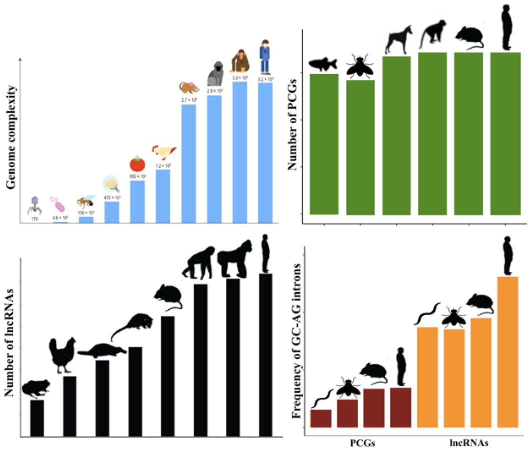
UNIVERSITA' DEGLI STUDI DI PAVIA

Dipartimento di Biologia e Biotecnologie

"L. Spallanzani"

# Genome-wide Characterization of the Genomic and Splicing Features of Long Non-coding RNAs Using Bioinformatic Approaches



**Monah Abou Alezz**

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
XXXIII Ciclo – A.A. 2017-2020

UNIVERSITÀ
DI PAVIA

*Dipartimento di Biologia e Biotecnologie*
*"L. Spallanzani"*

# Genome-wide Characterization of the Genomic and Splicing Features of Long Non-coding RNAs Using Bioinformatics Approaches

**Monah Abou Alezz**

**Supervised by *Prof. Silvia Bione***

Institute of Molecular Genetics L.L. Cavalli-Sforza
National Research Council

Dottorato di Ricerca in
Genetica, Biologia Molecolare e Cellulare
XXXIII Ciclo – A.A. 2017-2020

*On the cover:*

Bar graphs depicting that the increase of organismal complexity is accompanied by an increase in genome size, number of lncRNAs and frequency of GC-AG introns despite the number of protein-coding gene remains highly similar.

***The Road goes ever on and on***
***Down from the door where it began.***
***Now far ahead the Road has gone,***
***And I must follow, if I can,***
***Pursuing it with eager feet,***
***Until it joins some larger way***
***Where many paths and errands meet.***
***And wither then? I cannot say.***

***Still round the corner there may wait***
***A new road or a secret gate;***
***And though I oft have passed them by,***
***A day will come at last when I***
***Shall take the hidden paths that run***
***West of the Moon, East of the Sun.***

***J. R. R. Tolkien***

***For those who allow me to be a better person***
***and for me to know myself better.***

# Table of Contents

# Abstract

Long non-coding RNAs (lncRNAs) are recognized as a new class of regulatory molecules associated with organisms complexity despite very little is known about their functions in the cellular processes. Due to their overall low expression level and tissue-specificity, the identification and annotation of lncRNA genes still remains challenging. LncRNAs show a low level of sequence conservation, but an evolutionary constraint on lncRNA sequences is localized at splicing regulatory elements suggesting that the recognition of the intron boundaries and their splicing is a crucial step required for their function. We exploited recent annotations by the GENCODE compendium to characterize the splicing features of long non-coding genes, in comparison to protein-coding ones, in the human and mouse genome by using bioinformatics approaches. A significant difference in the splice sites usage was observed between the two gene classes. While the frequency of non-canonical GC-AG splice junctions represents about 0.8% of total splice sites in protein-coding genes, we identified a remarkable enrichment of the GC-AG splice sites in long non-coding genes, both in human (3.0%) and mouse (1.9%). In addition, we identified peculiar characteristics of the GC-AG introns in terms of their intron length, a positional bias in the first intron, their donor and acceptor splice sites strength, poly-pyrimidine tract, and alternative polyadenylation signaling. Genes containing GC-AG introns were found conserved in many species across large evolutionary distances and a functional analysis pointed toward their enrichment in specific biological processes such as DNA repair. Moreover, GC-AG introns appeared more prone to alternative splicing and enriched in a special alternative splicing mechanism termed wobble-splicing. Wobble-splicing appeared to be a rare mechanism, subjected to tissue-specific regulation and involved in inducing subtle changes in the expressed isoforms with a putative regulatory role. Taken together, our data suggests that GC-AG introns represent new regulatory elements mainly associated with lncRNAs, which could contribute to the evolution of complexity, adding a new layer in gene expression regulation.

# Abbreviations

| | |
|---|---|
| 3P-Seq | poly(A)-position profiling by sequencing |
| 3’ss | 3’ splice sites |
| 5’ss | 5’ splice sites |
| A | adenosine |
| A3 | alternative 3’ss |
| A5 | alternative 5’ss |
| ACC | anterior cingulate cortex |
| AMI | Amazon Machine Image |
| AMY | amygdala |
| APA | alternative polyadenylation |
| AWS | Amazon Web Services |
| bp | base pairs |
| BPS | branch-point sequence |
| C | cytosine |
| CAGE-seq | cap analysis gene expression sequencing |
| CAT | CAGE associated transcriptome |
| cDNA | complementary DNA |
| CER | cerebellum |
| circRNA | circular RNA |
| DEA | differential expression analysis |
| DEG | differentially expressed genes |
| df | degree of freedom |
| DHS | DNase hypersensitivity sites |
| DNA | deoxyribonucleic acid |
| EC2 | Elastic Compute Cloud |
| ENCODE | Encyclopedia of DNA elements |
| ES | exon skipping |
| ESE | exonic splicing enhancer |
| ESS | exonic splicing silencer |
| EST | expressed sequence tags |
| ETS | Erythroblast Transformation Specific |
| FDR | false-discovery rate |
| G | guanosine |
| GO | gene ontology |

| | |
|---|---|
| GRU | general research use |
| GTEx | Genotype-Tissue Expression |
| HBM | Human Body Map |
| HEA | heart |
| hnRNP | heterogenous nuclear ribonucleoprotein |
| ICGC | International Cancer Genome Consortium |
| Ins | insertion |
| IR | intron retention |
| ISE | intron splicing enhancer |
| ISS | intronic splicing silencer |
| kb | kilobases |
| KID | kidney |
| LE | last exon |
| lincRNA | long intergenic RNA |
| LIV | liver |
| lncRNAs | long non-coding RNAs |
| LUN | lung |
| Mb | megabases |
| miRNA | microRNA |
| mRNA | messenger RNA |
| MXE | mutually exclusive exons |
| NAT | natural antisense transcripts |
| NCBI | National Center for Biotechnology Information |
| ncRNAs | non-coding RNAs |
| NGS | next generation sequencing |
| NIH | National Institute of Health |
| NLS | nuclear localizing sequence |
| NMD | non-sense mediated decay |
| nt | nucleotides |
| ORF | open reading frame |
| PAS | polyadenylation site |
| PCA | principal component analysis |
| PCGs | protein-coding genes |
| PCR | polymerase chain reaction |
| PPT | polypyrimidine tract |
| PSI | percent-splice-in |
| RACE-Seq | Rapid Amplification of cDNA Ends Sequencing |
| RBP | RNA-binding protein |
| RefSeq | reference sequence database |

| | |
|---|---|
| RNA-Seq | RNA sequencing |
| RNA | ribonucleic acid |
| rRNA | ribosomal RNA |
| SF | splicing factor |
| siRNA | small interfering RNA |
| SKI | skin |
| sncRNA | short non-coding RNA |
| SNP | single-nucleotide polymorphism |
| snRNA | small nuclear RNA |
| snRNP | small nuclear ribonucleoprotein |
| SPL | spleen |
| spliRNA | splice sites RNA |
| SRE | splicing regulatory elements |
| T | thymine |
| TE | transposable elements |
| TES | testis |
| tiRNA | transcription initiation RNA |
| TPM | transcript per million |
| tRNA | transfer RNA |
| TSS | transcription start site |
| U | uracil |
| UTR | untranslated region |
| VPC | Virtual Private Cloud |
| WM | weight-matrix |
| WS | wobble splicing |
| Y | pyrimidine |

# I. Introduction

## 1. Genomes evolution and organism complexity

The development and wide-spread adoption of high throughput techniques in the field of functional genomics led to the discovery that mammalian genomes produce a large number of RNA molecules which do not encode for proteins. Even though we already have the genomic sequence of several complex organisms, our understanding of the complicated network that takes place inside eukaryotic cells is far from complete. Decades ago, it was appreciated that there were relatively few coding sequence changes between closely related species such as human and chimp [1][2]. It was therefore hypothesized that the evolution of gene expression regulation, rather than the protein sequence, would play a dominant role in driving evolutionary changes. Several questions still remain to be answered in regard to how all this genomic information is used by eukaryotic cells and the main players driving organismal complexity.

### 1.1. Evolution of regulatory non-coding genes

The genomes of distantly related species house remarkably similar numbers of protein-coding genes (PCGs) prompting the notion that many aspects of complex organisms arise from non-coding regions[3][4]. Non-coding RNAs (ncRNAs) are mature products of genes that are transcribed but not translated into proteins. The size of these ncRNAs can range from as small as 18-22 nucleotides (nt) to tens of kilobases (kb). Non-coding RNA genes can be found within introns of PCGs, proximally to the promoters of such genes, or in intergenic regions defined with reference to the PCGs. In recent years, ncRNA field has rapidly expanded with a fast increase in the number of newly identified and biologically relevant ncRNAs. Since then, the number of novel ncRNAs has increased dramatically and much more is known about their function, biogenesis, length, structural and sequence features.

The degree of organismal complexity among species better correlates with the proportion of each genome that is transcribed into non-coding RNAs than with the number of protein-coding genes [5]. Moreover, the analysis of sequenced genomes showed that the relative amount of non-protein-coding sequence increases consistently with complexity [6] (Figure 1).

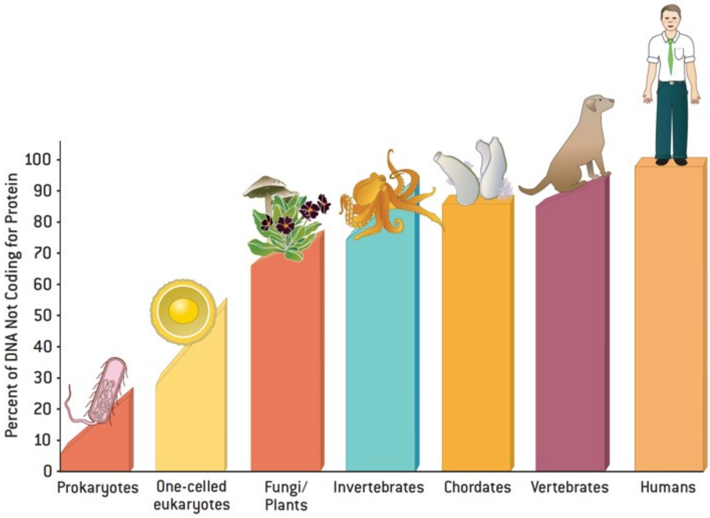


**Figure 1. Bar graph showing the percentage of non-coding DNA across species.** As the organism complexity increases, so does the proportion of their transcribed DNA that does not code for protein [7].

This suggests that RNA-based regulatory mechanisms had a relevant role in the evolution of developmental complexity in eukaryotes [4]. A substantial amount of research has been devoted to the identification and characterization of these non-coding RNAs, and the picture that has emerged indicates that they represent a broad and heterogeneous group of molecules with diverse roles in the regulation of biological processes [8].

Different classes of functional RNAs were discovered, which led to the realization that non-coding RNAs have a plethora of roles in gene regulation. NcRNAs can be divided into two major groups based on their nucleotides length namely short non-coding (sncRNA) and long non-coding RNAs (lncRNAs) (Figure 2). The sncRNAs include functional RNAs such as transfer RNAs (tRNAs), ribosomal RNAs (rRNAs) and small nuclear RNAs (snRNAs), which are involved in transcriptional and translational regulation, and regulatory sncRNAs. MicroRNAs (miRNAs) were one of the first classes of regulatory ncRNAs to be characterized in detail and it emerged that they represent a conserved family of small RNA molecules with important roles in the post-transcriptional regulation of gene expression [9]. These are further complemented by an ever-increasing catalog of small RNAs, including small interfering RNAs (siRNAs), PIWI and Argonaute-associated RNAs [10], species derived from structural or housekeeping RNAs such as snRNA-derived RNAs and tRNA-derived RNAs [11], and small RNAs derived from transcription start sites (tiRNAs) and splice sites (spliRNAs) [12]. Moreover, long non-coding RNAs (lncRNAs) recently emerged as new important regulators of gene expression which are implicated in various cellular processes and exhibit limited primary sequence conservation across species [13].

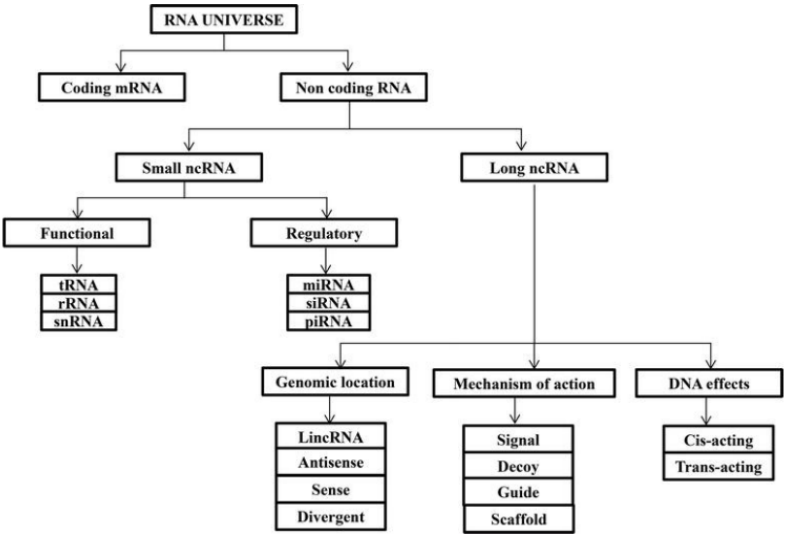


**Figure 2. Classification of non-coding RNAs (ncRNAs)**. The ncRNAs universe can be classified into short and long ncRNAs. sncRNAs include tRNAs, rRNAs, snRNAs involved in gene expression processes and miRNAs, siRNAs and piRNAs involved in gene expression regulation. LncRNAs forms a diverse class of molecules with different genomic locations and various mechanisms of action [14].

The widespread use of ncRNAs by eukaryotic cells in complex mechanisms of gene regulation might be intimately involved in the evolution, maintenance and development of complex life and indicates that evolution might have selected several of these transcripts for specific functions that can be associated with complexity of a molecular network [3][15].

## 1.2. Alternative splicing evolution and transcriptome complexity

It was proposed that alternative splicing could be accountable for complex gene regulation architectures. Alternative splicing greatly expands the coding capacities of genomes by allowing the generation of multiple mature mRNAs from a limited number of genes. Alternative splicing generates widespread differences between the proteomes of mammalian tissues, and several studies suggests that the evolutionary trajectories of mammalian proteins are strongly biased by the locations and phases of the introns that interrupt coding sequences [16]. It has been proposed that the unexpectedly high frequency of alternative splicing might provide an attractive explanation for increasing organismal complexity in higher eukaryotes [17].

The proportion of alternatively spliced genes has increased in line with the evolutionary diversification of cell types, suggesting that alternative splicing may contribute to the complexity of developmental programs [18]. Although other genomic features have been shown to correlate with cell-type number and may be important contributors to the evolution of complexity, alternative splicing appeared as a mechanism allowing transcript diversification in the absence of any increases in gene number [18]. Comparative studies have reported marked differences in the prevalence of alternative splicing across eukaryotic lineages as well as a significant correlation between alternative splicing and the number of cell types per species. These results are in principle consistent with an adaptive role of alternative splicing in determining genome's functional information capacity and in facilitating transcript diversification in species with greater numbers of cell types. Another way to assess the functional contribution of alternative splicing is to examine the conservation of alternatively spliced events across evolution. A base-by-base analysis of the genotype–phenotype relationship for an alternatively spliced human exon demonstrated that nucleotides which affect splicing regulation are dispersed through the entire length of the exon, and so any mutations could conceivably alter the accuracy of splicing [19]. This predicts the rapid divergence of alternative splicing events between species in which new species-specific alternative splicing events could arise from various mutations along the exon sequences, except if they are selectively constrained. Conserved patterns of differential exon usage, and in some cases, conserved expression profiles have been observed for a proportion of genes in species including nematodes and mammals [20][21]. Conserved alternative splicing profiles have also been identified in a lineage-, species- and tissue-specific manner among vertebrate. These studies suggest that in some of the species with the highest number of cell types, a significant number of the alternative splicing events are functional.

A significant amount of literature was dedicated to hypotheses concerning the origin of alternative splicing events, and accumulating evidence indicates that new exons are also constantly being added to evolving genomes. Several mechanisms contribute to the creation of novel exons in metazoan genomes including whole gene and single exon duplications. But perhaps among the most intriguing are events of exonization, where intronic sequences become exons de novo [22]. Exonizations of intronic sequences, particularly those originating from repetitive elements such as Alu sequences, are now widely documented in many genomes including human, mouse, dog, and fish (Figure 3). Alu sequences insert into the introns of primate genes by retro-transposition usually in the antisense orientation. The consensus Alu sequence carries multiple sites that are similar, but not identical, to real splice sites; thus a few mutations in the 3′ splice site (3′ss) or 5′ss are required to create a new exon [22].

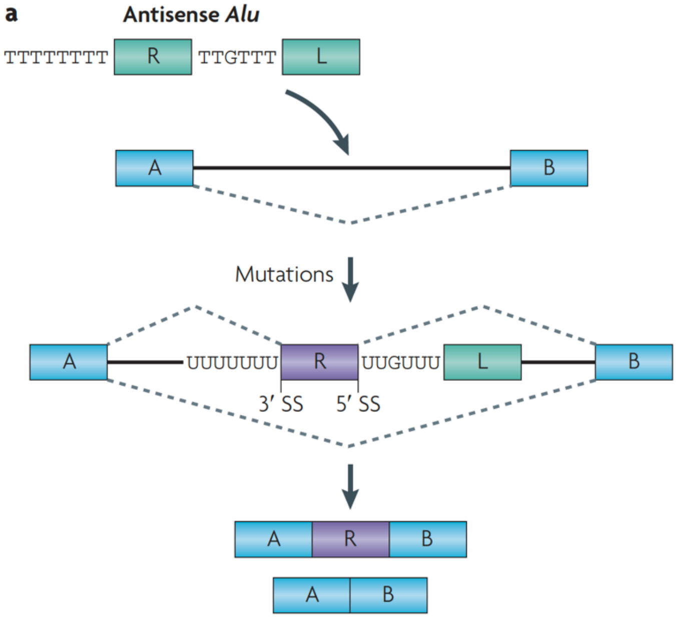


**Figure 3. Exonization from Alu elements**. A typical Alu is around 300 nucleotides (nt) long and contains two similar monomer segments (the right arm and the left arm, green R and L in the figure) joined by an A-rich linker and a poly(A) tail-like region. 85% of exonizations occur from the right arm in the antisense orientation. The poly(A) tract of this arm in the antisense orientation creates a strong polypyrimidine tract (PPT). Downstream from this PPT, a 3′ss is selected, and further downstream from that site (approximately 120 nt), a 5′ss is recognized [22].

Such de novo appearance of exons is very frequently associated with alternative splicing, with the new exon-containing variant typically being the rare one [23]. This allows exonized sequences to increase the coding and regulatory versatility of the transcriptome but at the same time maintain the intactness of the original proteomic repertoire [24]. As evolution proceeds, some of the newly emerged exons might be fixed, and their expression then becomes more pronounced. Indeed, it was shown that inclusion levels of new exons become higher for older exonizations [25], and that this increase of inclusion levels is correlated with mutations creating stronger splice sites [26].

## 2. Long non-coding RNAs

The advent of high-throughput sequencing technologies has provided an unprecedented opportunity to explore the complexity of mammalian transcriptome which its breadth and diversity still remains unrealized. New sequencing approaches brought the discovery of novel regions in the genome with transcriptional potential [27]. A large portion of mammalian genomes is transcribed to produce non-coding RNAs among which lncRNAs are the most prevalent [28]. Despite their lack of coding capacity and relatively low conservation, many lncRNAs molecules have been shown to be functional, adding a new level of complexity on the structural organization, function, and evolution of the genome [29]. The interest in lncRNAs is deeply rooted in biology's longstanding concern with evolution and function of genomes [30].

### 2.1. LncRNAs definition and genomic characteristics

LncRNAs are traditionally defined as transcripts of more than 200 nucleotides that lacks a coding potential [31]. This is an arbitrary definition and there is no definition of lncRNA that is based on biological argumentation and widely accepted in the community [32].

Most lncRNAs are regulated, transcribed, and processed in a similar fashion to mRNAs [28]. Similarly to mRNAs, many lncRNAs are transcribed by RNA polymerase II, regulated by conventional transcription factors, capped at their 5' ends, and polyadenylated at their 3' ends [33][34]. Moreover, lncRNAs are spliced, exhibit standard canonical splice site signals, and undergo alternative splicing [35]. The genomic structure of lncRNAs is very similar to that of PCGs, although surprisingly, it has been reported that a high proportion of lncRNAs are two-exon transcripts [35].

LncRNAs show a relatively lower level of expression than PCGs and a highly tissue specific expression reflecting a heightened spatiotemporal precision in their transcription [36]. LncRNAs transcription arises from recognizable promoters which are enriched for transcription factors binding sites and the canonical histone marks of active gene expression or repression such as H3K4me3, H3K9ac, H3K27ac and

H3K9me3 [37]. The transcription of lncRNAs was mainly studied in relationship to those of nearby PCGs. In many cases, lncRNAs were reported to be co-expressed and co-regulated with their neighbor PCGs especially when divergently transcribed from bidirectional promoters [38] [39]. In some cases, the direct involvement of lncRNAs in the transcription regulation of neighbor PCGs was demonstrated. In the work of Luo and colleagues [40], the correlation between the expression of some lncRNAs and of the neighbor PCGs was experimentally demonstrated and estimated to account for 75% of total lncRNAs. As an example, the lncRNA *Evx1as* (EVX1- antisense RNA) was reported to promote the transcription of the *Evx1* (even-skipped homeobox 1) gene during mesodermal differentiation by modifying chromatin accessibility [40].

Moreover, it was also reported that lncRNA transcription itself, rather than the RNA transcript, exerts regulatory effects on neighboring genes [41][42]. For example, the silencing of *Igf2r* (insulin like growth factor 2 receptor) gene expression was demonstrated to be due to the transcription of the lncRNA *Airn* (antisense of Igf2r non-protein coding RNA) that interfere with RNA polymerase II recruitment [43]. Similarly, Anderson and colleagues [44] reported that the lack of transcription of the lncRNA *Hand2os1* (Hand2, opposite strand 1), but not the knockdown of its mature transcript, abolished the expression of the *Hand2* (heart and neural crest derivatives expressed 2) gene leading to embryonic lethality in mice. Taken together, this evidence points toward the tight regulation of lncRNAs expression and the importance of lncRNAs transcription in regulating PCGs.

### 2.2. Classification of lncRNAs

LncRNAs can be classified into the following locus biotypes based on their location with respect to protein-coding genes (Figure 4):

i. *Antisense RNAs* which are transcripts that intersect any exon of a protein-coding locus on the opposite strand, or have a published evidence of antisense regulation of a coding gene. Abundant transcription appears to occur opposite the sense DNA strand of annotated PCGs. The overlap between these sense–antisense pairs can be complete, with either transcript nested within the other or tend mostly to be enriched around the 5’ promoter or 3’ terminator ends of the sense transcript. The latter transcripts are usually termed natural antisense transcripts (NAT) [30].

ii. *LincRNA* are transcripts that are intergenic noncoding RNA loci. These lncRNAs are distinct transcription units located in sequence space that do not overlap PCGs. Some of these have been referred to as “lincRNAs” for “large intergenic (or intervening) noncoding RNAs” [45][46][47]. A large number was identified through chromatin signatures for actively transcribed genes H3K4me3 at the promoter and H3K36me3 along the transcribed length.

iii. *Sense overlapping transcripts* contain a coding gene within an intron on the same strand. These transcripts cover the entire sequence of a PCG which is transcribed from within the lncRNA transcript itself [35].

iv. *Sense intronic transcripts* reside within introns of a coding gene, but do not intersect any exons. Many long transcripts have been reported, by large-scale transcriptomic or computational analyses, to be encoded within the introns of annotated genes [48][49]. Many of these are observed to have differential expression patterns, respond to stimuli, or be misregulated in cancer, but only a few have been studied in detail to date [50].

v. *Bidirectional promoter transcripts* which are abundant short transcripts (ranging from 200 to 2500 nt) that have been found to be produced from the vicinity of transcription start sites in both sense and antisense directions, corresponding to peaks of Pol II occupancy due to pausing. Several studies have reported an essential role of these lncRNAs in the gene expression regulation of their corresponding PCGs at the transcriptional, post-transcriptional and epigenetic level [51][52][53][54][55].

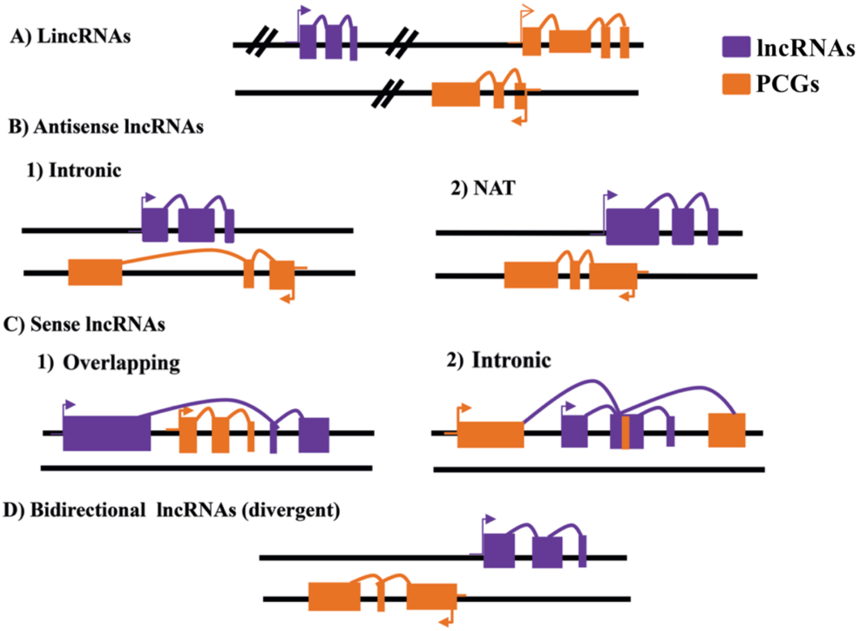


**Figure 4. Classification of lncRNAs.** The classification is based on the position relative to the nearest protein coding gene. (A): Long intergenic RNAs (lincRNAs). (B): Antisense lncRNAs (intronic and natural antisense transcripts). (C): Sense lncRNAs (intronic and sense overlapping). (D): Bidirectional promoter lncRNAs (divergent) (modified from [56]).

In addition to these main biotypes, lncRNAs can fall into other classifications such as *3′ overlapping ncRNA* which are transcripts where published experimental data strongly supports the existence of long (>200 bp) non-coding transcripts that overlap the 3'UTR of a protein-coding locus on the same strand, *macrolncRNAs* which are unspliced lncRNAs that are several kb in length, and *processed transcripts* which do not contain an open reading frame (ORF) and cannot be placed in any of the other categories. The classification of lncRNAs into the above mentioned biotypes was initially adopted to aid in the understanding of the functional role of the transcribed isoforms. However, little is known so far about the lncRNAs functions and their classification only with respect to PCGs could not reflect their functional properties completely. It is worth to mention here that the classification of lncRNAs into the different biotypes is usually tricky and complex, and in many cases a lncRNA transcript can fall into more than one biotype.

### 2.3. LncRNAs conservation and evolution

Across species, lncRNAs have markedly different sequence conservation patterns than PCGs. The evolutionary history of lncRNAs provides important insights into their identification and functionality. Overall, lncRNAs have a higher sequence conservation than the neutrally evolving regions of the genome, but lower conservation than protein coding genes. Indeed, although exceptionally highly conserved lncRNAs exist, such as the transcribed ultraconserved regions, lncRNAs in general are under more selective pressure than ancestral repeat sequences with neutral selection, and about one-third of lncRNAs seem to have arisen within the primate lineage. Furthermore, it has been shown that 81% of lncRNA families are primate specific, which is consistent with previous studies that propose that lncRNA transcription can evolve extremely rapidly, even between closely related mammals. Interestingly, the number of lncRNAs encoded by the genome has increased during animal evolution, suggesting that the presence of lncRNAs could be linked to organismal complexity (Figure 5) [29].

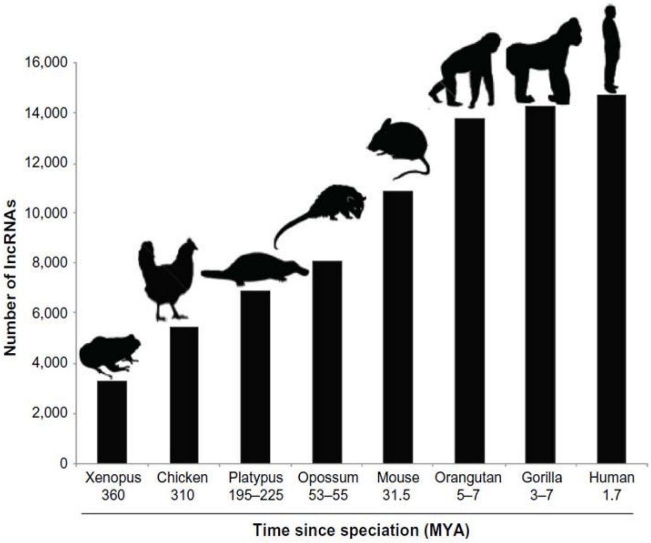


**Figure 5. Bar graph showing the number of encoded lncRNAs in different species.** During evolution, the number of lncRNA across large evolutionary distant species has been increasing [29].

Compared with protein-coding RNAs and other non-coding RNAs, mammalian lncRNAs have fewer invertebrate orthologues and have undergone rapid evolution. For example, approximately 5% of mammalian lncRNAs are conserved in zebrafish, and conservation is typically restricted to short polynucleotide stretches [47].

A relatively low level of conservation may be explained by the fact that structural and regulatory functions of the ncRNA molecules mostly depend on their 3D shape, as opposed to functions of protein-coding RNAs which mainly depend on their primary sequence. The likelihood of secondary-structure-driven evolution is supported by the overrepresentation of transposable elements (TE) sequences (known to maintain complex and stable RNA secondary structures) in lncRNAs as compared to protein coding genes. Another possible reason for low lncRNA sequence conservation among species could be the wide participation of TEs in their evolution [57]. TE abundance and composition in mouse and human genomes are substantially different. For example, primate-specific Alu elements are enriched in human lncRNA, while in mouse, where Alu are not present, the mouse-specific elements stand in, creating different conditions for lncRNA evolution in these species. Indeed, while the transmapping of human genes to the mouse genome showed that 88% of PCGs are conserved, only 15% of lncRNA genes appeared conserved in the two species (Figure 6) [58].

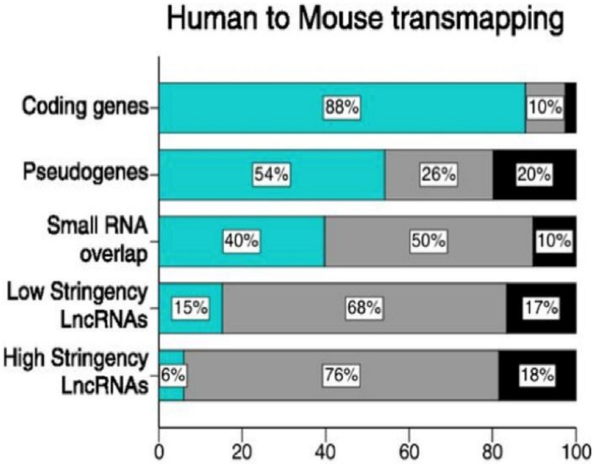


**Figure 6. Human to mouse genes transmapping**. Percentages of human genes transmapping to the mouse genome and expressed in mouse erythroblasts showing low levels of conserved lncRNA expression [58].

LncRNAs poor conservation has confounded efforts to predict their functions across species but has also helped to highlight focal areas of potential functional importance.

Despite showing a low level of conservation at the sequence level, lncRNA show high conservation in their regulatory elements such as in their promoters and enhancer sequences. Conservation of parts of their transcript structure constitutes compelling evidence for stabilizing selection, despite the often negligible constraints on the sequence itself [59]. Indeed, splice sites have much higher conservation than the rest of lncRNA sequence and can be used to trace the evolution of lncRNAs. It has been reported that mammalian lncRNAs show a high evolutionary conservation of the exon-intron structure and some lncRNA introns have been conserved for over 100 million years [60]. Thus, this predicts the important regulatory role of these elements and that the primary and/or secondary structure of these molecules is likely to be highly relevant for the correct function of lncRNAs.

### 2.4. LncRNAs subcellular localization and functions

LncRNAs show precise patterns of subcellular localization which could be informative regarding their functions. LncRNAs have been reported to be more nuclear on average than mRNAs [35][61], but the determinants of this difference are largely unknown. The majority of lncRNAs with an identified functional role has been reported to localize more in the nucleus than the cytoplasm. This could be attributed mainly to the high turnover rate of lncRNA transcripts making their subcellular detection a challenging task. Indeed, recent studies using high resolution transcriptome-wide maps suggested that the numbers of cytoplasmic lncRNAs transcripts could rival those in the nucleus [62]. Understanding lncRNAs subcellular localization is a powerful and necessary step toward understanding the nature and mechanisms of their functions in the cell.

LncRNAs are versatile molecules able to perform numerous tasks in the cell through binding of proteins, DNA or other RNA molecules. LncRNAs can be broadly classified into those that act in cis, influencing the expression and/or chromatin state of nearby genes, and those that execute an array of functions throughout the cell in trans (Figure 7) [63].

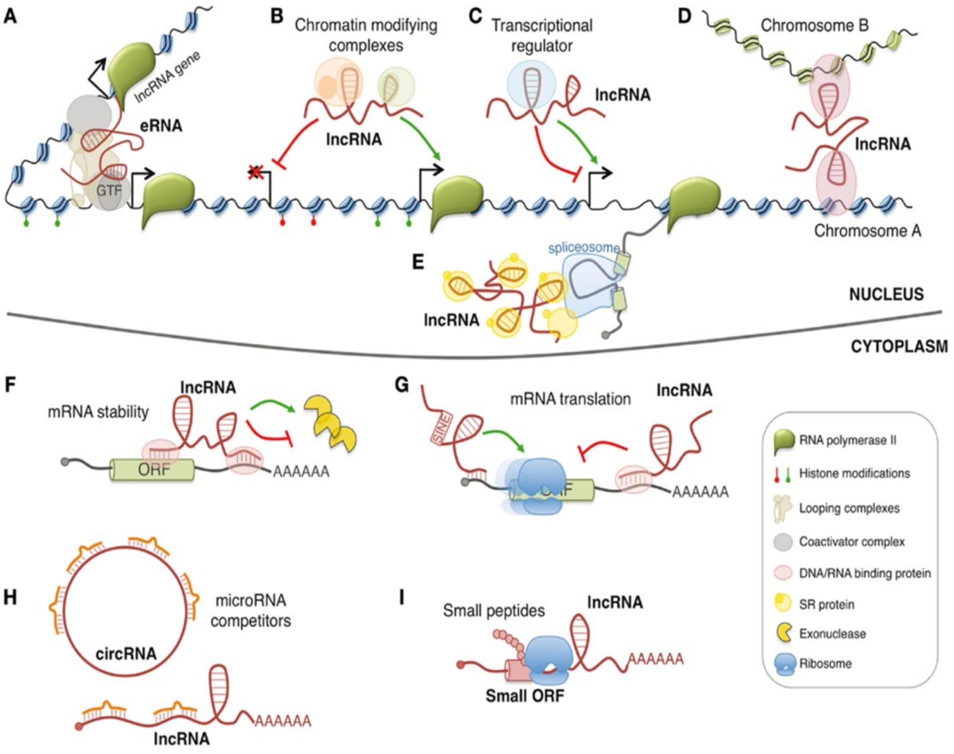


**Figure 7. Models of the cellular functions of lncRNAs.** Genomic location relative to regulatory mechanisms of lncRNAs in the nucleus, cytoplasm, and extracellular compartments. Nuclear-localized lncRNAs can act as enhancers to induce transcription in cis or in trans (A), regulate transcription by recruiting chromatin modifying complexes (B), or by regulating transcription factors activity (C). Moreover, they can regulate gene expression by acting on the spatial conformation of chromosomes (D) or by influencing pre-mRNA splicing (E). Cytoplasmic lncRNAs can regulate mRNA expression by regulating mRNA stability (F), mRNA translation (G), miRNAs sponge by competing for microRNA binding (H). In addition, few lncRNAs contain small open reading frames (ORFs) that can be translated in biological active small peptides (I) (adapted from [64]).

A significant number of nuclear lncRNAs have been suggested to interact with chromatin-remodeling complexes, driving them to specific genomic loci, while others have been implicated in the architectural conformation and activity of transcriptional enhancers, interfere with the transcriptional machinery and the spliceosome, or to maintain the structure of nuclear speckles. For instance, the lncRNA *Mistral* (*Mira*) was reported to recruit the H3K4 methyltransferase MLL1 (lysine methyltransferase 2A) and activates the expression of the *HOXA6* (homeobox A6) locus. Moreover, many reported chromatin-interacting lncRNAs are found to bind to repressive chromatin modifiers, such as PRC2 or H3K9 methyltransferases. In several cases, the interaction of the lncRNAs with the chromatin complexes is required for the repression of specific gene loci. For example, the lncRNA *HOTAIR* (HOX transcript antisense RNA) is expressed from the *HOXC* gene cluster and represses the *HOXD* locus by interacting with the chromatin complexes PRC2 and LSD1 [65][66].

An emerging theme is the role that lncRNAs may play as structural elements, contributing to the nuclear architecture. *FIRRE* (firre intergenic repeating RNA element), a lncRNA located on the X chromosome, regulates transgenomic regions in cooperation with hnRNPU, a protein also previously shown to be associated with proper localization of *XIST* (X inactive specific transcript) and the formation of highly structured chromatin territories [67]. Also, the highly expressed nuclear lncRNAs, *NEAT1* (nuclear paraspeckle assembly transcript 1) and *MALAT1* (metastasis associated lung adenocarcinoma transcript 1), have been related to nuclear architecture [29].

LncRNAs can regulate gene expression by acting as an enhancer in which they interact with DNA to upregulate gene transcription through two possible mechanisms such as enhancer-promoter looping and tracking of transcriptional machinery, forming a scaffold for a protein complex that bridges the enhancer-like element and the promoter of a coding gene [68]. A number of enhancer lncRNAs have been shown to be associated with the mediator complex which is involved in bridging promoters with enhancers; and depletion of this complex inhibits looping between the activating lncRNAs locus and its target gene.

While several of the best-known lncRNAs exert their functions in the nucleus of the cell, other functional lncRNAs are localized in the cytoplasm and regulate gene expression at the post-transcriptional level. Some of the cytoplasmic lncRNAs have been reported to work through RNA–RNA interactions. Sequence complementarity-mediated interactions allow for the regulation of mRNA stability, transport, translation, and miRNA decoys by lncRNAs. LncRNAs can mediate mRNAs degradation by promoting the recruitment of proteins that mediate this process. In contrast, other lncRNAs, such as *UCHL1-AS1* (UCHL1 antisense RNA 1), enhance mRNA translation. *UCHL1-AS1* expression is induced by mTOR and shuttled to the cytoplasm where, via an antisense complementary to the UCHL1 AUG initiation codon and combined inverted SINEB2 domains, it increases UCHL1 protein synthesis [69].

An interesting mechanism of action of lncRNAs that has come into focus is their activity as microRNA (miRNA) sponges. LncRNAs that are known to exert such an action harbor sequences complementary to miRNA sequences thereby sequestering them and preventing them from binding to their targets [31]. LncRNAs may also contribute to or inhibit the formation of macromolecular complexes by allowing or blocking protein–protein interactions. Additionally, some lncRNAs has been reported to act as regulators of protein post-translational modifications [70].

### 2.5. Roles of lncRNA in common diseases

LncRNAs have emerged as key regulators in a wide range of biological processes such as cell proliferation, cell cycle, metabolism, apoptosis, differentiation and maintenance of pluripotency [31]. Therefore, altered lncRNA function is identified as one of the causes for the dysregulation of gene expression which leads to several human diseases (Figure 8).

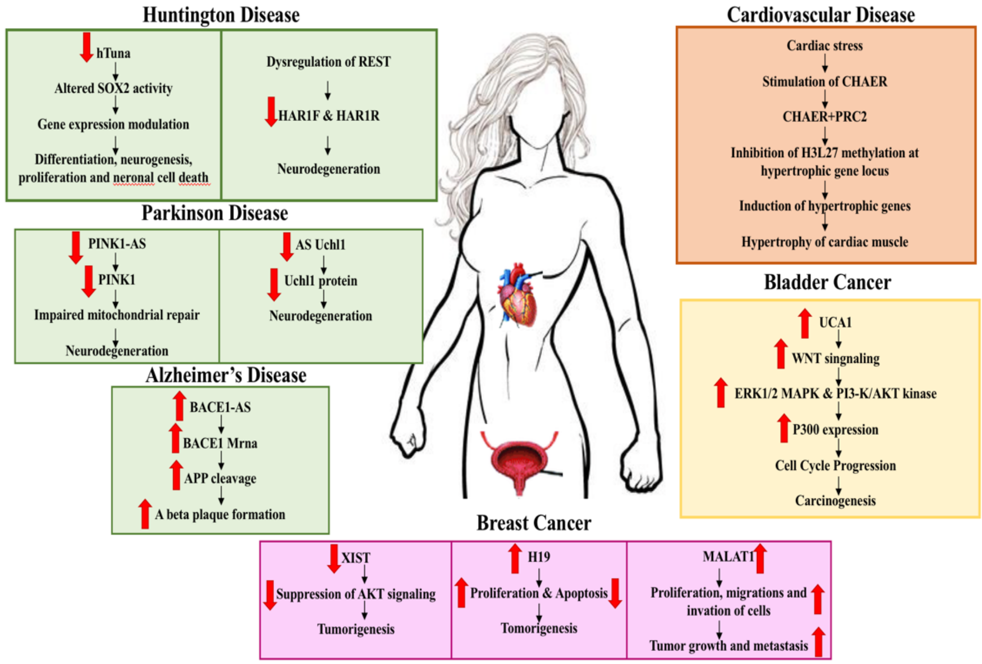


**Figure 8. The role of lncRNAs in human diseases**. Huntington Disease: Human lncRNA *hTUNA* negatively correlates with severity of Huntington Disease by regulating *SOX2* activity. Cardiovascular Disease: Heart-enriched lncRNA, *Chaer* is upregulated by cardiac stress and is involved in development of cardiac hypertrophy. Parkinson Disease: LncRNAs *PINK-As* and *AS Uch1* negatively correlate with severity of Huntington Disease. Bladder Cancer: LncRNA *UCA1* is highly expressed in bladder tumor tissues and promotes cell growth and tumorigenesis. Alzheimer's Disease: Upregulated lncRNA *BACE1-AS* led to a significant increase of *BACE1* mRNA in AD brains, subsequently exacerbating Aβ plaque formation. Breast Cancer: LncRNAs, *H19* and *MALAT1* are both significantly upregulated in breast cancer and are involved in tumorigenesis and tumor growth. LncRNA *XIST* is downregulated in breast tumor and acts as a tumor suppressor via regulating AKT signaling. (modified from [71])

The presence of lncRNA in biological fluids and their deep involvement in disease pathogenesis make them ideal candidates for the development of efficient diagnostic assays such as in neurological and neurodegenerative diseases where disease site is largely inaccessible [72]. For example, Alzheimer disease-specific lncRNA transcripts such as (*BACE1-AS* or *BC200*) were present at increased levels in the blood of affected individuals [73]. Furthermore, Liu and colleagues have identified several lncRNA in the blood which were differentially expressed in subjects with major depressive disorder compared to healthy controls and could be used for diagnostics [74].

LncRNAs have also been associated with different types of cancer [75]. For example, a recent RNA-Seq study in prostate cancer tissues and cell lines uncovered a lncRNA, *PCAT-1* (prostate cancer associated transcript 1), that promotes cell proliferation and is a target of PRC2 regulation while also possibly interacting with PRC2 itself [76]. *ANRIL* (CDKN2B antisense RNA 1), also upregulated in prostate cancer, is required for the repression of the tumor suppressors INK4a/p16 and INK4b/p15 [77][78]. *HOTAIR* (HOX transcript antisense RNA) overexpression is associated with poor prognosis in breast cancer [79], liver cancer [80], colorectal cancer [81], gastrointestinal cancer [82], and pancreatic cancer [83], and is proposed to increase tumor invasiveness and metastasis [84]. *MALAT1*, another lncRNA associated with various cancers and metastasis [85], is found to affect the transcriptional and post-transcriptional regulation of cytoskeletal and extracellular matrix genes [86]. *lincRNA-p21* (tumor protein p53 pathway corepressor 1), named for its vicinity to the CDKN1A/p21 locus, is upregulated by p53 upon DNA damage and implicated in downstream repressive effects of the p53 pathway, particularly on genes regulating apoptosis, possibly by directing the recruitment of hnRNP-K to its genomic targets [87]. Another DNA damage responsive, p53-induced lncRNA that lies upstream of p21, *PANDA* (P21 associated ncRNA DNA damage activated), is also implicated in the repression of pro-apoptotic genes, such as *FAS* (Fas cell surface death receptor) and *BIK* (BCL2 interacting killer), by acting as a decoy for the transcription factor NF-YA [88][89]. The above examples suggest that lncRNAs may be used as diagnostic markers or therapeutic targets in the treatment of cancer, but much work needs to be done before such applications become clinically practical.

Aging is a complex physiological phenomenon with a progressive decline in functional capacities and environmental adaptations. The expression of lncRNAs is known to be affected during aging process and in turn, many lncRNAs govern major senescent pathways and senescence-associated secretory phenotypes [90][91]. In human fibroblasts, senescence-associated *lncRNA-SAL-RNA1* (senescence associated long non-coding RNA 1) delays senescence and reduced levels of this lncRNA enhances senescence traits such as enlarged morphology, increased p53 levels and positive β-galactosidase activity [92].

LncRNAs also have roles in other diseases like neurogenetic Angelman syndrome and Beckwith-Wiedemann syndrome [73]. LncRNAs have also been associated with other neurological disorders such as *HAR1* (human accelerated region 1 lncRNA) in Huntington disease and *ATXN8OS* (Ataxin8 opposite strand lncRNA) in spinocerebellar ataxia type 8 [93][94].

### 2.6. LncRNA genes annotations resources

The methodologies applied for the detection and identification of lncRNAs have evolved hand in hand with the genomic technologies. More recently, the development of next-generation sequencing (NGS) applied to RNA has allowed for the direct detection and assembly of entire transcriptomes. RNA sequencing (RNA-seq) offers a more reliable and high-resolution method for measuring gene expression than previous techniques such as microarrays and PCR. Furthermore, thanks to computational methods for transcriptome reconstruction, novel transcribed regions, as well as alternative spliced forms of annotated genes, can be identified. LncRNA annotations lag considerably behind those of PCGs, for reasons that go beyond their more recent discovery. There are at least three factors that make lncRNA annotation challenging: (1) lncRNAs are relatively lowly expressed, (2) our understanding of the lncRNA sequence–function relationship is poor, (3) lncRNAs show a low level of sequence conservation [95].

