1471-2199-11-32.

Papasaikas P and Valcárcel J. “The Spliceosome: The Ultimate RNA Chaperone and Sculptor”. Trends Biochem Sci. 2016. 41:33-45. doi:10.1016/j.tibs.2015.11.003.

Parada GE, Munita R, Cerda CA and Gysling K. “A comprehensive survey of non-canonical splice sites in the human transcriptome”. Nucleic Acids Res. 2014. 42: 10564-10578. doi: 10.1093/nar/gku744.

Park SG, Hannenhalli S and Choi SS. “Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals”. BMC Genomics. 2014. 15: 526. doi: 10.1186/1471-2164-15-526.

Ponjavic J, Ponting CP and Lunter G. “Functionality or Transcriptional Noise? Evidence for Selection within Long Noncoding RNAs”. Genome Res. 2007. 17: 556–565. doi: 10.1101/gr.6036807.

Ransohoff JD, Wei Y and Khavari PA. “The functions and unique features of long intergenic non-coding RNA”. Nat Rev Mol Cell Biol. 2018. 19:143-157. doi:10.1038/nrm.2017.104.

Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, Okunishi R et al. “Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome”. Genome Res. 2006. 16:11-19. doi: 10.1101/gr.4200206.

Rose AB. “Introns as gene regulators: a brick on the accelerator”. Front Genet. 2019. 9:672. doi: 10.3389/fgene.2018.00672.

Samudyata, Castelo-Branco G and Bonetti A. "Birth, coming of age and death: the intriguing life of long noncoding RNAs". Semin Cell Dev Biol. 2018. 79:143-152. doi:10.1016/j.semcdb.2017.11.012.

Schaefke B, Sun W, Li YS, Fang L and Chen W. "The evolution of posttranscriptional regulation". Wiley Interdiscip Rev RNA. 2018. e1485. doi: 10.1002/wrna.1485.

Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR and Sachidanandam R. "Comprehensive splice-site analysis using comparative genomics". Nucleic Acids Res. 2006. 34: 3955-3967. DOI: 10.1093/nar/gkl556.

Signal B, Gloss BS, Dinger ME and Mercer TR. "Machine learning annotation of human branchpoints". Bioinformatics. 2018. 34:920-927. doi: 10.1093/bioinformatics/btx688.

Singh RN and Singh NN. "A novel role of U1 snRNP: Splice site selection from a distance". Biochim Biophys Acta Gene Regul Mech. 2019. 1862: 634-642. doi: 10.1016/j.bbagrm.2019.04.004.

Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J et al. "The BioMart Community Portal: An Innovative Alternative to Large, Centralized Data Repositories". Nucleic Acids Res. 2015. 43:W589-598. doi: 10.1093/nar/gkv350.

Sorek R, Lev-Maor G, Reznik M, Dagan T, Belinky F, Graur D et al. "Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons". Mol Cell. 2004. 14: 221-231. Doi: 10.1016/S1097-2765(04)00181-9.

Stamm S, Zhang MQ, Marr TG and Helfman DM. "A sequence compilation and comparison of exons that are alternatively spliced in neurons". Nucleic Acids Res. 1994. 22: 1515-1526. doi: 10.1093/nar/22.9.1515.

Taft RJ, Pheasant M and Mattick JS. "The relationship between non-protein-coding DNA and eukaryotic complexity." BioEssays. 2007. 29: 288–299. doi: 10.1002/bies.20544.

Thanaraj TA and Clark F. "Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions." Nucleic Acids Res. 2001. 29: 2581–2593. doi: 10.1093/nar/29.12.2581.

Tilgner H, Knowles D, Johnson R, Davis CA, Chakrabortty S, Djebali S et al. "Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs". Genome Res. 2012. 22: 1616-1625. doi: 10.1101/gr.134445.111.

Tsai KW, Tseng HC and Lin WC. "Two Wobble-Splicing Events Affect ING4 Protein Subnuclear Localization and Degradation". Exp Cell Res. 2008. 314: 3130–3141. doi: 10.1016/j.yexcr.2008.08.002.

Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigò R and Johnson R. "Towards a complete map of the human long non-coding RNA transcriptome". Nat Rev Genet. 2018. 19: 535–548. doi: 10.1038/s41576-018-0017-y.

Wang J, Xu W, He Y, Xia Q and Liu S. "LncRNA MEG3 impacts proliferation, invasion, and migration of ovarian cancer cells through regulating PTEN". Inflamm Res. 2018. 67:927-936. doi: 10.1007/s00011-018-1186-z.

Wang Z, Tan M, Chen G, Li Z and Lu X. "LncRNA SOX2-OT is a novel prognostic biomarker for osteosarcoma patients and regulates osteosarcoma cells proliferation and motility through modulating SOX2". IUBMB Life. 2017. 69:867-876. doi: 10.1002/iub.1681.

Xu DH, Chi GN, Zhao CH and Li DY. "Long noncoding RNA MEG3 inhibits proliferation and migration but induces autophagy by regulation of Sirt7 and PI3K/AKT/mTOR pathway in glioma cells". J Cell Biochem. 2018. doi: 10.1002/jcb.28026.

Yeo G and Burge CB. "Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals". J Comput Biol. 2004. 11:377-394. doi: 10.1089/1066527041410418.

You D and You H. “Repression of long non-coding RNA MEG3 restores nerve growth and alleviates neurological impairment after cerebral ischemia-reperfusion injury in a rat model”. Biomed Pharmacother. 2019. 111:1447-1457. doi: 10.1016/j.biopha.2018.12.067.

ACKNOWLEDGEMENTS

We thank Prof. C. Calvio and Dr. R. Alfieri for critical reading of the manuscript. Dr. M. Abou Alezz is enrolled in the PhD program in "Genetics, Molecular and Cellular Biology" of the University of Pavia, Italy.

Table 1 – Number of different splice junctions consensus

	Human			
	lnc	**%**	**pc**	**%**
GT-AG	54667	96.6	517730	98.6
GC-AG	1683	3.0	4351	0.8
AT-AC	9	0.0	583	0.1
Others	223	0.4	2485	0.5
Total	56582		525149	

	Mouse			
	lnc	**%**	**pc**	**%**
GT-AG	28586	96.5	388973	98.8
GC-AG	570	1.9	3217	0.8
AT-AC	6	0.0	363	0.1
Others	449	1.5	1235	0.3
Total	29611		393788	

Table 2 – Number of GC-AG introns in first or inner positions

	Human					
	lnc			pc		
	N	GC-AG	%	N	GC-AG	%
First	23997	1000	4.2	53776	665	1.2
Inner	32585	683	2.1	471373	3686	0.8
Total	56582	1683	3.0	525149	4351	0.8

	Mouse					
	lnc			pc		
	N	GC-AG	%	N	GC-AG	%
First	13079	309	2.4	40990	472	1.2
Inner	16532	61	0.4	352798	2745	0.8
Total	29611	570	1.9	393788	3217	0.8

Table 3 – Length characteristics of GC-AG and GT-AG introns

		Human					
		lnc			**pc**		
		N	**Mean (bp)**	**SEM(*)**	**N**	**Mean (bp)**	**SEM(*)**
GC-AG introns	**First**	1000	6700	600	665	8999	648
	Inner	683	8666	827	3686	4165	197
	Total	1683	7500	490	4351	4907	194
GT-AG introns	**First**	22904	12923	201	52587	15335	162
	Inner	31763	13995	194	465143	5411	25
	Total	54667	13367	141	517730	6420	28

		Mouse					
		lnc			**pc**		
		N	**Mean (bp)**	**SEM(*)**	**N**	**Mean (bp)**	**SEM(*)**
GC-AG introns	**First**	309	8619	1182	472	15933	1639
	Inner	261	5910	734	2745	3186	192
	Total	570	7049	726	3217	5056	301
GT-AG introns	**First**	12489	8507	183	40180	13606	187
	Inner	16097	7523	148	348793	4437	27
	Total	28586	7953	116	388973	5384	31

(*) Standard error of the mean

Table 4 – Weight-matrix scores of GC-AG and GT-AG splice sites

	5'ss				3'ss			
	Human		Mouse		Human		Mouse	
	lnc	pc	lnc	pc	lnc	pc	lnc	pc
GC-AG introns	0.50	2.76	1.63	3.38	5.00	8.75	7.76	9.25
GT-AG introns	7.82	8.00	7.68	8.12	8.90	9.44	8.78	9.42
First GC-AG introns	-0.93	1.67	0.78	2.68	2.50	8.44	6.34	9.70
First GT-AG introns	7.60	8.10	7.49	8.23	8.60	9.83	8.60	9.71
Inner GC-AG introns	2.60	2.96	2.65	3.50	8.70	8.80	9.44	9.17
Inner GT-AG introns	8.00	8.01	7.83	8.11	9.00	9.40	8.92	9.38