Systematic curation efforts have enabled the development of several lncRNA databases and a range of lncRNA annotation resources obtained by different methods are presently available:

i. *GENCODE*: The most widely used manual annotation is GENCODE which stands out thanks to its extensive experimental validation and integration into the Ensembl annotation set. GENCODE is one of the most exhaustive and comprehensive database for lncRNAs gene annotations [35][96][97]. GENCODE comprises gene annotations in the human and mouse genomes. Whereas the main GENCODE PCGs annotation is created by merging the output from two pipelines, one manual and one automated, the lncRNA annotation is almost entirely manual. GENCODE has adopted in the recent years a stringent lncRNAs annotation pipeline based on long-reads sequencing associated with CAGE-seq thus providing an accurate annotation of lncRNA genes. Where there is no evidence of coding potential from mass spectrometry data, orthologues or paralogues in reference databases such as UniProt, structural or functional protein domains identified by Pfam, or conservation data such as PhyloCSF, a locus is defined as noncoding. Gene annotations are regularly released as the Ensembl/GENCODE gene sets providing constant improvements and updates. The gene sets are comprehensive and include protein-coding and non-coding loci including alternatively spliced isoforms and pseudogenes [97]. Owing to the quality deriving from its manual annotation, regularly updated versions, long- term support, well- defined and consistent source data, identifier stability and integration into Ensembl, GENCODE has been adopted by most large-scale genomics projects, including the Encyclopedia of DNA Elements (ENCODE) (for which it was originally created), Genotype-Tissue Expression (GTEx) project, International Cancer Genome Consortium (ICGC), and Epigenome Roadmap. The use of stable Ensembl identifiers simplifies the integration of data across projects and releases. GENCODE serves as a major source of lncRNA annotations with a significant dynamism in gene annotations, reflective of the evolution and consensus in nomenclature of genes [98].

ii. *RefSeq*: Another manual gene annotation resource, Reference Sequence (RefSeq), was created and is maintained by the National Center for Biotechnology Information (NCBI) and covers multiple species, including human [99]. Consisting of a mixture of manual and automated annotations, RefSeq is created using a variety of evidence, including cDNAs, ESTs and RNA-seq. Entries carry unique and stable identifiers and are associated with metadata summarizing their annotation history. The RefSeq annotation process is similar to GENCODE, with the exception of usage of RNA-seq. Along with GENCODE, RefSeq is one of the most widely used database for lncRNA annotations.

iii. *FANTOM*: The Functional Annotation of the Mammalian genome (FANTOM) CAGE-associated transcriptome (CAT) meta-assembly combines both published sources and in-house short-read assemblies [100] [101]. What sets this collection apart is its use of CAGE tags, which mark transcript transcription start sites (TSS), to identify 5′-complete transcript models. The resulting gene loci are more complete at the 5′ end compared with other annotations, as judged by independent evidence, such as histone 3 lysine 4 trimethylation (H3K4me3) and DNase I hypersensitivity sites (DHSs).

iv. *BIGTranscriptome*: The BIGTranscriptome catalogue comprises transcripts that are complete at both the 5′ and 3′ ends [102]. It employs a new method, CAFE, which is capable of inferring strands of unstranded RNA-seq reads. Consequently, CAFE overcomes strand ambiguity, which particularly affects genic transcript models generated from unstranded data sets, such as those from the Human Body Map (HBM) or the Genotype-Tissue Expression (GTEx) project. CAGE and poly(A)-position profiling by sequencing (3P-seq) were used to assess 5′ end and 3′ end completeness, respectively. Combining 169 RNA-seq data sets, BIGTranscriptome comprises novel full-length lncRNA loci and represent one of the databases with the most accurate annotation of lncRNAs 5’ and 3’ ends.

v. *MiTranscriptome*: The MiTranscriptome annotation combines 6,503 data sets, heavily weighted to 27 cancer types, to automatically annotate lncRNA genes using a two-stage assembly strategy [103]. At the time of its creation, 54% of loci were not present in any other available resource. Several studies are taking steps to improve the completeness of annotations.

vi. *NONCODE*: the NONCODE dataset has integrated annotations from a mixture of manual literature searches and other annotations [104]. The latest version, NONCODE (version 5), is the single largest present collection describing lncRNA gene loci in human. It also has data for 15 species other than human and mouse.

vii. *RNACentral*: this dataset is a large-scale resource of non-coding RNA sequences from a broad range of species, integrating various other databases which lists lncRNA sequences [105]. It is based on sequences, rather than annotations, making the total number of lncRNA loci unclear.

viii. *LNCipedia* stands out in its usefulness for integrating functional data. LNCipedia holds a database of carefully filtered lncRNA genes from a range of sources [106]. LNCipedia provides access information on peptide mapping, coding potential, RNA folding and microRNA recognition, disease association and putative small peptides, and is an invaluable resource of manually curated functional information for hundreds of lncRNAs [95].

ix. *LncBook*: LncBook is a curated knowledgebase of human lncRNAs that features a comprehensive collection of human lncRNAs and systematic curation of lncRNAs by multi-omics data integration, functional annotation and disease association [107]. In the present version, LncBook houses a large number of lncRNAs and includes lncRNA–function associations. Also, it incorporates 3772 experimentally validated lncRNA-disease associations and further identifies lncRNAs that are putatively disease-associated.

Though there have been a number of databases systematically annotating various aspects of lncRNAs including their functions and interactions, most databases have been lacking continuous updates. GENCODE fills in this gap by covering and integrating the latest in terms of gene and transcript annotation, methodologies, and standards [98], and it is the database of choice for the analysis in this current research project.

## 3. Splicing and its regulation mechanism

### 3.1. Splicing definition and elements

Splicing can be defined as a process in which certain intronic regions are stripped off from precursor transcripts followed by joining of exons to form functional mature RNAs [108]. Splicing represents a main mechanism of post-transcriptional regulation of gene expression. It is not only involved in the maturation of pre-mRNAs, but can also influence the subcellular localization of mature transcripts and increase transcriptional rates by several folds [109][110].

Splicing involves 4 main elements: (1) the donor splice site, (2) the acceptor splice site, (3) the branch-point sequence (BPS), and (4) the polypyrimidine tract (PPT) (Figure 9) [111].

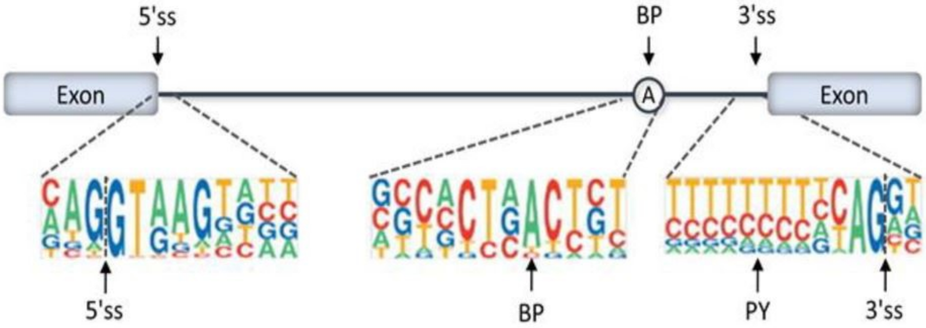


**Figure 9. Major splicing elements.** Exons are represented by boxes and introns by lines. The most conserved nucleotides at the 5' splice site (5'ss), branch point (BP), polypyrimidine tract (PY), and 3' splice site (3'ss) are indicated (adapted from [112]).

The donor splice site refers to the dinucleotide sequence characterized by a highly conserved GY (where Y is a pyrimidine) at the 5' end of the intron while the acceptor splice site refers to the dinucleotide sequence which is mostly AG at the 3' end of the intron. Typically, both BPS and PPT are located within the intron, before the 3' acceptor site [113]. The polypyrimidine tract is a short nucleotidic sequence, usually up to 20 nucleotides, rich in pyrimidines and especially uridines, recognized and bound by splicing factors, and thought to assist the correct positioning of the spliceosome complex [113][114]. The branch-point sequence, located few nucleotides upstream of the PPT, contains a highly conserved adenosine "A" that is crucial for the splicing reaction [115]. Moreover, especially in long introns, that may contain many cryptic sites, intronic signals may be not sufficient to correctly define exon-intron junctions. For this reason, exon definition and splicing factors positioning can be assisted by splicing enhancers that are either exonic (ESEs) or intronic (ISEs) short regulatory sequences [116]. On the other hand, the opposite effect can be promoted by exonic and intronic silencer sequences (ESSs and ISSs) which can reduce the splicing efficiency [8].

### 3.2. Types of RNA splicing and their mechanism

Splicing of the majority of introns in eukaryotes is accomplished through the actions of the major spliceosomes, that are large ribonucleoprotein complexes, found in eukaryotic nuclei, and composed by small nuclear RNAs (snRNAs) and several associated proteins [114]. Since the spliceosomal snRNAs are uridine-rich, they are commonly labelled with the letter "U". At least two different types of spliceosomes are known: the U2-dependent and the minor U12-dependent spliceosomes.

The U2-dependent spliceosome is the most common, being present in all eukaryotes, whereas the non-canonical U12-type is not ubiquitously found in all organisms [117]. From a molecular point of view, these two spliceosomes are composed by different subunits, and recognize different introns. The U2-dependent spliceosome is composed by U1, U2, U5, U4 and U6 snRNAs, assembled together with many proteins. On the contrary, the minor U12-spliceosome contains the U11, U12, U5, U4atac and U6atac snRNAs and corresponding proteins. The vast majority of the introns are recognized and spliced by the U2-dependent spliceosome, thanks to some distinctive elements, including particularly conserved GT-AG motifs found at the two edges of each intronic sequence [113][118][119]. In *Homo sapiens*, just a small fraction of introns are recognized by the U12-spliceosome [120]. The minor spliceosome is originally reported to recognize AT-AC splice sites although other canonical GT-AG motifs are also frequent in U12-type introns [113][121].

The splicing process is performed in two major steps, catalyzed by the spliceosome (Figure 10). The first reaction occurs between BPS and donor splice site: the hydroxyl group of the BPS adenosine performs a nucleophilic substitution and displaces the last nucleotide of the upstream exon, by forming a new covalent bond with the first nucleotide of the donor splice site. As a result, the phosphodiester bond between upstream exon and intron is broken.

The second transesterification is similar, and occurs between the free hydroxyl group of the upstream exon and the phosphate group of the downstream exon, right after the acceptor splice site. This latter nucleophilic substitution joins together the two exons, while the intron is released with a peculiar circular structure indicated as the lariat [113][122][123].

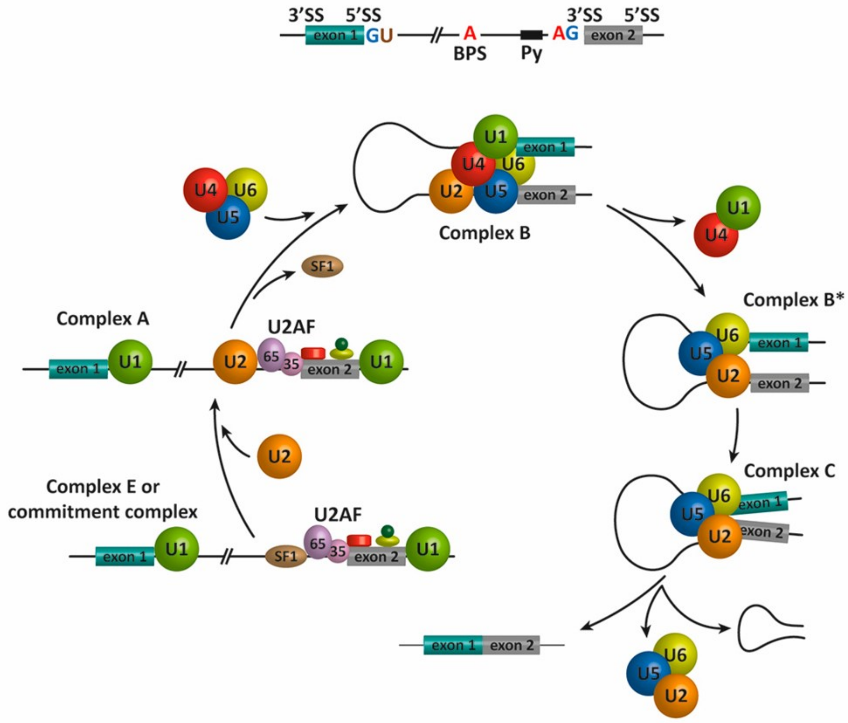


**Figure 10. Concerted activity of the spliceosomal components**. The U1 snRNP and other factors bind to the intron and form the complex E. Then, the U2 snRNP binds in proximity of the branch point, displacing the previously bound factors. This forms the complex A, in which U1 U2 attract each other, bringing closer 5' site and BPS. Later , U4, U5 and U6 bind to the intron, forming the complex B. There, conformational changes force the detachment of U1 and U4. This event also triggers the catalytic effect of U2 and U6, that act as ribozymes, to merge the contiguous exons and remove the intron lariat (adapted from [124]).

Interestingly, despite the abundance of proteins in snRNPs, spliceosomes are ribozymes, meaning that the catalytic activity is proper of the snRNAs [125][126]. The organization of the snRNPs is not random: splicing is initiated by the formation of the complex E, which includes the U1 snRNP and the splicing factors SF1, U2AF1 and U2AF2. The complex E binds to the 5' donor splice site through RNA-RNA interactions provided by the U1 snRNP, while the above mentioned splicing factors make contact with BPS, PPT and the acceptor splice site [8]. Subsequently, the U2 snRNP joins the branch point and forms an RNA duplex, displacing SF1: the complex E is converted into the pre-spliceosomal complex A, in which the two U1 and U2 snRNPs start attracting each other, reducing the distance between donor splice site and BPS [125][127].

In parallel, U1 and U2 are also crucial for another process, called exon definition, since they can recognize intron-exon boundaries [114][128]. After the formation of the complex E, preassembled U4, U5 and U6 snRNPs attach to the intron, forming the precatalytic spliceosome, also named complex B. These snRNPs force the 3' acceptor splice site to approach both the BPS and the donor site [8]. Eventually, U1 and U4 dissociate, and a conformational change forms the mature spliceosome. Then, the two splicing reactions are catalyzed by U2 and U6 [125]. Finally, the two flanking exons are ligated together, while the intron lariat is released [129]. After each splicing event, intron lariats are released, while still bound to the U2-U5-U6 snRNPs. Normally, spliced introns are quickly degraded to immediately recycle these spliceosomal components [130][131].

U12 minor spliceosomal RNA is formed from U12 snRNA, together with U4atac/U6atac, U5, and U11 snRNAs and associated proteins. The molecular mechanism of the U12-dependent spliceosome is similar, since U11, U12, U4atac and U6atac have analogous secondary structures as the canonical U1, U2, U4 and U6. Additionally, both systems utilize the U5 snRNP [120]. U12-type introns are present in most eukaryotes but only account for less than 0.5% of all introns in any given genome. U12-type introns are spliced somewhat less efficiently than the major introns, and it is believed that this limits the expression of the genes containing such introns. Recent findings on the role of U12-dependent splicing in development and human disease have shown that it can also affect multiple cellular processes not directly related to the functions of the host genes of U12-type introns.

### 3.3. Alternative splicing definition and types

In higher eukaryotes, entire exons or part of them can be included or not within the mature mRNAs, meaning that RNA splicing can occur in alternative ways. This process is indicated as alternative splicing [132][133]. As a consequence, different mature transcripts can be generated from the same gene. Eukaryotic genomes get a key advantage from this additional layer of gene regulation in the variety of isoforms and proteins that can be produced from the same genome, greatly expanding their information content [134].

Alternative splicing is based on different types of events, such as exon skipping, mutually exclusive exons, alternative splice sites utilization, and intron retention (Figure 11) [135].

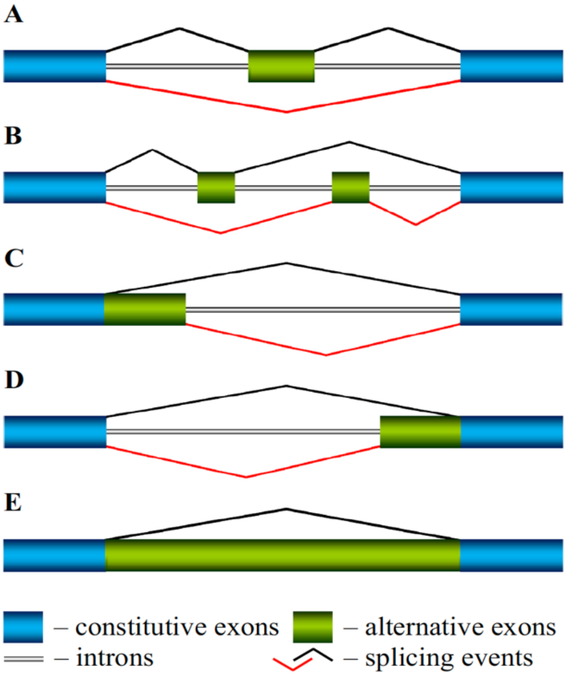


**Figure 11. Schematic representation of types of the alternative splicing events.** (A) Exon skipping, (B) mutually exclusive exons, (C) alternative 5′ donor splice site, (D) alternative 3′ acceptor splice site, (E) Intron retention [136].

*Exon skipping (ES)*, otherwise indicated as cassette exon, is the most common type of alternative splicing events. Differently from constitutive exons, the presence of cassette exons is not mandatory, as they may be excluded from the mature transcript [137]. These exons can be included or excluded, depending on regulatory mechanisms that may enhance or prevent the recognition between the spliceosomal components and splice sites [138][139].

Not only exons can be excluded through ES events, but also by the retention of other alternative exons. This splicing event is indicated as *mutually exclusive exons (MXEs)* [140]. These exons are clustered together within a relatively small region, probably originated through duplication events of a common ancestor sequence [141]. From cluster to cluster, the number of mutually exclusive exons can vary a lot, ranging from 2 to more than 40 exons [140][142]. As the event name suggests, the selection of an exon is sufficient to exclude all the others [141]. This is probably due to peculiar RNA secondary structures, each competing against the others for being recognized by the splicing apparatus. Alternatively, although less frequently, the co-existence of two or more exclusive exons is prevented by nonsense mediated decay (NMD) pathways [141]. MXEs may also occur within untranslated regions (UTRs), thus producing alternative 5' or 3' UTRs [133].

*Alternative 5' splice sites (A5)* and *Alternative 3' splice sites (A3)* involve the alternative usage of donor or acceptor splice sites found at 5' and 3' sites respectively [135][143]. In these situations, the usage of one alternative splice site will prevent the utilization of the other ones, in a reciprocally exclusive manner. As a consequence, alternative splice sites influence the process of exon and intron definition, resulting in the retention of differently-sized exons. The activation of cryptic splice sites can be the driver of this type of alternative splice sites events [144].

Another type of alternative splicing events is termed *intron retention* (IR) which occurs when one or more introns are not removed through splicing and are maintained in the mature RNA [145]. Since the sequence of transcripts with retained introns can overlap with that of unspliced precursor RNAs, the detection of IR events is typically more challenging [146]. Retained introns can be found in untranslated regions, but also within coding ORFs; in this latter case, due to the additional intronic sequence, the resulting protein may be affected by the introduction of extra amino acids, or by the presence of an early termination signal often leading to NMD [8][147]. In general, the mean length of introns subject to IR events tend to be shorter than that of constitutively removed ones [145].

As products of a novel type of alternative splicing, circular RNAs (circRNAs) are a recently discovered class of molecules with functional relevance for the regulation of gene expression. CircRNAs differ from other types of RNA by their structure in which the 3' and 5' ends of circRNAs are covalently joined (backsplicing) in a covalently closed loop structure and may consist of only a single or multiple exons (Figure 12) [148].

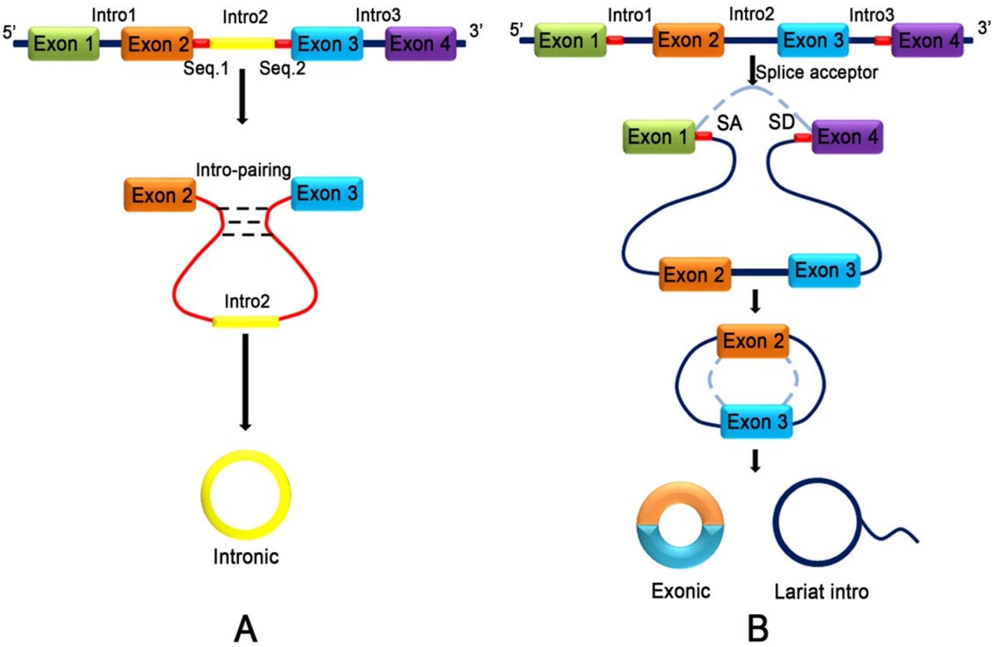


**Figure 12. Two different models of exon circularization of circRNA.** (A) Intron-pairing-driven circularization: during the formation of circRNA, an intron reverse complementary motif comprising GU-rich and C-rich elements is the key component to facilitate cyclization. (B) Lariat-driven circularization: the formation of circRNA is facilitated by the lariat structure. The complementary Alu flanking element which is repeated in the intron region competing for classical linear RNA splicing and the circularization is accelerated by reverse complementarity [149].

Many factors are involved in the circRNA biogenesis such as the spliceosome components and splicing factors, inverted repeats, RNA editing and RNA-binding proteins [150]. Multiple circRNAs can be produced from a single gene (alternative circularization) and by alternative splicing events (alternative 5’ and 3’ splice sites, exon skipping and intron retention) that alter the backsplice junction [151]. On a global scale, circRNAs display an appreciable degree of conservation between closely related species such as human and mouse with a subset of circRNAs conserved in evolutionary distant species such as Drosophila and even Archaea [152][153]. Generally, 5.8% to 23% of the human genes that are actively transcribed give rise to circRNAs in a way that is dynamically regulated between tissues and cell types as well as throughout differentiation [154][155]. CircRNA has been reported to be involved in various molecular mechanisms such as miRNAs sponges, transporting miRNAs throughout the cell, regulating mRNA through limited base pairing, and binding to RNA-binding proteins to form RNA-protein complexes [156].

### 3.4. Regulation of alternative splicing mechanisms

The splicing process is tightly regulated at different levels, by a plethora of mechanisms [157]. Some cis-acting regulatory elements are intrinsic features of the unspliced precursor transcripts; however, there are also trans-acting splicing modulators, that are RNA-binding proteins and even other RNA molecules [158].

#### 3.4.1. Roles of cis-elements in alternative splicing regulation

Regarding cis-acting elements, as previously described, specific consensus sequences, like donor and acceptor splice sites, polypyrimidine tract (PPT) and branch-point signals (BPS) play crucial roles in splicing [113]. Further surrounding regions, forming peculiar secondary structures, can work as enhancers or silencers too [159].

Depending on their sequence context, splice sites are characterized by a particular strength degree [160]. Stronger splice sites are often more conserved, and tend to be better recognized by the spliceosome [161]. The vast majority of the splice sites motifs are conserved: 5’ donor splice sites mostly utilize a GT consensus, whereas an AG motif is typically found at 3’ acceptor splice sites [121][162]. More rarely, non-canonical splice sites can be found; the most frequent alternative motif is that of GC-AG [118]. GT-AG motifs are highly dominant, while GC-AG account for about 0.8% of the total splice sites in human and mouse in PCGs [121]. On average, non-canonical splice sites are weaker than canonical ones, and their proper recognition seems to be more dependent on splicing regulatory elements [163]. The exact mechanism by which spliceosomes discriminate between canonical and non-canonical sites is unclear; still, many GT/GC pairs are present at alternative splice sites of highly expressed genes, suggesting a regulatory potential [163]. Interestingly, non-canonical GC donor splice sites were reported to be evolutionary enriched in mammals, reinforcing the hypothesis of their involvement in the alternative splicing modulation [164]. It is important to highlight that these consensus motifs are not the only determinants of the resulting splicing efficiency. For example, the length of the polypyrimidine tract has been reported to affect the splicing outcome as well [165]. This is perfectly in line with the function of such elements, involved in the proper recognition and assembly of the spliceosomal components (Figure 13).

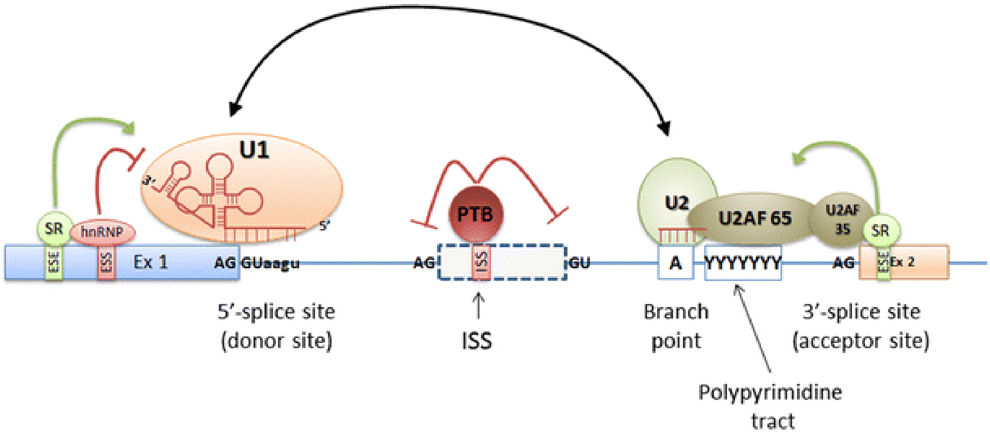


**Figure 13. Schematic localization of the cis and trans splicing elements.** The cis elements are the DNA sequences that include donor (5′) and acceptor (3′) splice sites, branch point and polypyrimidine tract sequences, and splicing silencers and enhancers. Donor and acceptor sites are evolutionary conserved and are usually defined by GT and AG nucleotides at the 5′ and 3′ ends of the intron, respectively. The branch site and the polypyrimidine tract sequences are highly degenerate and together with donor and acceptor sites are recognized by the elements of the splicing complex called spliceosome. Spliceosome proteins together with splicing repressors and activators recognize cis splicing elements and are called trans-acting elements [166].

In addition, four types of cis-acting splicing regulatory elements (SRE) are known to provide an additional control of splicing events. Those include exonic splicing enhancers (ESE) or silencers (ESS), and intronic splicing enhancers (ISE) or silencers (ISS). These elements influence splicing through the binding of specific RNA-binding non-spliceosomal regulatory proteins (RBP), which either promote or hinder the spliceosome activity on the adjacent splicing sites. It should be noted that, depending on the position of SRE in relation to the splicing sites, the same splicing factor may influence splicing differentially (sometimes oppositely) [167][168]. Particularly, position effect was found for motifs that are recognized by NOVA1, NOVA2, FOX1, FOX2, hnRNP L, hnRNP LL, hnRNP F, and hnRNP H splicing regulatory factors [136].

The presence of modified ribonucleotides, which have been detected in almost every kind of RNA, can affect splicing too; these modifications may occur on precursor transcripts, but also on spliceosomal snRNAs [161]. For instance, the pseudouridylation of certain regions of the U2 snRNA is crucial for the correct spliceosome assembly [169]. Not only splicing depends on RNA intrinsic properties, but also on the genomic context. In fact, splicing often occurs co-transcriptionally [170][171], hence it can be modulated by epigenetic and transcriptional regulations.

### 3.4.2. Roles of trans factors in alternative splicing regulation

Regarding trans-acting regulatory elements, the direct or indirect control of any biomolecule involved in splicing, epigenetic and transcriptional processes can finally result in the modulation of the splicing efficiency. Several splicing factors are members of the SR and hnRNP protein families (Figure 12) [158][172]. These proteins work as accessory splicing factors, not strictly required for spliceosome assembly and splicing catalysis and therefore they do not belong to the core splicing machinery (that is the set of essential spliceosomal components), but they can influence its binding to specific introns by synergistic or antagonistic interactions [173][174]. Given these combinatorial effects, the relative abundance of these regulators seems to be carefully controlled in different cell types. Interestingly, alternative splicing is one of the ways in which splicing factors are regulated [175]. Particularly important factors are U2AF1 and U2AF2, required for the proper functioning of the U2 snRNP, and whose regulation has been reported to influence the splicing process [176][177]. SR proteins not only regulate splice site utilization during alternative splicing, but they are also needed for constitutive splicing where the same splice sites are constitutively selected [178].

Several epigenetic layers are implicated in alternative splicing regulation, including methylation of cytosines in DNA molecules, nucleosome occupancy, histone variants and histone modifications. The main hypothesis assumes that the influence of epigenetic modifications on the process of splicing is mediated via remodeling of chromatin compaction, which predetermines changes in transcription elongation rate on certain regions of the transcribed genes [179]. For example, nucleosome occupancy is higher in exonic than in intronic regions [180][181][182]. Besides, the number of introns and exons in a gene positively correlates with nucleosome abundance [183]. Methylation of CpGs participates in splicing regulation of about 20% of all exons in human genes and its effect can either promote or inhibit exon inclusion (this presumably depends on the interacting proteins) [184]. Specific histone modifications were shown to differentially mark constitutive and alternatively spliced exons, introns, and their borders [185]. Histone modifications such as H3K36me3, H3K9me3 and H3K27me3 are enriched over exons. Liu and colleagues [186] showed that H3K36me3, H3K9me3, and H4K20me1 occupancy in cassette exons and their flanking regions are associated with more frequent inclusion of these exons into the mature transcripts in human cell lines. Histone acetylation on the exon–intron junctions usually promotes skipping of cassette exons, whereas deacetylation has the opposite effect [181].

### 3.5. Alternative polyadenylation and splicing

Alternative polyadenylation (APA) is emerging as a widespread mechanism used to control gene expression. Like alternative splicing, usage of alternative poly(A) sites allows a single gene to encode multiple transcripts. The mature 3′ ends of eukaryotic mRNAs and lncRNAs are created by a two-step reaction that involves an endonucleolytic cleavage of the RNA followed by synthesis of a polyadenylate tail onto the upstream cleavage product [187]. The assembly of the 3′ end processing complex on the pre-mRNA begins with the cooperative interaction of CPSF and CstF with specific sequences. The canonical poly(A) signal “AAUAAA” located upstream of the cleavage site, recognized by CPSF and a less defined downstream U/GU-rich region that constitutes the binding site for CstF [188]. Usage of one poly(A) site over another is often attributed to the relative “strength” of these core elements in addition to auxiliary sequences and protein factors that play a role in influencing poly(A) site choice in different contexts.

In some cases, APA changes the mRNA coding potential by altering the translated protein. In other cases, the 3′UTR length is altered, influencing the fate of the transcribed RNA in several ways, for example, by altering the availability of RNA binding protein sites, microRNA binding sites, stability, and subcellular localization [189][190][191].

U1 snRNP is emerging as a master regulator of APA through inhibiting the usage of intronic polyadenylation sites (PASs) [192]. Importantly, it could inhibit the cleavage step of intronic PAS processing to protect the RNA integrity for many genes. A recent study suggested that U1 snRNP may form a complex with canonical 3′ processing factors, called U1-CPAFs (U1-cleavage and polyadenylation factors), at intronic PASs [193]. In the study of Deng and colleagues [194], the authors provided experimental evidence that U1 is involved in inhibiting the cleavage of an endogenous PAS via protein-RNA interaction in the context of intact U1 snRNP complex. Through mapping the global U1A-RNA interactome, they showed that U1 may bind most of the U1-suppressed PASs, providing a possibility that U1-PAS interaction might play a general role in U1-mediated premature cleavage and polyadenylation (Figure 14).

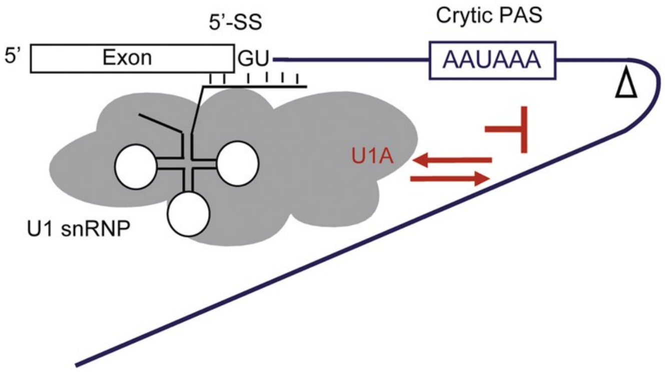


**Figure 14. Schematic representation of U1 snRNP-mediated PAS suppression.** U1 snRNP docking at 5′-ss might facilitate U1A binding near intronic PAS region, which might interfere with the recruitment of core 3′ processing factors, such as CstF64, on intronic PASs, thereby inhibiting the cleavage of PASs within pre-mRNA [194].

These studies provided evidence of the importance of the correct recognition and binding of U1 snRNP to the 5'ss during the splicing process which can affect the gene expression regulation through distant regulatory elements.

### 3.6. Splicing in long non-coding RNAs

Many lncRNAs contain introns and are spliced by the same splicing machinery as pre-mRNAs [125]. While the diversity of protein-coding isoforms is limited by the requirement to maintain an ORF, no such constraint is imposed on lncRNAs, allowing the spliceosome to explore the full range of noncoding exon combinations to generate an effectively inexhaustible noncoding isoform diversity [195]. Because cryptic splice sites are relatively abundant throughout the transcribed lncRNAs, the recognition of the correct exon boundaries is a crucial step during the splicing process. High fidelity of splice site recognition is mediated throughout a network of interactions that include snRNA base-pairing with sequences around splice sites and the binding of numerous splicing regulatory proteins such as U2 auxiliary factors, SR proteins and hnRNP proteins [196].

Several studies reported that lncRNAs are less efficiently spliced than pre-mRNAs of PCGs [197][198][199]. However, the precise molecular mechanism for this phenomenon has not been fully elucidated. The inefficiency in lncRNAs splicing was mildly correlated to weak U2AF65 binding to 3′ss, in addition to the 5′ss strength and a lower thymidine content in the polypyrimidine tract of lncRNA introns [196][198]. Nevertheless, efficient splicing was observed among lncRNAs with specific functions [198].

As well as for transcription, lncRNAs splicing can also affect the transcription of neighboring PCGs. In the study of Engreitz and colleagues [41], it was demonstrated that the first 5′splice site of the mouse lncRNA *Blustr* has a critical impact on its ability to regulate the upstream PCG *Sfmbt2* (Scm-like with four mbt domains 2). Thus, a better understanding of the mechanisms regulating lncRNAs splicing could contribute to understand their regulation and impact on PCGs transcription.

LncRNAs can undergo alternative splicing showing a complexity in their gene expression regulation as mRNAs. The human GENCODE annotations reports that, like PCGs, many lncRNAs are subject to alternative splicing although to a lesser extent [35]. In some examples the regulation can be extreme, as in the case of *GNG12-AS1* (*GNG12, DIRAS3* and *WLS* antisense RNA 1), a nuclear lncRNA that spans ten exons and can have 38 alternative isoforms [200]. Recent studies based on RACE-Seq and long-read RNA sequencing suggested that lncRNAs have as many exons and undergo as much alternative splicing as PCGs contrary to previous assumptions [201][202]. In addition, a recent study reported an increased level of complexity in lncRNA splicing regulation by using a high-resolution transcriptional cross-section of human and mouse chromosome 21 by targeted transcript enrichment, followed by single-molecule and saturating short-read RNA-Seq [195]. This approach revealed that, contrary to the impression from shallower RNA-seq studies [35][46], lncRNA loci were found highly prone to alternative splicing with their internal exons being almost universally alternatively spliced [195].

Despite showing a low level of conservation at their sequence level, an evolutionary constraint on lncRNA sequences is localized at lncRNA splice sites and splicing regulatory elements, suggesting that the recognition of the intron boundaries is a crucial step and the correct splicing of lncRNA introns is required for their function [59][60][203].

## 4. Wobble splicing

### 4.1. Definition and types of wobble splicing events

While much research focused on the main alternative splice events, it recently became clear that numerous alternative splicing events result in only subtle changes of the isoforms and their translated proteins. Among these alternative events is a special type of alternative 5’ss or 3’ss splicing events which involve tandem splice sites that are close to each other at a very short distance allowing the splicing factors to wobble between the proximal or the distal splice site [204][205]. Such alternative splicing event, termed “wobble splicing”, allows the production of alternative RNAs isoforms with subtle changes in their nucleotide sequence [206][207][208]. It is proposed that binding of splicing regulatory factors between the alternative splice sites or immediately adjacent to one site or the other can shift splicing toward the intron-proximal or distal splice site (Figure 15) [16].

Wobble splicing can occur at the donor splice sites in which the spliceosome can wobble between the two close donor splice sites having the following consensus “GYΔNGY” (where Y stands for pyrimidines, and ΔN stands for a number of separating nucleotides). Similarly, the spliceosome can wobble at the acceptor splice sites having the following consensus “NAGΔNAG” [209]. The distance between the alternative splice sites can vary over a wide range, from tens of bases [210] to as few as three bases in the case of NAGNAG alternative 3’ splice sites or even a single base [211].

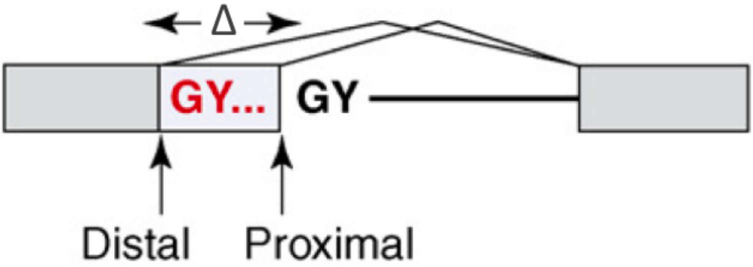


**Figure 15. Schematic presentation of a wobble splicing event at the donor splice site.** The two donor splice sites, distal and proximal, are in tandem and separated by a short distance termed (Δ) (modified from [205])

Wobble splicing has been extensively studied in protein-coding genes. However recently, computational evidence of wobble splicing at the 3' splice sites have been reported in human lncRNAs [212]. It is still unclear why particular splice sites are recognized while others are not utilized and the functional importance of the wobble splicing mechanisms.

### 4.2. Conservation and evolutionary dynamics of wobble splicing

The wobble splicing of donor and acceptor splice sites are not specific to the human genome, but they are abundant in other mammals, fruit flies, worms, and plants [206]. Evolutionary patterns suggest that purifying selection acts to maintain wobble splicing. Hiller and colleagues [208] reported the occurrence of alternative donor sites usage in eight investigated eukaryotic species. The existence of orthologous tandem donors that are confirmed in two or more species makes the wobble splice sites annotations errors unlikely. Moreover, the wobble donor splice sites have a higher conservation of the exonic and intronic flanking regions, a situation that is typical for conserved alternative splice events. In a subsequent study, it was reported a higher conservation of wobble splice sites in mouse, dog, chicken, zebrafish, and Fugu genomes [208].

Interestingly, a number of frameshifting tandems are under selection, suggesting a role in regulating mRNA and protein levels via nonsense-mediated decay (NMD). Multiple lines of evidence indicated that the human protein coding sequences are under selection against such in-frame tandem splice events, indicating that these events are often deleterious. The strength of selection is not homogeneous within the coding sequence, as protein regions that fold into a fixed 3D structure (intrinsically ordered) are under stronger selection, especially against sites with a strong minor splice site. Investigating structures of functional protein domains, it was found that tandem acceptors are preferentially located at the domain surface and outside structural elements such as helices and sheets.

For NAGNAG acceptors, initial studies reported that a fraction of tandem sites was found to be under purifying selection, even across large evolutionary distances such as between human and chicken or fish [208][213]. Applying comparative analysis to the NAGNAG sequence between the human and mouse genomes, it was found that both AGs at the tandem acceptors are highly conserved in all of the cases of EST-confirmed NAGNAG 3’ wobble splice sites [214].

### 4.3. Tissue-specificity of wobble splicing events

Several computational studies investigated the expression ratios of the different wobble splicing isoforms in different tissues. Different studies reached conflicting conclusions regarding whether the wobble splicing events are stochastic or show a tissue-specific regulation.

Initial studies reported that the relative ratio of each wobble-splicing isoform tends to be constant among various tissues [207]. Stochastic selection of either of the two splice sites was thought to explain the wobble splicing at most tandem splice sites [215][216]. Stochastic splicing events are expected to yield similar splice variant ratios in different tissues [206][217]. However, it was argued that stochastic splicing does not preclude functional importance of the alternative splicing event. Especially where both protein isoforms were required ubiquitously, stochastic splice site selection based only on spliceosomal core components offered the advantage of producing the two variant isoforms nearly independent of tissue type or other conditions that regulate alternative splicing [206].

On the other hand, in a more recent study based on RNA-seq data, it was reported that wobble splicing shows a tissue-specific expression and regulation. Bradley and colleagues [16] reported in their study that for more than 2,000 alternatively used NAGNAG motifs in human PCGs, 73% showed evidence for tissue-specific regulation with more than 40% displaying major changes between the two wobble spliced isoforms (>25% change between tissues). Thus, it was proposed that most wobble splicing events at the 3'splice sites are subject to some form of regulation. Moreover, Bradley and colleagues found a positive correlation between the magnitude of tissue-specific splicing differences and the conservation of the NAGNAG alternative event, suggesting that regulated NAGNAG wobble splicing has been evolutionarily fixed to retain an advantageous function for the different cells [218].

Despite these reports, there is still no universal clear idea of the tissue-specific regulation of the wobble splicing events, especially for those occurring at the donor wobble splice sites. Moreover, while most of these studies focused on PCGs only, a systematic study of the wobble splicing events in the lncRNAs is still lacking.

### 4.4. Regulatory roles of wobble splicing events

Since the discovery of wobble alternative splicing, debates about its functional implications and regulatory mechanisms have been controversial. Several lines of work focused on delineating whether alternative acceptor site selection is actively regulated or stochastically selected during the splicing process [219]. Analyzing splicing events covered by a large number of EST entries suggested that wobble splicing events may be tightly regulated, a proposal supported by high evolutionary conservation and an overabundance of cis-regulatory elements in proximity of alternative tandem splice sites [214].

Subtle alternative splice events are of interest since several cases are known to result in functionally different protein isoforms and wobble splicing in the untranslated region (UTR) can affect the translational efficiency [220]. The site choice between tandem splice sites pair can have severe functional consequences in terms of protein structure, subcellular localization, and isoform degradation. For example, the use of the upstream site in a $GT(\Delta)_2GT$ pair caused by a single nucleotide polymorphism (SNP) in the human *BTNL2* (butyrophilin like 2) gene yields to a truncated protein lacking the C-terminal IgC domain and the transmembrane helix, and results in predisposition to sarcoidosis [221]. In the study of Tsai and colleagues [222], it was demonstrated that the subcellular localization and degradation of *ING4* (inhibitor of growth family member 4), a tumor suppressor gene, is modulated by two wobble-splicing events at the exon 4–5 boundary involving alternative splicing at two tandem splice sites, GC$(\Delta)_7$GT and NAGNAG, which caused canonical (GT-AG) and non-canonical (GC-AG) splice site wobbling selection. The wobble splicing events caused the displacement of the ING4 transcript from the nucleolus to the nucleus through the disruption of a nuclear localizing signal (NLS) sequence. The nucleolar accumulation of *ING4* prolonged its half-life, but the lack of nucleolar targeting potentially increased *ING4* degradation. A wobble splicing event in the *RBM10* (RNA binding motif protein 10) gene containing a tandem donor splice sites of the configuration GT$(\Delta)_1$GT was identified in *RBM10* exon 10 [223]. The wobble splicing event resulted in the absence of a valine residue which altered the α-helical configuration associated with the RNA recognition motif tertiary structure.

These studies indicate a putative functional relevance of wobble splicing events in gene expression regulation. The choice between the competing splice sites could provide an additional level in the regulation of expression with subtle changes in the encoded isoforms. However, how the selection of splice sites in tandem occurs and the factors involved are still widely unknown.

# II. Aims of the Research

It is now clear that the larger portion of mammalian genomes is transcribed to produce non-coding RNAs which emerged as an important regulatory layer of the transcriptome, playing a role through various molecular mechanisms and in a variety of genomic contexts. LncRNAs gained an enormous attention in the past years given their important involvements in gene expression regulation and other several biological processes in addition to their significant implications in various pathological conditions.

While many studies have been devoted to understand the putative functions and roles of lncRNAs in gene expression regulation, how lncRNAs gene expression is regulated remains less understood. Moreover, previous studies have reached conflicting conclusions about the splicing extent and its efficiency in lncRNAs. Indeed, very few studies have addressed thoroughly the genome wide splicing features of lncRNAs in comparison to protein-coding genes.

Alternative splicing represents one of the main mechanisms influencing gene expression outcomes and have been broadly studied, however, only few works have focused on the genome-wide quantification and characterization of alternative tandem splice sites involved in wobble splicing. In most cases, only wobble splicing at acceptor splice sites was considered and very few studies have investigated wobble splicing at the donor splice sites. Furthermore, most studies focused on PCGs, while little is known about the wobble splicing prevalence and features in lncRNAs. Thus, a systematic analysis based on recent genomic annotations of wobble splicing at lncRNAs and PCGs is currently lacking.

At the beginning of my PhD internship, I started the investigation of the genetic characteristics of the newest annotations of lncRNAs. We noticed substantial differences in lncRNA genetic structure especially in their exons and introns features. From a preliminary observation of lncRNA intronic sequences, we noticed a high occurrence of the non-canonical “GC” splice sites in lncRNA transcripts in comparison to protein-coding ones. This led us to proceed with a further characterization of the GC-AG introns features in lncRNA and PCGs that showed multiple peculiar features in comparison to GT-AG ones suggesting a putative functional role. Moreover, by manually analyzing the alternative splicing events involving GC-AG introns, we observed in many cases the occurrence of donor splice sites in tandem involved in wobble splicing events which led us to further characterize the prevalence and functional mechanisms of this special type of alternative splicing events.

In this research project, we took advantage of the most recent lncRNA annotations provided by the GENCODE project [97] to:

i. characterize the genomic and splicing features of human and mouse lncRNAs in comparison to PCGs.

ii. Analyze the various splicing features of GC-AG introns in comparison to GT-AG ones in terms of their splicing strength, conservation, expression, alternative splicing and functional outcomes.

iii. quantify and assess the prevalence of wobble splicing in lncRNAs and PCGs at the donor splice sites in the human and mouse genomes

iv. characterize the involvement of GC-AG introns in wobble splicing and examine the putative functional impact of wobble splicing on the alternatively spliced isoforms.

# III. Materials and Methods

## 1. Data collection

The lists of lncRNAs and PCGs 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 mouse genes annotated on the genome sequence GRCm38 (gencode.vM16.long_noncoding_RNAs.gtf.gz; gencode.vM16.basic.annotation.gtf. gz). PCGs were selected from the basic annotation when both gene and transcript were indicated as "protein_coding".

Wobble splicing events were analyzed using a more recent version of annotations from GENCODE: release 30 for human annotated on the genome sequence GRCh38 (gencode.v30.long_noncoding_RNAs.gtf.gz; gencode.v30.basic.annotation.gtf.gz), and release M19 was used for mouse genes annotated on the genome sequence GRCm38 (gencode.vM19.long_noncoding_RNAs.gtf.gz; gencode.vM19.basic. annotation.gtf.gz).

The GTF files were downloaded and analyzed in RStudio version 1.1.456 (http://www.rstudio.com/) running under R version (3.6.3) using the "GencoDymo" R-package (version 0.2.1) developed by our group (https://github.com/ monahton/GencoDymo).

Single exon genes were excluded from the analysis as they are not subjected to splicing. Introns lists were retrieved using the "*extract_introns()*" command from GencoDymo. Introns sequences and introns splice sites were extracted from Table Browser tool from UCSC [224] using human GRCh38 and mouse GRCm38 genome sequences via the "*assign_ss()*" command from GencoDymo exploiting the "BSgenome.Hsapiens.UCSC.hg38" version (1.4.3) and "BSgenome.Mmusculus. UCSC.mm10" version (1.4.0) R-packages.

## 2. Data validation

An independent validation of the results from GENCODE was obtained by collecting data of human lncRNAs annotations from 6 different databases:

i. the FANTOM5 database (http://fantom.gsc.riken.jp/cat/) (Fantom CAT genes; FANTOM_CAT.lv3_robust.only_lncRNA.gtf) [100].

ii. the NONCODEv5 (version 5) database (http://noncode.org/datadownload/NONCODEv5_human_hg38_lncRNA.gtf.gz) [104].

iii. the BIGTranscriptome database (release 2016) lncRNA catalog (http://big.hanyang.ac.kr/UCSC/RNA-seq/hg19/CAFE/GTFs/ BIGTranscriptome/BIGTranscriptome_lncRNA_catalog.gtf) [102].

iv. the LncBook database (http://bigd.big.ac.cn/lncbook/index) [107].

v. the MiTranscriptome database (http://mitranscriptome.org/download /mitranscriptome.gtf.tar.gz) [103].

vi. the LNCipedia database (version 5.2) (https://lncipedia.org/downloads /lncipedia_5_2/full-database/lncipedia_5_2_hg38.gtf) [106].

A validation of results obtained from the mouse genome was performed using lncRNAs annotations from the NONCODEv5 database (http://noncode.org/datadownload/NONCODEv5_mouse_ mm10_lncRNA.gtf.gz).

For the validation of data in other species, the lists of lncRNAs and PCGs of *Drosophila melanogaster* and *Caenorhabditis elegans* were downloaded from the BioMart data mining tool (https://www.ensembl.org/biomart/martview/) [225] in the Ensembl genome database (release 91) and analyzed using the GencoDymo package.

## 3. Splicing elements analysis

The scores of splice junctions were calculated using the MaxEntScan web tool [226], a program for predicting the strength of the splicing sequences based on the maximum entropy model. In particular, MaxEntScan::score5ss (http://hollywood.mit.edu /burgelab/maxent/Xmaxentscan_scoreseq.html) scores the donor splice site strength 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 (http://hollywood.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html) scores the acceptor splice sites from a sequence motif of 23 nucleotide covering bases −20 to +3. Splice sites motifs sequences were extracted using the "*extract_5ss_motif()*" and "*extract_3ss_motif()*" commands from the GencoDymo package for the donor and acceptor splice sites respectively. The motifs were uploaded to the corresponding MaxEntScan tool in FASTA format. We evaluated the strength of 5′ and 3′ splice sites of human and mouse introns using the Weight Matrix Model as provided by the MaxEntScan tool. The output was downloaded and analyzed in R using custom scripts.

The evaluation of the polypyrimidine tract strength was performed using the "branchpointer" R package (version 1.10.0) [227]. The package predicted polypyrimidine tracts in query regions located at −18 to −44 nucleotides from the 3′ splice sites. The query regions were determined by custom scripts in R and saved in query files. Query files were read into branchpointer using "*readQueryFile()*" command. Branchpoint probability scores were evaluated by the branchpointer model using "*predictBranchpoints()*" command. We used a cut-off probability 0.52 to distinguish branchpoints and non-branchpoint sites as recommended by the package developers.

### 4. Conservation analysis

To evaluate the conservation of genes containing GC-AG introns, the list of orthologous genes in the human (GRCh38.p10) and mouse genomes (GRCm38.p5) were downloaded from the Ensembl genome database (release 91) by using multi-species comparison in the BioMart data mining tool [225].

Multi-species conservation of 5′ splice sites was assessed manually by aligning the sequences of corresponding introns in different organisms using the UCSC genome browser as data source [228]. Species considered in this analysis were: human, chimp, macaque, mouse, rat, dog, cow, pig, chicken, fugu, and zebrafish.

## 5. Alternative splicing analysis

The assignment of alternative splicing events involving GC-AG and GT-AG introns was performed using the SUPPA2 tool (version 2.3) (https://github.com/comprna/SUPPA) [229]. Splicing events were extracted from the gtf files of lncRNA and PCGs annotations from the GENCODE database using the "*generateEvents*" command in a python3 environment:

```
python3.4 suppa.py generateEvents -i <input-file.gtf> -o <output-file> -f ioe
-e <list-of-events>
```

The SUPPA2 tool classified alternative spliced events according to the following types: (i) exon skipping, (ii) intron retention, (iii) mutually exclusive exons, (iv) alternative 5′ss, (v) alternative 3′ss, (vi) alternative first exons and (vii) alternative last exons. Custom R scripts were used to extract introns involved in each type of alternative splicing event and to evaluate alternative last exons.

Alternative polyadenylation signals (PAS) were extracted according to the 16 PAS reported in the paper of [230] in a bin of 40 nucleotides at the end of each last exon using custom R scripts.

## 6. Identification of wobble splicing events

All possible wobble splicing events were enumerated using custom scripts in R. In particular, alternative splice sites that were localized in tandem and within a separating distance (Δ) equal or inferior to 20 nucleotides (nt) ($1 \leq \Delta \leq 20$ nt) were classified among wobble splicing (WS) events. This procedure was applied for both protein coding and long noncoding transcripts, and for both donor and acceptor splice junctions. When counting the number of 5' WS events, only the canonical "GT" and the non-canonical "GC" and "AT" donor splice sites were selected. All the other non-canonical donor splice sites were excluded as they were particularly rare and could represent annotation errors. The same approach was done for 3' WS events, where only canonical "AG" and non-canonical "AC" acceptor splice sites were evaluated. Introns having two simultaneous wobble splicing events at both 5' and 3' junctions, characterized by an identical reciprocal distance (Δ) were termed "compensating events".

Custom R scripts were used to generate subsets containing the scores of splice sites involved in wobble splicing. To classify each splicing junction score into a strength category (i.e. weak, medium or strong), the global scores were divided into quartiles (first, median and third); then, all values below the lower quartile threshold were classified as weak, while those above the upper quartile threshold were labelled as strong. All the remaining intermediate scores were assigned to the medium category. The entire pipeline was performed on lncRNAs and protein-coding genes separately in both human and mouse.

## 7. Expression analysis

RNA-Seq data of healthy individuals were obtained from the Genotype-Tissue Expression (GTEx) (version 8) data set (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424. v8.p2) (phs000424.v8.p2.c1, July 18, 2019) [231] and downloaded using dbGaP web site (https://www.ncbi.nlm.nih.gov/gap) (approved protocol #23403) using SRA Toolkit tool (version 2.10.8) provided by SRA [232]. Raw SRA files were downloaded from the dbGap database using the "*prefetch*" command:

```
prefetch <SRAfile.sra>
```

Separate paired-end FASTQ files were extracted from the downloaded SRA files using the "*fastq-dump*" command:

```
fastq-dump --split-files –gzip <SRAfile.sra>
```

Data were collected from 10 different tissues: anterior cingulate cortex, amygdala, cerebellum, heart left ventricle, kidney cortex, lung, liver, spleen, skin, and testis, only for male individuals using 8 samples per tissue for a total of 80 samples. Only male individuals were selected for the analysis to avoid any profound effects or biases that might affect the subsequent analyses. All sample files selected were sequenced using the Illumina HiSeq 2000 platform following a paired-end protocol (2x76 bp) thus sharing the sequencing features.

Quality control analyses on the raw sequence data in the FASTQ format were performed using the FastQC tool (version 0.11.7) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The aim of the FastQC tool is to check the quality of reads coming from high throughput sequencing experiments. The FastQC tool needs the reads in FASTQ format, that is a text file containing the nucleotide sequence and a Phred score assigned to each nucleotide in ASCII code. This score is calculated as an integer (Q) ranging from 2 to 40; the higher the score, the better the quality.

Reads passing quality control were quantified with the transcripts per million method implemented in the Salmon software (version 1.2.0) [233] using default parameters and the human hg38 reference transcriptome in FASTA format provided from GENCODE:

```
quant -i transcripts_index -l <LIBTYPE> -1 reads1.fq -2 reads2.fq --validateMappings -o transcripts_quant
```

The transcripts quantifications were then imported into R and summarized using custom scripts. The unexpressed lncRNA transcripts with TPM < 0.1 and protein-coding transcripts with TPM < 0.5 were filtered out in subsequent analyses.

## 9. Wobble splicing transcripts sub-cellular localization analysis

Analysis of the cellular localization of transcripts involved in wobble splicing was performed using RNA-Seq samples derived from the nucleus or cytoplasm separately. We analyzed 19 RNA-Seq samples from the nucleus and 19 RNA-Seq samples from the cytosol for a total of 38 samples obtained from 12 cell lines: A549, GM12878, GM12879 HeLa-S3, HepG2, HUVEC, h1-hESC, IMR90, K562, MCF7, NHEK, and SK-N-SH and provided by the ENCODE project [197].

SRA files were downloaded from the “E-GEOD-30567” dataset publicly available in the ArrayExpress repository (https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-30567/) using SRA Toolkit. Separate paired-end FASTQ files were extracted from the downloaded SRA files as previously described.

Quality control was performed using the FastQC tool and reads passing quality check were quantified with the transcripts per million method implemented in the Salmon software using default parameters and the human hg19 reference transcriptome in FASTA format provided from GENCODE.

Transcripts alternative splicing events were generated using the "*generateEvents*" command as previously described and the percent-splice-in (PSI) values per event were determined using the "*psiPerEvent*" command:

> *python3.4 suppa.py psiPerEvent --ioe-file <ioe-file> --expression-file <expression-file> -o <output-file>*

A differential alternative splicing analysis for the transcripts involved in wobble splicing was performed by evaluating the dPSI values of wobble splicing introns between the 2 different conditions (nucleus or cytosol) using the "*diffSplice*" command using default parameters:

> *python3.4 suppa.py diffSplice --method <empirical> --input <ioe-file> --psi <Cond1.psi> <Cond2.psi> --tpm <Cond1_expression-file> <Cond2_expression-file> --area <1000> --lower-bound <0.05> -gc -o <output-file>*

The differential splicing operation generates a dPSI file indicating the PSI difference between the 2 conditions and the statistical significance of the difference.

## 8. Functional enrichment analyses

Gene list functional enrichment analyses were performed using the DAVID (Database for Annotation, Visualization and Integrated Discovery) tool (version 6.8) [234] and the PANTHER (Protein ANalysis THrough Evolutionary Relationships) overrepresentation test (version 15.0; release 2020-02-14) [235] implemented in the Gene Ontology (GO) website [236][237].

The lists of PCGs containing a GC-AG intron from both human (n = 1934) and mouse (n = 1669) were subjected to an enrichment analysis on GO Biological Process terms and filtered applying a statistical significance threshold of 0.05 based on the multiple testing corrected p-values (i.e., Benjamini adjusted p-value in DAVID or false-discovery rates (FDR) in PANTHER).

## 9. Wobble splicing functional impact analysis

All wobble splicing events occurring in PCGs were visualized and manually analyzed using the UCSC Genome Browser (https://genome.ucsc.edu/) [238]. The position of each wobble splicing event in the corresponding isoforms was identified in the GENCODE track, as shown in Figure 16.

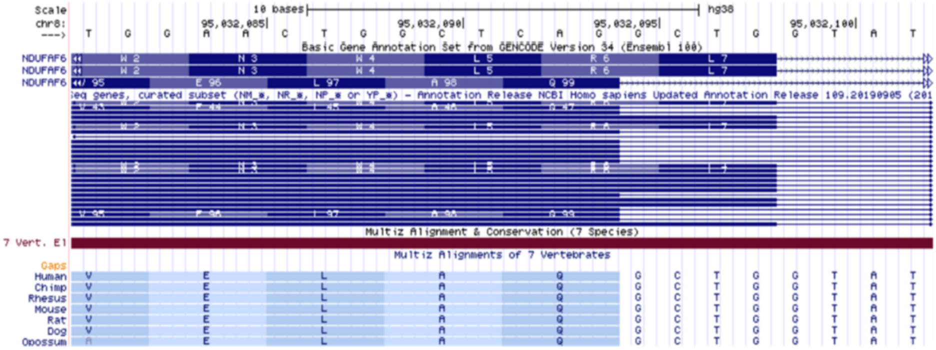

Scale
chr8:
10 bases
hg38
95,032,085|
95,032,090|
95,032,095|
95,032,100|
T G G A A C T G G C T C A G G C T G G T A T
Basic Gene Annotation Set from GENCODE Version 34 (Ensembl 100)
NDUFAF6
NDUFAF6
NDUFAF6
W 2
N 3
W 4
L 5
R 6
L 7
V 95
E 96
L 97
A 98
Q 99
Seq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release NCBI Homo sapiens Updated Annotation Release 109.20190905 (201
Multiz Alignment & Conservation (7 Species)
7 Vert. El
Multiz Alignments of 7 Vertebrates
Gaps
Human
Chimp
Rhesus
Mouse
Rat
Dog
Opossum
V
E
L
A
Q
G C T G G T A T

**Figure 16. UCSC genome browser graphical representation of 5' wobble splicing event.** The wobble splicing event occur in the intron of the NDUFAF6 gene.

Depending on its position, the wobble splicing impact was classified according to the following categories: (i) insertion, when wobble splicing resulted in the addition of a number of amino acids without inducing a frameshift in the open reading frame (ORF); (ii) frameshift, when the added nucleotides were triggering a frameshift downstream the wobble splicing event; (iii) 5' untranslated region (UTR), when wobble splicing events occurred within 5' UTRs, without affecting the translated protein; (iv) 3' UTR, when wobble splicing events affected the 3' UTR; (v) 5'UTR/ins and (vi) 3' UTR/ins, when wobble splicing events occurred in the 5'UTR and 3'UTR of one isoform, and in the ORF of another isoform, leading to the insertion of amino acids; (vii) alternative N-terminus, when the wobble splicing event altered the N-terminus of the encoded protein.

A functional analysis was performed to study if the wobble splicing events occurring inside ORFs lead to disruption of functional domains in the translated proteins. First the analysis was performed by identifying the protein region in which the wobble splicing event occurred using the “UnitProt” database (https://www.uniprot.org/). Then, the putative functional domains and their characteristics were identified by using the “InterPro” webtool (https://www.ebi.ac.uk/interpro/), as illustrated in Figure 17.

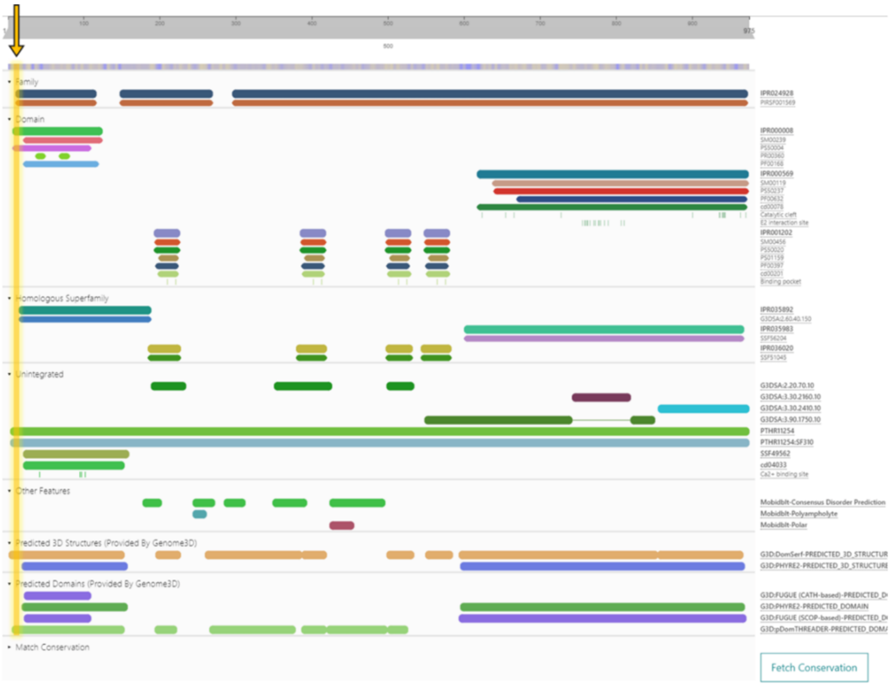


**Figure 17. InterPro graphical representation of protein domains**. The figure illustrates the position of the different domains and other features of the protein isoforms of a gene. The arrow indicates the site in which a wobble splicing event occurs affecting the amino acid sequence of its corresponding isoform.

Visualization of amino acids composition of different isoforms involved in the wobble splicing was performed using the Clustal Omega webtool (https://www.ebi.ac.uk/Tools/msa/clustalo/) for multiple sequence alignment.

## 10. Statistical analyses

Data analyses and descriptive statistics were performed using RStudio version 1.1.45611. 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.

## 11. Data use policy and availability of data

This research project was based on genomic data of lncRNAs and PCGs provided by the GENCODE Project. In particular, data of lncRNAs and PCGs were downloaded from the GENCODE web-pages for human and mouse as gtf files.

We only analyzed anonymized samples for which the corresponding donor consent information was available in the GTEx dataset (dbGaP:phs000424.v8.p2) at the time of the analysis. Samples were downloaded from the dbGap database according to the specified guidelines. All of the samples we analyzed were approved for General Research Use (GRU) and thus have no further limitations outside of those in the NIH model Data Use Certification Agreement.

The datasets supporting the conclusions of this article (i.e., introns data) are available in the GitHub repository (https://github.com/laBione).

## 12. Additional tools

All the R packages were used according to the R version 3.6.3.

Graphs and distribution plots were performed using the R package “ggplot2” (version 2.1.0) (http://ggplot2.org/), through the “*ggplot()*” command.

Data manipulations were performed using the “dplyr” (version 1.0.2) and “tidyr” (version 1.1.1) R-packages.

Aligned files conversions SAM/BAM and BAM/BED were performed using SAMtools (version 1.11) (http://samtools.sourceforge.net/) [239].

Wobble splicing events were visualized using IGV software (version 2.3.6) (http://software.broadinstitute.org/software/igv/) [240].

# IV. Results

## 1. Genomic characterization of lncRNAs in comparison to PCGs

### 1.1. Genomic features of lncRNAs

We performed a general analysis on a plethora of lncRNA genomic features of human and mouse lncRNAs in comparison with PCGs. Our analysis, based on recent gene annotations from the GENCODE database (human release 27 and mouse release M16), considered an increased number of genes with respect to previous studies [35][46]. The total number of genes, transcripts and exons are reported in Table 1.

**Table 1. Number of annotations from the GENCODE datasets used in the analyses**

| | **Human<br>V27** | | **Mouse<br>M16** | |
|---|---|---|---|---|
| | ***lncRNA*** | ***PCGs*** | ***lncRNA*** | ***PCGs*** |
| **genes** | 15778 | 19836 | 12374 | 21963 |
| **transcripts** | 27908 | 55358 | 17266 | 43798 |
| **exons** | 84490 | 580507 | 46877 | 437586 |
| **introns** | 56582 | 525149 | 29611 | 393788 |

The genomic organization of lncRNAs and PCGs appeared highly similar in both species. Human and mouse lncRNAs appeared equally transcribed from the forward and the reverse strand (forward: n=8015, 50.8%; reverse: n=7763, 49.2%) as PCGs (forward: n=10030, 50.6%; reverse: n=9806, 49.4%).

LncRNA genes appeared to be homogeneously interspersed in the genome. Although most lncRNAs are produced either within their target PCGs or in their vicinity, a few exceptions were observed as some particular regions larger than 1 Mb harbor only lncRNA genes and are devoid of any protein-coding ones. For example, the genomic region on chromosome 4 in human (chr4:174,515,115-183,985,775) spanning a distance of (~4 Mb) harbored a number of lncRNA genes far apart from any protein-coding ones (Figure 18). This genomic region appeared to be also conserved in mouse on chromosome 8 in which the lncRNA genes spanned a long genomic distance with no PCGs in their vicinity.

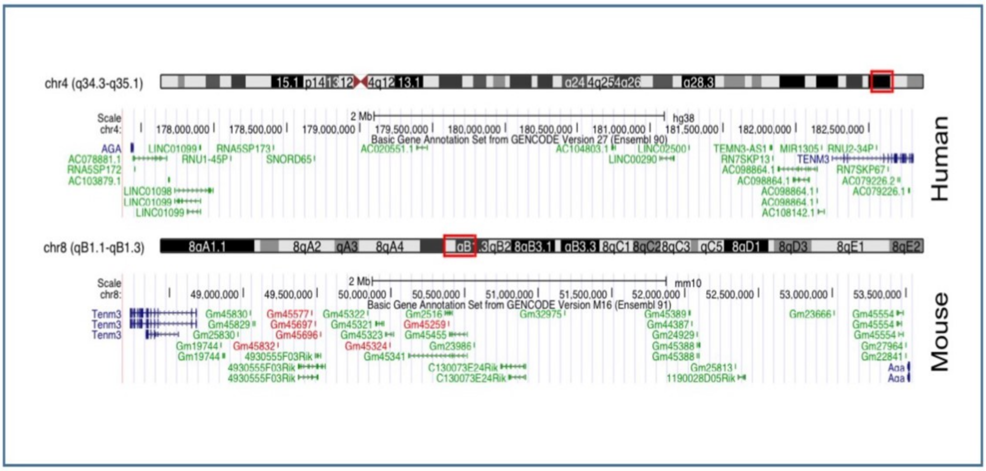


**Figure 18**. **UCSC genome browser representation of conserved genomic harboring only lncRNA genes in human and mouse.** The upper panel shows a human genomic region on chromosome 4 spanning a distance (~ 4 Mb) and containing a number of lncRNA genes with no neighboring PCGs. The lower panel shows the same genomic region conserved in mouse.

We next analyzed the distribution of lncRNA genes in the genome. The gene density resulted highly variable among chromosomes for lncRNAs and PCGs in both species. LncRNA genes appeared to be almost homogenously distributed on the different chromosomes although they were underrepresented on chromosomes 1, 11, 19, and X in human while overrepresented on chromosomes 13, 18, and 21 with respect to protein-coding genes (Figure 19). Similarly, lncRNAs appeared to be underrepresented on chromosomes 2, 7, 11, 17, 18, 19 and X in the mouse genome while overrepresented on chromosome 13 and Y. This could hint towards a possible precise distribution of lncRNAs along the genome and how lncRNAs gene positioning are involved in genome organization and control of gene expression regulation [241].

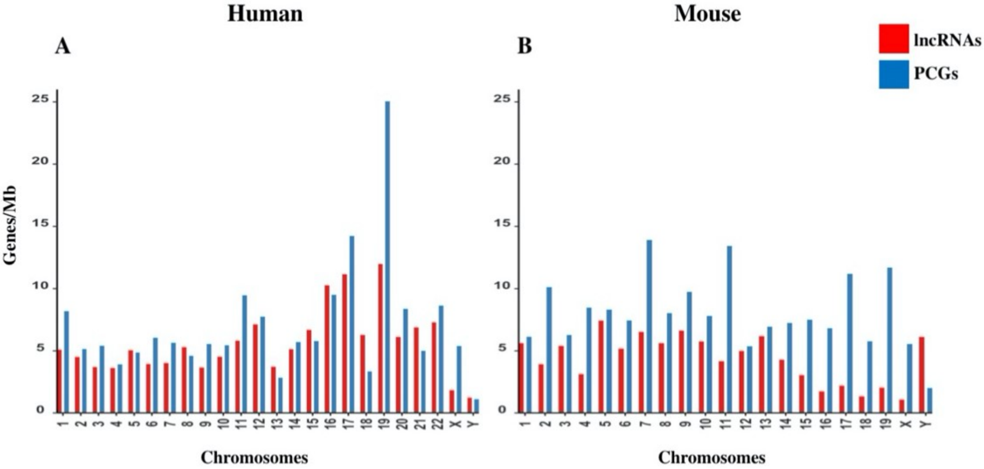


**Figure 19. Gene densities of lncRNAs and PCGs**. Gene densities in (A) human and (B) mouse across chromosomes. Densities are reported as number of genes per Megabase (Mb).

The genome coverage of long non-coding genes was found remarkably lower with respect to that of protein-coding ones. Indeed, long non-coding genes accounted for 12.5% of the human genome while 43.4% is occupied by PCGs (Chi-square test = 730.4, 1 df, p-value $< 2.2 \times 10^{-16}$) in human. Similarly, the genome coverage of mouse lncRNAs was lower (6.8%) than that of PCGs (39.2%) (Chi-square test = 802.5, 1 df, p-value $< 2.2 \times 10^{-16}$). The reduced genome coverage was not entirely due to the smaller number of lncRNAs, as they account for about 80% of protein-coding ones in human and 56% in mouse, but it appeared to be due to the lncRNAs length, that resulted significantly lower than that of PCGs in both species.

### 1.2. Gene structure and transcriptional complexity of lncRNAs

As the gene structure may affect gene expression regulation, we characterized the human and mouse lncRNAs genetic features and their transcriptional complexity in comparison with those of PCGs.

Human lncRNAs resulted, on average, almost three times shorter than protein-coding ones with an average length of about 24 kb versus 68 kb, respectively (Wilcoxon test p-value < 2.2 × $10^{-16}$) (Figure 20). Similarly, lncRNAs gene length in mouse resulted significantly shorter than that of PCGs with an average length of about 15 kb versus 49 kb, respectively (Wilcoxon test p-value < 2.2 × $10^{-16}$) (Figure 18).

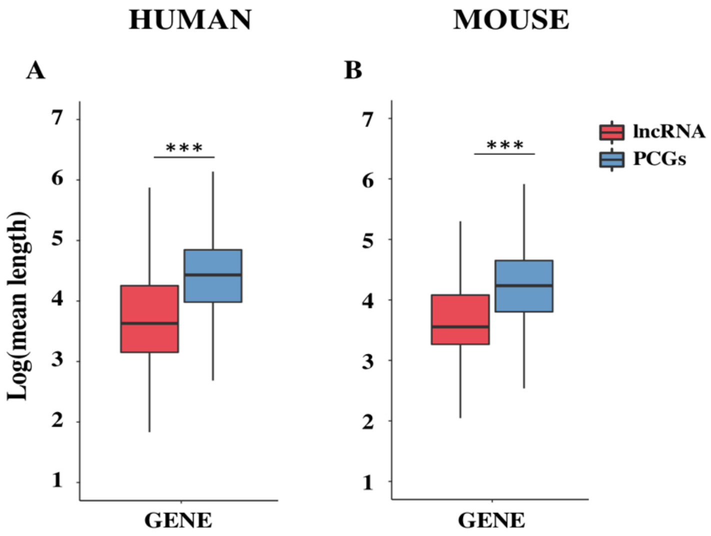


**Figure 20. Boxplot showing the mean gene length.** (A) human and (B) mouse. The mean length was calculated in base pairs (bp). The $\log_{10}$ of the mean length is represented on the y-axis. *** p-value < 0.01.

The shorter length of lncRNAs was attributable to the lower number of exons composing them (Figure 21). In human, more than 70% of lncRNA transcripts had 3 exons or less, compared with 16% of protein-coding transcripts bearing the same characteristics (Chi-squared test = 24407.0, 1 df, p-value < 2.2 × $10^{-16}$) (Figure 21A). A large proportion of lncRNA transcripts was composed of 2 exons (34%) as previously reported [35] and 14% are single-exon genes. In mouse, more than 75% of lncRNAs had 3 exons or less versus 23% in protein-coding transcripts (Chi-squared test = 14613.7, 1 df, p-value < 2.2 × $10^{-16}$) (Figure 21B) and 24% of lncRNAs were single-exon genes versus 6.4% in protein-coding ones. Also in the mouse genome, an enrichment of 2-exons transcripts in lncRNAs (30%) was observed.

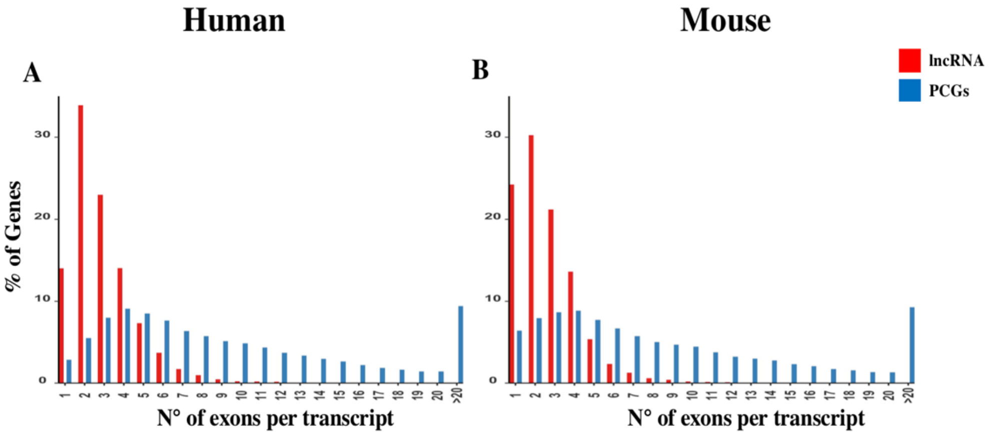


**Figure 21. Bar graph showing the distribution of the number of exons per transcript.** The data were reported in (A) human and (B) mouse.

A deeper characterization of exons and introns length allowed us to appreciate differences between lncRNAs and PCGs. Conversely to what was previously reported in the paper of Derrien and colleagues [35], our data revealed that first exons and especially last exons in lncRNAs are significantly shorter than protein-coding ones in both species (Figure 22). LncRNA introns were found longer than PCGs ones when they are inner introns; instead, they resulted slightly shorter when they are first introns. Interestingly, unlike what was described for PCGs in which first introns are generally longer than inner introns [242], lncRNAs first and inner introns appeared similar in length in both species.

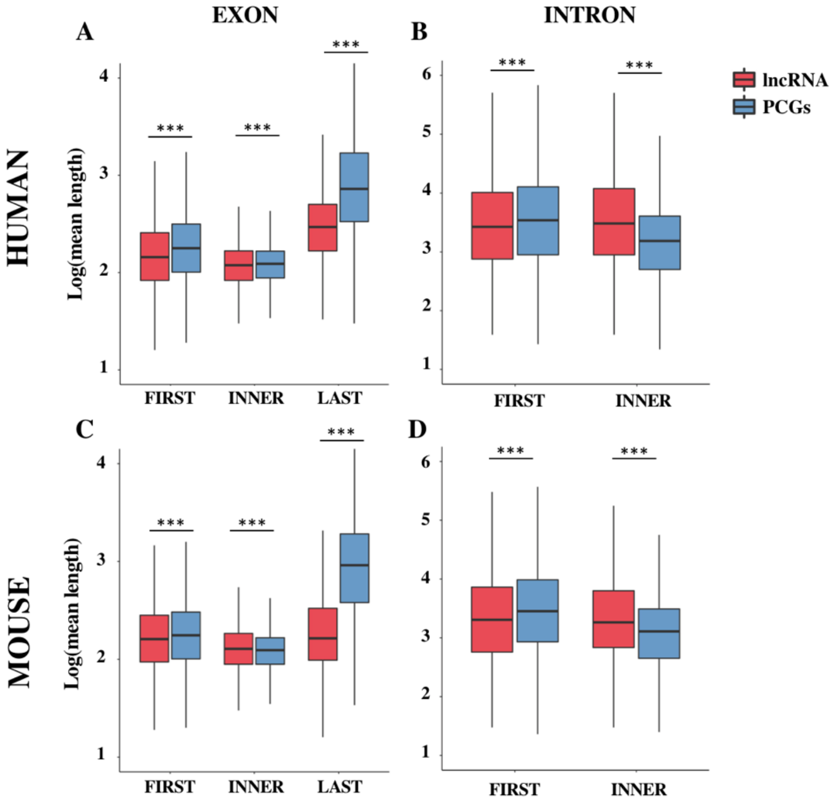


**Figure 22. Boxplot showing the mean exon and intron length**. (A) human and (B) mouse. Exons were divided according to their position into first, inner and last while introns were divided into first and inner. The mean length was calculated in base pairs (bp). The $\log_{10}$ of the mean length is represented on the y-axis. *** p-value < 0.01.

Although it is possible that these differences in length could be due to an incomplete annotation of lncRNAs, it is nevertheless interesting to note that the reduction in length affects those portions of the gene mainly involved in gene expression regulation.

## 2. Splicing features of lncRNAs

### 2.1. Enrichment of GC-AG splice sites in lncRNAs

As splicing is a main determinant of post-transcriptional gene expression regulation, we characterized the splicing features of lncRNA introns in comparison with those of protein-coding ones.

The splice junctions sequence analysis highlighted differences between lncRNAs and PCGs consensus sequences (Table 2). The GC-AG splice junctions appeared strongly enriched in human lncRNAs in which they represent 3.0% of the total splice junctions, thus almost four times more than in PCGs (0.8%) (Chi-square test = 2289.4, 1 df, p-value < $2.2 \times 10^{-16}$). The same enrichment was found in mouse, in which GC-AG splicing junctions were more than the double with respect to protein-coding ones (lncRNAs: 1.9%, PCGs 0.8%) (Chi-square test = 380.2, 1 df, p-value < $2.2 \times 10^{-16}$).

**Table 2. Number of different splice junctions consensus in human and mouse**

| | **Human** | | | |
|---|---|---|---|---|
| | ***lncRNAs*** | ***%*** | ***PCGs*** | ***%*** |
| **GT-AG** | 54667 | 96.6 | 517730 | 98.6 |
| **GC-AG** | 1683 | 3.0 | 4351 | 0.8 |
| **Others** | 232 | 0.4 | 3068 | 0.6 |
| **Total** | 56582 | | 525149 | |

| | **Mouse** | | | |
|---|---|---|---|---|
| | ***lncRNAs*** | ***%*** | ***PCGs*** | ***%*** |
| **GT-AG** | 28586 | 96.5 | 388973 | 98.8 |
| **GC-AG** | 570 | 1.9 | 3217 | 0.8 |
| **Others** | 455 | 1.5 | 1598 | 0.4 |
| **Total** | 29611 | | 393788 | |

GC-AG introns showed a preferential location in the first intron of both lncRNAs and PCGs (Table 3). Indeed, in the human genome, their percentage resulted higher in the first intron (lncRNAs: 4.2%; PCGs: 1.2%) with respect to inner introns (lncRNAs: 2.1%; PCGs: 0.8%) and the same trend was observed in mouse (first: lncRNAs 2.4%, PCGs: 1.2%; inner: lncRNAs 0.4%, PCGs 0.8%). In all cases, differences were statistically significant (Chi-square tests = 204.7 and 120.9, 1 df, p-value < $2.2 \times 10^{-16}$, respectively, for human lncRNAs and PCGs; Chi-square tests = 233.6 and 62.7, 1 df, p-value < $2.2 \times 10^{-16}$ and < $2.4 \times 10^{-15}$, respectively, for mouse lncRNAs and PCGs).

**Table 3. Number of GC-AG introns in first or inner position**

| | **Human** | | | | | |
|---|---|---|---|---|---|---|
| | **lncRNA** | | | **PCGs** | | |
| | ***Total*** | ***GC-AG*** | ***%*** | ***Total*** | ***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** | | | | | |
|---|---|---|---|---|---|---|
| | **lncRNA** | | | **PCGs** | | |
| | ***Total*** | ***GC-AG*** | ***%*** | ***Total*** | ***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 |

### 2.2. Confirmation of GC-AG splice sites enrichment in other datasets

A validation of these results was obtained by investigating six alternative source of lncRNA annotations: (1) the FANTOM5 dataset, (2) the NONCODE, (3) the BIGTranscriptome dataset, (4) the LncBook dataset, (5) the MITranscriptome dataset, and (6) the LNCipedia dataset.

In all datasets, the frequency of GC-AG splice junctions was found higher with respect to that in PCG introns (Table 4). The prevalence of GC-AG introns among lncRNAs new datasets ranged from 2.3 to 3.5%, resulting in all cases significantly higher respect to PCGs (all comparisons p-values < $2.2 \times 10^{-16}$).

**Table 4. Percentage of splice sites in different human lncRNAs datasets**

| Source | Type | %GT-AG | %GC-AG | %others | p-value |
|---|---|---|---|---|---|
| **GENCODE** | PCGs | 98.6 | 0.8 | 0.6 | NA |
| **GENCODE** | lncRNA | 96.6 | 3.0 | 0.4 | $<2.2\text{x}10^{-16}$ |
| **FANTOM5** | lncRNA | 94.4 | 2.6 | 3.0 | $<2.2\text{x}10^{-16}$ |
| **NONCODE** | lncRNA | 64.7 | 2.3 | 33.0 | $<2.2\text{x}10^{-16}$ |
| **BIGTranscriptome** | lncRNA | 90.5 | 2.5 | 6.9 | $<2.2\text{x}10^{-16}$ |
| **LncBook** | lncRNA | 81.2 | 2.8 | 16.0 | $<2.2\text{x}10^{-16}$ |
| **MITranscriptome** | lncRNA | 93.2 | 3.5 | 3.3 | $<2.2\text{x}10^{-16}$ |
| **LNCipedia** | lncRNA | 83.6 | 2.3 | 14.1 | $<2.2\text{x}10^{-16}$ |

In all datasets, GC-AG introns showed their preferential localization in the first introns in which their prevalence is constantly almost the double with respect to inner introns (all comparisons p-values $< 2.2 \times 10^{-16}$) (Table 5).

**Table 5. Percentage of GC-AG introns in first or inner position in different human datasets**

| Source | Type | % in first | % in inner | p-value |
|---|---|---|---|---|
| **GENCODE** | PCGs | 1.2 | 0.8 | $<2.2\text{x}10^{-16}$ |
| **GENCODE** | lncRNA | 4.2 | 2.1 | $<2.2\text{x}10^{-16}$ |
| **FANTOM5** | lncRNA | 3.5 | 2.0 | $<2.2\text{x}10^{-16}$ |
| **NONCODE** | lncRNA | 3.3 | 1.5 | $<2.2\text{x}10^{-16}$ |
| **BIGTranscriptome** | lncRNA | 3.7 | 1.8 | $<2.2\text{x}10^{-16}$ |
| **LncBook** | lncRNA | 3.9 | 1.9 | $<2.2\text{x}10^{-16}$ |
| **MITranscriptome** | lncRNA | 4.8 | 2.6 | $<2.2\text{x}10^{-16}$ |
| **LNCipedia** | lncRNA | 2.9 | 1.7 | $<2.2\text{x}10^{-16}$ |

In mouse, data was replicated in the NONCODE dataset in which both the enrichment of GC-AG splice junction (1.7% with respect to 0.8% in PCGs; p-values $< 2.2 \times 10^{-16}$) and their preferential location in the first intron (fist introns 2.3%, inner introns 1.1%; p-values $< 2.2 \times 10^{-16}$) was confirmed.

### 2.3. Confirmation of GC-AG splice sites enrichment in lower species

To evaluate the enrichment of GC-AG introns in lncRNAs during evolution, we analyzed the frequency of the different splice junctions in lower organisms as *D. melanogaster* and *C. elegans* (Table 6).

**Table 6. Number of GC-AG introns in different species**

| | **lncRNAs** | | **PCGs** | |
|---|---|---|---|---|
| | ***GC-AG*** | ***%*** | ***GC-AG*** | ***%*** |
| ***H. sapiens*** | 1683 | 3.0 | 4351 | 0.8 |
| ***M. musculus*** | 570 | 1.9 | 3217 | 0.8 |
| ***D. melanogaster*** | 21 | 1.7 | 1063 | 0.7 |
| ***C. elegans*** | 9 | 1.9 | 1189 | 0.5 |

The ratio of GC-AG splice sites in lncRNAs of *D. melanogaster* was found significantly higher than in PCGs (GC-AG in lncRNAs: 1.7% of total splice junctions with respect to GC-AG in PCGs: 0.7%; Chi-square test = 57.0, 1 df, p-value = $4.3 \times 10^{-14}$). In *C. elegans*, GC-AG splice junctions account for 2.0% of total splice junctions in lncRNAs thus confirming the enrichment with respect to the 0.6% in PCGs (Chi-square test = 12.7, 1 df, p-value = $3.5 \times 10^{-4}$). A preferential location of GC-AG splice sites in the first intron was also observed in lncRNA and PCGs of both *D. melanogaster* and *C. elegans* but due to their small number their statistical relevance could not be appreciated.

### 3. Peculiar features of GC-AG splice sites

The enrichment of GC-AG junctions in lncRNAs together with their preferential localization in first introns in both lncRNAs and PCGs suggested that they could play a particular role in gene expression regulation leading us to a deeper characterization of their features.

### 3.1. Introns length

In human, GC-AG introns resulted shorter both in lncRNAs and PCGs and they showed the same trend whether they are first or inner introns (Figure 23).

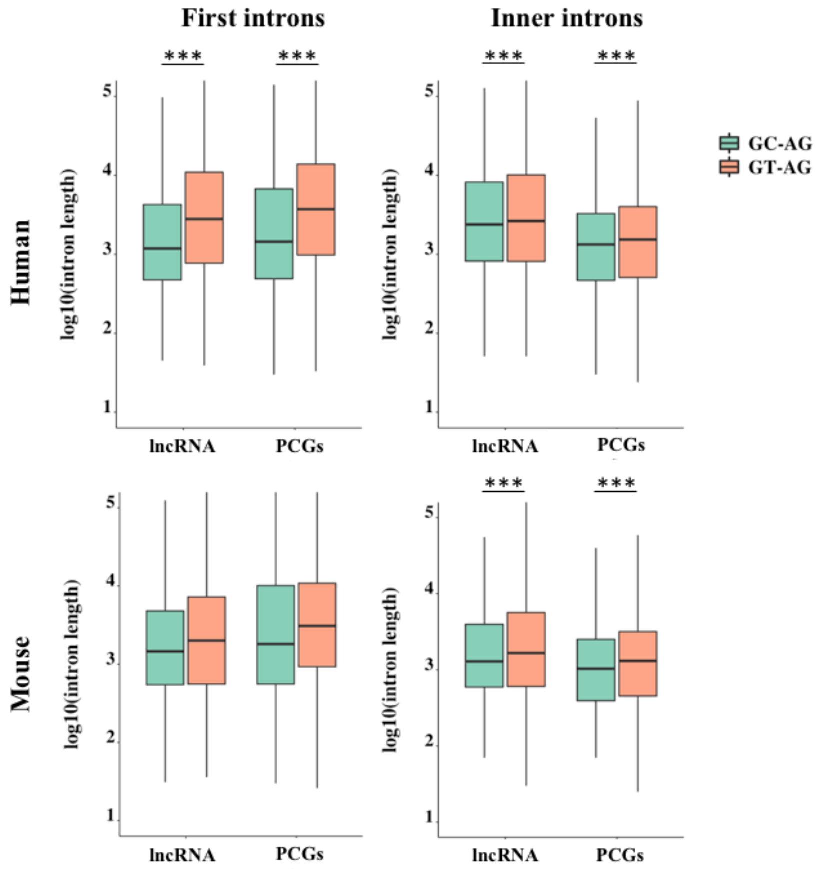


**Figure 23. Boxplot showing GC-AG and GT-AG mean introns length in human and mouse.** GC-AG and GT-AG introns were divided between first and inner in lncRNAs and PCGs. The mean length was calculated in base pairs (bp). *** p-value < 0.01.

For GC-AG first introns, the average length resulted almost halved with respect to GT-AG first introns in both human lncRNAs and PCGs (lncRNAs: GC 6700 ±600 bp, GT 12923 ±201 bp, Wilcoxon tests p-value $< 2.2 \times 10^{-16}$; PCGs: GC 8999 ±648 bp, GT 15335 ±162 bp, Wilcoxon tests p-value $< 2.2 \times 10^{-16}$). Human GC-AG inner introns showed the same decrease in length, albeit to a lesser extent (lncRNAs: GC 8666 ±827 bp, GT 13995 ±194 bp, Wilcoxon tests p-value = 0.012; PCGs: GC 4165 ±197 bp, GT 5411 ±25 bp, Wilcoxon tests p-value $= 6.3 \times 10^{-10}$).
In mouse, GC-AG introns appeared shorter but only when they are inner introns (lncRNAs: GC 5190 ±734 bp, GT 7523 ±148 bp, Wilcoxon tests p-value = 0.0302; PCGs: GC 3186 ±192 bp, GT 4437 ±27 bp, Wilcoxon tests p-value $= 9.5 \times 10^{-14}$).

The shorter length of human GC-AG introns was also confirmed in the FANTOM5 dataset as both GC-AG first and inner introns of lncRNAs were significantly shorter than GT-AG ones (first intron: GC 8169 ±600 bp, GT 14516 ±137 bp, Wilcoxon tests p-value $< 2.2 \times 10^{-16}$; inner introns: GC 8648 ±544 bp, GT 15784 ±119 bp, Wilcoxon tests p-value $< 2.2 \times 10^{-16}$).

## 3.2. Splicing elements features

### 3.2.1. Splice sites scores

To evaluate the splicing efficiency of GC-AG junctions, we computed their strength using the standard position weight-matrix (WM) model implemented in the MaxEntScan tool [226], which assigns a computationally predicted score for 5′ and 3′ splice sites. Overall, the strength of 5′ and 3′ss resulted lower in lncRNAs than in PCGs both in human and mouse and it was presumably one of the causes of the previously reported inefficiency of lncRNAs splicing (Figure 24).