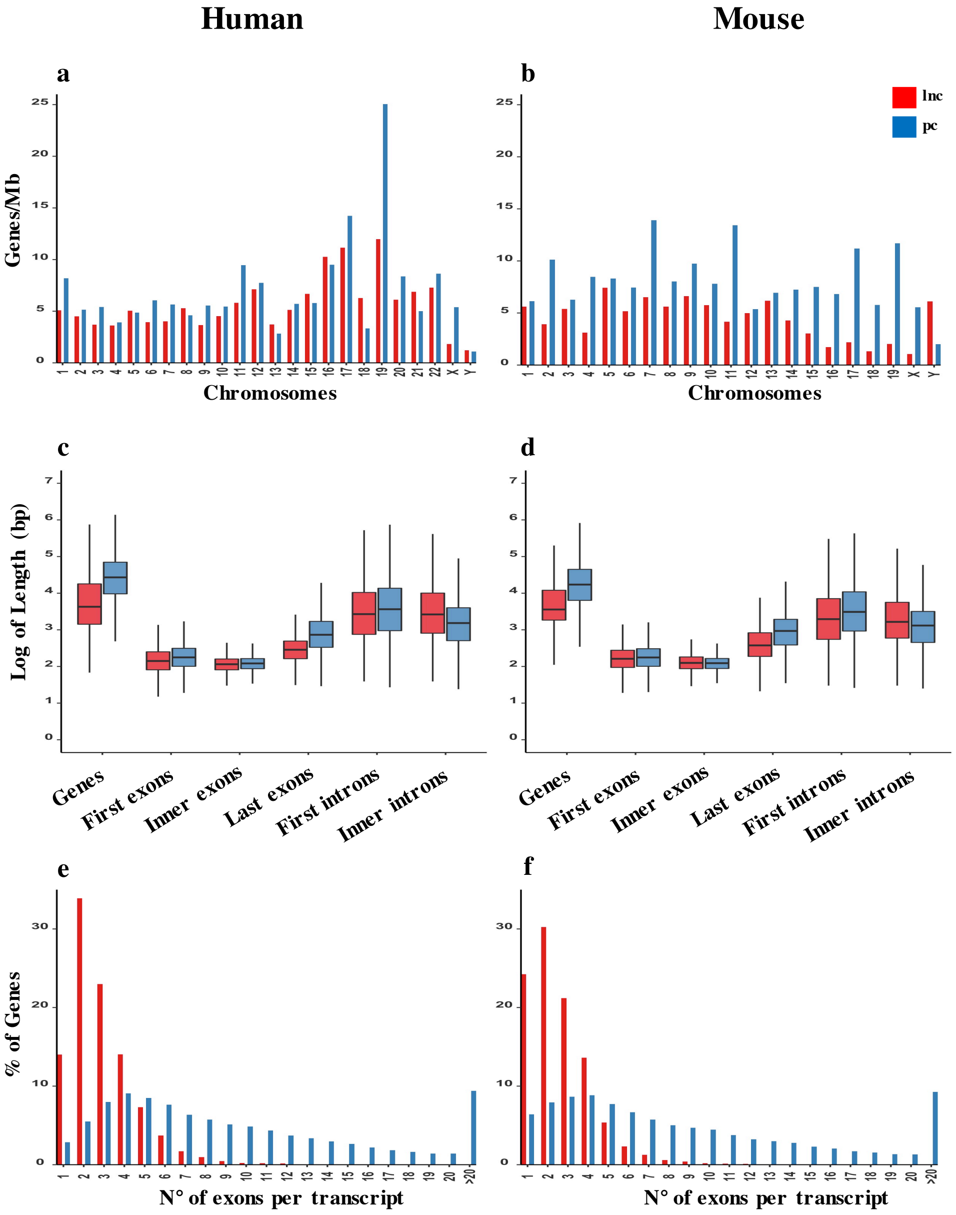

Figure 1. Genomic features of long non-coding and protein-coding genes in human and mouse. Gene densities in human (a) and mouse (b) chromosomes. Densities were reported as number of genes per Megabase (Mb). Length characteristics of genes, exons and introns in human (c) and mouse (d). Data were presented as $\log_{10}$ of basepairs (bp) lengths. Number of exons per transcripts in human (e) and mouse (f).