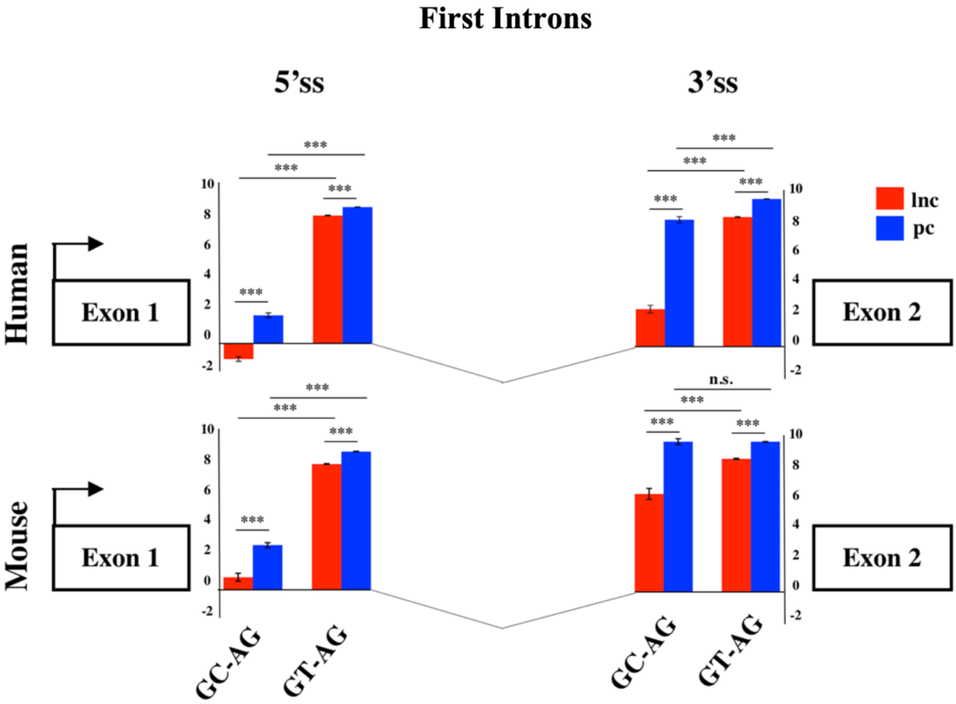


**Figure 24. Splice junction strengths of the first introns.** Schematic representation of the average scores of 5′ and 3′ss strengths of long non-coding and PCGs in human and mouse. The strengths of 5′ and 3′ ss were calculated as weight matrix scores for GC-AG and GT-AG first introns. *** p < 0.001.

Despite lower weight-matrix scores for 5′ss-GC were expected, due to their imperfect pairing with the U1 snRNA, 5′ss-GC scores of lncRNAs resulted strongly reduced with respect to 5′ss-GC of PCGs in both species (human: lncRNAs 5′ss-GC WM = 0.50, PCGs 5′ss-GC WM = 2.76, Wilcoxon test p-value < $2.2 \times 10^{-16}$; mouse: lncRNAs 5′ss-GC WM = 1.63, PCGs 5′ss-GC WM = 3.38, Wilcoxon test p-value < $2.2 \times 10^{-16}$). The reduced strength of lncRNAs 5′ss-GC appeared to be attributable almost exclusively to first intron junctions, whose scores resulted lower compared to those of inner introns, both in human and mouse (human: lncRNAs first intron 5′ss-GC WM = −0.93, inner intron 5′ss-GC WM = 2.60, Wilcoxon test p-value < $2.2 \times 10^{-16}$; mouse: lncRNAs first intron 5′ss-GC WM = 0.78, inner intron 5′ss-GC WM = 2.65, Wilcoxon test p-value < $2.2 \times 10^{-16}$) (Figure 25).

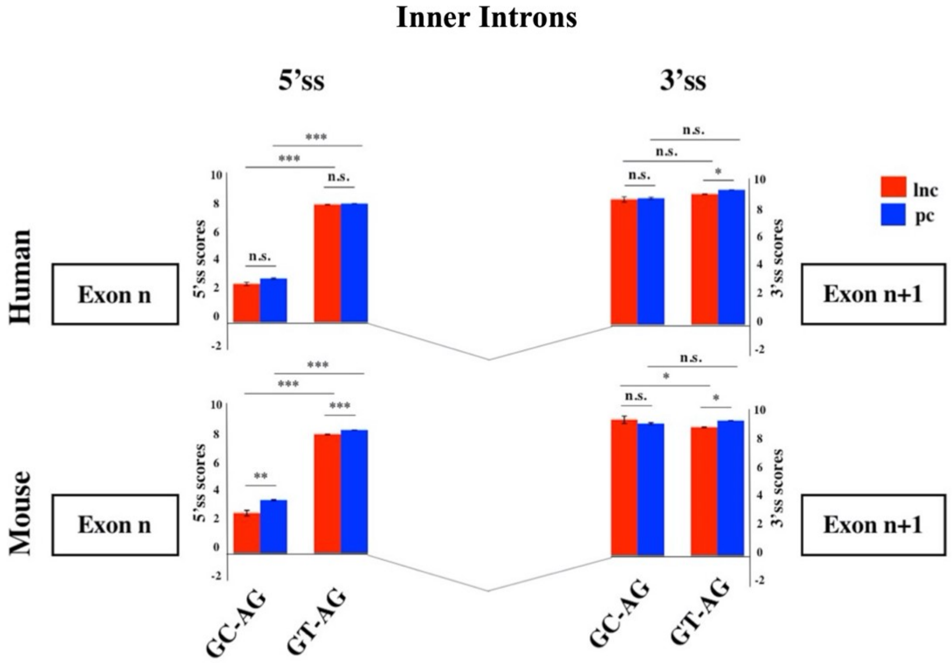


**Figure 25. Splice junction strengths of inner introns.** Schematic representation of the average scores of 5' and 3'ss strengths of lncRNAs and PCGs in human and mouse for GC-AG and GT-AG introns. *** p < 0.001.

To test whether the 5′ss and 3′ss weight-matrix scores and the introns length showed any correlation, the Spearman test was applied. The strength of 5′ss and 3′ss was found positively correlated when located in the first intron of human lncRNAs ($r$ = 0.58, p-value < 2.2 × $10^{-16}$) and PCGs ($r$ = 0.22, p-value = 1 × $10^{-16}$). In mouse, the correlation was significant only in lncRNAs (lncRNAs: $r$ = 0.51, p-value < 2.2 × $10^{-16}$; PCGs: $r$ = 0.04, p-value = 0.34). The strengths of both 5′ss and 3′ss were positively correlated to intron length and this correlation was found more pronounced in the first intron in both species.

### 3.2.2. Polypyrimidine tract

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 and this appeared attributable to their shorter polypyrimidine tracts (PPT) (Figure 26).

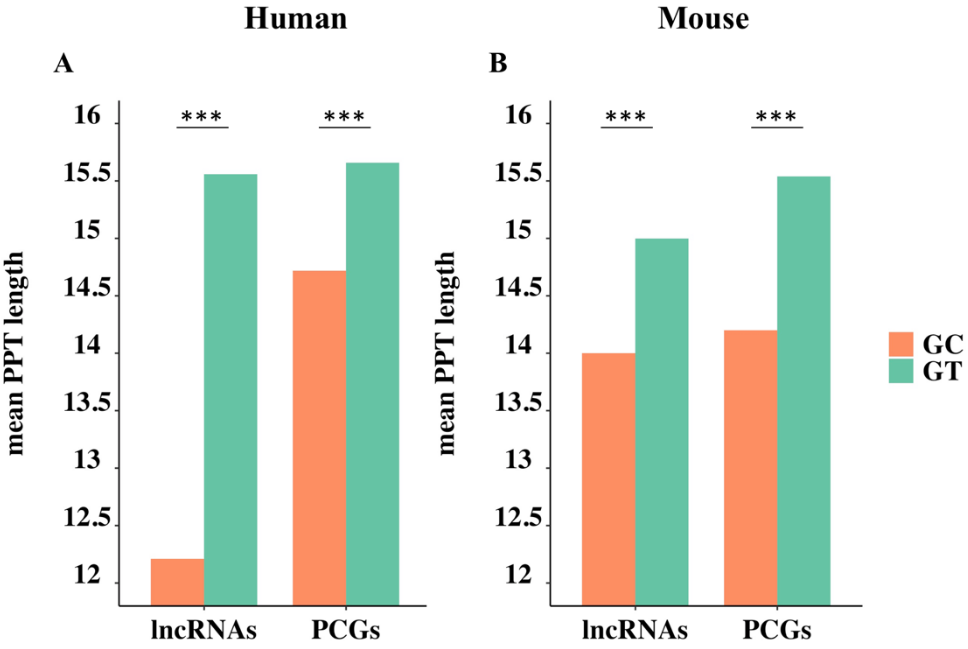


**Figure 26. Bar graph showing the mean PPT length in GC and GT introns.** The mean length of polypyrimidine tract was determined in the GC and GT containing introns of lncRNAs and PCGs in the human and mouse genome. *** p < 0.001.

In human, the mean length of PPT of GC-introns resulted significantly shorter than GT ones in both lncRNAs and PCGs (lncRNAs: GT-introns PPT mean = 16 bp, GC-introns PPT mean = 12 bp, Wilcoxon tests p-value $< 2.2 \times 10^{-16}$; PCGs: GT-introns PPT mean = 16 bp, GC-introns PPT mean = 15 bp, Wilcoxon tests p-value $< 2.2 \times 10^{-16}$). The same trend was observed in mouse for both gene classes (lncRNAs: GT-introns PPT mean = 15 bp, GC-introns PPT mean = 14 bp, Wilcoxon tests p-value = 0.001; PCGs: GT-introns PPT mean = 16 bp, GC-introns PPT mean = 15 bp, Wilcoxon tests p-value = 0.021). As it occurred for 5′ss, very weak 3′ss seemed preferentially located in the lncRNAs first intron in both human and mouse (data not shown).

Differently from what was reported for PCGs, in which weak donor sites appeared flanked by stronger consensus at the acceptor sites [243][244], our analysis demonstrated that lncRNAs contained a class of very weak introns, preferentially located as first.

### 3.3. Expression level of GC transcripts

In order to evaluate a putative effect of the presence of a GC-AG intron on the expression level of the corresponding transcripts, we analyzed a panel of ten different human tissues (i.e., anterior cingulate cortex, amygdala, cerebellum, heart, kidney, liver, lung, skin, spleen, and testis) obtained from the GTEx project [231]. For each tissue, raw RNA-seq data from eight samples were processed using the Salmon tool [233] which provide an accurate quantification of transcripts expression. Transcript per million (TPM) of each single transcript, were calculated in each tissue and expressed transcripts were defined based on a threshold of TPM > 0.1 for lncRNAs and of TPM > 0.5 for PCGs to account for highly different level of expression between the two classes of genes. The mean TPM of the expressed transcripts in each tissue were reported distinguishing between GC-AG- or GT-AG-intron containing transcripts and between transcripts containing a GC-AG intron in the first or inner position (Figure 27).

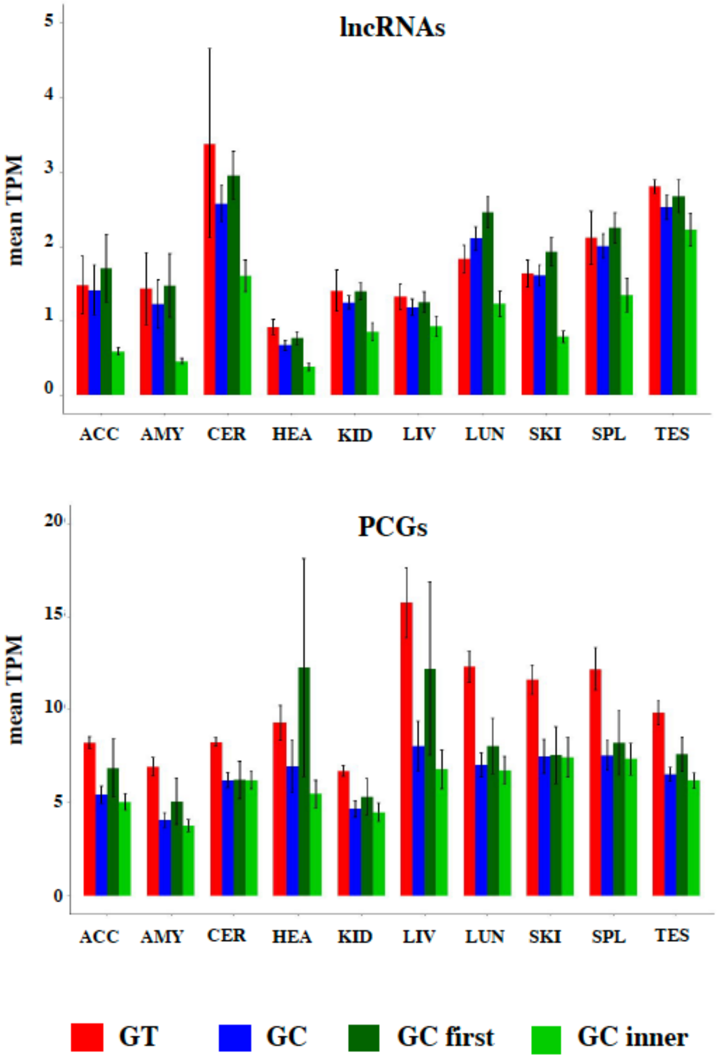


**Figure 27. Expression of GC-AG and GT-AG containing transcripts.** Bar graph representing the expression of lncRNAs and PCGs transcripts in different human tissues (ACC: anterior cingulate cortex; AMY: amygdala; CER: cerebellum; HEA: heart; KID: kidney; LIV: liver; LUN: lung; SKI: skin; SPL: spleen; TES: testis). Transcripts were divided as containing GC-AG or GT-AG-introns and between transcripts containing a GC-AG intron in the first or inner position. The expression of transcripts was calculated as mean TPM combining expression data from 10 different tissues together.

The mean TPM of GC-AG-containing transcripts appeared always lower with respect to GT-AG containing ones (with the exception of TPM values for lncRNAs in lung) and in the majority of the cases the difference resulted statistically significant. In addition, we calculated the mean TPM of all tissues combined together in order to provide an overall estimation of expression data. Combining all tissues together, the mean TPM of lncRNAs resulted significantly lower with respect to GT-AG-containing transcripts in both lncRNAs and PCGs (lncRNAs: 1.79 for GC-AG containing transcripts vs. 2.00 for GT-AG containing ones, Wilcoxon test p-value = $3.2 \times 10^{-15}$; PCGs: 8.40 for GC-AG containing transcripts vs. 11.10 for GT-AG containing ones, Wilcoxon test p-value < $2.2 \times 10^{-16}$) (Figure 28).

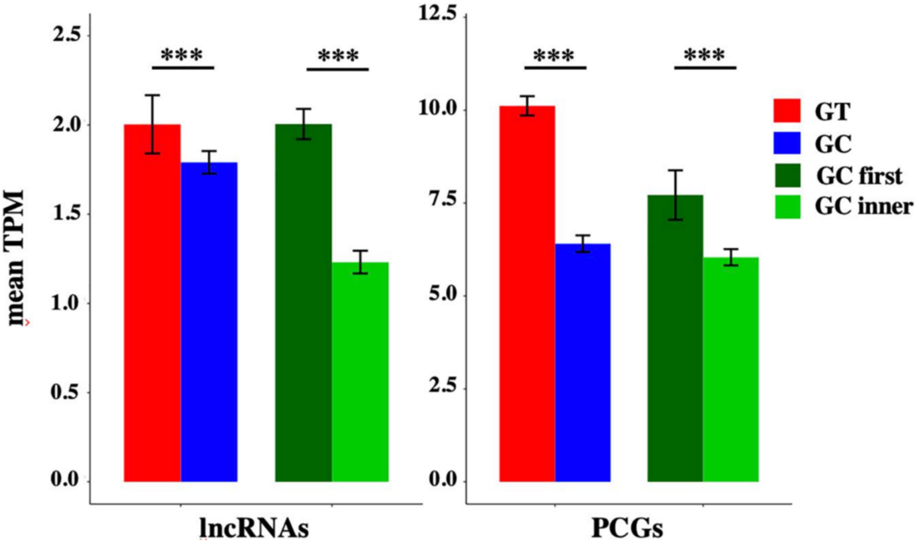


**Figure 28. Expression of GC-AG- and GT-AG-containing transcripts.** Bar graph representing the expression of lncRNAs and PCGs transcripts containing GC-AG- or GT-AG-introns and between transcripts containing a GC-AG intron in the first or inner position. The expression of transcripts was calculated as mean TPM combining expression data from 10 different tissues together. ***p < 0.001.

The mean TPM of transcripts containing a GC-AG intron in the first position appeared always higher with respect to transcripts having a GC-AG intron in inner positions, both in lncRNAs and PCGs. Considering the combination of all tissues, the mean TPM of GC-first introns lncRNAs resulted significantly higher with respect to GC-inner introns (lncRNAs: GC-first mean TPM 2.00 vs. GC-inner mean TPM 0.58, Wilcoxon test p-value < $2.2 \times 10^{-16}$; PCGs: GC-first mean TPM 7.71 vs. GC-inner

mean TPM 6.04, Wilcoxon test p-value = $5.4 \times 10^{-11}$) (Figure 28). Interestingly, in some cases the expression levels of transcripts with the GC-AG intron located as the first resulted higher than GT-AG-containing transcripts especially in lncRNAs (i.e., in anterior cingulate cortex, amygdala, lung, skin, and spleen in lncRNAs and in heart for PCGs).
These results suggest that the presence of a GC-AG intron may affect transcripts expression by reducing their overall transcription levels, both in lncRNAs and PCGs. Moreover, GC-AG introns may have a different effect on transcript expression levels depending on where they are located as transcripts harboring a GC-AG intron in their first intron showed overall higher expression levels with respect to transcripts with an inner GC-AG intron.

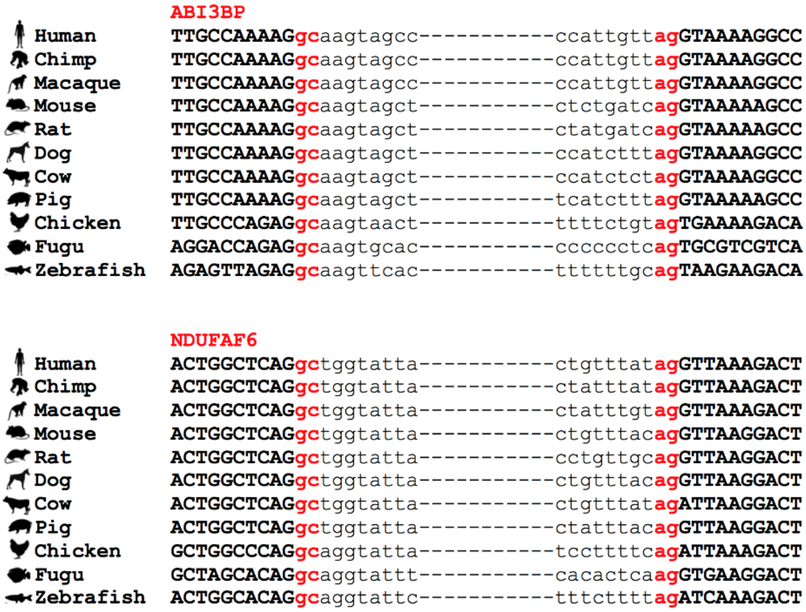
Nevertheless, these data must be taken with caution as: (i) the high variability in expression profiles, that is a common feature of both lncRNAs and PCGs, could affect mean TPM calculation especially for those categories containing a small number of transcripts, and (ii) transcripts containing a GC-AG intron often differ from GT-AG ones for other alternative splicing events which could possibly make differences in expression levels not univocally attributed to the presence of a GC intron.

### 3.4. Conservation of GC splice sites in mouse and other species

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

Based on the human-mouse ortholog information provided by the Ensembl project (https://www.ensembl.org/index.html), a total of 908 PCGs were conserved between the two species, thus accounting for a considerable fraction of total GC-AG containing genes (47% of human GC-AG containing genes; 54% of mouse GC-AG containing genes). Remarkably, in more than 75% of cases the GC-AG introns also shared the same ordinal position in the homologous genes.

Interestingly, 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 mouse but it extended across evolutionary distant species. For example, the GC-AG splice sites of human *ABI3BP* (ABI family member 3 binding protein) and *NDUFAF6* (NADH:ubiquinone oxidoreductase complex assembly factor 6) genes were shown to be conserved in chimp, macaque, mouse, rat, dog, cow, pig, chicken, fugu, and zebrafish (Figure 29).



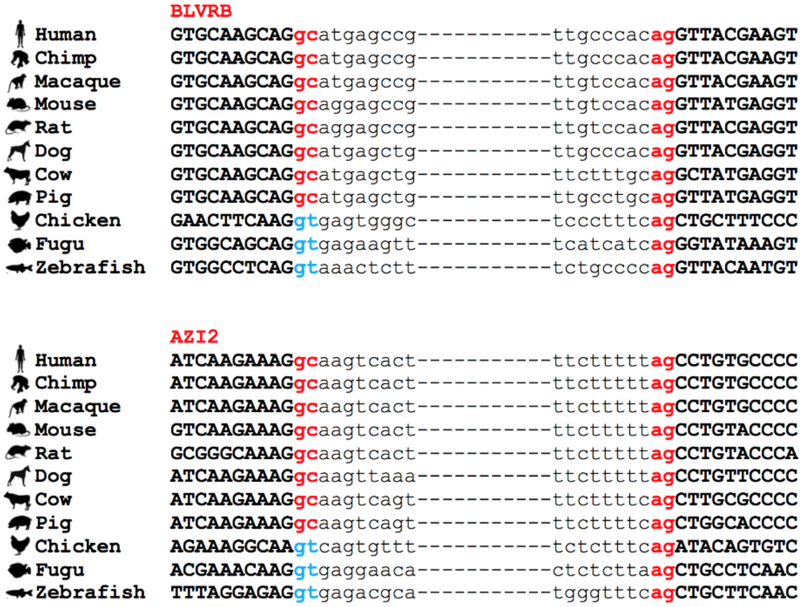
```
           ABI3BP
Human      TTGCCAAAAGgcaagtagcc------------ccattgttagGTAAAAGGCC
Chimp      TTGCCAAAAGgcaagtagcc------------ccattgttagGTAAAAGGCC
Macaque    TTGCCAAAAGgcaagtagcc------------ccattgttagGTAAAAGGCC
Mouse      TTGCCAAAAGgcaagtagct------------ctctgatcagGTAAAAAGCC
Rat        TTGCCAAAAGgcaagtagct------------ctatgatcagGTAAAAAGCC
Dog        TTGCCAAAAGgcaagtagct------------ccatctttagGTAAAAGGCC
Cow        TTGCCAAAAGgcaagtagct------------ccatctctagGTAAAAGGCC
Pig        TTGCCAAAAGgcaagtagct------------tcatctttagGTAAAAAGCC
Chicken    TTGCCCAGAGgcaagtaact------------ttttctgtagTGAAAAGACA
Fugu       AGGACCAGAGgcaagtgcac------------cccccctcagTGCGTCGTCA
Zebrafish  AGAGTTAGAGgcaagttcac------------ttttttgcagTAAGAAGACA

           NDUFAF6
Human      ACTGGCTCAGgctggtatta------------ctgtttatagGTTAAAGACT
Chimp      ACTGGCTCAGgctggtatta------------ctatttatagGTTAAAGACT
Macaque    ACTGGCTCAGgctggtatta------------ctatttgtagGTTAAAGACT
Mouse      ACTGGCTCAGgctggtatta------------ctgtttacagGTTAAGGACT
Rat        ACTGGCTCAGgctggtatta------------cctgttgcagGTTAAGGACT
Dog        ACTGGCTCAGgctggtatta------------ctgtttacagGTTAAGGACT
Cow        ACTGGCTCAGgctggtatta------------ctgtttatagATTAAGGACT
Pig        ACTGGCTCAGgctggtatta------------ctatttacagGTTAAGGACT
Chicken    GCTGGCCCAGgcaggtatta------------tcctttcagATTAAAGACT
Fugu       GCTAGCACAGgcaggtattt------------cacactcaagGTGAAGGACT
Zebrafish  ACTGGCACAGgcaggtattc------------tttcttttagATCAAAGACT
```

**Figure 29. Conservation of GC-AG introns across multiple species.** Multiple sequence alignment of GC-AG splice sites in the first intron of ABI3BP gene and the intron 6 of NDUFAF6 gene across the 11 species indicated.

Moreover, the ordinal position of the GC-AG intron was also conserved: in the *ABI3BP* gene, GC-AG introns was always the first intron in all cases and in the *NDUFAF6* gene, the GC-AG intron conserved its position in intron 6 in all species.

In other cases, the GC-AG splice sites appeared to be not conserved among lower species but only in mammals. For example, the GC-AG splice sites of the human genes *BLVRB* (biliverdin reductase B) and *AZI2* (5-azacytidine induced 2) were shown to be conserved in first and inner introns of mammals, respectively, while the canonical GT was found in chicken, fugu and zebrafish (Figure 30).



```
           BLVRB
Human      GTGCAAGCAGgcatgagccg------------ttgcccacagGTTACGAAGT
Chimp      GTGCAAGCAGgcatgagccg------------ttgcccacagGTTACGAAGT
Macaque    GTGCAAGCAGgcatgagccg------------ttgtccacagGTTACGAAGT
Mouse      GTGCAAGCAGgcaggagccg------------ttgtccacagGTTATGAGGT
Rat        GTGCAAGCAGgcaggagccg------------ttgtccacagGTTACGAGGT
Dog        GTGCAAGCAGgcatgagctg------------ttgcccacagGTTACGAGGT
Cow        GTGCAAGCAGgcatgagctg------------ttctttgcagGCTATGAGGT
Pig        GTGCAAGCAGgcatgagctg------------ttgcctgcagGTTATGAGGT
Chicken    GAACTTCAAGgtgagtgggc------------tccctttcagCTGCTTTCCC
Fugu       GTGGCAGCAGgtgagaagtt------------tcatcatcagGGTATAAAGT
Zebrafish  GTGGCCTCAGgtaaactctt------------tctgccccagGTTACAATGT

           AZI2
Human      ATCAAGAAAGgcaagtcact------------ttcttttt agCCTGTGCCCC
Chimp      ATCAAGAAAGgcaagtcact------------ttctttttagCCTGTGCCCC
Macaque    ATCAAGAAAGgcaagtcact------------ttctttttagCCTGTGCCCC
Mouse      GTCAAGAAAGgcaagtcact------------ttctttttagCCTGTACCCC
Rat        GCGGGCAAAGgcaagtcact------------ttctttttagCCTGTACCCA
Dog        ATCAAGAAAGgcaagttaaa------------ttctttttagCCTGTTCCCC
Cow        ATCAAGAAAGgcaagtcagt------------ttcttttcagCTTGCGCCCC
Pig        ATCAAGAAAGgcaagtcagt------------ttcttttcagCTGGCACCCC
Chicken    AGAAAGGCAAgtcagtgttt------------tctctttcagATACAGTGTC
Fugu       ACGAAACAAGgtgaggaaca------------ctctcttaagCTGCCTCAAC
Zebrafish  TTTAGGAGAGgtgagacgca------------tgggtttcagCTGCTTCAAC
```

**Figure 30. Conservation of GC-AG introns across multiple species.** Multiple sequence alignment of GC-AG splice sites in the first intron of BLVRB gene and the inner intron of AZI2 gene across 11 species.

Despite the assessment of the conservation of lncRNAs was hindered by the lack of annotation in most species, a number of conserved GC-AG splice junctions between human and mouse was determined. Indeed, the *TMEM51-AS1* (TMEM51 antisense RNA 1), the *MALAT*1 (metastasis associated lung adenocarcinoma transcript 1) and the *NEAT1* (nuclear paraspeckle assembly transcript 1) genes contained a first GC-AG intron in both species whereas the *JPX* (JPX transcript, XIST activator) gene contained an inner GC-AG intron in both human and mouse.

The high conservation of the GC-AG introns between human and mouse and across multiple species could hint toward their functional importance and suggest their involvement in specific biological processes.

### 3.5. Functional enrichment analysis

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 PCGs. By means of the DAVID Functional Annotation Tool [234] and the PANTHER Overrepresentation Test [235], we selected only those terms that resulted significantly enriched in both species and by both tools (Figure 31).

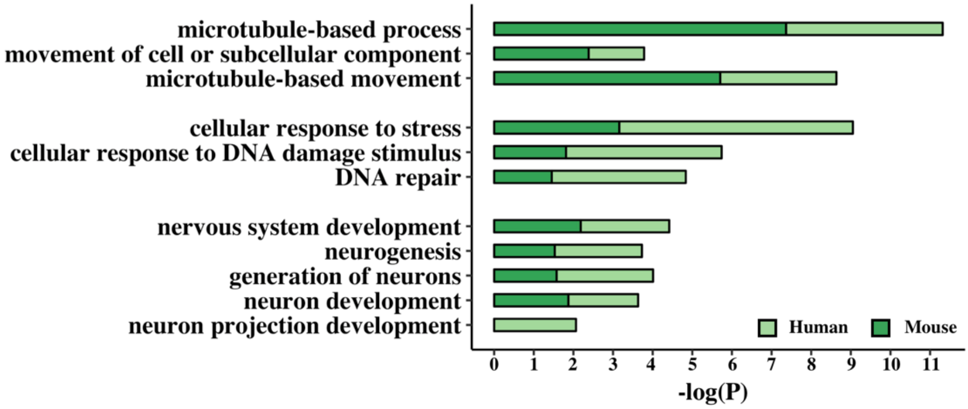


**Figure 31. Functional enrichment analysis of GC-AG-containing genes.** Bar graph representing the GO terms found significantly enriched in GC-AG containing PCGs. The GO term name is indicated on the Y-axis, and the (–$\log_{10}$) of the p-values is indicated on the X-axis.

This resulted in the identification of three groups of related terms in the biological process ontology. The first group comprised the GO term “microtubule-based movement” and its ancestors “movement of cell or subcellular component” and “microtubule-based process” and included 221 human and 176 mouse genes. Despite very little is known about the biological processes in which lncRNAs are involved, at least two of the GC-AG-containing lncRNAs were described to have a role in the regulation of the movement of cells or subcellular components: the *MEG3* (maternally expressed 3) gene [245][246] and the *SOX2-OT* (SOX2 overlapping transcript) gene [247]. The second group contained the GO term “DNA Repair” and its ancestors “cellular response to DNA damage stimulus” and “cellular response to stress” 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 [69] and the *NEAT1* gene [248]. In the third group, the GO term "neuron projection development" with its ancestors "neuron development," "generation of neurons," "neurogenesis," and "nervous system development" were included and contained 273 and 220 human and mouse genes. Several lncRNAs with a GC-AG intron were described to play a role in neuron development and growth like the *MEG3* gene [249], the *NEAT1* gene [250], the *SOX2-OT* gene, the *GDNF-AS1* (GDNF antisense RNA 1) gene and the *MIAT* (myocardial infarction associated transcript) [251]. All the reported GO terms resulted significantly enriched after correction for multiple testing.

### 3.6. Alternative splicing analysis of GC introns and polyA sites

As the presence of a GC-AG intron was proposed to increase the level of alternative splicing [164], we compared the transcriptional diversity of both lncRNAs and PCGs owning at least one GC-AG intron with respect to the ones containing only GT-AG introns (Table 7). In human, both long non-coding and protein-coding GC-AG-containing genes being transcribed in more than one isoform exceeded the number of GT-AG-containing genes with the same features [lncRNAs-GC n = 471 (38.5%) vs. lncRNAs-GT n = 3204 (28.9%), Chi-square test = 47.7, 1 df, p-value = $4.8 \times 10^{-12}$; PCGs-GC n = 1642 (84.9%) vs. PCGs-GT n = 11469 (68.9%), Chi-square test = 212.5, 1 df, p-value < $2.2 \times 10^{-16}$].

**Table 7. Number of genes containing GC-AG or GT-AG introns transcribed in different number of transcripts**

| | **Human** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **GC-AG** | | | | **GT-AG** | | | |
| ***N. Transcripts*** | ***lncRNAs*** | ***%*** | ***PCGs*** | ***%*** | ***lncRNAs*** | ***%*** | ***PCGs*** | ***%*** |
| **1** | 753 | 61.5 | 292 | 15.1 | 7879 | 71.1 | 5174 | 31.1 |
| **>1** | 471 | 38.5 | 1642 | 84.9 | 3204 | 28.9 | 11469 | 68.9 |
| **Total** | **1224** | | **1934** | | **11083** | | **16643** | |

| | **Mouse** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **GC-AG** | | | | **GT-AG** | | | |
| ***N. Transcripts*** | ***lncRNAs*** | ***%*** | ***PCGs*** | ***%*** | ***lncRNAs*** | ***%*** | ***PCGs*** | ***%*** |
| **1** | 285 | 60.3 | 552 | 33.1 | 6157 | 74.7 | 9380 | 49.8 |
| **>1** | 188 | 39.7 | 1117 | 66.9 | 2085 | 25.3 | 9463 | 50.2 |
| **Total** | **473** | | **1669** | | **8242** | | **18843** | |

The same trend was confirmed in mouse, where long non-coding and protein-coding GC-AG-containing genes with more than one isoform resulted more abundant than their GT-AG counterpart [lncRNAs-GC n = 188 (39.7%) vs. lncRNAs-GT n = 2085 (25.3%), Chi-square test = 47.7, 1 df, p-value = 4.9 × $10^{-12}$; PCGs-GC n = 1117 (66.9%) vs. PCGs-GT n = 9463 (50.2%), Chi-square test = 170.6, 1 df, p-value < 2.2 × $10^{-16}$].

To evaluate if the increase of alternative splicing could be attributed to some particular splicing events, we used the SUPPA2 tool [229] to perform a quantitative profiling of alternative splicing events involving GC-AG introns in comparison with GT-AG ones.

The analysis revealed that human GC-AG introns were preferentially involved in the alternative 5′ss events in both lncRNAs and PCGs [lncRNAs: n = 150 (18.9%) of GC-AG introns, n = 3494 (9.7%) of GT-AG introns, Chi-square test = 54.9, 1 df, p-value = 1.2 × $10^{-13}$; PCGs: n = 389 (31.6%) of GC-AG introns, n = 10500 (10.1%) of GT-AG introns, Chi-square test = 415.6, 1 df, p-value < 2.2 × $10^{-16}$]. The same trend was also observed in mouse [lncRNAs: n = 41 (34.7%) of GC-AG introns, n = 1,000 (11.1%) of GT-AG introns, Chi-square test = 40.9, 1 df, p-value = 1.5 × $10^{-10}$; PCGs: n = 188 (32.9%) of GC-AG introns, n = 5902 (12.5%) of GT-AG introns, Chi-square test = 136.7, 1 df, p-value < 2.2 × $10^{-16}$].

As alternative polyadenylation regulation is a process directly linked to 5′ss recognition and splicing, we analyzed the variability of last exon (LE) defining the total number of alternative last exon for each gene.

Interestingly, we observed a significant increase of LE variability in GC-AG-containing genes compared to GT-AG ones in both gene classes. In human lncRNAs, 37.6% of GC-AG-containing genes had more than one alternative last exon compared to 27.7% of GT-AG-containing genes (Chi-square test = 52.4, 1 df, p-value = $4.5 \times 10^{-13}$). The same difference was established for human PCGs (80% of GC-AG genes with alternative last exon versus 64.1% of GT-AG genes, Chi-square test = 151.4, 1 df, p-value $< 2.2 \times 10^{-16}$). The same significant enrichment were confirmed in mouse (lncRNAs: 38.3% of GC-AG genes with alternative last exon versus 23.5% of GT-AG genes, Chi-square test = 52.4, 1 df, p-value = $4.5 \times 10^{-13}$; PCGs: 59.3% of GC-AG genes with alternative last exon versus 43.7% of GT-AG genes, Chi-square test = 151.4, 1 df, p-value $< 2.2 \times 10^{-16}$) (Figure 32).

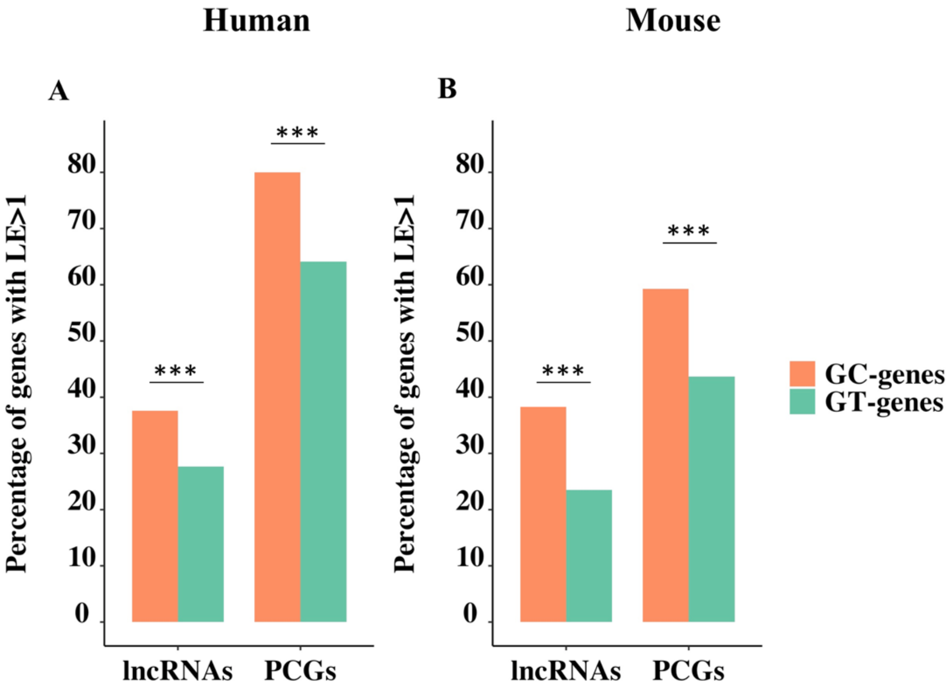


**Figure 32. Bar graph showing the percentage of genes with alternative last exon.** The percentage of alternative last exon was determined in GC and GT containing genes of lncRNAs and PCGs in human and mouse. *** $p < 0.001$.

Furthermore, the increased of LE variability in GC-AG-containing genes was strengthened by a higher mean of alternative last exons per gene respect to GT-AG-containing genes in human and mouse lncRNAs and PCGs (Table 8).

**Table 8. Mean number of alternative last exons (LE>1) in GC and GT-genes**

| | **Human** | | | | | |
|---|---|---|---|---|---|---|
| | **lncRNAs** | | | **PCGs** | | |
| | ***N*** | ***Mean*** | ***SEM*** | ***N*** | ***Mean*** | ***SEM*** |
| **GC-genes** | 460 | 7.86 | 0.59 | 1547 | 4.84 | 0.07 |
| **GT-genes** | 3069 | 3.72 | 0.09 | 10676 | 3.70 | 0.02 |

| | **Mouse** | | | | | |
|---|---|---|---|---|---|---|
| | **lncRNAs** | | | **PCGs** | | |
| | ***N*** | ***Mean*** | ***SEM*** | ***N*** | ***Mean*** | ***SEM*** |
| **GC-genes** | 181 | 3.65 | 0.17 | 990 | 3.47 | 0.07 |
| **GT-genes** | 1935 | 2.92 | 0.05 | 8226 | 2.98 | 0.02 |

As differences in polyadenylation regulation could result from the different assortment of polyadenylation signals (PAS), we analyzed the last 40 nucleotides of each last exon for their content in the 16 different PAS reported in the paper of Beaudoing and colleagues [230]. Our results highlighted a higher ratio of lncRNAs lacking any of the 16 PAS considered compared with PCGs in both species (human: lncRNAs PAS = 0 48.0% versus PCGs PAS = 0 29.7%, Chi-square test = 2200.3, 1 df, p-value $< 2.2 \times 10^{-16}$; mouse: lncRNAs PAS = 0 42.4% versus PCGs PAS = 0 20.3%, Chi-square test = 2370.9, 1 df, p-value $< 2.2 \times 10^{-16}$) (Table 9).

Table 9. Number of lncRNAs and PCGs with at least one or without any PAS

| | Human | | | |
|---|---|---|---|---|
| | lncRNAs | | PCGs | |
| | *N* | *%* | *N* | *%* |
| **PAS = 0** | 10857 | 48.0 | 13646 | 29.7 |
| **PAS = 1** | 11778 | 52.0 | 32289 | 70.3 |

| | Mouse | | | |
|---|---|---|---|---|
| | lncRNAs | | PCGs | |
| | *N* | *%* | *N* | *%* |
| **PAS = 0** | 5334 | 42.4 | 7320 | 20.3 |
| **PAS = 1** | 7238 | 57.6 | 28745 | 79.7 |

Considering GC-AG- and GT-AG-containing genes separately, we observed that the higher ratio of PAS = 0 was more evident in GC-AG transcripts but the difference was statistically significant only in human (Table 10).

Table 10. Number of GC or GT lncRNAs or PCGs without any PAS (PAS=0)

| | | lncRNAs | | PCGs | |
|---|---|---|---|---|---|
| | | *PAS = 0* | *%* | *PAS = 0* | *%* |
| **Human** | **GC transcripts** | 2210 | 56.9 | 1890 | 31.6 |
| | **GT transcripts** | 8647 | 46.1 | 11756 | 29.4 |
| **Mouse** | **GC transcripts** | 387 | 41.6 | 797 | 22.0 |
| | **GT transcripts** | 4947 | 42.5 | 7320 | 22.6 |

Looking at the assortment of different PAS, we observed a preferential usage of non-canonical PAS in lncRNAs with respect to PCGs in both species (human: lncRNAs non-canonical PAS 65.4% versus PCGs non-canonical PAS 57.6%, Chi-square test = 346.1, 1 df, p-value $< 2.2 \times 10^{-16}$; mouse: lncRNAs non-canonical PAS 65.8% versus PCGs non-canonical PAS 56.7%, Chi-square test = 309.2, 1 df, p-value $< 2.2 \times 10^{-16}$) (Table 11).

**Table 11. Number of the canonical AATAAA or of the other polyadenylation signals in lncRNAs and PCGs**

| | lncRNAs | | | | PCGs | | | |
|---|---|---|---|---|---|---|---|---|
| | *AATAAA* | *%* | *Other* | *%* | *AATAAA* | *%* | *Other* | *%* |
| **Human** | 6408 | 34.6 | 12144 | 65.4 | 21462 | 42.4 | 29187 | 57.6 |
| **Mouse** | 3849 | 34.2 | 7415 | 65.8 | 18967 | 43.3 | 24807 | 56.7 |

No differences between GC-AG- and GT-AG-containing genes were observed in the usage of different PAS.

Our results highlighted differences in alternative splicing and polyadenylation sites and signals between lncRNAs and PCGs which appeared more evident in GC-AG-containing genes thus suggesting that this 5′ss could contribute to gene expression regulation.

## 4. Enrichment of GC splice sites in wobble splicing mechanism

Given the peculiar features of GC-AG introns and their enrichment in alternative splicing mechanisms, in particular alternative 5'ss, we next aimed to further study the characteristics of GC-AG introns involved in alternative splicing. Interestingly, through the manual analysis of each intron using the UCSC Genome Browser, we observed that in many cases GC-AG introns were involved in a wobble splicing event, i.e. alternatively spliced from an alternative donor splice site located at a very short distance. Consequently, this made us hypothesize that wobble splicing events might be enriched with GC-AG introns in which they could have a functional role. Thus, we next aimed to study the prevalence and putative role of GC-AG introns in the wobble splicing mechanism.

### 4.1. Wobble splicing features at 5' splice sites

We characterized the wobble splicing features of human and mouse lncRNAs and PCGs considering a more recent gene annotations from GENCODE: human release 30 (16193 lncRNAs and 19986 PCGs) and mouse release M21 (13374 lncRNAs and

21951 PCGs) (Table 12) and in which we confirmed the previously determined enrichment of GC-AG introns in lncRNAs (Table 13).

**Table 12. Number of annotations from the GENCODE datasets used in the analyses.**

| | **Human<br>v30** | | **Mouse<br>M21** | |
|---|---|---|---|---|
| | ***lncRNA*** | ***PCGs*** | ***lncRNA*** | ***PCGs*** |
| **genes** | 16193 | 19986 | 13374 | 21951 |
| **transcripts** | 30369 | 57883 | 18930 | 45165 |
| **exons** | 96215 | 614179 | 52021 | 450480 |
| **introns** | 65846 | 556296 | 33091 | 405315 |

**Table 13. Number of different splice junctions consensus in GENCODE v30 and M21.**

| | **Human** | | | |
|---|---|---|---|---|
| | ***lncRNAs*** | ***%*** | ***PCGs*** | ***%*** |
| **GT-AG** | 63773 | 96.85 | 549176 | 98.72 |
| **GC-AG** | 1905 | 2.89 | 4780 | 0.86 |
| **AT-AC** | 39 | 0.06 | 871 | 0.16 |
| **Others** | 129 | 0.20 | 1469 | 0.26 |
| **Total** | **65846** | | **556296** | |

| | **Mouse** | | | |
|---|---|---|---|---|
| | ***lncRNAs*** | ***%*** | ***PCGs*** | ***%*** |
| **GT-AG** | 32088 | 96.97 | 400792 | 98.88 |
| **GC-AG** | 673 | 2.03 | 3363 | 0.83 |
| **AT-AC** | 28 | 0.08 | 535 | 0.13 |
| **Others** | 302 | 0.91 | 625 | 0.15 |
| **Total** | **33091** | | **405315** | |

The wobble splicing events were defined when two donor splice sites were separated by a sequence termed (Δ) which ranged from 1 to 18 nucleotide (i.e. 5’-GY$\Delta_{(1\text{-}18)}$GY-3’). Donor splice sites having only GT, GC, or AT dinucleotides have been considered for the analysis of 5’ss wobble splicing events as non-canonical splice sites could be mainly attributable to mis-annotations due to mis-alignments.

We identified 253 wobble splicing events in 221 lncRNA gene and 1090 event in 1023 PCGs in human for a total of 1343 event in 1244 gene (Table 14).

**Table 14. Number of 5'ss wobble splicing events in human and mouse.**

| | | N. of events | N. of genes involved by WS |
|---|---|---|---|
| **Human** | ***lncRNAs*** | 253 | 221 |
| | ***PCGs*** | 1090 | 1023 |
| | ***Total*** | 1343 | 1244 |
| **Mouse** | ***lncRNAs*** | 83 | 77 |
| | ***PCGs*** | 780 | 746 |
| | ***Total*** | 863 | 823 |

In mouse, 83 wobble splicing event appeared in 77 lncRNA gene and 780 event in 746 PCGs for a total of 863 event in 823 gene. Wobble splicing events at the donor splice site occurred in a relatively low number of genes in both species (Human: 3.44% of total genes; Mouse: 2.33% of total genes). This suggests that 5'ss wobble splicing appears to be not a common widespread mechanism. Nevertheless, the percentage of wobble splicing events were significantly higher in PCGs than in lncRNAs (Human: the percentage of lncRNA genes having at least one wobble splicing event was 1.36%, while for PCGs the percentage was 5.12%; in mouse: the percentage of lncRNA genes having at least one wobble splicing event was 0.58%, while for PCGs the percentage was 3.40%) (Table 15).

**Table 15. Percentage of genes involved in wobble splicing**

| | | Total N. genes | N. of genes involved by WS | % |
|---|---|---|---|---|
| **Human** | ***lncRNAs*** | 16193 | 221 | 1.36 |
| | ***PCGs*** | 19986 | 1023 | 5.12 |
| | ***Total*** | 36179 | 1244 | 3.44 |
| **Mouse** | ***lncRNAs*** | 13374 | 77 | 0.58 |
| | ***PCGs*** | 21951 | 746 | 3.40 |
| | ***Total*** | 35325 | 823 | 2.33 |

We quantified the ratio of 5'ss wobble splicing events from the total number of alternative splicing events determined by the SUPPA2 tool in lncRNAs and PCGs in both species. In human, 5'ss wobble splicing accounted for 1.1% and 2.7% of the total alternative splicing events in lncRNAs and PCGs respectively; in mouse, 5'ss wobble splicing accounted for 1.8% and 4.0% of the total alternative splicing events in lncRNAs and PCGs respectively (Table 16).

**Table 16. Percentage of 5' wobble splicing events among total alternative splicing events**

| | | N. alt. splicing events | N. 5' WS events | % |
|---|---|---|---|---|
| **Human** | ***lncRNAs*** | 22161 | 253 | 1.1 |
| | ***PCGs*** | 39924 | 1090 | 2.7 |
| | ***Total*** | 62085 | 1343 | 2.2 |
| **Mouse** | ***lncRNAs*** | 4656 | 83 | 1.8 |
| | ***PCGs*** | 19605 | 780 | 4.0 |
| | ***Total*** | 24261 | 863 | 3.6 |

Among 5'ss alternative splicing events, the percentage of 5'ss wobble splicing events in human was 19.2% for lncRNAs and 29.6% for PCGs. Similarly, in mouse, the percentage of 5'ss wobble splicing events was 17.9% for lncRNAs and 31.7% for PCGs (Table 17).

**Table 17. Percentage of 5' wobble splicing events among 5'ss alternative splicing events**

| | | N. 5'ss alt. splicing events | N. 5' WS events | % |
|---|---|---|---|---|
| **Human** | ***lncRNAs*** | 1316 | 253 | 19.2 |
| | ***PCGs*** | 3676 | 1090 | 29.6 |
| | ***Total*** | 4992 | 1343 | 26.9 |
| **Mouse** | ***lncRNAs*** | 463 | 83 | 17.9 |
| | ***PCGs*** | 2460 | 780 | 31.7 |
| | ***Total*** | 2923 | 863 | 29.5 |

We next evaluated the prevalence of wobble splicing events at 3'ss in comparison to wobble splicing events at the 5'ss. Similarly to 5'ss, wobble splicing events at the 3'ss were defined in lncRNAs and PCGs of both species when two acceptor splice sites were separated by a Δ which ranged from 1 to 18 nucleotide (i.e. 5'-AGΔ $_{(1-18)}$AG-3'). We identified 500 wobble splicing event in 416 lncRNA genes and 2034 event in 1779 PCGs in human; in mouse, 360 wobble splicing event in 348 lncRNA genes and 1518 event in 1373 PCGs. As expected and in accordance to previous studies, wobble splicing events at 3’ss appeared significantly more frequent than those at the 5’ss and in a higher number of genes (Human: 6.1% of total lncRNA and PCGs vs 3.44% for 5’ss; Mouse: 4.9% of total lncRNA and PCGs vs 2.33% for 5’ss) (Table 18).

**Table 18. Number of 3’ss WS events and their percentage**

| | | Total N. genes | N. 3’ss WS events | N. 3’ss WS genes | % |
|---|---|---|---|---|---|
| **Human** | ***lncRNAs*** | 16193 | 500 | 416 | 2.6 |
| | ***PCGs*** | 19986 | 2034 | 1779 | 8.9 |
| | ***Total*** | 36179 | 2534 | 2195 | 6.1 |
| **Mouse** | ***lncRNAs*** | 13374 | 360 | 348 | 2.6 |
| | ***PCGs*** | 21951 | 1518 | 1373 | 6.3 |
| | ***Total*** | 35325 | 1878 | 1721 | 4.9 |

In human, 3’ss wobble splicing accounted for 2.3% and 5.1% of the total alternative splicing events in lncRNAs and PCGs respectively; in mouse, 3’ss wobble splicing accounted for 7.7% and 7.7% of the total alternative splicing events in lncRNAs and PCGs respectively. Among 3’ss alternative splicing events, the percentage of 3’ss wobble splicing events in human was 32.7% for lncRNAs and 44.7% for PCGs; in mouse, the percentage of 3’ss wobble splicing events was 55.6% for lncRNAs and 48.4% for PCGs (Table 19).

**Table 19. Percentage of 3’ss WS among total and 3’ss alternative splicing events**

| | | N. 3’ss WS events | Tot. N. alt. spl. events | % | N. 3’ss alt. spl. events | % |
|---|---|---|---|---|---|---|
| **Human** | ***lncRNAs*** | 500 | 22161 | 2.3 | 1527 | 32.7 |
| | ***PCGs*** | 2034 | 39924 | 5.1 | 4551 | 44.7 |
| | ***Total*** | 2534 | 62085 | 4.1 | 6078 | 41.7 |
| **Mouse** | ***lncRNAs*** | 360 | 4656 | 7.7 | 647 | 55.6 |
| | ***PCGs*** | 1518 | 19605 | 7.7 | 3138 | 48.4 |
| | ***Total*** | 1878 | 24261 | 7.7 | 3785 | 49.6 |

Interestingly, a number of introns showed a co-occurrence of wobble splicing events at both the donor and acceptor site of the same intron (in human: 14 events in lncRNAs and 113 event in PCGs; in mouse: 6 events in lncRNAs and 94 events in PCGs). Moreover, in few cases among these wobble splicing events, a compensation mechanism was observed (i.e. Δ of added nucleotides at 5'ss = Δ of deleted nucleotides at 3'ss). We noted 2 compensated wobble splicing event in lncRNAs and 38 events in PCGs in human while in mouse 1 compensated event was observed in lncRNAs and 46 events in PCGs.

### 4.2. Validation of wobble splicing events in other datasets

A validation of the prevalence and ratios of wobble splicing events identified in the GENCODE datasets was obtained by investigating the wobble splicing events in two other gene annotation datasets: (1) the NONCODEv5 dataset for lncRNAs gene annotations (number of genes= 96308) and (2) the RefSeq dataset for PCGs annotations (number of genes= 19367). From the NONCODE dataset, 1675 wobble splicing event was identified in 1347 lncRNA genes accounting for 1.4% of total genes (Table 20).

**Table 20. Number of 5'ss WS events from the NONCODE and RefSeq datasets**

| | Dataset | Total genes | WS events | WS gene | % WS genes |
|---|---|---|---|---|---|
| **lncRNAs** | NONCODE | 96308 | 1675 | 1347 | 1.40 |
| **PCGs** | RefSeq | 19367 | 1077 | 971 | 5.01 |

Similarly to our results from the GENCODE dataset, the NONCODE wobble splicing events accounted for 2.68% of total alternative splicing events and 19.8% of the total 5'ss alternative splicing event. In RefSeq, 1077 wobble splicing events appeared in 971 genes accounting for 5.01% of the total number of genes. Similarly, the RefSeq wobble splicing events accounted for 2.7% of total alternative splicing events and 19.7% of the total 5'ss alternative splicing event (Table 21).

**Table 21. Percentage of 5' wobble splicing events among the total and 5'ss alternative splicing events**

| | Dataset | WS events | N. alt. spl. events | % | N. 5'ss alt. spl. events | % |
|---|---|---|---|---|---|---|
| **lncRNAs** | NONCODE | 1675 | 62585 | 2.68 | 8447 | 19.8 |
| **PCGs** | RefSeq | 1077 | 39415 | 2.70 | 5474 | 19.7 |

### 4.3. Enrichment of wobble splicing at GC splice sites

To get further insights on the donor splice sites of 5'ss wobbling events, we analyzed the prevalence of the different 5'ss sites in wobble splicing in both gene classes in human and mouse. The quantification of the splice sites involved in 5'ss wobble splicing showed that GC-AG were significantly enriched and more prone to undergo wobble splicing than GT-AG introns in lncRNAs and PCGs in both human and mouse (Table 22).

**Table 22. Enrichment of the different splice junctions involved in wobble splicing**

| | **Human** | | | | | |
|---|---|---|---|---|---|---|
| | **lncRNAs** | | | **PCGs** | | |
| | ***expected*** | ***observed*** | ***p-value*** | ***expected*** | ***observed*** | ***p-value*** |
| **GT** | 490 | 463 | n.s. | 2152 | 1905 | $1.1 \times 10^{-4}$ |
| **GC** | 15 | 43 | $3.5 \times 10^{-4}$ | 20 | 218 | $< 2.2 \times 10^{-16}$ |
| **AT** | 1 | 0 | n.s. | 4 | 56 | $2.1 \times 10^{-11}$ |

| | **Mouse** | | | | | |
|---|---|---|---|---|---|---|
| | **lncRNAs** | | | **PCGs** | | |
| | ***expected*** | ***observed*** | ***p-value*** | ***expected*** | ***observed*** | ***p-value*** |
| **GT** | 161 | 157 | n.s. | 1543 | 1408 | $1.1 \times 10^{-2}$ |
| **GC** | 3 | 8 | $2.1 \times 10^{-2}$ | 13 | 113 | $< 2.2 \times 10^{-16}$ |
| **AT** | 0 | 1 | n.s. | 2 | 39 | $1.1 \times 10^{-8}$ |

Moreover, the U12-type donor splice sites AT appeared also enriched in wobble splicing events of both species in PCGs only. The enrichment of AT splice sites was not observed in lncRNAs as they are quite rare in the transcribed isoforms.

By considering the different 5'ss pairs involved in wobble splicing, an excess of GC ss was observed in lncRNAs and PCGs of both species. An enrichment of GT/GC and GC/GC was reported in lncRNA and PCGs of human and an excess of GC/GC wobble splicing pairs was observed in mouse. Moreover, an excess of AT donor splice sites was observed in the PCGs of both species in which a significantly high number of GT/AT and AT/AT wobble splicing pair was observed (Table 23).

**Table 23. Enrichment of the different splice junctions pairs involved in wobble splicing**

| | **Human** | | | | | |
|---|---|---|---|---|---|---|
| | **lncRNAs** | | | **PCGs** | | |
| | ***expected*** | ***observed*** | ***p-value*** | ***expected*** | ***observed*** | ***p-value*** |
| **GT/GT** | 248 | 209 | n.s. | 1054 | 844 | $3.3 \times 10^{-4}$ |
| **GT/GC** | 4 | 31 | $5.2 \times 10^{-6}$ | 36 | 199 | $<2.2 \times 10^{-16}$ |
| **GT/AT** | 0 | 0 | NA | 0 | 18 | $5.7 \times 10^{-5}$ |
| **GC/GC** | 1 | 13 | $2.8 \times 10^{-6}$ | 0 | 9 | $7.5 \times 10^{-3}$ |
| **GC/AT** | 0 | 0 | NA | 0 | 1 | n.s. |
| **AT/AT** | 0 | 0 | NA | 0 | 19 | $3.4 \times 10^{-5}$ |
| **Total** | | 253 | | | 1090 | |

| | **Mouse** | | | | | |
|---|---|---|---|---|---|---|
| | **lncRNAs** | | | **PCGs** | | |
| | ***expected*** | ***observed*** | ***p-value*** | ***expected*** | ***observed*** | ***p-value*** |
| **GT/GT** | 82 | 74 | n.s. | 765 | 645 | n.s. |
| **GT/GC** | 1 | 8 | $3.1 \times 10^{-2}$ | 15 | 108 | $< 2.2 \times 10^{-16}$ |
| **GT/AT** | 0 | 1 | n.s. | 0 | 10 | $4.3 \times 10^{-3}$ |
| **GC/GC** | 0 | 0 | NA | 0 | 2 | n.s. |
| **GC/AT** | 0 | 0 | NA | 0 | 1 | 1 |
| **AT/AT** | 0 | 0 | NA | 0 | 14 | $4.8 \times 10^{-4}$ |
| **Total** | | 83 | | | 780 | |

### 4.4. Wobble splice sites scores

To evaluate the splicing efficiency of wobble splicing donor sites, we calculated their strength using the standard position weight-matrix (WM) model implemented in the MaxEntScan tool. In general, wobbling 5’ss appeared weaker compared to all donor splice sites (Figure 33).

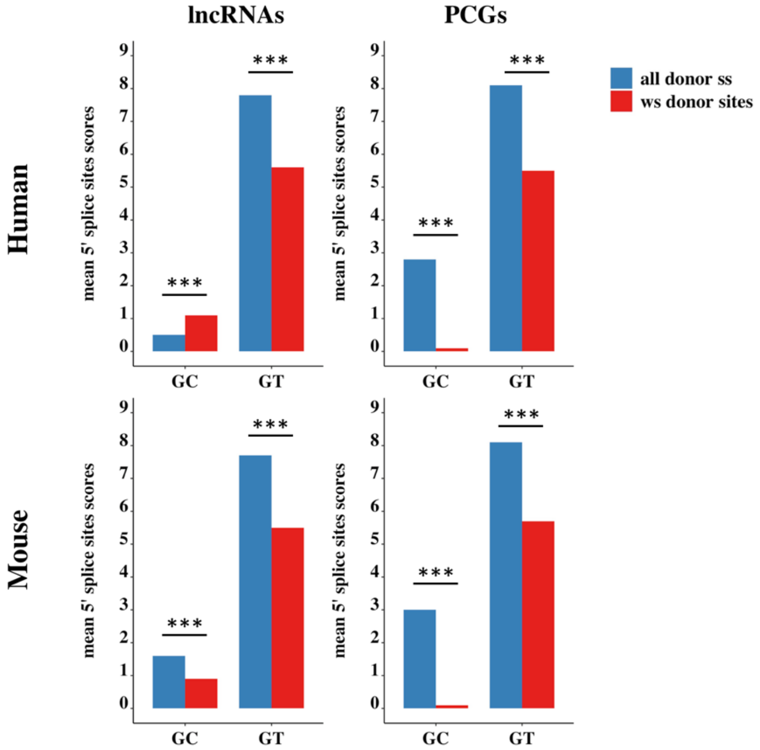


**Figure 33. Bar graph showing the scores of wobble 5’ss compared to all donor ss.** The average splice sites scores were calculated for GC and GT introns involved in wobble splicing in comparison to all other GC and GT introns in lncRNAs and PCGs of human and mouse. *** $p < 0.001$.

GT donor splice sites appeared significantly weaker in wobble splicing events in comparison to all other GT splice sites in both human (lncRNA: WS 5′ss-GT WM =

5.6, all 5′ss-GT WM = 7.8, Wilcoxon test p-value < $2.2\times10^{-16}$; PCGs: WS 5′ss-GT WM = 5.5, all 5′ss-GT WM = 8.1, Wilcoxon test p-value < $2.2\times10^{-16}$) and mouse (lncRNA: WS 5′ss-GT WM = 5.5, all 5′ss-GT WM = 7.7, Wilcoxon test p-value < $2.2\times10^{-16}$; PCGs: WS 5′ss-GT WM = 5.7, all 5′ss-GT WM = 8.1, Wilcoxon test p-value < $2.2\times10^{-16}$). The same trend was observed for GC splice sites except in human lncRNAs in which wobble-spliced GC donor sites appeared stronger than all GC ss (Human: PCGs: WS 5′ss-GC WM = 0, all 5′ss-GC WM = 2.8, Wilcoxon test p-value < $2.2\times10^{-16}$; mouse: PCGs: WS 5′ss-GC WM = 0, all 5′ss-GC WM = 3.0, Wilcoxon test p-value < $1.6\times10^{-12}$).
However, it is worth to note that in the majority of cases the splice sites pair in a wobble splicing events showed weak splice site strength (data not shown). This suggest that the selection of a donor splice site among the tandem splice sites in a wobble splicing event could be controlled and regulated through a different mechanism than the binding of the U1 snRNP.

To test whether the difference in the splice sites scores of wobble splicing splice sites pairs is affected by the separating distance (Δ), we performed a Spearman correlation analysis between the variance in scores of splice sites pairs of a wobble splicing event and Δ. The strength of donor splice sites and their separating distance showed a negative correlation in human lncRNAs ($r$ = -0.18, p-value = 0.004) and PCGs ($r$ = -0.19, p-value = $3.0 \text{ x } 10^{-10}$) and in mouse PCGs ($r$ = -0.19, p-value = $5.0\text{x}10^{-08}$) showing that as the two wobbling splice sites are closer to each other, the variation in their splice sites strength increase.

### 4.5. Expression of wobble splicing transcripts

#### 4.5.1. Expression levels of wobble splicing transcripts

In order to study the putative effect of the presence a wobble splicing events on the expression level of the corresponding transcripts, we analyzed the transcripts expression data in a panel of ten different human tissues (i.e., anterior cingulate cortex, amygdala, cerebellum, heart, kidney, liver, lung, skin, spleen, and testis) obtained from the GTEx project [231]. We analyzed 8 samples per tissue for a total of 80 RNA-seq samples using a bioinformatics pipeline devised in our group and described before (see Materials and Methods). We analyzed the expression of all lncRNAs and PCGs transcripts involved in wobble splicing. As lncRNAs show a substantial lower level of expression in comparison to PCGs, transcripts showing a mean TPM < 0.5 were defined as not expressed.

The mean TPM of the transcripts involved in wobble splicing appeared always lower with respect to the other transcripts and in all the cases the difference resulted statistically significant (Figure 34).

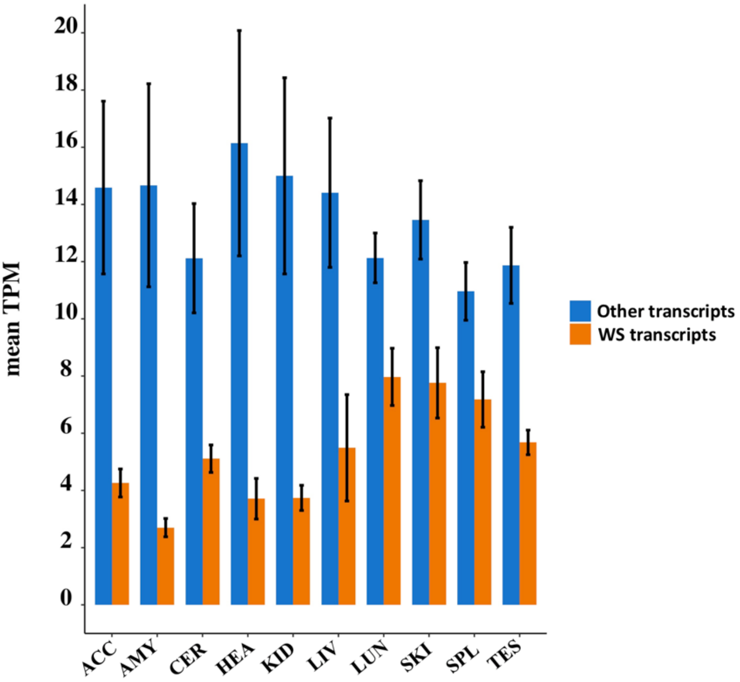


**Figure 34. Bar graph showing the expression of wobble splicing transcripts.** The average TPM was calculated for wobble splicing transcripts in comparison to the other transcripts in 10 different tissues ((ACC: anterior cingulate cortex; AMY: amygdala; CER: cerebellum; HEA: heart; KID: kidney; LIV: liver; LUN: lung; SKI: skin; SPL: spleen; TES: testis)). All comparisons were statistically significant (p-value < 0.001)

The mean TPM ranged between 2.5 and 8 for wobble splicing transcripts while for the remaining transcripts the TPM was between 11 and 16 TPM. We further analyzed the expression level of the transcripts involved in wobble splicing in terms of having GC or GT introns. The mean TPM of wobble splicing transcripts in each tissue was reported distinguishing between GC-AG- or GT-AG-intron containing transcripts (Figure 35).

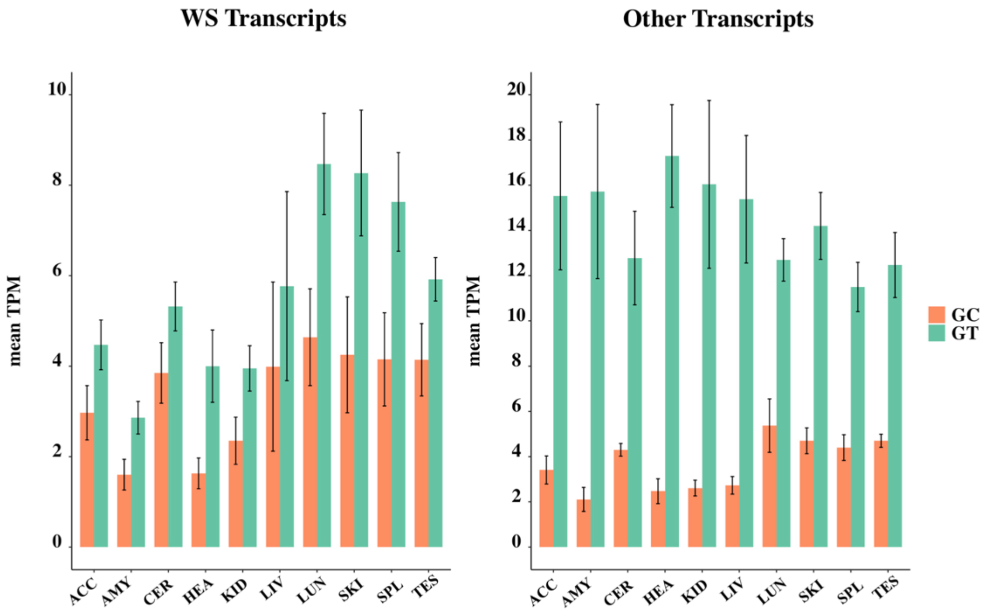


**Figure 35. Bar graph showing the expression of wobble splicing transcripts containing GC or GT splice sites.** The average TPM was calculated for GC transcripts in comparison to GT containing ones for those involved in wobble splicing and the other transcripts in 10 different tissues (ACC: anterior cingulate cortex; AMY: amygdala; CER: cerebellum; HEA: heart; KID: kidney; LIV: liver; LUN: lung; SKI: skin; SPL: spleen; TES: testis). All comparisons were statistically significant (p-value < 0.001)

In all cases, the GC containing transcripts showed a lower level of expression than GT ones in all studied tissues but to a lesser extent in the wobble spliced transcripts.

### 4.5.2. Tissue specificity of wobble splicing transcripts expression

As different previous studies reported conflicting observations whether wobble splicing events are tissue-specific or show tissue-specific regulation, we next analyzed the tissue specificity of wobble splicing mechanism in the 10 different studied tissues (Figure 36). We counted the expression of the transcript in the different tissues and we applied am arbitrary threshold to determine if a wobble splicing event is tissue specific. A transcript was identified as "housekeeping" if it was expressed in more than 8 tissues and "tissue-specific" if it showed expression in 3 tissues or less. The remaining transcripts were categorized as "intermediate".

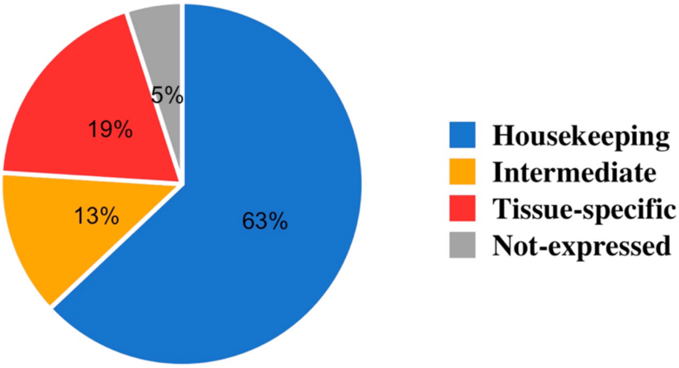


**Figure 36. Pie chart showing the tissue specificity of wobble splicing transcripts.** The wobble splicing events were categorized as housekeeping, intermediate, tissue-specific or not expressed.

Our results showed that in 63% of cases the transcripts appeared to be expressed in the majority of studied tissues, in 19% of cases the transcripts appeared to be tissue-specific and in 13% the transcripts were identified as intermediate. This suggests that wobble splicing could show a tissue-specific expression or regulation.

We next analyzed the expression level of the transcripts in the wobble splicing events in the different classes (Figure 37). As expected, in all different tissues, the housekeeping transcripts showed a significant higher expression than tissue-specific transcripts.

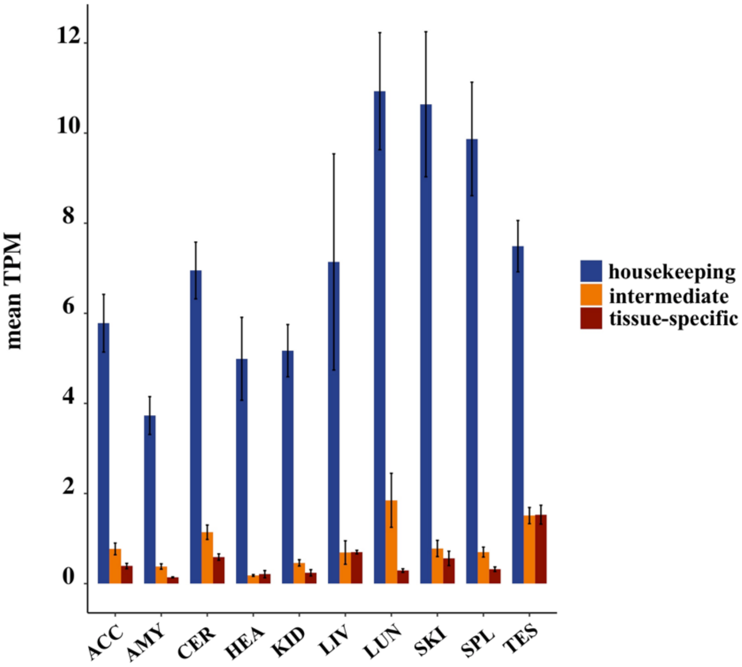


**Figure 37. Bar graph showing the expression of transcripts according to their tissue-specificity category.** The transcripts were categorized as housekeeping, intermediate or tissue specific in the 10 different analysed tissues (ACC: anterior cingulate cortex; AMY: amygdala; CER: cerebellum; HEA: heart; KID: kidney; LIV: liver; LUN: lung; SKI: skin; SPL: spleen; TES: testis). All comparisons appeared statistically significant (p-value < 0.001).

Moreover, we observed many examples in which the tissue specificity appeared between GC and GT containing transcripts involved in wobble splicing (Figure 38). For example, in the *TNS2* gene, both the GT and GC containing transcripts appeared to be expressed in all tissues, however showing different expression level in different tissues. In general, the GT transcript appears to be equally or more expressed than the GC transcript except in lung, skin and testis in which the GC transcript showed higher expression. In the *MAN1C1* gene, only the GC transcript appeared to be expressed. In the *AAK1* gene, the expression of the GC transcript was observed in heart, lung and testis while no expression for the GT transcript was observed in the other tissues. In the *KLDR1* gene, expression was observed in lung, spleen and testis. The GC transcript was more expressed than the GT one in lung and testis while the GT transcript was more expressed in spleen.

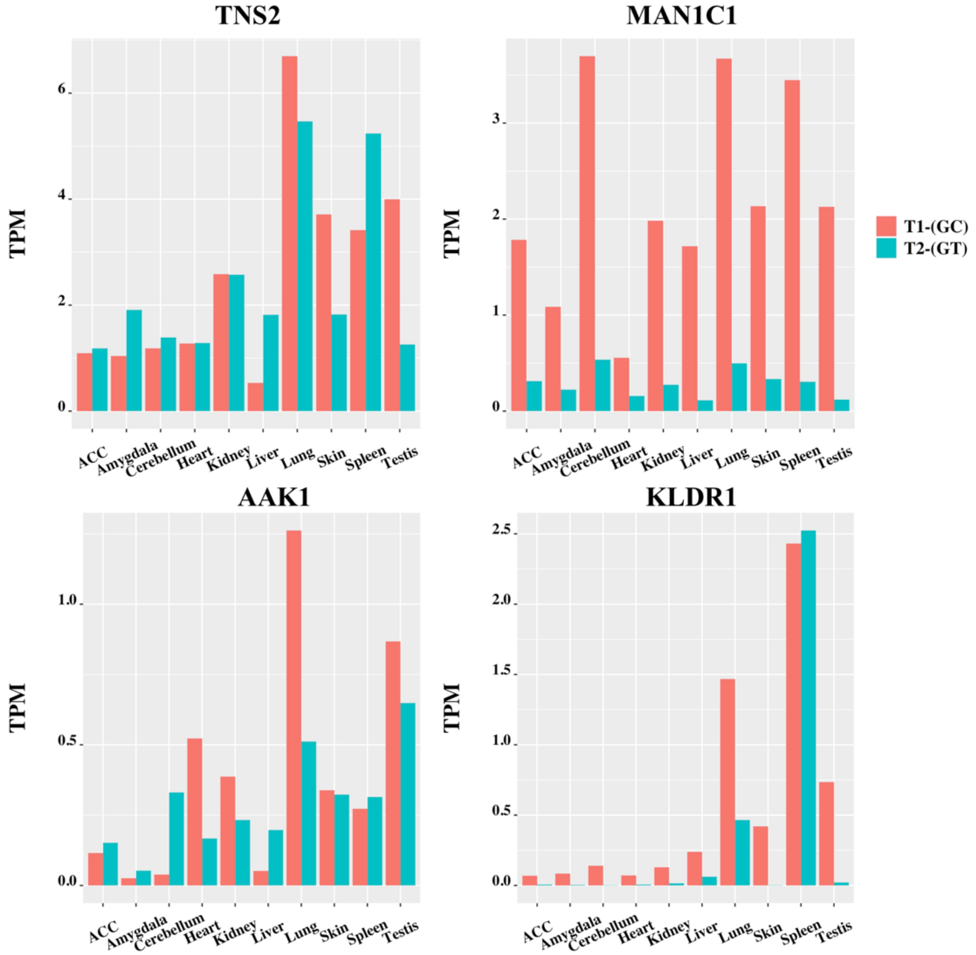


**Figure 38. Bar graph showing the expression levels of *TNS2*, *MAN1C1*, *AAK*1 and *KLDR1* genes.** The expression values of the transcripts involved in the wobble splicing mechanism in these genes is reported distinguishing between GC and GT containing transcripts.

Taken together, these data suggest that wobble-splicing transcripts have a tissue-specific expression and show expression regulation in a tissue-specific manner.

## 4.6. Sub-cellular localization

In order to study whether the wobble splicing mechanism could affect the subcellular localization of the transcripts, we performed a differential alternative splicing analysis for transcripts obtained from the nuclear and cytoplasmic compartments separately. We analyzed 38 nucleus and cytosol RNA-Seq samples obtained from 12 cell lines (A549, GM12878, GM12879 HeLa-S3, HepG2, HUVEC, h1-hESC, IMR90, K562, MCF7, NHEK, and SK-N-SH) provided by the ENCODE project [197]. The differential alternative splicing analysis for the transcripts involved in wobble splicing was performed using the SUPPA2 tool by evaluating the dPSI values of wobble splicing introns between the 2 different conditions (nucleus or cytosol). We identified 9 lncRNA genes 53 PCGs having a wobble splicing events and showing a differential alternative splicing in the wobble splicing transcripts (Appendix1).

Among lncRNAs, 4 genes (*LINC00265, AC138035.1, AL354733.2, and GUSBP11*) showed the differential alternative splicing between transcripts having a GC and a GT intron. In *AC138035.1, AL354733.2*, and *GUSBP*, by analyzing the Nuclear/Cytoplasm (Nuc/Cyto) dPSI ratios, the GC transcript appeared to be significantly more in the nucleus than the cytoplasm while the opposite was true for LINC00265 in which the GC transcript appeared more cytoplasmic.

Among PCGs, 14 genes (*FXN, ZNF559, APEX1, DIDO1, PKM, RNF220, NDUFAF6, HEXD, RAD52, APOPT1, FAM104B, ISCA2, ACADSB, U2AF2*) showed the differential alternative splicing between transcripts having a GC and a GT intron. In the *APEX1, DIDO1, RNF220, RAD52* and *U2AF* genes, the GC transcript appeared more significantly more nuclear while in *FXN, ZNF559, PKM, NDUFAF6, HEXD, APOPT1, FAM104B, ISCA2*, and *ACADSB*, the GC transcript was more cytoplasmic. Among the GC containing genes, 9 genes had the wobble splicing event in the 5'UTR. For example, in the *APEX1* (apurinic/apyrimidinic endodeoxyribonuclease 1) gene, for which 3 distinct isoforms were annotated by GENCODE and only differing in the 5'UTR, a wobble splicing event was identified between a GC/GT pair of alternative splice sites interspaced by a $\Delta = 10$ nucleotides and found in the 5' UTR. Expression data from the different studied tissues showed that all isoforms are constitutively expressed and can be considered housekeeping. The first isoform, having the GT motif, can be indicated as the major isoform, since it is the most represented (TPM $\approx$15), while the second isoform, containing the GC motif, was reported to be significantly less expressed (TPM $\approx$ 5). The major isoform, together with the third isoform not participating in wobbling appeared to be cytoplasmic while the GC containing transcript appeared to be significantly more nuclear. Unexpectedly, despite both splice sites involved in wobbling were rather weak, the GC junction was slightly stronger (WMM = 1.38) than its GT counterpart (WMM = 0.14).

These data suggest that wobble splicing could have a regulatory role in the subcellular localization of the expressed transcripts

### 4.7. Regulatory effects of wobble splicing on alternative isoforms

In order to gain further insights regarding the functionality of wobbling at donor splice sites, the impact of 5' WS events on human mRNAs and corresponding protein isoforms was investigated. Since lncRNAs lack coding potential and are generally scarcely annotated, the effects of wobble splicing in this category was not analyzed. For this analysis, all introns involved in 5' WS and containing donor GT, GC or AT motifs were considered. All events (1090) were classified according to the type of effect occurring on the corresponding transcripts. Wobble splicing events were categorized according to the following classes: insertion, frameshift, 5' UTR, 5' UTR/ins, 3' UTR, 3' UTR/ins and alternative N-terminus (Figure 39).

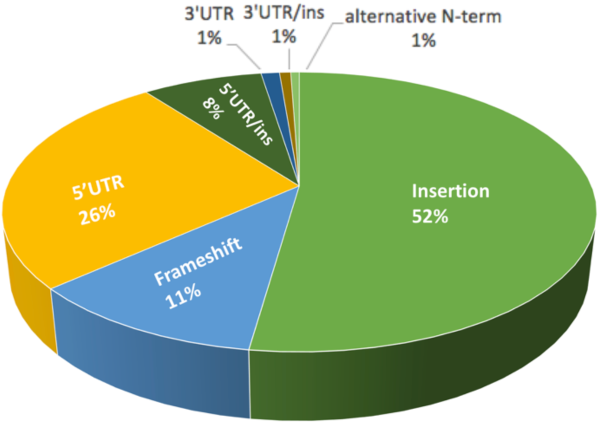


**Figure 39. Pie short of the different categories of wobble splicing effects on the encoded protein isoforms.** The effect of wobble splicing on protein-coding isoforms was classified as: insertion, frameshift, 5'UTR, 3'UTR, 5'UTR/ins, 3'UTR/ins, and alternative N terminal.

The insertion group (ins) contained introns in which wobble splicing was affecting the open reading frame without causing frameshifts. Depending on the Δ value of wobble splicing, 3 or multiple of 3 nucleotides were inserted in the proximal isoform with respect to the distal one, leading to the incorporation of additional codons. Of the 1090 considered events, the "ins" group had 569 instances (52.2%). The frameshift category (fs) included all those insertion events affecting ORFs and altering the preexisting reading frame; in total, 121 WS events were reported in the "fs" group, corresponding to 11.1% of the total. The two categories 5' UTR and 3' UTR were similar: all the contained introns were involved in wobbling exclusively within untranslated regions, having no direct impact on the ORFs. The 5' UTR group contained 287 events (26.4%), while only 13 events (1.2%) were reported in the 3' UTR group. Instead, those events in which the two junctions participating in WS were simultaneously present in UTRs and ORFs were put into the 5' UTR/ins and 3' UTR/ins groups; the former was composed by 85 events (7.8%), while the latter had just 8 events (0.7%). Only 6 WS events (0.6%) were considered in the alternative N-terminus class; this category included those event that, due to wobbling occurring at the very beginning of ORFs, resulted in the production of different N-termini. Insertion, frameshift and 5' UTR categories were the most abundant, accounting for approximately 90% of all WS events. As most wobble splicing events were classified in the insertions group (52.2%), this shows that wobble splicing is frequently associated with little changes, typically involving the variation of very few amino acids. Instead, major changes in the protein sequences were caused by frameshift effects; however, this category was less frequent, weighting for about 11.1% of the total. This last group was characterized by a larger number of stop codons insertions as well.

To better understand the consequences of the previously reported sequence variations, it was necessary to observe whether such WS events were affecting fundamental protein domains or other regulatory elements present in untranslated regions. For those events occurring inside ORFs, investigations with “UniProt” and “InterPro” were carried out to assess whether the affected positions were included in functional domains. Although the identification of all domains could be ambiguous due to the lack of complete annotations, approximately 400 events were predicted to occur inside annotated domains. One good example of wobbling-mediated effects occurring within protein domains was found in the *SPIB* (Spi-B transcription factor) gene, a transcription factor for which 4 distinct isoforms are known; this event was classified in the “fs” category (Figure 40).

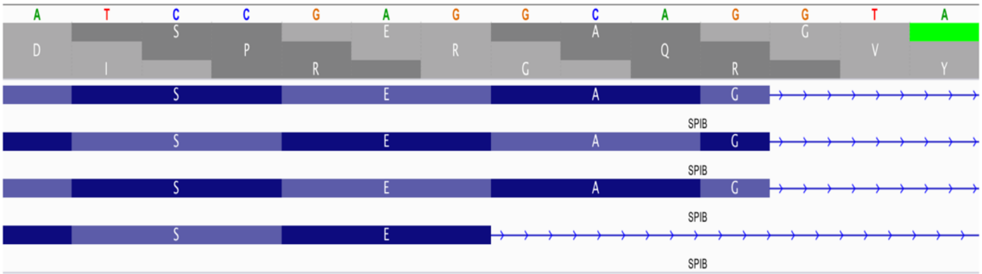


**Figure 40. IGV graphical representation of the wobble splicing event in *SPIB* transcripts.** A wobble splicing event occur in the exon4-exon5 boundary of *SPIB* transcripts. The wobble splicing event can result in three isoforms containing a GT splice site and a fourth isoform containing a GC splice site. The GC/GT splice sites are separated by Δ = 4 nucleotides

In this case, WS involves a GC/GT pair of splice sites, found at the end of the exon number five of the GC isoform, and separated by a Δ of 4 nucleotides. Interestingly, this WS event was reported to be conserved in mouse as well. The two splice sites were showing high divergence in their strength; the GC splice site was characterized by weak strength (WMM = 1.7), while the GT splice site was much stronger (WMM = 8.9). By investigation the expression data of the *SPIB* isoforms, we observed that in the majority of tissues, the GT-containing transcripts were expressed while GC-containing transcript showed a very low level of expression except in the spleen in which the GC-containing transcript was the predominantly transcribed one. (Figure 41).

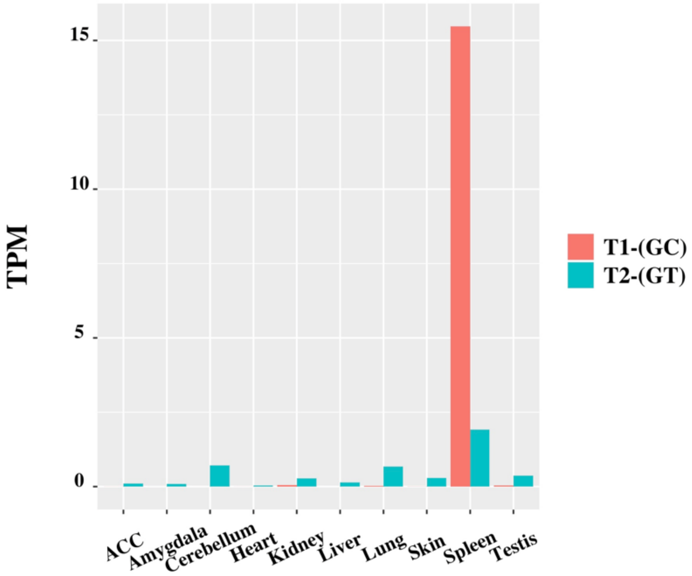
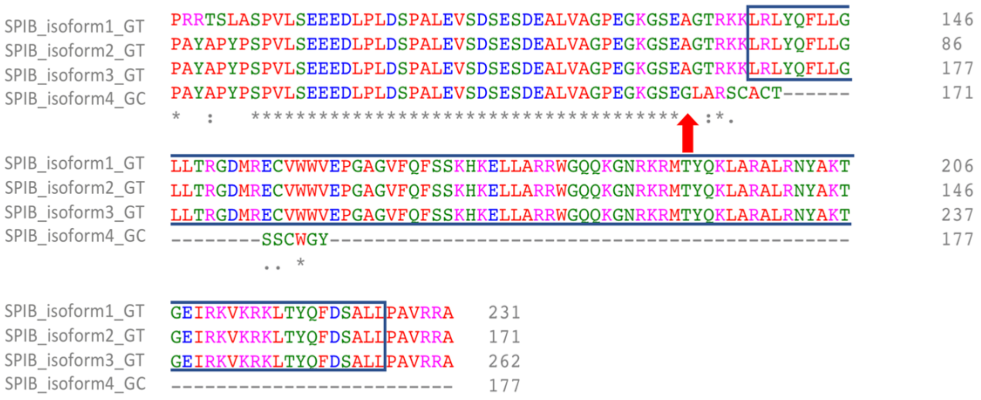


**Figure 41. Bar graph showing the expression levels of *SPIB* isoforms in different tissues.**

Regarding the effect of WS, when the GC splice site is selected, a premature stop codon is introduced in the ORF due to frameshift, producing a truncated protein (Figure 42).



```
SPIB_isoform1_GT    PRRTSLASPVLSEEEDLPLDSPALEVSDSESDEALVAGPEGKGSEAGTRKKLRLYQFLLG    146
SPIB_isoform2_GT    PAYAPYPSPVLSEEEDLPLDSPALEVSDSESDEALVAGPEGKGSEAGTRKKLRLYQFLLG    86
SPIB_isoform3_GT    PAYAPYPSPVLSEEEDLPLDSPALEVSDSESDEALVAGPEGKGSEAGTRKKLRLYQFLLG    177
SPIB_isoform4_GC    PAYAPYPSPVLSEEEDLPLDSPALEVSDSESDEALVAGPEGKGSEGLARSCACT------    171
                    *  :   **************************************  :*.

SPIB_isoform1_GT    LLTRGDMRECVWWVEPGAGVFQFSSKHKELLARRWGQQKGNRKRMTYQKLARALRNYAKT    206
SPIB_isoform2_GT    LLTRGDMRECVWWVEPGAGVFQFSSKHKELLARRWGQQKGNRKRMTYQKLARALRNYAKT    146
SPIB_isoform3_GT    LLTRGDMRECVWWVEPGAGVFQFSSKHKELLARRWGQQKGNRKRMTYQKLARALRNYAKT    237
SPIB_isoform4_GC    --------SSCWGY----------------------------------------------    177
                            .. *

SPIB_isoform1_GT    GEIRKVKRKLTYQFDSALLPAVRRA    231
SPIB_isoform2_GT    GEIRKVKRKLTYQFDSALLPAVRRA    171
SPIB_isoform3_GT    GEIRKVKRKLTYQFDSALLPAVRRA    262
SPIB_isoform4_GC    -------------------------    177
```

**Figure 42. A graphical representation of multiple sequence alignment of the *SPIB* protein isoforms.** The red arrow indicates the position in which the wobble splicing event occurred. The blue box indicates the ETS domain which is present in the first three protein isoforms but lack in the fourth isoform.

The GC protein isoform was reported to lack an EST domain, which was instead found in all the other isoforms having the GT splice site.

For what concerns coding regions, in various cases, frameshifts that would be induced by WS events (Δ distance ≠ 3 or multiples), were prevented by other types of co-occurring AS events at subsequent acceptor sites, including compensatory 3' WS events with matching Δ distance. Indeed, the existence of mechanisms that could revert the effects caused by wobbling further increases the complexity of this regulatory network. Instead, if no compensation events occur, the introduction of frameshifts can have dramatic effects on the resulting amino acid sequence, like in the case of SPIB, where it determines the introduction of a premature stop codon.

The events annotated in the "ins" group can also alter the protein sequence, although the evaluation of the short in-frame insertions impact was definitely harder. For what regards events affecting untranslated regions, although not directly associated with coding sequences changes, these could still play important roles, for example by influencing important regulatory motifs or by altering the stability of the spliced transcript such as in the *APEX1* gene.

The simultaneous presence of WS events at both donor and acceptor splice sites can lead to even more sophisticated scenarios. In general, a lot of different consequences may arise due to wobbling, and even events belonging to the same category may show dissimilar and complex features. Undeniably, these events are very complex and could play multiple regulatory roles.

# V. Discussion and Conclusions

In this work, different bioinformatics approaches were deployed for the characterization and the genome-wide comparison of the genomic and splicing features of lncRNAs in comparison to PCGs from recent gene annotations in the human and mouse genome. While the regulatory mechanisms exerted by lncRNAs on gene expression regulation is extensively studied, our understanding of how lncRNA gene expression is regulated still remains relatively vague. Being based on GENCODE releases 27 and M16, our analysis considered a higher number of genes with respect to previous studies [35][46] and it was strengthened by the comparison between the human and mouse species. Our analysis highlighted important differences between lncRNAs and PCGs and focused on a particular type of introns flanked by GC-AG splice sites that appeared to be more enriched in lncRNAs and to have a putative regulatory role.

The characterization of the genomic features revealed differences in the genetic architecture between lncRNAs and PCGs both in human and mouse. We found that lncRNAs were shorter than protein-coding ones in both species in agreement with previous studies [35][46][252]. This appeared to be due to the lower number of exons in lncRNAs and their shorter length. It didn't escape our mind that our results could be biased by the incomplete annotation of lncRNAs in the dataset we used. As previously suggested by the studies of Lagarde and colleagues [201][202], lncRNAs low expression level and high tissue specificity hinders their complete characterization. Thus, the shorter length and the limited number of exons in lncRNA genes might be attributed to their incomplete annotations. However, our results did not seem to be affected by this bias since in this study we exploited a more recent and complete release from GENCODE, whose annotation was based on both a stronger experimental and computational evidence [97]. Moreover, to rule out the possibility of incomplete annotations by GENCODE, we repeated our analysis in other different datasets. Our results were confirmed in six more lncRNAs annotation datasets: FANTOM5, NONCODEv5, BIGTranscriptome, MiTranscriptome, LNCipedia, and LncBook. In particular, the FANTOM CAT dataset of robust lncRNA annotations explicitly provide accurate annotations of transcripts' TSS and 5′ ends through the Cap Analyses of Gene Expression (CAGE) protocol, and the BIGTranscriptome dataset employ both CAGE and poly(A)-position profiling by sequencing (3P-seq) to assess 5′ and 3′ end completeness. This suggests that it is unlikely that our results are subjected or strongly affected by the bias of incompleteness in gene annotations.

The decrease in length of lncRNAs (i.e. shorter first and last exons) appeared to affect the regions of the gene mainly involved in gene expression regulation. First exons length was reported to be related to gene transcription efficiency as they harbor transcription factors binding sites and regulatory sequences such as enhancers. It was reported that short first exons could enhance transcriptional accuracy as they harbor a more concerted assembly of transcription factors near transcription start sites [253]. In addition, due to the presence of 3′UTR sequences, last exons tend to be longer than first and inner exons. These sequences are important for the regulation of numerous aspects such as nuclear export, cytoplasmic localization, stability, and translational efficiency [254][255]. Therefore, our results hints toward a difference in the regulatory potential contained in the first and last exons of lncRNAs. Moreover, lncRNAs did not show a significant difference between the first and inner introns unlike PCGs in which first introns tend to be longer. Taken together, our data suggested that the difference in gene architecture between lncRNAs and PCGs could denote their involvement in different mechanisms of genomic control and gene expression regulation.