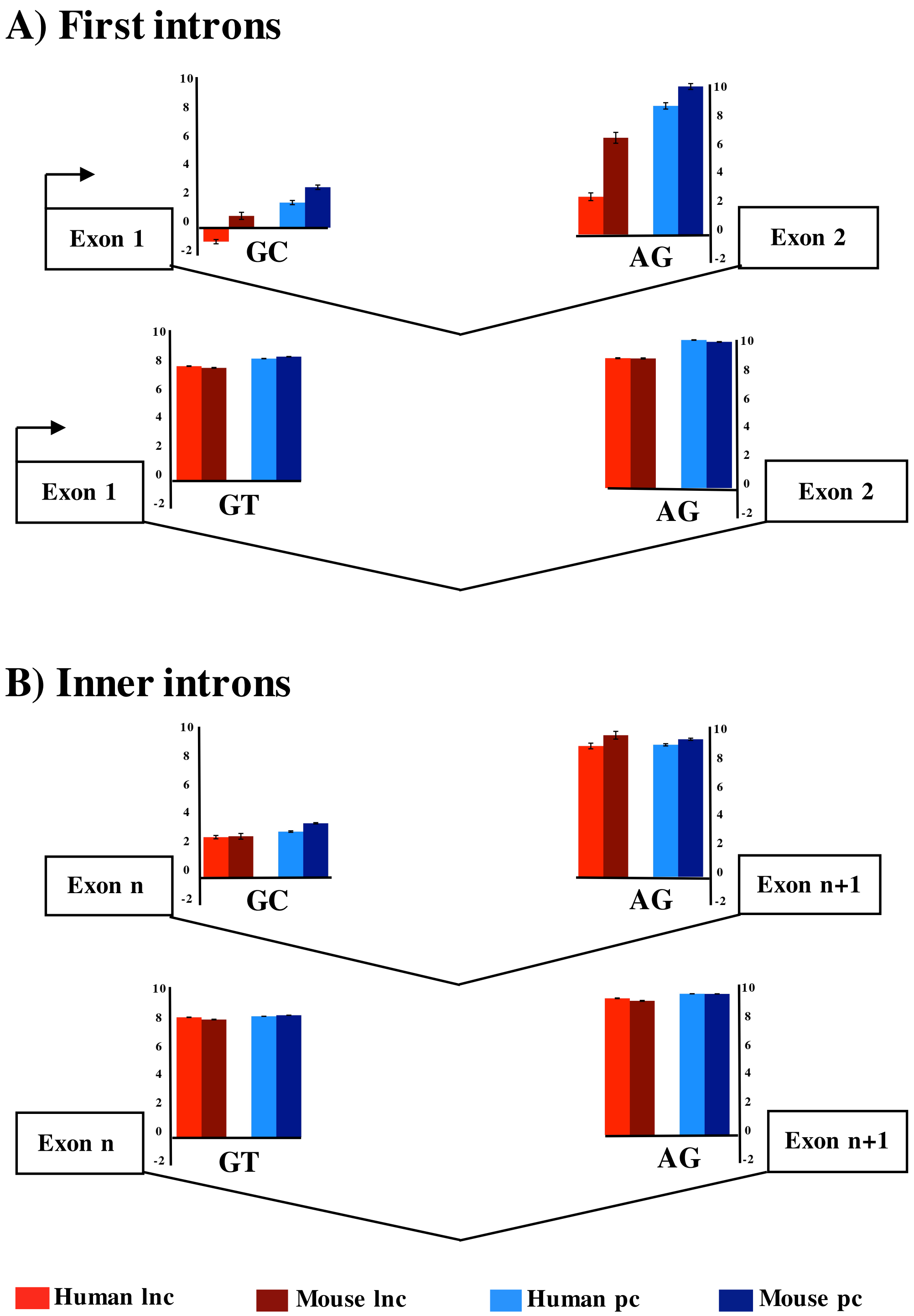

Figure 2. Splice junctions strengths. Schematic representation of average strengths of donor and acceptor splice sites of long non-coding and protein-coding genes in human and mouse**.** In panel A, GC-AG or GT-AG introns located as first were represented; in panel B, GC-AG or GT-AG inner introns are shown.

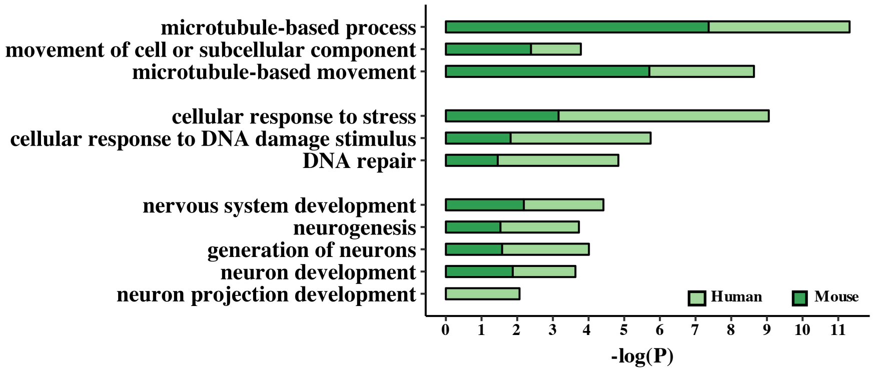

Figure 3. Enrichment analysis of GC-AG-containing genes. Bar graph representing the GO terms found significantly enriched. The GO term name is indicated in the Y axis, the –log of the p-value is indicated on the X-axis.