The characterization of splicing features showed significant differences in splice junction usage between lncRNAs and PCGs. An enrichment of introns harboring GC-AG splice sites in lncRNAs was observed in both human and mouse. GC-AG splice sites are commonly considered as a non-canonical variant of the major U2-type GT-AG splice junctions, accounting for 0.865 and 0.817% in human and mouse genomes respectively [121][256]. In agreement with previous reports, we observed the same frequency of GC-AG introns in both species when considering only PCGs (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, GC-AG introns showed a preferential localization in the transcripts. 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. GC-AG first introns in human accounted for 4.2% and 1.2% of total first introns of lncRNAs and PCGs respectively. The same trend was observed in mouse in which a higher ratio of GC-AG splice sites was found in the first intron in both lncRNAs and PCGs accounting for 2.4% and 1.2% respectively.

The enrichment of GC-AG introns in lncRNAs and their preferential position in the first intron did not appear to be driven by a mis-annotation bias in lncRNAs. The enrichment of GC-AG introns was confirmed in the six other studied datasets despite the variation in their total lncRNAs gene annotations. Furthermore, the same enrichment was also assessed in the lower organisms *D. melanogaster* and *C.*

*elegans*, despite this analysis could not be conclusive due to their incomplete annotations and limited number. The significant increase of GC-AG introns in lncRNAs, together with their non-random distribution along the gene, led us to hypothesize that they may represent unique regulatory elements. Moreover, the preferential localization of GC-AG splice sites in the first intron provided a clear indication of their role in gene expression 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 [253][257], 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 [258]. 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 suggesting that the non-canonical GC 5′ss could in some way perturb this mechanism of action [259][260][261].

GC-AG introns showed distinctive splicing features in comparison with GT-AG introns, in particular when located in the first intron of lncRNAs. Introns harboring GC-AG splice sites appeared significantly shorter than GT-AG introns, in both lncRNA and PCGs. This trend was more evident in human GC-AG first introns being significantly shorter than GT-AG first introns. In addition to their shorter length, GC-AG splice sites appeared significantly weaker than GT-AG ones. The reduction in strength of 5’ss of GC-AG introns was expected due to the mismatch at position +2 with the U1 snRNA consensus. Nevertheless, the reduction of 5′ss strength was more prominent in GC splice sites of lncRNAs rather than in PCGs and it was more evident in the first intron rather than in inner ones. Similar results were obtained for 3′ss, whose average weight-matrix scores for GC-AG introns appeared significantly lower compared to GT-AG junction, especially when located in lncRNAs first introns. Interestingly, the Spearman correlation test demonstrated a positive correlation among intron length and 5′/3′ss strength for the first intron of lncRNAs, thus implying the enrichment of short and very weak first introns in this class of molecules.

A preliminary analysis of RNA-seq data of 10 different human tissues from the GTEx project, allowed us to highlight a putative effect of GC-AG introns on gene expression profiles. Indeed, the overall expression of GC-AG introns containing transcripts appeared lower with respect to GT-AG ones thus suggesting they may have a reduction effect on gene expression. More interestingly, our data suggested that GC-AG introns located as first behave differently as they demonstrated higher level of expression with respect of transcripts containing an inner GC-AG intron thus underlining their peculiar regulatory role depending on the position. Despite we are aware that these data must be taken with caution as they may be biased in many ways, they represent a first experimental evidence of the effect of GC-AG introns at gene expression level.

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 lncRNAs in human (in mouse: about 8% of PCGs and 4% of lncRNAs). The relevance of GC-AG-containing genes emerged also from the analysis of their conservation: about 50% of GC-AG containing PCGs resulted conserved between human and mouse which could also be related to late intron gain events. Furthermore, in the majority of conserved PCGs (75%), the ordinal position of GC-AG introns was also conserved. As 25% of GC-AG introns do not have the same ordinal position, this could also argue about recent intron gain occurrence. 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 still could 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 PCGs 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 harboring a GC-AG intron 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, reasonably needing a concerted regulation.

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 [262][263]. In agreement with previously reported data

[244], our analysis at the gene level confirmed that GC-AG containing genes were more prone to alternative splicing than genes harboring GT-AG introns. Churbanov and colleagues [164] 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 favoring alternative splicing. Moreover, GC-AG introns were found to be strongly overrepresented in recent intron gain events occurring in segments associated with repetitive sequences that are highly alternatively spliced [264].

Alternative splicing is essential for the isoforms diversification and has been suggested to be mechanistically linked to organism complexity [17]. As GC-AG introns appeared highly prone to undergo alternative splicing, this provide a further evidence that their accumulation could be related to a regulatory role in higher organisms. Moreover, GC-AG introns appeared to be enriched in a particular type of alternative donor splice sites that are in tandem termed wobble splicing. Our analysis showed that this splicing mechanism appeared to be a rare event in the human and mouse genome occurring in a relatively small number of genes (3.5% of total genes in human and 2.5% of total genes in mouse). While one cannot rule out that a number of identified wobble splicing events could be attributed to mis-annotation or mis-alignment errors, this was unlikely in our analysis as our results appeared to be confirmed in both human and mouse and validated in another independent dataset (NONCODE) considering higher number of annotated genes. While wobble splicing could be mainly influenced by the coding pressure in PCGs as to maintain the reading frame avoiding drastic frameshifts in the coding sequence and the activation of the NMD pathway, this is not the case in lncRNA transcripts suggesting a different mechanism regulating the wobble splicing. In the majority of protein-coding transcripts, wobble splicing resulted in small changes in the encoded protein or between the alternative splicing isoforms. This could suggest that wobble splicing could have a tight regulatory role required at certain conditions. Moreover, this could be compatible with the selection and evolution of new GC splice sites arising from cryptic splice sites close to the dominant one thus promoting the evolution of novel functional isoforms across species. Indeed, the analysis of the splice sites scores of tandem splice sites involved in wobble splicing, whether GC or GT, showed a relatively lower strength than the total donor splice sites. This suggests a competitive binding of the U1 snRNP to the tandem splice sites that could select the proximal or the distal splice site. This was observed in the *SPIB* gene in which the GC splice site had a significantly weaker strength than the GT one, yet more spliced in spleen tissue. It is still not clear whether there are specific splicing factors aiding in the regulation of wobble splicing at the donor splice sites. In the study of Dou and colleagues [209], it was shown that the U1 snRNA binding at the 5’ss is the sole responsible for the 5’ tandem splice sites usage. In another study by Hiller and colleagues [206], it was reported that the U6 snRNA rather than U1 determines the splice site choice in wobble splicing.

The investigation of expression of transcripts involved in wobble splicing using RNA-Seq data from 10 different tissues showed that although a large proportion of transcripts involved in wobble splicing are housekeeping, a number of transcripts appeared to have tissue-specific expression. Moreover, in many cases, transcripts involved in wobble splicing showed regulation in their level of expression in a tissue-specific manner. This is in accordance with a previous study exploring the tissue specificity of wobble splicing events at acceptor splice sites NAGNAG. We have identified several examples in which we observed a tissue-specific splicing patterns and selection between GC and GT splice sites. How this regulation across various tissues is achieved still remains a challenge. Moreover, several examples showed the putative role of wobble splicing between GC and GT tandem splice sites in the transcript subcellular localization suggesting a further level of regulation by the wobble splicing mechanism.

The functional relevance of GC-AG introns was already illustrated in few previous studies. In the *ING4* (inhibitor of growth family member 4) gene, a wobble splicing event in which 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 [265]. In the study of Farrer and colleagues [266] it was demonstrated that the weak GC 5′ss located in intron 10 of the *let-2* (Collagen alpha-2(IV) chain) gene in *C. elegans* was essential for developmentally regulated alternative splicing, and that its substitution with a stronger GT splice site suppressed the alternative splicing regulation occurring during embryos development. Palaniswamy and colleagues [267] showed that a single nucleotide polymorphism converting a donor splice site from GT to GC, present with varying frequencies in different mouse strains, was responsible for an alternative splicing event affecting the length and the translational efficiency of the *Gli1* (GLI-Kruppel family member) gene in mouse. Moreover, for the *PRDM* (PR/SET domain) gene family in human [268] and for the starch synthase (SS) gene family in rice [269], the activation of a GC donor splice site was reported to be related to the evolution and the diversification of both gene families.

Organism complexity does not correlate with genome size or gene content, but it is instead more related to the level of gene expression regulation. Despite the fact that the number of PCGs is similar in evolutionary distant species, a higher level of gene regulation is thought to ensure the development of more sophisticated capabilities of higher organisms. The amount of alternative splicing, which permit the production of various isoforms starting from a smaller number of genes [18][270], and the amount of transcribed non-coding DNA resulting in the production of a large collection of ncRNAs mainly involved in the regulation of gene expression, is known to be positively correlated with eukaryotic complexity [3][271]. 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 [256], hence supporting the idea 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 a favorable evolutionary driven mechanism, putatively due to the increased number of alternative splicing events occurring at weak GC-AG introns [162][164].

In conclusion, our data suggested that GC-AG introns represent new regulatory elements with a preferential location in the first intron and primarily associated with lncRNAs. The increased occurrence of GC-AG splice sites in higher organisms suggested that they could contribute to the evolution of complexity, adding a new level in the regulation of gene expression. The regulatory role of GC-AG introns remains to be further investigated despite preliminary evidence suggested that they could favor alternative splicing and in particular wobble splicing. The elucidation of their mechanisms of action of could contribute to a deeper and 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.

# Project II: Identification of Sexual Dimorphism in Long Non-coding RNAs Genes Expression

During my PhD, I carried out a second research project developed in our laboratory, aimed to the characterization of the differential expression of lncRNAs between human males and females and described in the following section.

# I. Introduction

Despite females and males share highly similar genomes, there are numerous sex specific traits in phenotype, physiology, and pathology [272]. Sexual dimorphism, established by primary and secondary sexual hallmarks that differ between males and females, represents a key concept in the comprehension of molecular processes that guide several sex-specific physiological and pathological mechanisms [273]. Given the complexity of gene networks, it is unlikely that only genes on sexual chromosome can account for all the phenotypic differences between the sexes. Sexually dimorphic traits can be influenced by sex chromosome genes or by sex hormones but may also extend to alterations in autosomal genes regulation. Genes involved in sexual dimorphism can be classified as: (i) sex-specific genes that are expressed exclusively in males or in females and (ii) sex-biased genes those that are expressed at a higher level in either males (male-biased) or females (female-biased). Up to now very little is known about how differences in the expression levels of genes between the two sexes might affects the final phenotypic traits.

Sex divergence in health and disease is partially driven by the inherent inequality of the sex chromosomes, such as the effects of the expression of Y chromosome genes, differences in dosage of X chromosome genes, and epigenetic effects [274]. The sex-specific expression of Y-linked transcriptional regulator genes in males can interfere with rate of expression of the autosomal genes. Historically, females have been excluded for years from biomedical studies, due to the common accepted theory that hormonal cycles could be confounding factors [275][276]. What pushed scientists during the years to deeply investigate sex-biases is the different incidence of some diseases between the two sexes. These findings come from the fact that, besides the reproductive system, many other tissues show sex bias, such as the circulatory, immune or nervous system. The consequence of such diversity reflects in organ functions, disease susceptibility and drug response [277]. These sex differences have been variously attributed to hormones, sex chromosomes, differences in behavior, and differences in environmental exposures [278], but the mechanisms and underlying biology of the sex differences remain largely unknown.

Disease features such as prevalence, progression, age of onset, and response to treatment often differ by sex. Recently, the NIH has officially recognized the importance of sex-based research by suggesting to explicitly include sex as a biological variable in each research and clinical study [279]. As the severity and prevalence of many diseases are known to vary greatly between the sexes, tissue-specific sex-biased gene expression may underpin these and other sexually dimorphic traits. For example, brain is well known to be a highly sexually dimorphic tissue and sex differences in human brain structure, neurochemistry, behavior and susceptibility to neurodegenerative and neuropsychiatric disease have all been reported. Sex differences have been identified in a number of brain diseases such as schizophrenia, anxiety depressive disorders, multiple sclerosis, epilepsy and Parkinson's disease [273]. The differences in the incidence, prevalence, severity and response to treatment between the sexes can complicate the understanding of the biological processes guiding the progression of diseases. Important sex-by-treatment interactions can be masked when investigators only analyze data sets in which the two sexes are pooled (Figure 43) [280]. Disaggregating the data and considering the sexes separately can unmask important sex influences in studied subjects.

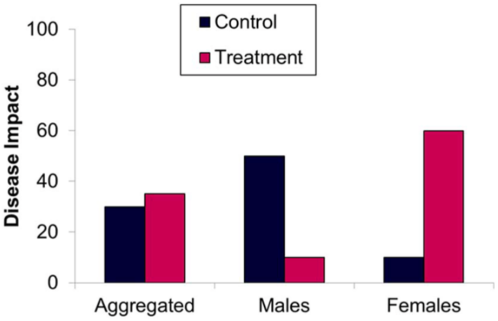


**Figure 43. Bar graph showing the disease treatment impact among patients.** The patients were aggregated or divided between males and females (adapted from [280]).

Hence understanding the molecular basis behind the sexual dimorphism will allow to address people to the most effective and safe treatments and to move towards the era of personalized medicine.

# II. Aims of the Research

Sexual dimorphism in gene expression represents an important concept for the understanding of the different physiological and pathological mechanisms that differs between males from females. The identification of sex-biased genes and the comprehension of their roles could allow the introduction of personalized medicine based on different sex-related biomarkers.

Sexual dimorphism occurs in the majority of human tissues, but the brain still remains the organ that manifests it at a greater level. Studying sex-biased genes in the brain may help in the disentangling of complex processes that could eventually contribute to the etiology of brain disorders.

Until recent years, most studies have addressed sexual dimorphism by using a small number of samples in array experiments and by focusing only on PCGs while very few studies have identified the sex-biased expression of lncRNAs. To further our understanding of sex differences in regulation of the expression of lncRNA and protein-coding genes, we performed a differential expression analysis of sex biased genes in the brain and whole blood tissues. Because the impaired tissue is usually unknown for many diseases and disorders, analysis of this diverse tissue set can serve as a powerful resource for investigations into the basis of sex-differentiated phenotypes.

The Genotype-Tissue Expression (GTEx) (The GTEx Consortium, 2015) project provides an opportunity to investigate the prevalence and genetic mechanisms of sex differences in transcriptomes and to identify how sex and genetics interact to influence complex traits and disease. By exploiting RNA sequencing samples provided by the GTEx dataset, in this study we aimed to:

- Identify differences in the expression profile of lncRNA genes in different brain tissues in addition to the whole blood tissue using RNA-seq data.
- Describe and characterize the differentially expressed genes, identifying the sex specific and sex-biased genes in the studied tissues.
- Characterize the putative functions of the differentially expressed genes and their possible involvement in sex-related regulatory pathways or diseases.

# III. Materials and Methods

## 1. Data collection

RNA-Seq samples of healthy individuals were downloaded from the dbGap database (approved project #20055: Investigation of sex-biased long non-coding RNAs and protein-coding RNAs expression) using the GTEx dataset according to the specified guidelines. Data were collected from 4 different tissues: (i) cerebellum (32 females and 32 males), (ii) amygdala (19 females and 19 males), (iii) anterior cingulate cortex (ACC) (24 females and 24 males), and (iv) whole blood tissue (21 females and 21 males) for a total of 192 samples.
All the selected samples were paired-end (2x76 bp) sequenced using the Illumina HiSeq 2000 apparatus of a sequencing coverage between 60 and 80 million reads, thus sharing the same sequencing features. Equal number of males and females samples were selected in each tissue in order to avoid profound effects or biases in subsequent analyses.

The lists of lncRNAs and PCGs 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). PCGs were selected from the basic annotation when both gene and transcript were indicated as “protein_coding”.

## 2. Gene expression analysis pipeline

### 2.1. Computational resources

The RNA-seq data analysis was performed exploiting the Amazon Web Services (AWS) (http://aws.amazon.com) resources of cloud-computing. This service provided us the computational power and the space for quality control, reads alignment, and reads quantification. A logically isolated section of the AWS Cloud, called Amazon Virtual Private Cloud (Amazon VPC) allowed us to launch resources in a virtual network, having the complete control over the virtual networking environment. AWS Elastic Compute Cloud (EC2) is a web service that provides compute capacity, designed to make web-scale cloud computing. Among the five different EC2 instance types, we selected the M4 instance class, which provides a good balance of compute, memory and network resources. After setting the optimal conditions, in order to have the tools to analyze RNA-Seq data, the *MolBioCloud* Desktop Service (https://molbiocloud.com) was used. It is an Amazon Machine Image (AMI) that contains hundreds of open source computational tools for molecular biology and runs in association with the AWS EC2 instances. We used the MolBioCloud service since it provides all the software and the technical support for our RNA-seq analysis pipeline.

### 2.2. Sequencing files retrieval

The SRA files for each sample were downloaded using the "*prefetch*" command from the SRA Toolkit tool as previously described. The separate FASTQ files of the paired-end reads for each sample were extracted from the SRA files using the "*fastq-dump*" command.

### 2.3. Quality check

The quality check of the raw reads in each FASTQ file was performed using FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The fastqc script was run in the Java environment using the default parameters. Files containing raw reads, in the FASTQ format, were provided to the tools. As output the tools produced several plots and tables measuring different parameters such as the GC content, N base content, overrepresented sequences, distribution of sequence length. One of the main parameter evaluated is represented by the Phred, a score related to the base-calling error probabilities and usually ranging from 0 to 40. All the tested parameters suggested a high quality of reads, so that no additional filters were applied.

## 2.4. Alignment analysis

For the alignment of the reads to the reference genome, the STAR (Spliced Transcripts Alignment to a Reference) software (version 2.5.4b) was used [281]. This tool is ultra-fast in its performance, having at the same time an improvement in alignment sensitivity and precision. For each sample, alignment with STAR was performed using default parameters providing the 2 paired-end FASTQ files and the human reference genome (GRCh38) according to the following command:

```
STAR --genomeDir <dir_name> --runThreadN 8 --readFilesIn <fastq1.fastq.gz> <fastq2.fastq.gz> --readFilesCommand zcat --genomeLoad LoadAndKeep –outFileNamePrefix <name> --outSAMstrandField intronMotif --outFilterIntronMotifs RemoveNoncanonicalUnannotated --outSAMtype BAM SortedByCoordinate --limitBAMsortRAM 20000000000
```

As an output, STAR generated BAM files containing coordinates for each sample and a summary data for the alignment characteristics of mapped reads.

## 2.5. Gene expression quantification

Following the alignment to the reference genome, the mapped reads were quantified using the featureCounts tool [282]. The featureCounts tool performed the quantification against the lncRNAs and protein-coding GTF files separately downloaded from the GENCODE database release 27 according to the following command:

```
featureCounts -p -t exon -g gene_id -a <gtf_file> -o <outout.txt> <bam_file>
```

In each step, we obtained as an output the gene-count table and a summary file of all the percentage of counts assigned to each gene.

### 3. Differential expression analysis

A differential expression analysis (DEA) step was performed in order to get insights to the differences in gene expression in males and females. For this analysis, we used the "DESeq2" (version 1.12.4) R-package [283].
Gene-counts tables from each tissue and for lncRNAs and PCGs were uploaded together with a file containing the list of all the individuals, labelled with the condition. In this case, the condition is sex and DEA was performed female versus males. In this way the log2FoldChange, that is the ratio of the expression in the two conditions, with a value major than 0 indicates the genes that are female biased and minor than 0 are for the genes that are male-biased. Genes were considered differentially expressed having an FDR value < 0.1 and an adjusted p-value < 0.05. After obtaining the output table with the gene expression information, Principal Component Analysis (PCA) was performed to check for the homogeneity among the samples and for presence of any extreme outliers which could be due to mislabeling errors.

### 4. Functional enrichment analysis

A functional enrichment analysis for the differentially expressed protein-coding genes was performed using DAVID [234]. The lists of differentially expressed genes (DEGs) from each tissue were subjected to an enrichment analysis on GO Biological Process, Molecular Function, and Cellular Component terms and KEGG pathway, and filtered applying a statistical significance threshold of 0.05 based on the multiple testing corrected p-values.

The lists of differentially expressed lncRNAs were interrogated for their association with common diseases using the "LncRNADisease 2.0" database [284].

### 5. Data use policy

We only analyzed anonymized samples for which the corresponding donor consent information was available in the GTEx dataset (dbGaP:phs000424.v8.p2) at the time of the analysis. Samples were downloaded from the dbGap database according to the specified guidelines. All of the samples we analyzed were approved for General

Research Use (GRU) and thus have no further limitations outside of those in the NIH model Data Use Certification Agreement.

## 6. Additional tools

All the R packages were used according to the R version 3.6.3.

Graphs and distribution plots were performed using the “ggplot2” (version 2.1.0) R-package (http://ggplot2.org/) through the “*ggplot()*” command.

Data manipulations were performed using the “dplyr” (version 1.0.2) and “tidyr” (version 1.1.1) R-packages.

FASTQ files were loaded on the AWS instances using FileZilla software (version 3.50).

Venn diagrams were plotted using the “Venny” (version 2.1.0) webtool (https://bioinfogp.cnb.csic.es/tools/venny/).

# IV. Results

## 1. Identification of sex-biased genes

For this analysis, we exploited the RNA-Seq samples of healthy human individuals provided by the GTEx database. The analysis was performed on samples from 3 brain tissues: anterior cingulate cortex (ACC), amygdala, and cerebellum in addition to the whole blood tissue. Our analysis employed a large number of samples to increase the statistical power for the detection of differential gene expression as sex differences in many traits are often small and require large sample sizes to be sufficiently powered [273][285] (Table 24).

**Table 24. Summary of demographic details of the individuals sampled**

| Tissue | Male samples | Female samples | Age (years) | Modal Cause of Death |
|---|---|---|---|---|
| **ACC** | 24 | 24 | 20 - 69 | Natural Causes |
| **Amygdala** | 19 | 19 | 20 - 70 | Natural Causes |
| **Cerebellum** | 32 | 32 | 20 - 70 | Unknown |
| **Whole Blood** | 21 | 21 | 20 - 69 | Natural Causes |

The analysis was performed on the samples from each tissue using an RNA-seq data analysis pipeline prepared in our laboratory using cloud computing services and including the following steps: reads quality control, genome mapping, lncRNA and protein-coding gene expression quantification, differential expression analysis for both gene classes and downstream functional and pathway enrichment analyses.

Our analyses showed that each tissue had different numbers of differentially expressed lncRNA and protein-coding genes indicating that sex-biased expression is wide-spread among tissues. The total number of differentially expressed lncRNA genes was 218 DEG in ACC, 220 DEG in amygdala, 81 DEG in cerebellum, and 100

DEG in whole blood and the larger proportion of sex-biased genes in all four tissues were autosomal (Figure 44).

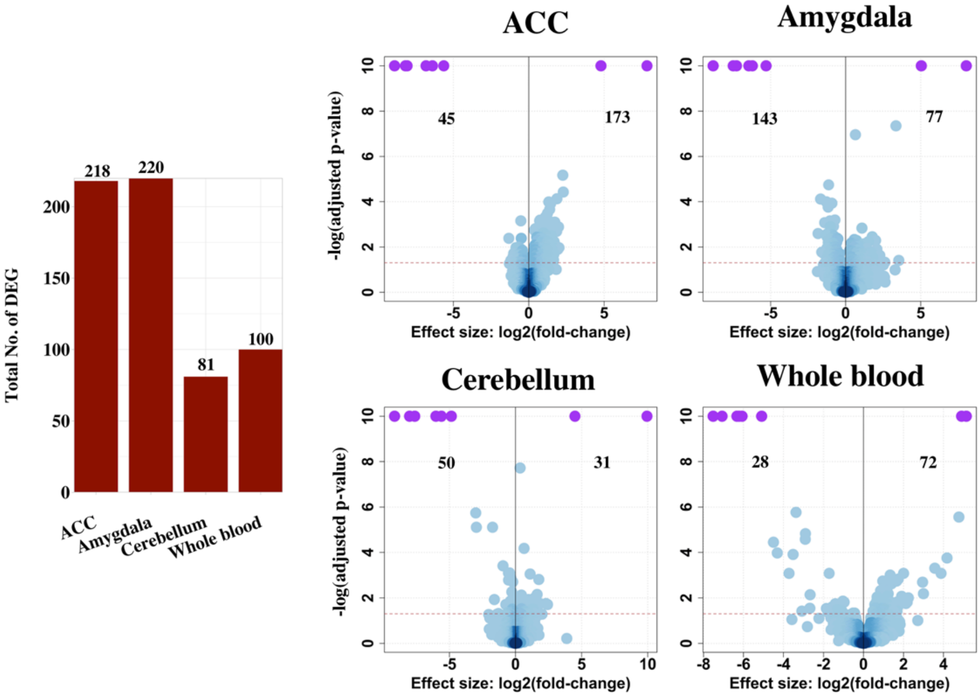


**Figure 44. LncRNAs differentially expressed between males and females.** Right: bar graph showing the number of DEGs in the different tissues. Left: Volcano plot showing statistical significance vs $\log_2$(fold-change) of lncRNAs in different tissues. Control genes (i.e. genes on the Y chromosome, *XIST* and *TSIX*) are shown as purple dots.

Regarding protein-coding genes, 686 DEG resulted in ACC, 1237 DEG in amygdala, 738 in cerebellum, and 907 DEG in whole blood (Figure 45).
As an internal control, it was noted as expected that all lncRNA and protein-coding DEG on the Y chromosomes were male biased in addition to *XIST* and *TSIX* lncRNAs on the X chromosome that were female-biased.

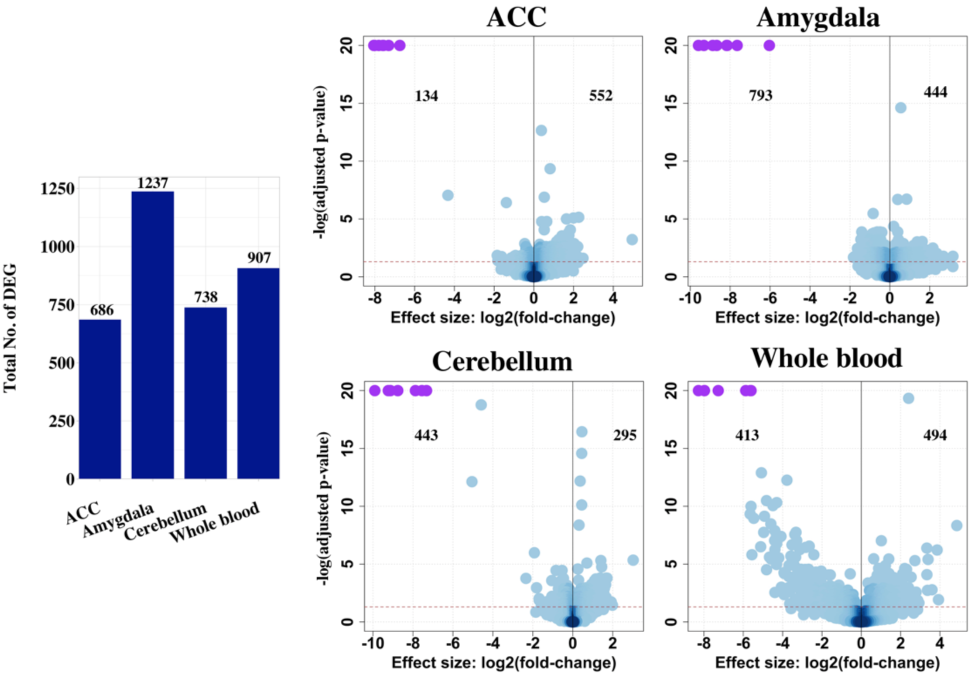


**Figure 45. PCGs differentially expressed between males and females.** Right: bar graph showing the number of DEGs in the different tissues. Left: Volcano plot showing statistical significance vs $\log_2$(fold-change) of PCGs in different tissues. Control genes (i.e. genes on the Y chromosome) are shown as purple dots.

Of the total differentially expressed genes, female-biased genes resulted higher in amygdala and cerebellum accounting for 65% and 61.7% of total lncRNA DEGs respectively and 64.1% and 60% of total protein-coding DEGs respectively. Conversely, we observed a greater proportion of male-biased DEGs in ACC and blood accounting for 79.4% and 72.0% of total lncRNA DEGs respectively and 80.5% and 54.4% of protein-coding DEGs respectively.

Moreover, we also found that apart from genes on the Y chromosome, *XIST* and *TSIX*, each tissue was unique in its proportion of sex-biased lncRNA and protein-coding genes with as many sex-biased genes in one tissue that were not sex-biased in another indicating a high-tissue specificity of these sex-biased lncRNA and protein coding genes (Figure 46).

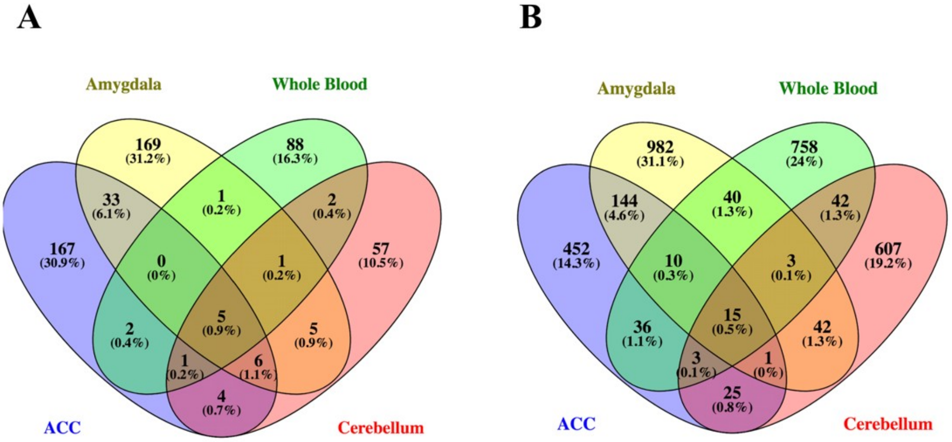


**Figure 46. Venn diagram showing the common DEGs among tissues.** (A) DE lncRNAs and (B) DE PCGs among the four tissues studied.

Only 5 differentially expressed lncRNAs and 15 protein-coding genes were common between the 4 different tissues in which they all appeared to be located on X chromosome. All the common genes on the X chromosome appeared to be female-biased. Being more expressed in females than in males, we could speculate that the differential expression of these genes could be due to the mechanism of escape from X-inactivation. Among brain tissues, 11 lncRNA appeared to be common including 9 genes on the sex chromosomes and 2 autosomal genes (*AC233266.2* and *AC087241.1*) whereas 16 protein-coding genes appeared common between brain tissues including 15 genes on the X chromosome and 1 autosomal gene (*CHI3L2*). Among the 15 genes on the X chromosome, 12 genes appeared to be female-biased while 3 genes appeared to be male-biased. While the female-biased genes could be related to the mechanism of escape of X-inactivation, the male-biased genes on the X chromosome suggest a different mechanism in the regulation of expression of these genes in male individuals than in female ones.

## 2. Functional enrichment analysis

In order to get insights on the functional role of the differentially expressed genes between males and females, a functional enrichment analysis was performed using the DAVID database [234]. Since the function of most of the lncRNA genes is still unknown, the analysis was done taking into consideration only protein-coding genes. In blood tissue, male-biased protein-coding genes were found according to DAVID to be strongly enriched for GO terms related to muscle contraction (Figure 47A).

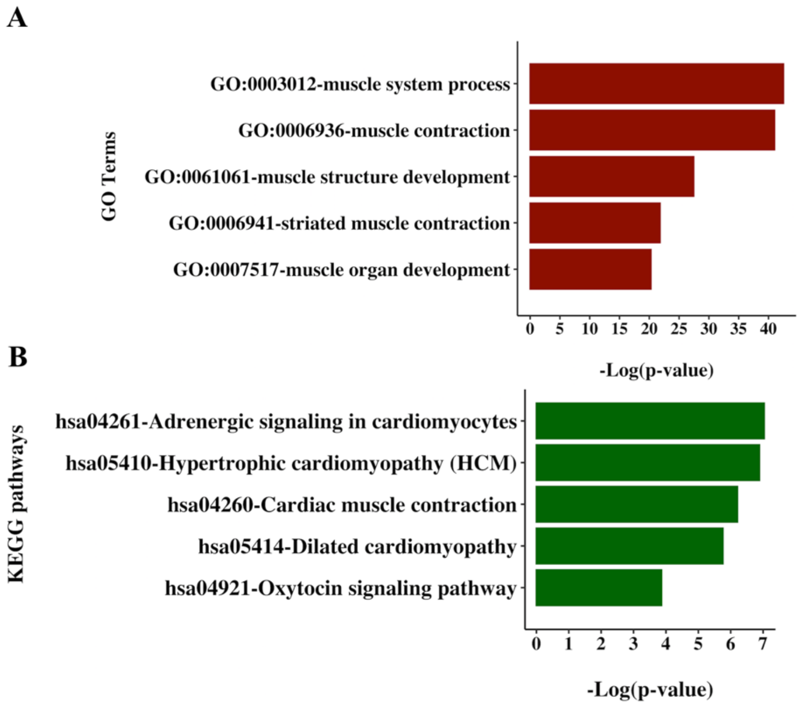


**Figure 47. Bar graphs showing significantly enriched GO and KEGG Pathway terms from the whole blood tissue.** (A) Biological process GO terms and (B) KEGG pathways significantly enriched among DE PCGs

According to KEGG pathways, male-biased genes showed an enrichment for pathways related to cardiac muscle and cardiomyopathy. (Figure 47B). Conversely female biased genes showed no significant enrichment.

Male-biased protein-coding genes from cerebellum appeared to be enriched in biological processes GO-terms related to nervous system development and synaptic signaling (Figure 48A) and male-biased protein-coding genes from amygdala were enriched in GO-terms related to nervous system development and myelination

(Figure 48B). No significant functional terms appeared enriched for female-biased genes from these tissues and both male and female sex-biased genes from ACC didn't show any significant enrichment in both GO and KEGG pathway.

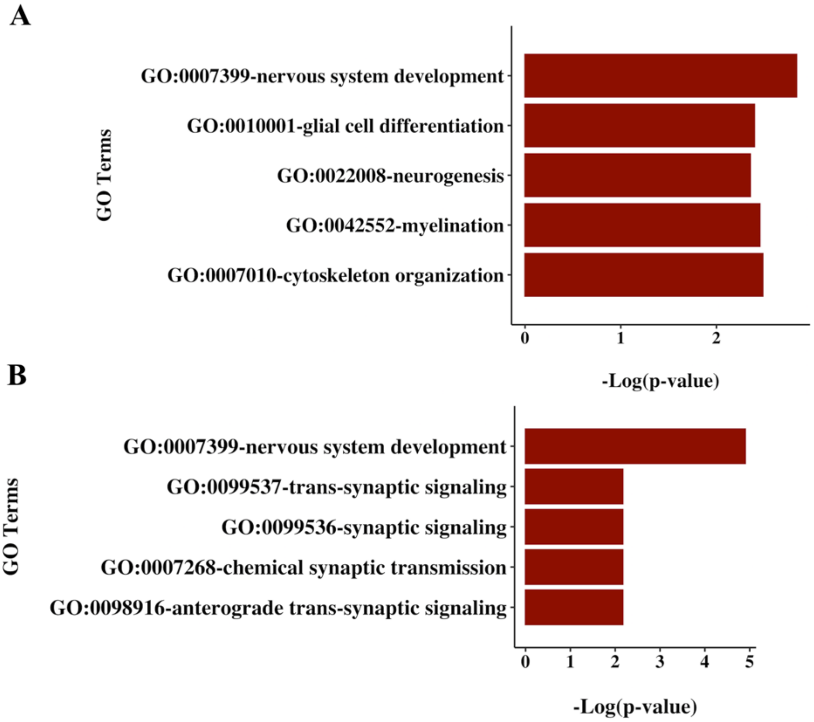


**Figure 48. Bar graphs showing significantly enriched GO terms in cerebellum and amygdala tissues.** Biological process GO terms significantly enriched among male biased differentially expressed PCGs of (A) cerebellum and (B) amygdala tissues.

## 3. Identification of sex-biased lncRNAs associated to diseases

We next analyzed the autosomal differentially expressed lncRNA genes in the different tissues for lncRNA-disease associations using the LncRNADisease database [284] (Table 25).

**Table 25. Autosomal differentially expressed lncRNAs showing disease associations**

| Gene name | Disease | Tissue | Sex-bias | Reference |
|---|---|---|---|---|
| ***BDNF-AS*** | Depression | ACC | female | [286] |
| | Schizophrenia | ACC | female | [286] |
| | Huntington | ACC | female | [287] |
| ***AFAP1-AS1*** | Alzheimer | ACC | male | [288] |
| ***BDNF-AS1*** | Obesity | ACC | female | [289] |
| ***HAGLR*** | Neuroblastoma | ACC | female | [290] |
| ***SOX2-OT*** | Alzheimer | ACC | female | - |
| | Neurodevelopmental syndrome | ACC | female | [291] |
| ***CRNDE*** | Glioma | Amygdala | female | [292] |
| | Medulloblastoma | Amygdala | female | [293] |
| ***EPB41L4A-AS1*** | Cancer | Amygdala | Male | [294] |
| ***LOXL1-AS1*** | Exfoliation syndrome | Amygdala | female | [295] |
| ***DGCR5*** | DiGeorge syndrome | Cerebellum | male | [296] |
| | Huntington | Cerebellum | male | [94] |
| ***AGAP2-AS1*** | Glioma | Cerebellum | male | [297] |
| ***LINC00115*** | Lung adenocarcinoma | Blood | male | [298] |
| ***SBF2-AS1*** | Non-small cell lung cancer | Blood | male | [299] |
| ***LINC00346*** | Hepatocellular carcinoma | Blood | female | [300] |
| ***NPTN-IT1*** | Cancer | Blood | male | [301] |
| ***NEAT1*** | Chronic lymphocytic leukemia | Blood | male | [302] |

From ACC tissue, 5 DEG have been described in the LncDisease database to be related to brain diseases. *BDNF-AS* was described to be involved in depression, schizophrenia, and Huntington disease, *BDNF-AS1* was associated with obesity, *HAGLR* was described to regulate the expression levels of clinically significant protein-coding genes involved in neuroblastoma, and *SOX2-OT* was described to be related to Alzheimer's disease and neurodevelopmental syndrome. All these genes appeared to be female biased. *AFAP1-AS1* was described to be related to Alzheimer's disease and appeared to be a male-biased DEG. From amygdala, 2 female-biased genes were noted: the *CRNDE* gene was described to be involved in glioma and medulloblastoma and *LOXL1-AS*1 was associated with exfoliation syndrome. *EPB41L4A-AS1* was described to inhibit tumor cell proliferation and appeared to be a male-biased gene. From cerebellum, 2 male-biased genes were described: the *DGCR5* gene that was associated to DiGeorge syndrome and Huntington and the *AGAP2-AS1* gene associated to glioma. From the whole blood tissue, 4 male-biased genes were described: *LINC00115* was associated with lung adenocarcinoma, *SBF2-AS1* with non-small cell lung cancer, an aberrant expression of *NPTN-IT1* was described in tumor tissues and *NEAT1* was associated with chronic lymphocytic leukemia. *LINC00346* was associated to hepatocellular carcinoma and appeared to be female-biased.

# V. Discussion and Conclusions

Sex differences have been recognized among the most significantly understudied aspects of human disease. Historically, sex has not been properly taken into account, many experimental studies have been done only on males, the sex chromosomes have been excluded from analyses, and sequence mapping protocols have not account for sex chromosome biases [278]. For years, protein coding genes have been considered the only molecules responsible for sexual dimorphism in humans, a condition that embraces not only the most obvious physical differences between male and female but also differences in development, behavior, longevity, morbidity and metabolism. However, the recent discovery of long non-coding RNAs has reignited the discussion on the complexity of this phenomenon. Our analyses have revealed substantial differences in the differential expression landscape between sexes across a range of human tissues and identified a number of lncRNAs and protein-coding genes that may contribute to sexually dimorphic traits. Moreover, we describe the biological processes in which the sex-biased protein-coding genes appeared to be involved and the association of sex-biased lncRNAs with a number of diseases.

The majority of differentially expressed genes were autosomal confirming that regulation beyond the sexual chromosomes is implied in concerting complex phenotypic traits. The fact that males and females share the vast majority of their genome [303] means also that traits showing sex-differences result from differences in the expression of genes common to both sexes [304][305]. It can also imply that, during the evolution, mutations occurring at the level of autosomal genes may be subjected to different selective pressures in male and females. This phenomenon is known as sexual antagonism [306] and describes the case in which a mutation is harmful in one sex, but beneficial in the other. This will lead to positive selection if the positive effect that the mutation has on one sex are much more than the negative on the other [307]. Most importantly, we found regulatory differences between the sexes in every tissue, indicating that sex-biased regulation of cellular processes is systemic, rather than isolated to particular tissues in which one might argue that differences in biological function are driving differences in regulation. As expected, the number of detected lncRNAs is moderately low. The causes can be various but certainly what must be taken into account is the intrinsic low expression of lncRNAs which is also tissue-related. Moreover, not only lncRNAs showed a differential expression bias in the different tissues but also PCGs suggesting that the sex-biased expression feature is a tissue specific one.

Using gene functional analysis, we observed that male and females show significant differences in their biological processes and pathways suggesting a sex related regulation of these genes. Male-biased genes in the whole blood tissue were enriched in biological processes terms related to adrenergic signaling in cardiomyocytes, hypertrophic cardiomyopathy, cardiac muscle contraction, and dilated cardiomyopathy. Indeed, it has been previously reported that cardiovascular diseases and coronary heart disease tend to affect more male individuals than female ones [308]. These results suggest that male-biased genes may potentially regulate mechanism involved in cardiac diseases and cardiovascular system and give insights on putative biomarkers related to these processes.
The analysis of the GO terms enriched among the male-biased protein coding genes in brain tissues, highlighted the role of central nervous system, neurotransmitter secretion and axon formation. These terms are concordant with the previous findings that men show a higher density of synapses [309] in some brain regions. In fact, even though the pattern of gene expression is always tissue-specific, the fact that the most enriched terms in male-biased genes are always related to muscle system processes and muscle development from whole blood, and synapses and nervous system development in cerebellum and amygdala tissues is a proof of the already known tendency to develop cardiovascular diseases and psychiatric disorders among males. Moreover, this could help in highlighting some sex-biased details of the latter, in order to get new insight in cerebellum tied disorders such as schizophrenia, autism, and ADHD.

Since the function of the majority of annotated lncRNAs is still unknown, a functional enrichment analysis for lncRNAs was not feasible. However, by exploiting the LncDisease database, we identified a number of differentially expressed lncRNAs that have been allocated to a number of diseases. Interestingly, many of these allocated diseases are sex-biased such as Alzheimer’s, obesity, Schizophrenia, and depression. The functional role of the lncRNAs in the manifestation of these diseases is not known for the majority of the identified genes. However, various evidence have showed a differential expression and regulation of these lncRNAs between the controls and patients in the studies done. Our analysis sheds the light on the importance of taking the sex as an important confounding factor in these kind of studies given that these lncRNAs show differential expression and regulation between healthy male and female individuals.

A further characterization of the differentially expressed lncRNAs should be carried out in order to get insights into their putative roles in sex-related mechanisms by performing a network and co-modulation analysis to identify their putative targets and pathways in which they could be involved. Improved understanding of these genes and the mechanisms in which they are involved is fundamental to understanding diseases with different prevalence between the sexes and provide potential targets for novel sex-dependent treatments. Males and females often do not manifest disease in the same way, or respond identically to treatment, and the exploration of the causes of these differences should remain an area of active study. Moreover, fully exploring sex-biased patterns of gene regulation is crucial not only for understanding how sex-specific biological processes drive health and disease but also for the development of precision therapeutics that will best treat disease in an individual, accounting for sex.

# Appendix 1

**Supplementary Table 1.** List of lncRNAs and PCGs differentially expressed between nuclear and cytoplasmic sub-compartments

| Gene Type | Gene Name | Transcript Name | 5'ss | Nuc/Cyt dPSI | p-value |
|---|---|---|---|---|---|
| **lncRNA** | *GUSBP11* | GUSBP11-205 | GC | 0.611 | 0.014 |
| **lncRNA** | *GUSBP11* | GUSBP11-210 | GT | -0.66 | 0.014 |
| **lncRNA** | *UBL7-AS1* | UBL7-AS1-201 | GT | 0.41 | 0.029 |
| **lncRNA** | *UCA1* | UCA1-222 | GT | 0.30 | 0.032 |
| **lncRNA** | *SPAG5-AS1* | SPAG5-AS1-201 | GT | 0.445 | 0.033 |
| **lncRNA** | *AL354733.2* | AL354733.2-210 | GC | -0.24 | 0.040 |
| **lncRNA** | *AL162586.1* | AL162586.1-202 | GT | 0.19 | 0.045 |
| **lncRNA** | *AC138035.1* | AC138035.1-201 | GC | -0.03 | 0.047 |
| **lncRNA** | *AC011476.3* | AC011476.3-202 | GT | 0.16 | 0.048 |
| **lncRNA** | *LINC00265* | LINC00265-202 | GC | 0.10 | 0.049 |
| **PCGs** | *FXN* | FXN-205 | GC | 0.84 | 0.001 |
| **PCGs** | *RNF7* | RNF7-202 | GT | 0.58 | 0.006 |
| **PCGs** | *ZNF559* | ZNF559-202 | GT | -0.40 | 0.009 |
| **PCGs** | *ZNF559* | ZNF559-201 | GC | 0.43 | 0.009 |
| **PCGs** | *MAPKAPK5* | MAPKAPK5-207 | GT | -0.47 | 0.009 |
| **PCGs** | *APEX1* | APEX1-210 | GT | -0.22 | 0.010 |
| **PCGs** | *APEX1* | APEX1-202 | GC | -0.14 | 0.010 |
| **PCGs** | *JADE1* | JADE1-202 | GT | -0.49 | 0.026 |
| **PCGs** | *JADE1* | JADE1-202 | GT | -0.24 | 0.010 |
| **PCGs** | *JADE1* | JADE1-202 | GT | -0.20 | 0.010 |
| **PCGs** | *MBD4* | MBD4-201 | GT | 0.26 | 0.011 |
| **PCGs** | *TMEM189* | TMEM189-201 | GT | 0.47 | 0.012 |
| **PCGs** | *ZNF235* | ZNF235-204 | GT | -0.25 | 0.013 |
| **PCGs** | *ZNF235* | ZNF235-204 | GT | 0.19 | 0.013 |
| **PCGs** | *TMEM217* | TMEM217-202 | GT | 0.18 | 0.017 |
| **PCGs** | *GPANK1* | GPANK1-202 | GT | -0.34 | 0.017 |
| **PCGs** | *PIGP* | PIGP-202 | GT | -0.61 | 0.017 |

| | | | | | |
|---|---|---|---|---|---|
| **PCGs** | *RAD52* | RAD52-217 | GC | 0.41 | 0.018 |
| **PCGs** | *SRSF5* | SRSF5-205 | GT | -0.17 | 0.020 |
| **PCGs** | *DIDO1* | DIDO1-207 | GC | -0.24 | 0.020 |
| **PCGs** | *DIDO1* | DIDO1-202 | GT | 0.22 | 0.020 |
| **PCGs** | *RPL29* | RPL29-209 | GT | -0.26 | 0.022 |
| **PCGs** | *RPL29* | RPL29-204 | GT | -0.19 | 0.022 |
| **PCGs** | *RPL29* | RPL29-203 | GT | -0.15 | 0.022 |
| **PCGs** | *RPL29* | RPL29-203 | GT | 0.025 | 0.022 |
| **PCGs** | *RPL29* | RPL29-203 | GT | 0.24 | 0.022 |
| **PCGs** | *RPL29* | RPL29-202 | GT | 0.30 | 0.022 |
| **PCGs** | *PKM* | PKM-201 | GC | 0.16 | 0.023 |
| **PCGs** | *PKM* | PKM-201 | GC | 0.40 | 0.023 |
| **PCGs** | *RNF220* | RNF220-202 | GC | -0.13 | 0.023 |
| **PCGs** | *YIF1B* | YIF1B-216 | GT | -0.52 | 0.023 |
| **PCGs** | *OARD1* | OARD1-203 | GT | 0.46 | 0.023 |
| **PCGs** | *NF2* | NF2-201 | GT | 0.31 | 0.027 |
| **PCGs** | *MEF2B* | MEF2B-201 | GT | 0.32 | 0.030 |
| **PCGs** | *NDUFAF6* | NDUFAF6-203 | GC | 0.16 | 0.031 |
| **PCGs** | *PPT2* | PPT2-244 | GT | -0.19 | 0.032 |
| **PCGs** | *SRSF7* | SRSF7-201 | GT | -0.17 | 0.032 |
| **PCGs** | *HEXD* | HEXD-216 | GC | 0.33 | 0.032 |
| **PCGs** | *C1D* | C1D-201 | GT | 0.30 | 0.033 |
| **PCGs** | *SMUG1* | SMUG1-213 | GT | 0.65 | 0.033 |
| **PCGs** | *PAK4* | PAK4-208 | GT | -0.61 | 0.034 |
| **PCGs** | *PAK4* | PAK4-208 | GT | 0.49 | 0.034 |
| **PCGs** | *PIGP* | PIGP-202 | GT | -0.33 | 0.036 |
| **PCGs** | *CNOT4* | CNOT4-207 | GT | 0.22 | 0.037 |
| **PCGs** | *RAD52* | RAD52-217 | GC | -0.14 | 0.037 |
| **PCGs** | *GTF2IRD2B* | GTF2IRD2B-210 | GT | 0.17 | 0.037 |
| **PCGs** | *GTF2IRD2B* | GTF2IRD2B-203 | GT | 0.28 | 0.037 |
| **PCGs** | *MYD88* | MYD88-216 | GT | -0.45 | 0.038 |
| **PCGs** | *MYD88* | MYD88-203 | GT | -0.23 | 0.038 |
| **PCGs** | *MYD88* | MYD88-203 | GT | 0.37 | 0.038 |

| | | | | | |
|---|---|---|---|---|---|
| **PCGs** | *MYD88* | MYD88-203 | GT | 0.44 | 0.038 |
| **PCGs** | *MYD88* | MYD88-216 | GT | 0.59 | 0.038 |
| **PCGs** | *FBXL6* | FBXL6-201 | GT | -0.18 | 0.038 |
| **PCGs** | *APOPT1* | APOPT1-207 | GC | 0.31 | 0.038 |
| **PCGs** | *KRBOX4* | KRBOX4-201 | GT | -0.34 | 0.039 |
| **PCGs** | *FAM104B* | FAM104B-201 | GC | 0.19 | 0.039 |
| **PCGs** | *YIF1B* | YIF1B-216 | GT | -0.45 | 0.040 |
| **PCGs** | *ISCA2* | ISCA2-204 | GC | 0.21 | 0.040 |
| **PCGs** | *DDX3X* | DDX3X-225 | GT | 0.15 | 0.041 |
| **PCGs** | *NDUFA3* | NDUFA3-203 | GT | -0.22 | 0.041 |
| **PCGs** | *NDUFV1* | NDUFV1-202 | GT | -0.22 | 0.041 |
| **PCGs** | *PLA2G12A* | PLA2G12A-202 | GT | 0.30 | 0.041 |
| **PCGs** | *POLR2E* | POLR2E-204 | GT | 0.35 | 0.042 |
| **PCGs** | *ARPP19* | ARPP19-204 | GT | 0.55 | 0.043 |
| **PCGs** | *PSRC1* | PSRC1-205 | GT | -0.48 | 0.043 |
| **PCGs** | *SMUG1* | SMUG1-213 | GT | -0.46 | 0.043 |
| **PCGs** | *ENY2* | ENY2-208 | GT | -0.42 | 0.043 |
| **PCGs** | *LYRM2* | LYRM2-204 | GT | -0.21 | 0.044 |
| **PCGs** | *ACADSB* | ACADSB-202 | GC | 0.15 | 0.044 |
| **PCGs** | *WNK1* | WNK1-201 | GT | 0.20 | 0.044 |
| **PCGs** | *POLA1* | POLA1-202 | GT | 0.46 | 0.044 |
| **PCGs** | *GALK2* | GALK2-202 | GT | 0.45 | 0.044 |
| **PCGs** | *POLR2E* | POLR2E-214 | GT | 0.39 | 0.046 |
| **PCGs** | *POLR2E* | POLR2E-204 | GT | 0.47 | 0.046 |
| **PCGs** | *FXN* | FXN-203 | GT | -0.21 | 0.047 |
| **PCGs** | *U2AF2* | U2AF2-201 | GC | -0.10 | 0.047 |
| **PCGs** | *SLC25A3* | SLC25A3-203 | GT | -0.36 | 0.047 |
| **PCGs** | *RNASEK* | RNASEK-202 | GT | 0.18 | 0.049 |

# References

1. Costa FF. Non-coding RNAs, epigenetics and complexity. Gene. 2008;410:9–17.

2. Marques-Bonet T, Ryder OA, Eichler EE. Sequencing Primate Genomes: What Have We Learned? Annu Rev Genom Hum Genet. 2009;10:355–86.

3. Liu G, Mattick J, Taft RJ. A meta-analysis of the genomic and transcriptomic composition of complex life. Cell Cycle. 2013;12:2061–72.

4. Fatica A, Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet. 2014;15:7–21.

5. Lekka E, Hall J. Noncoding RNA s in disease. FEBS Lett. 2018;592:2884–900.

6. Taft RJ, Pheasant M, Mattick JS. The relationship between non-protein-coding DNA and eukaryotic complexity. BioEssays. 2007;29:288–99.

7. Mattick JS. The Hidden Genetic Program of Complex Organisms. Sci Am. 2004;291:60–7.

8. Brown T. Genomes 4. 4th ed. New York (NY): Garland Science; 2018.

9. Huntzinger E, Izaurralde E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat Rev Genet. 2011;12:99–110.

10. Ross RJ, Weiner MM, Lin H. PIWI proteins and PIWI-interacting RNAs in the soma. Nature. 2014;505:353–9.

11. Malone CD, Hannon GJ. Small RNAs as Guardians of the Genome. Cell. 2009;136:656–68.

12. Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, et al. Tiny RNAs associated with transcription start sites in animals. Nat Genet. 2009;41:572–8.

13. Nitsche A, Stadler PF. Evolutionary clues in lncRNAs: Evolutionary clues in lncRNAs. Wiley Interdisciplinary Reviews: RNA. 2017;8:e1376.

14. Srijyothi L, Ponne S, Prathama T, Ashok C, Baluchamy S. Roles of Non-Coding RNAs in Transcriptional Regulation. In: Ghedira K, editor. Transcriptional and Post-transcriptional Regulation. InTech; 2018.

15. Pheasant M, Mattick JS. Raising the estimate of functional human sequences. Genome Research. 2007;17:1245–53.

16. Bradley RK, Merkin J, Lambert NJ, Burge CB. Alternative Splicing of RNA Triplets Is Often Regulated and Accelerates Proteome Evolution. PLoS Biol. 2012;10:e1001229.

17. Xing Y, Lee C. Alternative splicing and RNA selection pressure — evolutionary consequences for eukaryotic genomes. Nat Rev Genet. 2006;7:499–509.

18. Bush SJ, Chen L, Tovar-Corona JM, Urrutia AO. Alternative splicing and the evolution of phenotypic novelty. Philosophical Transactions of the Royal Society B: Biological Sciences. 2017;372:20150474.

19. Julien P, Miñana B, Baeza-Centurion P, Valcárcel J, Lehner B. The complete local genotype–phenotype landscape for the alternative splicing of a human exon. Nat Commun. 2016;7:11558.

20. Yeo GW, Van Nostrand E, Holste D, Poggio T, Burge CB. Identification and analysis of alternative splicing events conserved in human and mouse. Proceedings of the National Academy of Sciences. 2005;102:2850–5.

21. Irimia M, Rukov JL, Roy SW, Vinther J, Garcia-Fernandez J. Quantitative regulation of alternative splicing in evolution and development. BioEssays. 2009;31:40–50.

22. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet. 2010;11:345–55.

23. Sorek R. The birth of new exons: Mechanisms and evolutionary consequences. RNA. 2007;13:1603–8.

24. Blencowe BJ. The Relationship between Alternative Splicing and Proteomic Complexity. Trends in Biochemical Sciences. 2017;42:407–8.

25. Zhang XH-F, Chasin LA. Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. Proceedings of the National Academy of Sciences. 2006;103:13427–32.

26. Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, Ast G. Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. Genome Biol. 2007;8:R127.

27. Morozova O, Hirst M, Marra MA. Applications of New Sequencing Technologies for Transcriptome Analysis. Annu Rev Genom Hum Genet. 2009;10:135–51.

28. Deveson IW, Hardwick SA, Mercer TR, Mattick JS. The Dimensions, Dynamics, and Relevance of the Mammalian Noncoding Transcriptome. Trends in Genetics. 2017;33:464–78.

29. Huarte M, Marín-Béjar O. Long noncoding RNAs: from identification to functions and mechanisms. AGG. 2015;:257.

30. Kung JTY, Colognori D, Lee JT. Long Noncoding RNAs: Past, Present, and Future. Genetics. 2013;193:651–69.

31. Akhade VS, Pal D, Kanduri C. Long Noncoding RNA: Genome Organization and Mechanism of Action. Adv Exp Med Biol. 2017;1008:47-74.

32. Ma L, Bajic VB, Zhang Z. On the classification of long non-coding RNAs. RNA Biology. 2013;10:924–33.

33. Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. Nature. 2012;482:339–46.

34. Ransohoff JD, Wei Y, Khavari PA. The functions and unique features of long intergenic non-coding RNA. Nature Reviews Molecular Cell Biology. 2018;19:143–57.

35. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Research. 2012;22:1775–89.

36. Washietl S, Kellis M, Garber M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. Genome Research. 2014;24:616–28.

37. Zhang X, Wang W, Zhu W, Dong J, Cheng Y, Yin Z, et al. Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels. IJMS. 2019;20:5573.

38. Sigova AA, Mullen AC, Molinie B, Gupta S, Orlando DA, Guenther MG, et al. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proceedings of the National Academy of Sciences. 2013;110:2876–81.

39. Uesaka M, Nishimura O, Go Y, Nakashima K, Agata K, Imamura T. Bidirectional promoters are the major source of gene activation-associated non-coding RNAs in mammals. BMC Genomics. 2014;15:35.

40. Luo S, Lu JY, Liu L, Yin Y, Chen C, Han X, et al. Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells. Cell Stem Cell. 2016;18:637–52.

41. 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;539:452–5.

42. Long Y, Wang X, Youmans DT, Cech TR. How do lncRNAs regulate transcription? Science Advances. 2017;3:eaao2110.

43. Latos PA, Pauler FM, Koerner MV, Senergin HB, Hudson QJ, Stocsits RR, et al. Airn Transcriptional Overlap, But Not Its lncRNA Products, Induces Imprinted Igf2r Silencing. Science. 2012;338:1469–72.

44. Anderson KM, Anderson DM, McAnally JR, Shelton JM, Bassel-Duby R, Olson EN. Transcription of the non-coding RNA upperhand controls Hand2 expression and heart development. Nature. 2016;539:433–6.

45. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458:223–7.

46. 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 & Development. 2011;25:1915–27.

47. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution. Cell. 2011;147:1537–50.

48. Louro R, Smirnova AS, Verjovski-Almeida S. Long intronic noncoding RNA transcription: Expression noise or expression choice? Genomics. 2009;93:291–8.

49. Rearick D, Prakash A, McSweeny A, Shepard SS, Fedorova L, Fedorov A. Critical association of ncRNA with introns. Nucleic Acids Research. 2011;39:2357–66.

50. Guil S, Soler M, Portela A, Carrère J, Fonalleras E, Gómez A, et al. Intronic RNAs mediate EZH2 regulation of epigenetic targets. Nat Struct Mol Biol. 2012;19:664–70.

51. Buratowski S. Gene Expression--Where to Start? Science. 2008;322:1804–5.

52. Core LJ, Waterfall JJ, Lis JT. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. Science. 2008;322:1845–8.

53. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW. The Antisense Transcriptomes of Human Cells. Science. 2008;322:1855.

54. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, et al. Divergent Transcription from Active Promoters. Science. 2008;322:1849–51.

55. Preker P, Nielsen J, Schierup MH, Heick Jensen T. RNA polymerase plays both sides: Vivid and bidirectional transcription around and upstream of active promoters. Cell Cycle. 2009;8:1105–11.

56. Hrdlickova B, de Almeida RC, Borek Z, Withoff S. Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. 2014;1842:1910–22.

57. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 2012;13:R107.

58. Paralkar VR, Mishra T, Luan J, Yao Y, Kossenkov AV, Anderson SM, et al. Lineage and species-specific long noncoding RNAs during erythro-megakaryocytic development. Blood. 2014;123:1927–37.

59. Nitsche A, Rose D, Fasold M, Reiche K, Stadler PF. Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved. RNA. 2015;21:801–12.

60. Chernikova D, Managadze D, Glazko G, Makalowski W, Rogozin I. Conservation of the Exon-Intron Structure of Long Intergenic Non-Coding RNA Genes in Eutherian Mammals. Life. 2016;6:27.

61. Mukherjee N, Calviello L, Hirsekorn A, de Pretis S, Pelizzola M, Ohler U. Integrative classification of human coding and noncoding genes through RNA metabolism profiles. Nat Struct Mol Biol. 2017;24:86–96.

62. Carlevaro-Fita J, Johnson R. Global Positioning System: Understanding Long Noncoding RNAs through Subcellular Localization. Molecular Cell. 2019;73:869–83.

63. Kopp F, Mendell JT. Functional Classification and Experimental Dissection of Long Noncoding RNAs. Cell. 2018;172:393–407.

64. Morlando M, Ballarino M, Fatica A. Long Non-Coding RNAs: New Players in Hematopoiesis and Leukemia. Front Med. 2015;2.

65. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, et al. Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. Cell. 2007;129:1311–23.

66. Tsai M-C, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, et al. Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. Science. 2010;329:689–93.

67. Hasegawa Y, Brockdorff N, Kawano S, Tsutui K, Tsutui K, Nakagawa S. The Matrix Protein hnRNP U Is Required for Chromosomal Localization of Xist RNA. Developmental Cell. 2010;19:469–76.

68. Li W, Notani D, Rosenfeld MG. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. Nat Rev Genet. 2016;17:207–23.

69. Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, et al. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature. 2012;491:454–7.

70. Dykes IM, Emanueli C. Transcriptional and Post-transcriptional Gene Regulation by Long Non-coding RNA. Genomics, Proteomics & Bioinformatics. 2017;15:177–86.

71. Hu G, Niu F, Humburg BA, Liao K, Bendi S, Callen S, et al. Molecular mechanisms of long noncoding RNAs and their role in disease pathogenesis. Oncotarget. 2018;9:18648–63.

72. Khorkova O, Hsiao J, Wahlestedt C. Basic biology and therapeutic implications of lncRNA. Advanced Drug Delivery Reviews. 2015;87:15–24.

73. Batista PJ, Chang HY. Long Noncoding RNAs: Cellular Address Codes in Development and Disease. Cell. 2013;152:1298–307.

74. Liu Z, Li X, Sun N, Xu Y, Meng Y, Yang C, et al. Microarray Profiling and Co-Expression Network Analysis of Circulating lncRNAs and mRNAs Associated with Major Depressive Disorder. PLoS ONE. 2014;9:e93388.

75. Gutschner T, Diederichs S. The hallmarks of cancer: A long non-coding RNA point of view. RNA Biology. 2012;9:703–19.

76. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol. 2011;29:742–9.

77. Yap KL, Li S, Muñoz-Cabello AM, Raguz S, Zeng L, Mujtaba S, et al. Molecular Interplay of the Noncoding RNA ANRIL and Methylated Histone H3 Lysine 27 by Polycomb CBX7 in Transcriptional Silencing of INK4a. Molecular Cell. 2010;38:662–74.

78. Kotake Y, Nakagawa T, Kitagawa K, Suzuki S, Liu N, Kitagawa M, et al. Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15INK4B tumor suppressor gene. Oncogene. 2011;30:1956–62.

79. Mozdarani H, Ezzatizadeh V, Rahbar Parvaneh R. The emerging role of the long non-coding RNA HOTAIR in breast cancer development and treatment. J Transl Med. 2020;18:152.

80. Li H, An J, Wu M, Zheng Q, Gui X, Li T, et al. LncRNA HOTAIR promotes human liver cancer stem cell malignant growth through downregulation of SETD2. Oncotarget. 2015;6:27847–64.

81. Lin K, Jiang H, Zhang L-L, Jiang Y, Yang Y-X, Qiu G-D, et al. Down-Regulated LncRNA-HOTAIR Suppressed Colorectal Cancer Cell Proliferation, Invasion, and Migration by Mediating p21. Dig Dis Sci. 2018;63:2320–31.

82. Xue M, Chen L, Wang W, Su T, Shi L, Wang L, et al. HOTAIR induces the ubiquitination of Runx3 by interacting with Mex3b and enhances the invasion of gastric cancer cells. Gastric Cancer. 2018;21:756–64.

83. Jiang D, Xu L, Ni J, Zhang J, Cai M, Shen L. Functional polymorphisms in LncRNA HOTAIR contribute to susceptibility of pancreatic cancer. Cancer Cell Int. 2019;19:47.

84. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010;464:1071–6.

85. Li Z-X, Zhu Q-N, Zhang H-B, Hu Y, Wang G, Zhu Y-S. MALAT1: a potential biomarker in cancer. CMAR. 2018;10:6757–68.

86. Tano K, Mizuno R, Okada T, Rakwal R, Shibato J, Masuo Y, et al. MALAT-1 enhances cell motility of lung adenocarcinoma cells by influencing the expression of motility-related genes. FEBS Letters. 2010;584:4575–80.

87. Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, et al. A Large Intergenic Noncoding RNA Induced by p53 Mediates Global Gene Repression in the p53 Response. Cell. 2010;142:409–19.

88. Hung T, Wang Y, Lin MF, Koegel AK, Kotake Y, Grant GD, et al. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nat Genet. 2011;43:621–9.

89. Thapar R. Regulation of DNA Double-Strand Break Repair by Non-Coding RNAs. Molecules. 2018;23:2789.

90. Grammatikakis I, Panda AC, Abdelmohsen K, Gorospe M. Long noncoding RNAs (lncRNAs) and the molecular hallmarks of aging. Aging. 2014;6:992–1009.

91. Costa MC, Leitão AL, Enguita FJ. Noncoding Transcriptional Landscape in Human Aging. Curr Top Microbial Immunol. 2016;394:177-202.

92. Abdelmohsen K, Panda A, Kang M-J, Xu J, Selimyan R, Yoon J-H, et al. Senescence-associated lncRNAs: senescence-associated long noncoding RNAs. Aging Cell. 2013;12:890–900.

93. Daughters RS, Tuttle DL, Gao W, Ikeda Y, Moseley ML, Ebner TJ, et al. RNA Gain-of-Function in Spinocerebellar Ataxia Type 8. PLoS Genet. 2009;5:e1000600.

94. Johnson R, Richter N, Jauch R, Gaughwin PM, Zuccato C, Cattaneo E, et al. Human accelerated region 1 noncoding RNA is repressed by REST in Huntington's disease. Physiological Genomics. 2010;41:269–74.

95. Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R. Towards a complete map of the human long non-coding RNA transcriptome. Nature Reviews Genetics. 2018;19:535–48.

96. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Research. 2012;22:1760–74.

97. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research. 2019;47:D766–73.

98. Jalali S, Gandhi S, Scaria V. Navigating the dynamic landscape of long noncoding RNA and protein-coding gene annotations in GENCODE. Hum Genomics. 2016;10:35.

99. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733–45.

100. Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J, et al. An atlas of human long non-coding RNAs with accurate 5′ ends. Nature. 2017;543:199–204.

101. Imada EL, Sanchez DF, Collado-Torres L, Wilks C, Matam T, Dinalankara W, et al. Recounting the FANTOM CAGE-Associated Transcriptome. Genome Res. 2020;30:1073–81.

102. You B-H, Yoon S-H, Nam J-W. High-confidence coding and noncoding transcriptome maps. Genome Res. 2017;27:1050–62.

103. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet. 2015;47:199–208.

104. Fang S, Zhang L, Guo J, Niu Y, Wu Y, Li H, et al. NONCODEV5: a comprehensive annotation database for long non-coding RNAs. Nucleic Acids Research. 2018;46:D308–14.

105. The RNAcentral Consortium, Sweeney BA, Petrov AI, Burkov B, Finn RD, Bateman A, et al. RNAcentral: a hub of information for non-coding RNA sequences. Nucleic Acids Research. 2019;47:D221–9.

106. Volders P-J, Lefever S, Baute S, Nuytens J, Vanderheyden K, Menten B, et al. Targeted Genomic Screen Reveals Focal Long Non-Coding RNA Copy Number Alterations in Cancer Cell Lines. ncRNA. 2018;4:21.

107. Ma L, Cao J, Liu L, Du Q, Li Z, Zou D, et al. LncBook: a curated knowledgebase of human long non-coding RNAs. Nucleic Acids Research. 2019;47:D128–34.

108. Papasaikas P, Valcárcel J. The Spliceosome: The Ultimate RNA Chaperone and Sculptor. Trends in Biochemical Sciences. 2016;41:33–45.

109. Fong YW, Zhou Q. Stimulatory effect of splicing factors on transcriptional elongation. Nature. 2001;414:929–33.

110. Yoshimoto R, Kaida D, Furuno M, Burroughs AM, Noma S, Suzuki H, et al. Global analysis of pre-mRNA subcellular localization following splicing inhibition by spliceostatin A. RNA. 2017;23:47–57.

111. Wachutka L, Caizzi L, Gagneur J, Cramer P. Global donor and acceptor splicing site kinetics in human cells. eLife. 2019;8:e45056.

112. Ribeiro M, Furtado M, Martins S, Carvalho T, Carmo-Fonseca M. RNA Splicing Defects in Hypertrophic Cardiomyopathy: Implications for Diagnosis and Therapy. IJMS. 2020;21:1329.

113. Wongpalee SP, Sharma S. The Pre-mRNA Splicing Reaction. Methods Molecular Biology. 2014;1126:3-12.

114. Wahl MC, Will CL, Lührmann R. The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell. 2009;136:701–18.

115. Zhang Q, Fan X, Wang Y, Sun M, Shao J, Guo D. BPP: a sequence-based algorithm for branch point prediction. Bioinformatics. 2017;33:3166–72.

116. Hertel KJ. Combinatorial Control of Exon Recognition. J Biol Chem. 2008;283:1211–5.

117. Will CL, Luhrmann R. Spliceosome Structure and Function. Cold Spring Harbor Perspectives in Biology. 2011;3:a003707–a003707.

118. Sibley CR, Blazquez L, Ule J. Lessons from non-canonical splicing. Nat Rev Genet. 2016;17:407–21.

119. Naito T. Human Splice-Site Prediction with Deep Neural Networks. Journal of Computational Biology. 2018;25:954–61.

120. Turunen JJ, Niemelä EH, Verma B, Frilander MJ. The significant other: splicing by the minor spliceosome: Splicing by the minor spliceosome. WIREs RNA. 2013;4:61–76.

121. Parada GE, Munita R, Cerda CA, Gysling K. A comprehensive survey of non-canonical splice sites in the human transcriptome. Nucleic Acids Research. 2014;42:10564–78.

122. Alberts B. Molecular biology of the cell. Sixth edition. New York, NY: Garland Science, Taylor and Francis Group; 2015.

123. Huang W, Huang Y, Xu J, Liao J-L. How Does the Spliceosome Catalyze Intron Lariat Formation? Insights from Quantum Mechanics/Molecular Mechanics Free-Energy Simulations. J Phys Chem B. 2019;123:6049–55.

124. Suñé-Pou M, Prieto-Sánchez S, Boyero-Corral S, Moreno-Castro C, El Yousfi Y, Suñé-Negre J, et al. Targeting Splicing in the Treatment of Human Disease. Genes. 2017;8:87.

125. Matera AG, Wang Z. A day in the life of the spliceosome. Nat Rev Mol Cell Biol. 2014;15:108–21.

126. Zhang L, Vielle A, Espinosa S, Zhao R. RNAs in the spliceosome: Insight from cryoEM structures. WIREs RNA. 2019;10:e1523.

127. Shi Y. The Spliceosome: A Protein-Directed Metalloribozyme. Journal of Molecular Biology. 2017;429:2640–53.

128. Schneider M, Will CL, Anokhina M, Tazi J, Urlaub H, Lührmann R. Exon Definition Complexes Contain the Tri-snRNP and Can Be Directly Converted into B-like Precatalytic Splicing Complexes. Molecular Cell. 2010;38:223–35.

129. Lee Y, Rio DC. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annu Rev Biochem. 2015;84:291–323.

130. Yang L. Splicing noncoding RNAs from the inside out: Splicing noncoding RNAs from the inside out. WIREs RNA. 2015;6:651–60.

131. Talhouarne GJS, Gall JG. Lariat intronic RNAs in the cytoplasm of vertebrate cells. Proc Natl Acad Sci USA. 2018;115:E7970–7.

132. Li H-D, Menon R, Omenn GS, Guan Y. Revisiting the identification of canonical splice isoforms through integration of functional genomics and proteomics evidence. Proteomics. 2014;14:2709–18.

133. Wagner SD, Berglund JA. Alternative Pre-mRNA Splicing. Methods Molecular Biology. 2014;1126:45-54.

134. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. Nature. 2010;463:457–63.

135. Park E, Pan Z, Zhang Z, Lin L, Xing Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. The American Journal of Human Genetics. 2018;102:11–26.

136. Ramanouskaya TV, Grinev VV. The determinants of alternative RNA splicing in human cells. Mol Genet Genomics. 2017;292:1175–95.

137. Zhang X, Peng Q, Li L, Li X. Recognition of alternatively spliced cassette exons based on a hybrid model. Biochemical and Biophysical Research Communications. 2016;471:368–72.

138. Cho S, Moon H, Loh TJ, Jang HN, Liu Y, Zhou J, et al. Splicing inhibition of U2AF 65 leads to alternative exon skipping. Proc Natl Acad Sci USA. 2015;112:9926–31.

139. Izquierdo JM, Majós N, Bonnal S, Martínez C, Castelo R, Guigó R, et al. Regulation of Fas Alternative Splicing by Antagonistic Effects of TIA-1 and PTB on Exon Definition. Molecular Cell. 2005;19:475–84.

140. Yue Y, Yang Y, Dai L, Cao G, Chen R, Hong W, et al. Long-range RNA pairings contribute to mutually exclusive splicing. RNA. 2016;22:96–110.

141. Ivanov T, Pervouchine D. An Evolutionary Mechanism for the Generation of Competing RNA Structures Associated with Mutually Exclusive Exons. Genes. 2018;9:356.

142. Hatje K, Rahman R, Vidal RO, Simm D, Hammesfahr B, Bansal V, et al. The landscape of human mutually exclusive splicing. Mol Syst Biol. 2017;13:959.

143. Sammeth M, Foissac S, Guigó R. A General Definition and Nomenclature for Alternative Splicing Events. PLoS Comput Biol. 2008;4:e1000147.

144. Nevo Y, Sperling J, Sperling R. Heat shock activates splicing at latent alternative 5′ splice sites in nematodes. Nucleus. 2015;6:225–35.

145. Schmitz U, Pinello N, Jia F, Alasmari S, Ritchie W, Keightley M-C, et al. Intron retention enhances gene regulatory complexity in vertebrates. Genome Biol. 2017;18:216.

146. Jacob AG, Smith CWJ. Intron retention as a component of regulated gene expression programs. Hum Genet. 2017;136:1043–57.

147. Nishida A, Minegishi M, Takeuchi A, Niba ETE, Awano H, Lee T, et al. Tissue- and case-specific retention of intron 40 in mature dystrophin mRNA. J Hum Genet. 2015;60:327–33.

148. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature. 2013;495:333–8.

149. Xu S, Zhou L, Ponnusamy M, Zhang L, Dong Y, Zhang Y, et al. A comprehensive review of circRNA: from purification and identification to disease marker potential. PeerJ. 2018;6:e5503.

150. Ebbesen KK, Hansen TB, Kjems J. Insights into circular RNA biology. RNA Biology. 2017;14:1035–45.

151. Zhang X-O, Dong R, Zhang Y, Zhang J-L, Luo Z, Zhang J, et al. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. Genome Res. 2016;26:1277–87.

152. Danan M, Schwartz S, Edelheit S, Sorek R. Transcriptome-wide discovery of circular RNAs in Archaea. Nucleic Acids Research. 2012;40:3131–42.

153. Rybak-Wolf A, Stottmeister C, Glažar P, Jens M, Pino N, Giusti S, et al. Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. Molecular Cell. 2015;58:870–85.

154. Venø MT, Hansen TB, Venø ST, Clausen BH, Grebing M, Finsen B, et al. Spatio-temporal regulation of circular RNA expression during porcine embryonic brain development. Genome Biol. 2015;16:245.

155. Dang Y, Yan L, Hu B, Fan X, Ren Y, Li R, et al. Tracing the expression of circular RNAs in human pre-implantation embryos. Genome Biol. 2016;17:130.

156. Wilusz JE, Sharp PA. A Circuitous Route to Noncoding RNA. Science. 2013;340:440–1.

157. Ule J, Blencowe BJ. Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. Molecular Cell. 2019;76:329–45.

158. Coelho MB, Smith CWJ. Regulation of Alternative Pre-mRNA Splicing. Methods Molecular Biology. 2014;1126:55-82.

159. Bartys N, Kierzek R, Lisowiec-Wachnicka J. The regulation properties of RNA secondary structure in alternative splicing. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 2019;1862:194401.

160. Kaisers W, Ptok J, Schwender H, Schaal H. Validation of Splicing Events in Transcriptome Sequencing Data. IJMS. 2017;18:1110.

161. Shenasa H, Hertel KJ. Combinatorial regulation of alternative splicing. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 2019;1862:194392.

162. Abril JF. Comparison of splice sites in mammals and chicken. Genome Research. 2005;15:111–9.

163. Erkelenz S, Theiss S, Kaisers W, Ptok J, Walotka L, Müller L, et al. Ranking noncanonical 5′ splice site usage by genome-wide RNA-seq analysis and splicing reporter assays. Genome Res. 2018;28:1826–40.

164. Churbanov A, Winters-Hilt S, Koonin EV, Rogozin IB. Accumulation of GC donor splice signals in mammals. Biology Direct. 2008;3:30.

165. Iwata H, Gotoh O. Comparative analysis of information contents relevant to recognition of introns in many species. BMC Genomics. 2011;12:45.

166. Abramowicz A, Gos M. Splicing mutations in human genetic disorders: examples, detection, and confirmation. J Appl Genetics. 2018;59:253–68.

167. Chen M, Manley JL. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat Rev Mol Cell Biol. 2009;10:741–54.

168. Wang Y, Ma M, Xiao X, Wang Z. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. Nat Struct Mol Biol. 2012;19:1044–52.

169. Wu G, Adachi H, Ge J, Stephenson D, Query CC, Yu Y. Pseudouridines in U2 snRNA stimulate the ATPase activity of Prp5 during spliceosome assembly. EMBO J. 2016;35:654–67.

170. Herzel L, Ottoz DSM, Alpert T, Neugebauer KM. Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. Nat Rev Mol Cell Biol. 2017;18:637–50.

171. Hoffmann T, Valcárcel J. Splicing Calls Back. Cell. 2019;179:1446–7.

172. Long JC, Caceres JF. The SR protein family of splicing factors: master regulators of gene expression. Biochemical Journal. 2009;417:15–27.

173. Braunschweig U, Gueroussov S, Plocik AM, Graveley BR, Blencowe BJ. Dynamic Integration of Splicing within Gene Regulatory Pathways. Cell. 2013;152:1252–69.

174. Scotti MM, Swanson MS. RNA mis-splicing in disease. Nat Rev Genet. 2016;17:19–32.

175. Lareau LF, Brenner SE. Regulation of Splicing Factors by Alternative Splicing and NMD Is Conserved between Kingdoms Yet Evolutionarily Flexible. Molecular Biology and Evolution. 2015;32:1072–9.

176. Sutandy FXR, Ebersberger S, Huang L, Busch A, Bach M, Kang H-S, et al. In vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by cofactors. Genome Res. 2018;28:699–713.

177. Chang J-W, Yeh H-S, Park M, Erber L, Sun J, Cheng S, et al. mTOR-regulated U2af1 tandem exon splicing specifies transcriptome features for translational control. Nucleic Acids Research. 2019;47:10373–87.

178. Shepard PJ, Hertel KJ. The SR protein family. Genome Biol. 2009;10:242.

179. Jimeno-González S, Payán-Bravo L, Muñoz-Cabello AM, Guijo M, Gutierrez G, Prado F, et al. Defective histone supply causes changes in RNA polymerase II

elongation rate and cotranscriptional pre-mRNA splicing. Proc Natl Acad Sci USA. 2015;112:14840–5.

180. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature. 2011;479:74–9.

181. Maunakea AK, Chepelev I, Cui K, Zhao K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. Cell Res. 2013;23:1256–69.

182. Patrick KL, Ryan CJ, Xu J, Lipp JJ, Nissen KE, Roguev A, et al. Genetic Interaction Mapping Reveals a Role for the SWI/SNF Nucleosome Remodeler in Spliceosome Activation in Fission Yeast. PLoS Genet. 2015;11:e1005074.

183. Davis-Turak JC, Allison K, Shokhirev MN, Ponomarenko P, Tsimring LS, Glass CK, et al. Considering the kinetics of mRNA synthesis in the analysis of the genome and epigenome reveals determinants of co-transcriptional splicing. Nucleic Acids Research. 2015;43:699–707.

184. Yearim A, Gelfman S, Shayevitch R, Melcer S, Glaich O, Mallm J-P, et al. HP1 Is Involved in Regulating the Global Impact of DNA Methylation on Alternative Splicing. Cell Reports. 2015;10:1122–34.

185. Zhu S, Wang G, Liu B, Wang Y. Modeling Exon Expression Using Histone Modifications. PLoS ONE. 2013;8:e67448.

186. Liu H, Jin T, Guan J, Zhou S. Histone modifications involved in cassette exon inclusions: a quantitative and interpretable analysis. BMC Genomics. 2014;15:1148.

187. Di Giammartino DC, Nishida K, Manley JL. Mechanisms and Consequences of Alternative Polyadenylation. Molecular Cell. 2011;43:853–66.

188. Shi Y, Di Giammartino DC, Taylor D, Sarkeshik A, Rice WJ, Yates JR, et al. Molecular Architecture of the Human Pre-mRNA 3′ Processing Complex. Molecular Cell. 2009;33:365–76.

189. Ji X, Kong J, Liebhaber SA. An RNA-protein complex links enhanced nuclear 3′ processing with cytoplasmic mRNA stabilization: A novel upstream enhancer of 3′ processing. The EMBO Journal. 2011;30:2622–33.

190. Richard P, Manley JL. Transcription termination by nuclear RNA polymerases. Genes & Development. 2009;23:1247–69.

191. Zhang X, Virtanen A, Kleiman FE. To polyadenylate or to deadenylate: That is the question. Cell Cycle. 2010;9:4437–49.

192. Kaida D, Berg MG, Younis I, Kasim M, Singh LN, Wan L, et al. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature. 2010;468:664–8.

193. So BR, Di C, Cai Z, Venters CC, Guo J, Oh J-M, et al. A Complex of U1 snRNP with Cleavage and Polyadenylation Factors Controls Telescripting, Regulating mRNA Transcription in Human Cells. Molecular Cell. 2019;76:590-599.e4.

194. Deng Y, Shi J, Ran Y, Xiang AP, Yao C. A potential mechanism underlying U1 snRNP inhibition of the cleavage step of mRNA 3' processing. Biochemical and Biophysical Research Communications. 2020;530:196–202.

195. Deveson IW, Brunck ME, Blackburn J, Tseng E, Hon T, Clark TA, et al. Universal Alternative Splicing of Noncoding Exons. Cell Systems. 2018;6:245-255.e5.

196. Krchňáková Z, Thakur PK, Krausová M, Bieberstein N, Haberman N, Müller-McNicoll M, et al. Splicing of long non-coding RNAs primarily depends on polypyrimidine tract and 5′ splice-site sequences due to weak interactions with SR proteins. Nucleic Acids Research. 2019;47:911–28.

197. Tilgner H, Knowles DG, 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 Research. 2012;22:1616–25.

198. Melé M, Mattioli K, Mallard W, Shechner DM, Gerhardinger C, Rinn JL. Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. Genome Research. 2017;27:27–37.

199. Schlackow M, Nojima T, Gomes T, Dhir A, Carmo-Fonseca M, Proudfoot NJ. Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs. Molecular Cell. 2017;65:25–38.

200. Samudyata, Castelo-Branco G, Bonetti A. Birth, coming of age and death: The intriguing life of long noncoding RNAs. Seminars in Cell & Developmental Biology. 2018;79:143–52.

201. 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). Nature Communications. 2016;7.

202. 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. Nature Genetics. 2017;49:1731–40.

203. Ponjavic J, Ponting CP, Lunter G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Research. 2007;17:556–65.

204. Bortfeldt R, Schindler S, Szafranski K, Schuster S, Holste D. Comparative analysis of sequence features involved in the recognition of tandem splice sites. BMC Genomics. 2008;9:202.

205. Hiller M, Platzer M. Widespread and subtle: alternative splicing at short-distance tandem sites. Trends in Genetics. 2008;24:246–55.

206. Hiller M, Huse K, Szafranskzi K, Rosenstiel P, Schreiber S, Backofen R, et al. Phylogenetically widespread alternative splicing at unusual GYNGYN donors. Genome Biol. 2006;7:R65.

207. Tsai K-W, Tarn W-Y, Lin W. Wobble Splicing Reveals the Role of the Branch Point Sequence-to-NAGNAG Region in 3′ Tandem Splice Site Selection. MCB. 2007;27:5835–48.

208. Hiller M, Szafranski K, Sinha R, Huse K, Nikolajewa S, Rosenstiel P, et al. Assessing the fraction of short-distance tandem splice sites under purifying selection. RNA. 2008;14:616–29.

209. Dou Y, Fox-Walsh KL, Baldi PF, Hertel KJ. Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. RNA. 2006;12:2047–56.

210. Hicks MJ, Mueller WF, Shepard PJ, Hertel KJ. Competing Upstream 5′ Splice Sites Enhance the Rate of Proximal Splicing. MCB. 2010;30:1878–86.

211. Hiller M, Szafranski K, Huse K, Backofen R, Platzer M. Selection against tandem splice sites affecting structured protein regions. BMC Evol Biol. 2008;8:89.

212. Sun X, Lin SM, Yan X. Computational Evidence of NAGNAG Alternative Splicing in Human Large Intergenic Noncoding RNA. BioMed Research International. 2014;2014:1–7.

213. Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, et al. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. Nat Genet. 2004;36:1255–7.

214. Akerman M, Gutfreund YM. Alternative splicing regulation at tandem 3' splice sites. Nucleic Acids Research. 2006;34:23-31.

215. Chern T-M, van Nimwegen E, Kai C, Kawai J, Carninci P, Hayashizaki Y, et al. A Simple Physical Model Predicts Small Exon Length Variations. PLoS Genet. 2006;2:e45.

216. Tsai K-W, Lin W. Quantitative analysis of wobble splicing indicates that it is not tissue specific. Genomics. 2006;88:855–64.

217. Szafranski K, Kramer M. It’s a bit over, is that ok? The subtle surplus from tandem alternative splicing. RNA Biology. 2015;12:115–22.

218. Hujovà P, Grodeckà L, Soucek P, Freiberger T. Impact of acceptor splice site NAGTAG motif on exon recognition. Molecular Biology Reports. 2019;46:2877-2884.

219. Busch A, Hertel KJ. Extensive regulation of NAGNAG alternative splicing: new tricks for the spliceosome? Genome Biol. 2012;13:143.

220. Tadokoro K, Yamazaki-Inoue M, Tachibana M, Fujishiro M, Nagao K, Toyoda M, et al. Frequent occurrence of protein isoforms with or without a single amino acid residue by subtle alternative splicing: the case of Gln in DRPLA affects subcellular localization of the products. J Hum Genet. 2005;50:382–94.

221. Valentonyte R, Hampe J, Huse K, Rosenstiel P, Albrecht M, Stenzel A, et al. Sarcoidosis is associated with a truncating splice site mutation in BTNL2. Nat Genet. 2005;37:357–64.

222. Tsai KW, Chan WC, Hsu CN, Lin WC. Sequence features involved in the mechanism of 3' splice junction wobbling. 2010;11:34.

223. Tessier SJ, Loiselle JJ, McBain A, Pullen C, Koenderink BW, Roy JG, et al. Insight into the role of alternative splicing within the RBM10v1 exon 10 tandem donor site. BMC Research Notes. 2015;8:46.

224. Karolchik D. The UCSC Table Browser data retrieval tool. Nucleic Acids Research. 2004;32:493D – 496.

225. 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 Research. 2015;43:W589–98.

226. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. 2004;11:377–394.

227. Signal B, Gloss BS, Dinger ME, Mercer TR. Machine learning annotation of human branchpoints. Bioinformatics. 2018;34:920–7.

228. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Research. 2002;12:996-1006.

229. Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. 2018;19:40.

230. Beaudoing E. Patterns of Variant Polyadenylation Signal Usage in Human Genes. Genome Research. 2000;10:1001–10.

231. The GTEx Consortium, Ardlie KG, Deluca DS, Segre AV, Sullivan TJ, Young TR, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015;348:648–60.

232. Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. Nucleic Acids Research. 2011;39:D19–21.

233. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.

234. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols. 2009;4:44–57.

235. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Research. 2019;47:D419–26.

236. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature Genetics. 2000;25:25–9.

237. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research. 2019;47:D330–8.

238. Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2019 update. Nucleic Acids Research. 2019;47:D853–8.

239. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

240. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011; 29:24-26.

241. Maass PG, Barutcu AR, Rinn JL. Interchromosomal interactions: A genomic love story of kissing chromosomes. Journal of Cell Biology. 2019;218:27–38.

242. Bradnam KR, Korf I. Longer First Introns Are a General Property of Eukaryotic Gene Structure. PLoS ONE. 2008;3:e3093.

243. Thanaraj TA. Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. Nucleic Acids Research. 2001;29:2581–93.

244. Kralovicova J, Hwang G, Asplund AC, Churbanov A, Smith CIE, Vorechovsky I. Compensatory signals associated with the activation of human GC 5′ splice sites. Nucleic Acids Research. 2011;39:7077–91.

245. Wang J, Xu W, He Y, Xia Q, Liu S. LncRNA MEG3 impacts proliferation, invasion, and migration of ovarian cancer cells through regulating PTEN. Inflamm Res. 2018;67:927–36.

246. Xu D, Chi G, Zhao C, Li D. Long noncoding RNA MEG3 inhibits proliferation and migration but induces autophagy by regulation of Sirt7 and PI3K/AKT/mTOR pathway in glioma cells. Journal of Cellular Biochemistry. 2019;120:7516–26.

247. Wang Z, Tan M, Chen G, Li Z, Lu X. LncRNA SOX2-OT is a novel prognostic biomarker for osteosarcoma patients and regulates osteosarcoma cells proliferation and motility through modulating SOX2: The Role of LNC RNA SOX2-OT in Osteosarcoma. IUBMB Life. 2017;69:867–76.

248. 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. Nature Medicine. 2016;22:861–8.

249. You D, 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. Biomedicine & Pharmacotherapy. 2019;111:1447–57.

250. Barry G, Briggs JA, Hwang DW, Nayler SP, Fortuna PRJ, Jonkhout N, et al. The long non-coding RNA NEAT1 is responsive to neuronal activity and is associated with hyperexcitability states. Scientific Reports. 2017;7.

251. Clark BS, Blackshaw S. Long non-coding RNA-dependent transcriptional regulation in neuronal development and disease. Frontiers in Genetics. 2014;5.

252. Ravasi T. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Research. 2005;16:11–9.

253. Bieberstein NI, Carrillo Oesterreich F, Straube K, Neugebauer KM. First Exon Length Controls Active Chromatin Signatures and Transcription. Cell Reports. 2012;2:62–8.

254. Kalari KR, Casavant M, Bair TB, Keen HL, Comeron JM, Casavant TL, et al. First Exons and Introns – A Survey of GC Content and Gene Structure in the Human Genome. In Silico Biol. 2006;6:237-242.

255. Andreassi C, Riccio A. To localize or not to localize: mRNA fate is in 3′UTR ends. Trends in Cell Biology. 2009;19:465–74.

256. Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Research. 2006;34:3955–67.

257. Park S, Hannenhalli S, Choi S. 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.

258. Rose AB. Introns as Gene Regulators: A Brick on the Accelerator. Frontiers in Genetics. 2019;9.

259. 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.

260. Almada AE, Wu X, Kriz AJ, Burge CB, Sharp PA. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature. 2013;499:360–3.

261. Singh RN, Singh NN. A novel role of U1 snRNP: Splice site selection from a distance. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 2019;1862:634–42.

262. Stamm S, Zhang MQ, Marr TG, Helfman DM. A sequence compilation and comparison of exons that are alternatively spliced in neurons. Nucleic Acids Research. 1994;22:1515–26.

263. 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. Molecular Cell. 2004;14:221-231.

264. Zhuo D, Madden R, Elela SA, Chabot B. Modern origin of numerous alternatively spliced human introns from tandem arrays. PNAS. 2007;104:882–6.

265. Tsai K, Tseng H, Lin W. Two wobble-splicing events affect ING4 protein subnuclear localization and degradation. Experimental Cell Research. 2008;314:3130–41.

266. Farrer T. Analysis of the role of Caenorhabditis elegans GC-AG introns in regulated splicing. Nucleic Acids Research. 2002;30:3360–7.

267. Palaniswamy R, Teglund S, Lauth M, Zaphiropoulos PG, Shimokawa T. Genetic variations regulate alternative splicing in the 5' untranslated regions of the mouse glioma-associated oncogene 1, Gli1. BMC Molecular Biology. 2010;11:32.

268. Fumasoni I, Meani N, Rambaldi D, Scafetta G, Alcalay M, Ciccarelli FD. Family expansion and gene rearrangements contributed to the functional specialization of PRDM genes in vertebrates. BMC Evolutionary Biology. 2007;7:187.

269. 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;31:1093–9.

270. Schaefke B, Sun W, Li Y-S, Fang L, Chen W. The evolution of posttranscriptional regulation. WIREs RNA. 2018;9:e1485.

271. Jandura A, Krause HM. The New RNA World: Growing Evidence for Long Noncoding RNA Functionality. Trends in Genetics. 2017;33:665–76.

272. Williams TM, Carroll SB. Genetic and molecular insights into the development and evolution of sexual dimorphism. Nat Rev Genet. 2009;10:797–804.

273. Mayne BT, Bianco-Miotto T, Buckberry S, Breen J, Clifton V, Shoubridge C, et al. Large Scale Gene Expression Meta-Analysis Reveals Tissue-Specific, Sex-Biased Gene Expression in Humans. Front Genet. 2016;7.

274. Arnold AP. A general theory of sexual differentiation: A General Theory of Sexual Differentiation. Journal of Neuroscience Research. 2017;95:291–300.

275. Zucker I, Beery AK. Males still dominate animal studies. Nature. 2010;465:690–690.

276. Beery AK, Zucker I. Sex bias in neuroscience and biomedical research. Neuroscience & Biobehavioral Reviews. 2011;35:565–72.

277. Zakiniaeiz Y, Cosgrove KP, Potenza MN, Mazure CM. Balance of the Sexes: Addressing Sex Differences in Preclinical Research. 2016;89:255-259.

278. Khramtsova EA, Davis LK, Stranger BE. The role of sex in the genomics of human complex traits. Nat Rev Genet. 2019;20:173–90.

279. Clayton JA. Applying the new SABV (sex as a biological variable) policy to research and clinical care. Physiology & Behavior. 2018;187:2–5.

280. Danska JS. Sex Matters for Mechanism. Science Translational Medicine. 2014;6:258fs40-258fs40.

281. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

282. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.

283. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

284. Bao Z, Yang Z, Huang Z, Zhou Y, Cui Q, Dong D. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. Nucleic Acids Research. 2019;47:D1034–7.

285. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. 2016;22:839–51.

286. Ng S-Y, Lin L, Soh BS, Stanton LW. Long noncoding RNAs in development and disease of the central nervous system. Trends in Genetics. 2013;29:461–8.

287. Vucicevic D, Schrewe H, Orom UA. Molecular mechanisms of long ncRNAs in neurological disorders. Front Genet. 2014;5:48.

288. Huo Z, Zhu Y, Yu L, Yang J, De Jager P, Bennett DA, et al. DNA methylation variability in Alzheimer's disease. Neurobiology of Aging. 2019;76:35–44.

289. MAGIC, on behalf of Procardis Consortium, Speliotes EK, Willer CJ, Berndt SI, Monda KL, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010;42:937–48.

290. Yarmishyn AA, Batagov AO, Tan JZ, Sundaram GM, Sampath P, Kuznetsov VA, et al. HOXD-AS1 is a novel lncRNA encoded in HOXD cluster and a marker of neuroblastoma progression revealed via integrative analysis of noncoding transcriptome. BMC Genomics. 2014;15 Suppl 9:S7.

291. Shi X, Sun M, Liu H, Yao Y, Song Y. Long non-coding RNAs: A new frontier in the study of human diseases. Cancer Letters. 2013;339:159–66.

292. Zhang F, Gao C, Ma X-F, Peng X-L, Zhang R-X, Kong D-X, et al. Expression Profile of Long Noncoding RNAs in Peripheral Blood Mononuclear Cells from Multiple Sclerosis Patients. CNS Neuroscience & Therapeutics. 2016;22:298–305.

293. Song H, Han LM, Gao Q, Sun Y. Long non-coding RNA CRNDE promotes tumor growth in medulloblastoma. 2016;20:2588-2597.

294. Yabuta N, Onda H, Watanabe M, Yoshioka N, Nagamori I, Funatsu T, et al. Isolation and characterization of the TIGA genes, whose transcripts are induced by growth arrest. Nucleic Acids Research. 2006;34:4878–92.

295. Hauser MA, Aboobakar IF, Liu Y, Miura S, Whigham BT, Challa P, et al. Genetic variants and cellular stressors associated with exfoliation syndrome modulate promoter activity of a lncRNA within the LOXL1 locus. Hum Mol Genet. 2015;24:6552–63.

296. Singh M. Dysregulated A to I RNA editing and non-coding RNAs in neurodegeneration. Front Gene. 2013;3:326.

297. Wang W, Yang F, Zhang L, Chen J, Zhao Z, Wang H, et al. LncRNA profile study reveals four-lncRNA signature associated with the prognosis of patients with anaplastic gliomas. Oncotarget. 2016;7:77225-77236.

298. Li X, Li B, Ran P, Wang L. Identification of ceRNA network based on a RNA-seq shows prognostic lncRNA biomarkers in human lung adenocarcinoma. Oncol Lett. 2018;16:5697-5708.

299. Lv J, Qiu M, Xia W, Liu C, Xu Y, Wang J, et al. High expression of long non-coding RNA SBF2-AS1 promotes proliferation in non-small cell lung cancer. J Exp Clin Cancer Res. 2016;35:75.

300. Zhang J, Fan D, Jian Z, Chen GG, Lai PBS. Cancer Specific Long Noncoding RNAs Show Differential Expression Patterns and Competing Endogenous RNA Potential in Hepatocellular Carcinoma. PLoS ONE. 2015;10:e0141042.

301. Yang F, Huo X, Yuan S, Zhang L, Zhou W, Wang F, et al. Repression of the Long Noncoding RNA-LET by Histone Deacetylase 3 Contributes to Hypoxia-Mediated Metastasis. Molecular Cell. 2013;49:1083–96.

302. Blume CJ, Hotz-Wagenblatt A, Hüllein J, Sellner L, Jethwa A, Stolz T, et al. p53-dependent non-coding RNA networks in chronic lymphocytic leukemia. Leukemia. 2015;29:2015–23.

303. Ober C, Loisel DA, Gilad Y. Sex-specific genetic architecture of human disease. Nat Rev Genet. 2008;9:911–22.

304. Connallon T, Knowles LL. Intergenomic conflict revealed by patterns of sex-biased gene expression. Trends in Genetics. 2005;21:495–9.

305. Rinn JL, Snyder M. Sexual dimorphism in mammalian gene expression. Trends in Genetics. 2005;21:298–305.

306. Connallon T, Clark AG. Evolutionary inevitability of sexual antagonism. Proc R Soc B. 2014;281:20132123.

307. Ellegren H, Parsch J. The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet. 2007;8:689–98.

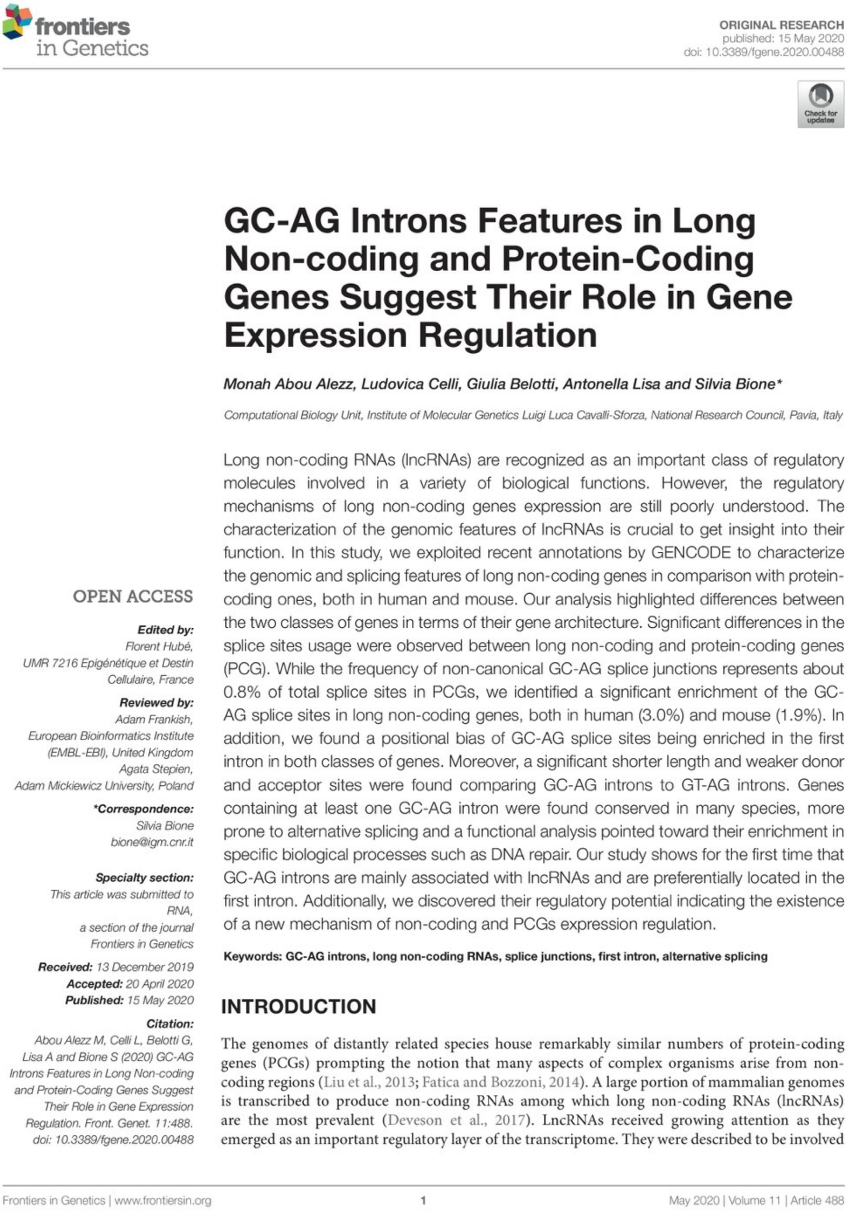
308. Galiuto L, De Caterina AR, Porfidia A, Paraggio L, Barchetta S, Locorotondo G, et al. Reversible coronary microvascular dysfunction: a common pathogenetic mechanism in Apical Ballooning or Tako-Tsubo Syndrome. European Heart Journal. 2010;31:1319–27.

309. Alonso-Nanclares L, Gonzalez-Soriano J, Rodriguez JR, DeFelipe J. Gender differences in human cortical synaptic density. Proceedings of the National Academy of Sciences. 2008;105:14615–9.

# List of Original Manuscripts

**Abou Alezz M**, Celli L, Belotti G, Lisa A and Bione S. 2020. GC-AG Introns Features in Long Non-coding and Protein-Coding Genes Suggest Their Role in Gene Expression Regulation. Front. Genet. 11:488. doi:10.3389/fgene.2020.00488.

*Pdf of the paper have been included.*



frontiers in Genetics

ORIGINAL RESEARCH
published: 15 May 2020
doi: 10.3389/fgene.2020.00488

# GC-AG Introns Features in Long Non-coding and Protein-Coding Genes Suggest Their Role in Gene Expression Regulation

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

*Computational Biology Unit, Institute of Molecular Genetics Luigi Luca Cavalli-Sforza, National Research Council, Pavia, Italy*

Long non-coding RNAs (lncRNAs) are recognized as an important class of regulatory molecules involved in a variety of biological functions. However, the regulatory mechanisms of long non-coding genes expression are still poorly understood. The characterization of the genomic features of lncRNAs is crucial to get insight into their function. In this study, we exploited recent annotations by GENCODE to characterize the genomic and splicing features of long non-coding genes in comparison with protein-coding ones, both in human and mouse. Our analysis highlighted differences between the two classes of genes in terms of their gene architecture. Significant differences in the splice sites usage were observed between long non-coding and protein-coding genes (PCG). While the frequency of non-canonical GC-AG splice junctions represents about 0.8% of total splice sites in PCGs, we identified a significant enrichment of the GC-AG splice sites in long non-coding 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 donor and acceptor sites were found comparing GC-AG introns to GT-AG introns. Genes containing at least one GC-AG intron were found conserved in many species, more prone to alternative splicing and a functional analysis pointed toward their enrichment in specific biological processes such as DNA repair. Our study shows for the first time that GC-AG introns are mainly associated with lncRNAs and are preferentially located in the first intron. Additionally, we discovered their regulatory potential indicating the existence of a new mechanism of non-coding and PCGs expression regulation.

**Keywords: GC-AG introns, long non-coding RNAs, splice junctions, first intron, alternative splicing**

OPEN ACCESS

***Edited by:***
*Florent Hubé,*
*UMR 7216 Epigénétique et Destin Cellulaire, France*

***Reviewed by:***
*Adam Frankish,*
*European Bioinformatics Institute (EMBL-EBI), United Kingdom*
*Agata Stepien,*
*Adam Mickiewicz University, Poland*

****Correspondence:***
*Silvia Bione*
*bione@igm.cnr.it*

***Specialty section:***
*This article was submitted to RNA,*
*a section of the journal Frontiers in Genetics*

***Received:*** *13 December 2019*
***Accepted:*** *20 April 2020*
***Published:*** *15 May 2020*

***Citation:***
*Abou Alezz M, Celli L, Belotti G, Lisa A and Bione S (2020) GC-AG Introns Features in Long Non-coding and Protein-Coding Genes Suggest Their Role in Gene Expression Regulation. Front. Genet. 11:488. doi: 10.3389/fgene.2020.00488*

## INTRODUCTION

The genomes of distantly related species house remarkably similar numbers of protein-coding genes (PCGs) prompting the notion that many aspects of complex organisms arise from non-coding regions (Liu et al., 2013; Fatica and Bozzoni, 2014). A large portion of mammalian genomes is transcribed to produce non-coding RNAs among which long non-coding RNAs (lncRNAs) are the most prevalent (Deveson et al., 2017). LncRNAs received growing attention as they emerged as an important regulatory layer of the transcriptome. They were described to be involved in transcriptional regulation, splicing, mRNA translation, chromatin modifications, and spatial conformation of chromosomes (Jandura and Krause, 2017; Mattick, 2018). Despite several studies reported their role in regulating the expression of other genes, how the transcription of lncRNAs is regulated remains less understood.

Similarly to PCGs, the majority of lncRNAs are transcribed by RNA Polymerase II and undergo the same RNA processing steps including capping, splicing, and polyadenylation. In comparison to PCGs, lncRNAs show lower levels of expression and higher tissue-specificity. The transcription of lncRNAs was mainly studied in relationship to those of nearby PCGs. In many cases, lncRNAs were reported to be co-expressed and co-regulated with their neighbor PCGs especially when divergently transcribed from bidirectional promoters (Sigova et al., 2013; Uesaka et al., 2014). In some cases, the direct involvement of lncRNAs in the transcription regulation of neighbor PCGs was demonstrated: in the work of Luo et al. (2016), the correlation between the expression of some lncRNAs and of the neighbor PCGs was experimentally demonstrated and estimated to account for 75% of total lncRNAs. As an example, the lncRNA *EVX1-AS* (EVX1-antisense RNA) was reported to promote the transcription of the *EVX1* (even-skipped homeobox 1) gene during mesodermal differentiation by modifying chromatin accessibility (Luo et al., 2016). It was also reported that lncRNA transcription itself, rather than the RNA transcript, exerts regulatory effects on neighboring genes (Long et al., 2017). For example, the silencing of *Igf2r* (insulin like growth factor 2 receptor) gene expression was demonstrated to be due to the transcription of the lncRNAs *Airn* (antisense of Igf2r non-protein coding RNA) that interfere with RNA polymerase II recruitment (Latos et al., 2012). Similarly, Anderson et al. (2016) reported that the lack of transcription of the lncRNA *Hand2os1* (Hand2, opposite strand 1), but not the knockdown of its mature transcript, abolished the expression of the *Hand2* (heart and neural crest derivatives expressed 2) gene leading to embryonic lethality in mice. Taken together, this evidence points toward the tight regulation of lncRNAs expression and the importance of lncRNAs transcription in regulating PCGs.

Splicing represents a main mechanism of post-transcriptional regulation of gene expression (Papasaikas and Valcárcel, 2016). It is not only involved in the maturation of pre-mRNAs, but can also influence the subcellular localization of mature transcripts and increase transcriptional rates by several folds (Fong and Zhou, 2001). The majority of lncRNAs are processed by the splicing machinery and they can undergo alternative splicing showing a complexity in their gene expression regulation as mRNAs. Despite lncRNAs are less conserved than PCGs due to the absence of constraints on coding sequences (Hezroni et al., 2015), they exhibit conservation and selective constraints at their exon–intron structures and splicing regulatory elements (Schüler et al., 2014; Nitsche et al., 2015; Chernikova et al., 2016). Thus, the recognition of lncRNAs intron boundaries and the correct splicing of their introns is a crucial step for their functional maturation (Ponjavic et al., 2007; Nitsche and Stadler, 2017). Initial studies reported that lncRNAs show an overall splicing inefficiency compared with PCGs (Derrien et al., 2012; Tilgner et al., 2012; Melé et al., 2017). The inefficiency in lncRNAs splicing was mildly correlated to weak U2AF65 binding to 3′splice site (ss), in addition to the 5′ss strength and a lower thymidine content in the polypyrimidine tract of lncRNA introns (Melé et al., 2017; Krchòáková et al., 2019). Nevertheless, efficient splicing was observed among lncRNAs with specific functions (Melé et al., 2017). As well as for transcription, lncRNAs splicing can also affect the transcription of neighboring PCGs. In the study of Engreitz et al. (2016), it was demonstrated that the first 5′splice site of the mouse lncRNA *Blustr* has a critical impact on its ability to regulate the upstream PCG *Sfmbt2* (Scm-like with four mbt domains 2). Thus, a better understanding of the mechanisms regulating lncRNAs splicing could contribute to understand their regulation and impact on PCGs transcription.

In this study, we took advantage of lncRNA annotations provided by the GENCODE project (Frankish et al., 2019) to characterize the genomic and splicing features of human and mouse lncRNAs in comparison to PCGs. At the genomic level, our analysis revealed differences in gene architecture between lncRNAs and PCGs, mainly in genic regions involved in gene expression regulation. The characterization of splicing features revealed a significant enrichment of GC-AG splice junctions in lncRNAs of human and mouse. Moreover, the GC-AG introns were preferentially found located in the first intron in both lncRNAs and mRNAs of both species. Based on the evidence that the frequency of 5′ss-GC was reported to increase with organisms complexity (Sheth et al., 2006) and that an accumulation of 5′ss-GC was previously described in mammals (Churbanov et al., 2008), we hypothesized that GC-AG introns may represent new key regulatory elements. Further analyses demonstrated that GC-AG introns differ from GT-AG introns in terms of length and donor and acceptor splice sites strength especially in lncRNAs. Interestingly, GC-AG introns appeared more prone to alternative splicing in both lncRNAs and mRNAs and in particular in alternative donor splice sites. In addition, GC-AG introns in PCGs appeared highly conserved and significantly enriched in specific biological processes such as DNA repair and neurogenesis. Taken together, our results highlighted unique features of GC-AG introns thus supporting their role as specific transcription regulators.

## MATERIALS AND METHODS

### Data Collection

The lists of lncRNAs and PCGs were downloaded from the GENCODE website[1]. 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 mouse genes annotated on the genome sequence GRCm38 (gencode.vM16.long_noncoding_RNAs.gtf.gz; gen code.vM16.basic.annotation.gtf.gz). PCGs were selected from the basic annotation when both gene and transcript were indicated as "protein_coding". The total number of genes,

[1] https://www.gencodegenes.org/

transcripts and exons considered in both species are reported in **Supplementary Table S1**.

An independent validation of the results from GENCODE was obtained by collecting human lncRNAs annotations data from 6 different databases: the FANTOM5 database (Fantom CAT genes[2]; FANTOM_CAT.lv3_robust.only_lncRNA.gtf) (Hon et al., 2017), the NONCODE v.5 database[3] (Fang et al., 2018), the BIGTranscriptome database release 2016 lncRNA catalog[4] (You et al., 2017), the LncBook database[5] (Ma et al., 2019), the MiTranscriptome database[6] (Iyer et al., 2015), and the LNCipedia database version 5.2[7] (Volders et al., 2013). A validation of results obtained from the mouse genome was performed using lncRNAs annotations from the NONCODEv5 database[8].

The lists of lncRNAs and PCGs of *Drosophila melanogaster* and *Caenorhabditis elegans* were downloaded from the BioMart data mining tool (Smedley et al., 2015) in the Ensembl genome database (release 91).

## Conservation Analysis

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 comparison in the BioMart data mining tool (Smedley et al., 2015). Multi-species conservation of 5′splice sites was assessed manually by aligning the sequences of corresponding introns in different organisms using the UCSC genome browser as data source (Kent et al., 2002). Species considered in this analysis were: human, chimp, macaque, mouse, rat, dog, cow, pig, chicken, fugu, and zebrafish.

## Introns Analysis

Intron sequences were retrieved using the Table Browser tool from UCSC using human GRCh38 and mouse GRCm38 genome sequences (Karolchik et al., 2004). We excluded from the analysis all single-exon genes as they are not subjected to splicing: this resulted in a total of 56582 lncRNAs and 525149 PCGs introns in human and 29611 lncRNAs and 393788 PCGs 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′ and 3′ splice sites of human and mouse introns using the 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.

## Alternative Splicing Analysis

The assignment of alternative splicing events involving GC-AG and GT-AG introns was performed using the SUPPA2 tool (Trincado et al., 2018). Splicing events were extracted from the gtf files of lncRNA and PCGs annotations from the GENCODE database. The SUPPA2 tool classified alternative spliced events according to the following types: exon skipping, intron retention, mutually exclusive exons, alternative 5′ss, alternative 3′ss, alternative first exons and alternative last exons. Custom R scripts were used to extract introns involved in each type of alternative splicing event and to evaluate alternative last exons. Polyadenylation signals (PAS) were extracted according to the 16 PAS reported in the paper of Beaudoing et al. (2000) in a bin of 40 nucleotides at the end of each last exon.

## Expression Analysis

RNA-Seq data of healthy individuals were obtained from the Genotype-Tissue Expression (GTEx) version 8 data set[9] (phs000424.v8.p2.c1, July 18, 2019) (GTEx Consortium, 2015) and downloaded using dbGaP web site (approved protocol #23403). Data were collected from 10 different tissues (anterior cingulate cortex, amygdala, cerebellum, heart left ventricle, kidney cortex, lung, liver, spleen, skin, and testis) of male individuals using 8 samples per tissue for a total of 80 samples. Quality control analyses on the raw sequence data were performed using the FastQC tool[10]. Reads in FASTQ format passing quality control were quantified with the transcripts per million method implemented in the Salmon software (version 1.2.0) (Patro et al., 2017) using default parameters and the human hg38 reference transcriptome from GENCODE v27. The transcripts quantifications were then imported into R and summarized using custom scripts. The unexpressed lncRNA transcripts with TPM $< 0.1$ and protein-coding transcripts with TPM $< 0.5$ were filtered out in subsequent analyses.

## Statistical Analysis

Data analyses and descriptive statistics were performed using RStudio version 1.1.456[11]. 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.

## Functional Enrichment Analysis

Gene list functional enrichment analyses were performed using the DAVID (Database for Annotation, Visualization and

[2] http://fantom.gsc.riken.jp/cat/

[3] http://noncode.org/datadownload/NONCODEv5_human_hg38_lncRNA.gtf.gz

[4] http://big.hanyang.ac.kr/UCSC/RNA-seq/hg19/CAFE/GTFs/BIGTranscriptome/BIGTranscriptome_lncRNA_catalog.gtf

[5] http://bigd.big.ac.cn/lncbook/index

[6] http://mitranscriptome.org/download/mitranscriptome.gtf.tar.gz

[7] https://lncipedia.org/downloads/lncipedia_5_2/full-database/lncipedia_5_2_hg38.gtf

[8] http://noncode.org/datadownload/NONCODEv5_mouse_mm10_lncRNA.gtf.gz

[9] https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v8.p2

[10] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[11] http://www.rstudio.com/

Integrated Discovery; version 6.8) tool (Huang et al., 2009) and the PANTHER ("Protein ANalysis THrough Evolutionary Relationships"; release 20181113) overrepresentation test (Mi et al., 2019) implemented in the Gene Ontology (GO) website (Ashburner et al., 2000; The Gene Ontology Consortium, 2019). The lists of PCGs containing a GC-AG intron from both human ($n$ = 1934) and mouse ($n$ = 1669) were subjected to an enrichment analysis on GO Biological Process terms and filtered applying a statistical significance threshold of 0.05 based on the multiple testing corrected $p$-values [i.e., Benjamini adjusted $p$-value in DAVID or false-discovery rates (FDR) in PANTHER].

## Availability of Data and Materials

This study was based on genomic data of lncRNAs and PCGs provided by the GENCODE Project. In particular, data of lncRNAs and PCGs were downloaded from the GENCODE web-pages for human[12] and mouse[13] as gtf files. We only analyzed anonymized samples for which the corresponding donor consent information was available in the GTEx dataset (dbGaP:phs000424.v8.p2) at the time of the analysis. Samples were downloaded from the dbGap database[14] according to the specified guidelines. All of the samples we analyzed were approved for General Research Use (GRU) and thus have no further limitations outside of those in the NIH model Data Use Certification Agreement.

The datasets supporting the conclusions of this article (i.e., introns data) are available in the GitHub repository[15].

# RESULTS

## Long Non-coding and Protein-Coding Genes Showed Differences in Their Gene Structure

As genomic organization and gene structure may affect gene expression regulation, we characterized the genomic features of human and mouse lncRNAs in comparison with PCGs. Our analysis, based on GENCODE human release 27 (15778 lncRNAs and 19836 PCGs) and mouse release M16 (12374 lncRNAs and 21963 PCGs), considered an increased number of genes with respect to previous studies (Cabili et al., 2011; Derrien et al., 2012). The total number of genes, transcripts and exons are reported in **Supplementary Table S1**.

The genomic organization of lncRNAs and PCGs appeared highly similar in both species. Human and mouse lncRNAs appeared equally transcribed from the forward and the reverse strand as PCGs (**Supplementary Table S2**) and almost homogeneously interspersed along chromosomes. Gene density resulted highly variable among chromosomes for lncRNAs and PCGs in both species (**Supplementary Table S3** and **Supplementary Figure S1**).

The genome coverage of long non-coding genes was found remarkably lower with respect to that of protein-coding ones. Indeed, long non-coding genes accounted for 12.5% of the human genome while 43.4% is occupied by PCGs (Chi-square test = 730.4, 1 df, $p$-value $< 2.2 \times 10^{-16}$). The reduced genome coverage was not entirely due to the smaller number of lncRNAs, as they account for about 80% of protein-coding ones, but it appeared to be due to the lncRNAs length, that resulted significantly lower than that of PCGs. Human lncRNAs resulted, on average, almost three times shorter than protein-coding ones with an average length of about 24 kb versus 68 kb, respectively (Wilcoxon test $p$-value $< 2.2 \times 10^{-16}$) (**Figure 1A** and **Supplementary Table S4**). Similarly, the genome coverage of mouse lncRNAs was lower (6.8%) than that of PCGs (39.2%) (Chi-square test = 802.5, 1 df, $p$-value $< 2.2 \times 10^{-16}$). The lower genome coverage was not only due to the smaller number of lncRNAs (accounting for 56% of PCGs) but also to their gene length that, as in human, resulted significantly shorter than that of PCGs with an average length of about 15 kb versus 49 kb, respectively (Wilcoxon test $p$-value $< 2.2 \times 10^{-16}$) (**Figure 1B** and **Supplementary Table S4**).

The shorter length of lncRNAs was attributable to the lower number of exons composing them (**Supplementary Table S5**). In human, more than 70% of lncRNA transcripts had 3 exons or less, compared with 16% of protein-coding transcripts bearing the same characteristics (Chi-squared test = 24407.0, 1 df, $p$-value $< 2.2 \times 10^{-16}$). 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 (Chi-squared test = 14613.7, 1 df, $p$-value $< 2.2 \times 10^{-16}$) and 24% of lncRNAs were single-exon genes versus 6.4% in protein-coding ones. Also in the mouse genome, an enrichment of 2-exons transcripts (30%) was observed (**Supplementary Table S5**). These results were confirmed using the FANTOM5 collection of human lncRNAs, an independent source for the annotation of lncRNAs (Hon et al., 2017) in which we observed the same trend for lncRNAs length (mean = 28.2 kb, $SEM$ = 458.4 bp) and lower number of exons (less than 3 exons: 56%).

A deep characterization of exons and introns length allowed us to appreciate differences between lncRNAs and PCGs. Conversely to what was previously reported in Derrien et al. (2012), our data revealed that first exons and especially last exons in lncRNAs are significantly shorter in both species (**Supplementary Table S6** and **Figures 1C,D**). LncRNA introns were found longer than PCGs ones when they are inner introns; instead, they resulted slightly shorter when they are first introns (**Supplementary Table S6** and **Figures 1E,F**).

Although it is possible that these differences were due to an incomplete annotation of lncRNAs, it is nevertheless

[12] https://www.gencodegenes.org/human/release_27.html
[13] https://www.gencodegenes.org/mouse/release_M16.html
[14] https://www.ncbi.nlm.nih.gov/gap
[15] https://github.com/laBione/introns_data

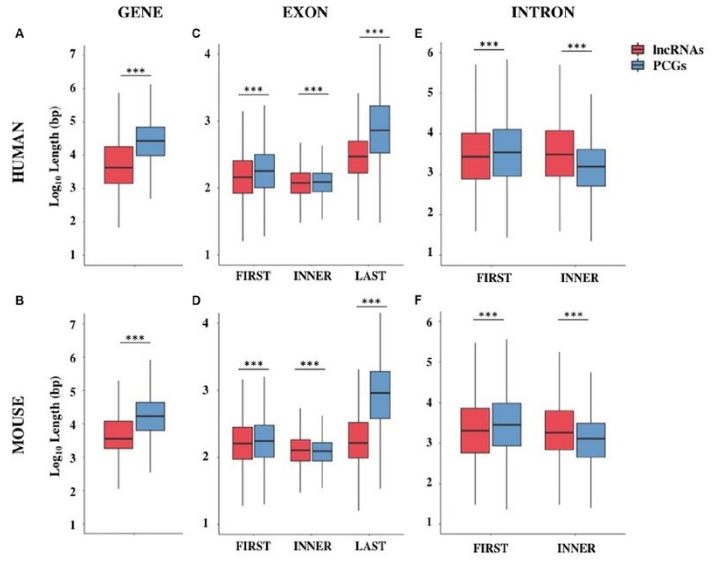


**FIGURE 1 |** Gene structure features of long non-coding and protein-coding genes (PCGs) in human and mouse. Boxplots showing the gene length **(A,B)**, exon length **(C,D)**, and intron length **(E,F)** in human and mouse, respectively. Data were presented as $\log_{10}$ of length in base pairs (bp). Exons were classified as first, inner or last and introns were classified as first or inner. ***$p < 0.001$.

interesting to note that the reduction in size affects those portions of the gene mainly involved in gene expression regulation.

## Different Assortment of Splicing Junctions Consensus in Long Non-coding RNAs

As splicing is a main determinant of post-transcriptional gene expression regulation, we characterized the splicing features of lncRNA introns in comparison with those of protein-coding ones.

The splice junctions sequence analysis highlighted differences between lncRNAs and PCGs consensus sequences (**Table 1**). The GC-AG splice junctions appeared strongly enriched in lncRNAs in which they represent 3.0% of the total splice junctions, thus almost four times more than in PCGs (0.8%) (Chi-square test = 2289.4, 1 df, *p*-value $< 2.2 \times 10^{-16}$). The same enrichment was found in mouse, in which GC-AG splicing junctions were more than the double with respect to protein-coding ones (lncRNAs: 1.9%, pc 0.8%) (Chi-square test = 380.2, 1 df, *p*-value $< 2.2 \times 10^{-16}$).

**TABLE 1 |** Number of different splice junctions consensus.

| | Human | | | |
|---|---|---|---|---|
| | #lncRNAs | % | #PCGs | % |
| GT-AG | 54667 | 96.6 | 517730 | 98.6 |
| GC-AG | 1683 | 3.0 | 4351 | 0.8 |
| Others | 232 | 0.4 | 3068 | 0.6 |
| Total | 56582 | | 525149 | |
| | **Mouse** | | | |
| | #lncRNAs | % | #PCGs | % |
| GT-AG | 28586 | 96.5 | 388973 | 98.8 |
| GC-AG | 570 | 1.9 | 3217 | 0.8 |
| Others | 455 | 1.5 | 1598 | 0.4 |
| Total | 29611 | | 393788 | |

GC-AG introns showed a preferential location in the first intron of both lncRNAs and PCGs (**Table 2**). Indeed, in the human genome, their percentage resulted higher in the first intron (lncRNAs: 4.2%; PCGs: 1.2%) with respect to inner

**TABLE 2** | Number of GC-AG introns in first or inner positions.

| | Human | | | | | |
|---|---|---|---|---|---|---|
| | lncRNAs | | | PCGs | | |
| | # | #GC-AG | % | # | #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 | | | | | |
| | lncRNAs | | | PCGs | | |
| | # | #GC-AG | % | # | #GC-AG | % |
| First | 13079 | 309 | 2.4 | 40990 | 472 | 1.2 |
| Inner | 16532 | 261 | 0.4 | 352798 | 2745 | 0.8 |
| Total | 29611 | 570 | 1.9 | 393788 | 3217 | 0.8 |

introns (lncRNAs: 2.1%; PCGs: 0.8%) and the same trend was observed in mouse (first: lncRNAs 2.4%, PCGs: 1.2%; inner: lncRNAs 0.4%, PCGs 0.8%). In all cases, differences were statistically significant (Chi-square tests = 204.7 and 120.9, 1 df, $p$-value $< 2.2 \times 10^{-16}$, respectively, for human lncRNAs and PCGs; Chi-square tests = 233.6 and 62.7, 1 df, $p$-value $< 2.2 \times 10^{-16}$ and $< 2.4 \times 10^{-15}$, respectively, for mouse lncRNAs and PCGs).

A validation of these results was obtained by investigating six alternative source of lncRNA annotations: (1) the FANTOM5 dataset (number of transcripts = 161340), (2) the NONCODE dataset (number of transcripts = 257020), (3) the BIGTranscriptome dataset (number of transcripts = 61018), (4) the LncBokk dataset (number of transcripts = 410630), (5) the MITranscriptome dataset (number of transcripts = 364544), and (6) the LNCipedia dataset (number of transcripts = 229818). In all datasets, the frequency of GC-AG splice junctions was found higher with respect to that in PCG introns together with their preferential location as the first intron (**Supplementary Table S7**). The prevalence of GC-AG introns among lncRNAs new datasets ranged from 2.3 to 3.5%, resulting in all cases significantly higher respect to PCGs (all comparisons $p$-values $< 2.2 \times 10^{-16}$). In all datasets, GC-AG introns showed their preferential localization in the first introns in which their prevalence in constantly the double with respect to inner introns (all comparisons $p$-values $< 2.2 \times 10^{-16}$). In mouse, data was replicated in the NONCODE dataset (**Supplementary Table S8**) in which both the enrichment of GC-AG splice junction (1.7% with respect to 0.8% in PCGs; $p$-values $< 2.2 \times 10^{-16}$) and their preferential location in the first intron (fist introns 2.3%, inner introns 1.1%; $p$-values $< 2.2 \times 10^{-16}$) was confirmed.

To evaluate the enrichment of GC-AG introns in lncRNAs during evolution, we analyzed the frequency of the different splice junctions in lower organisms as *D. melanogaster* and *C. elegans*. The ratio of GC-AG splice sites in lncRNAs of *D. melanogaster* was found significantly higher than in PCGs (GC-AG in lncRNAs: 1.7% of total splice junctions with respect to GC-AG in PCGs: 0.7%; Chi-square test = 57.0, 1 df, $p$-value = $4.3 \times 10^{-14}$). In *C. elegans*, GC-AG splice junctions account for 2.0% of total splice junctions in lncRNAs thus confirming the enrichment with respect to the 0.6% in PCGs (Chi-square test = 12.7, 1 df, $p$-value = $3.5 \times 10^{-4}$) (**Supplementary Table S9**). A preferential location of GC-AG splice sites in the first intron was also observed in lncRNA and PCGs of both *D. melanogaster* and *C. elegans* but due to their small number their statistical relevance could not be appreciated.

## Peculiar Features of GC-AG Introns in Long Non-coding and Protein-Coding Genes

The enrichment of GC-AG junctions in lncRNAs together with their preferential localization in first introns in both lncRNAs and PCGs suggested that they could play a particular role in gene expression regulation leading us to a deeper characterization of their features.

In human, GC-AG introns resulted shorter both in lncRNAs and PCGs and they showed the same trend whether they are first or inner introns (**Supplementary Table S10**). For GC-AG first introns, the average length resulted almost halved with respect to GT-AG first introns in both human lncRNAs and PCGs (lncRNAs: GC 6700 ±600 bp, GT 12923 ±201 bp, Wilcoxon tests $p$-value $< 2.2 \times 10^{-16}$; PCGs: GC 8999 ±648 bp, GT 15335 ±162 bp, Wilcoxon tests $p$-value $< 2.2 \times 10^{-16}$). Human GC-AG inner introns showed the same decrease in length, albeit to a lesser extent (lncRNAs: GC 8666 ±827 bp, GT 13995 ±194 bp, Wilcoxon tests $p$-value = 0.012; PCGs: GC 4165 ±197 bp, GT 5411 ±25 bp, Wilcoxon tests $p$-value = $6.3 \times 10^{-10}$). In mouse, GC-AG introns appeared shorter but only when they are inner introns (lncRNAs: GC 5190 ±734 bp, GT 7523 ±148 bp, Wilcoxon tests $p$-value = 0.0302; PCGs: GC 3186 ±192 bp, GT 4437 ±27 bp, Wilcoxon tests $p$-value = $9.5 \times 10^{-14}$). The shorter length of human GC-AG introns was also confirmed in the FANTOM5 dataset as both GC-AG first and inner introns of lncRNAs were significantly shorter than GT-AG ones (first intron: GC 8169 ±600 bp, GT 14516 ±137 bp, Wilcoxon tests $p$-value $< 2.2 \times 10^{-16}$; inner introns: GC 8648 ±544 bp, GT 15784 ±119 bp, Wilcoxon tests $p$-value $< 2.2 \times 10^{-16}$) (**Supplementary Table S11**).

To evaluate the splicing efficiency of GC-AG junctions, we computed their strength using the standard position weight-matrix (WM) model implemented in the MaxEntScan tool (Yeo and Burge, 2004), which assigns a computationally predicted score for 5′ and 3′ splice sites. Overall, the strength of 5′and 3′ss resulted lower in lncRNAs than in PCGs both in human and mouse (**Figure 2**, **Supplementary Table S12**, and **Supplementary Figure S2**) and it was presumably one of the causes of the previously reported inefficiency of lncRNAs splicing (Tilgner et al., 2012; Melé et al., 2017). Despite lower weight-matrix scores for 5′ss-GC were expected, due to their imperfect pairing with the U1 snRNA, 5′ss-GC scores of lncRNAs resulted strongly reduced with respect to 5′ss-GC of PCGs in both species (human: lncRNAs 5′ss-GC WM = 0.50, PCGs 5′ss-GC WM = 2.76,

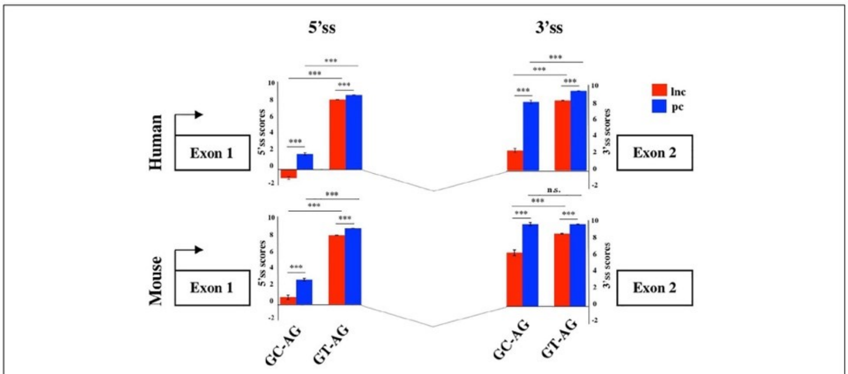


**FIGURE 2 |** Splice junctions strengths of the first introns. Schematic representation of the average scores of 5′ and 3′ss strengths of long non-coding and PCGs in human and mouse. The strengths of 5′ and 3′ ss were calculated as weight matrix scores for GC-AG and GT-AG first introns. ***$p < 0.001$.

Wilcoxon test $p$-value $< 2.2 \times 10^{-16}$; mouse: lncRNAs 5′ss-GC WM = 1.63, PCGs 5′ss-GC WM = 3.38, Wilcoxon test $p$-value $< 2.2 \times 10^{-16}$). The reduced strength of lncRNAs 5′ss-GC appeared to be attributable almost exclusively to first intron junctions, whose scores resulted lower compared to those of inner introns, both in human and mouse (human: lncRNAs first intron 5′ss-GC WM = −0.93, inner intron 5′ss-GC WM = 2.60, Wilcoxon test $p$-value $< 2.2 \times 10^{-16}$; mouse: lncRNAs first intron 5′ss-GC WM = 0.78, inner intron 5′ss-GC WM = 2.65, Wilcoxon test $p$-value $< 2.2 \times 10^{-16}$).

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 and this appeared attributable to their shorter polypyrimidine tracts (PPT) (**Supplementary Table S13**). In human, the mean length of PPT of GC-introns resulted significantly shorter than GT ones in both lncRNAs and PCGs (lncRNAs: GT-introns PPT mean = 16 bp, GC-introns PPT mean = 12 bp, Wilcoxon tests $p$-value $< 2.2 \times 10^{-16}$; PCGs: GT-introns PPT mean = 16 bp, GC-introns PPT mean = 15 bp, Wilcoxon tests $p$-value $< 2.2 \times 10^{-16}$). The same trend was observed in mouse for both gene classes (lncRNAs: GT-introns PPT mean = 15 bp, GC-introns PPT mean = 14 bp, Wilcoxon tests $p$-value = 0.001; PCGs: GT-introns PPT mean = 16 bp, GC-introns PPT mean = 15 bp, Wilcoxon tests $p$-value = 0.021). As it occurred for 5′ss, very weak 3′ss appeared preferentially located in the lncRNAs 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 (**Supplementary Table S14**). The strength of 5′ss and 3′ss was found positively correlated when located in the first intron of human lncRNAs ($r$ = 0.58, $p$-value $< 2.2 \times 10^{-16}$) and PCGs ($r$ = 0.22, $p$-value = $1 \times 10^{-16}$). In mouse, the correlation was significant only in lncRNAs (lncRNAs: $r$ = 0.51, $p$-value $< 2.2 \times 10^{-16}$; PCGs: : $r$ = 0.04, $p$-value = 0.34). The strengths of both 5′ss and 3′ss were positively correlated to intron length and this correlation was found more pronounced in the first intron in both species.

Differently from what was reported for PCGs, in which weak donor sites appeared flanked by stronger consensus at the acceptor sites (Thanaraj and Clark, 2001; Kralovicova et al., 2011), our analysis demonstrated that lncRNAs contained a class of very weak introns, preferentially located as first.

### GC-AG Introns in Alternative Splicing and Polyadenylation Regulation

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 diversity of both lncRNAs and PCGs owning at least one GC-AG intron with respect to the ones containing only GT-AG introns (**Supplementary Table S15**). In human, both long non-coding and protein-coding GC-AG-containing genes being transcribed in more than one isoform exceeded the number of GT-AG-containing genes [lncRNAs-GC n = 471 (38.5%) vs. lncRNAs-GT $n$ = 3204 (28.9%), Chi-square test = 47.7, 1 df, $p$-value = $4.8 \times 10^{-12}$; PCGs-GC $n$ = 1642 (84.9%) vs. PCGs-GT $n$ = 11469 (68.9%), Chi-square test = 212.5, 1 df, $p$-value $< 2.2 \times 10^{-16}$]. The same trend was confirmed in mouse (**Supplementary Table S15**), where long non-coding and protein-coding GC-AG-containing genes with more than one isoform resulted more abundant than their GT-AG counterpart [lncRNAs-GC $n$ = 188 (39.7%) vs. lncRNAs-GT $n$ = 2085 (25.3%), Chi-square test = 47.7, 1 df, $p$-value = $4.9 \times 10^{-12}$; PCGs-GC $n$ = 1117 (66.9%) vs. PCGs-GT $n$ = 9463 (50.2%), Chi-square test = 170.6, 1 df, $p$-value $< 2.2 \times 10^{-16}$]. To evaluate if the increase of alternative splicing could be attributed to some particular splicing events, we used the SUPPA2 tool (Trincado et al., 2018) to perform a quantitative profiling of alternative splicing events involving GC-AG introns in comparison with GT-AG ones (**Supplementary Table S16**). The analysis revealed that human GC-AG introns were preferentially involved in the alternative 5′ss events in both lncRNAs and PCGs [lncRNAs: $n = 150$ (18.9%) of GC-AG introns, $n = 3494$ (9.7%) of GT-AG introns, Chi-square test = 54.9, 1 df, $p$-value = $1.2 \times 10^{-13}$; PCGs: $n = 389$ (31.6%) of GC-AG introns, $n = 10500$ (10.1%) of GT-AG introns, Chi-square test = 415.6, 1 df, $p$-value $< 2.2 \times 10^{-16}$]. The same trend was also observed in mouse [lncRNAs: $n = 41$ (34.7%) of GC-AG introns, $n = 1{,}000$ (11.1%) of GT-AG introns, Chi-square test = 40.9, 1 df, $p$-value = $1.5 \times 10^{-10}$; PCGs: $n = 188$ (32.9%) of GC-AG introns, $n = 5902$ (12.5%) of GT-AG introns, Chi-square test = 136.7, 1 df, $p$-value $< 2.2 \times 10^{-16}$].

As alternative polyadenylation regulation is a process directly linked to 5′ss recognition and splicing, we analyzed the variability of last exon (LE) defining the total number of alternative last exon for each gene. In this analysis, we assessed an overall enrichment of LE variability in PCGs respect to lncRNAs in both species (**Supplementary Table S17**A). Interestingly, we observed a significant increase of LE variability in GC-AG-containing genes compared to GT-AG ones in both gene classes. In human lncRNAs, 37.6% of GC-AG-containing genes had more than one alternative last exon compared to 27.7% of GT-AG-containing genes (Chi-square test = 52.4, 1 df, $p$-value = $4.5 \times 10^{-13}$). The same difference was established for human PCGs (80% of GC-AG genes with alternative last exon versus 64.1% of GT-AG genes, Chi-square test = 151.4, 1 df, $p$-value $< 2.2 \times 10^{-16}$). The same significant enrichment were confirmed in mouse (lncRNAs: 38.3% of GC-AG genes with alternative last exon versus 23.5% of GT-AG genes, Chi-square test = 52.4, 1 df, $p$-value = $4.5 \times 10^{-13}$; PCGs: 59.3% of GC-AG genes with alternative last exon versus 43.7% of GT-AG genes, Chi-square test = 151.4, 1 df, $p$-value $< 2.2 \times 10^{-16}$) (**Supplementary Table S17B**). Furthermore, the increased of LE variability in GC-AG-containing genes was strengthened by a higher mean of alternative last exons per gene respect to GT-AG-containing genes in human and mouse lncRNAs and PCGs (**Supplementary Table S17C**). As differences in polyadenylation regulation could result from the different assortment of polyadenylation signals (PAS), we analyzed the last 40 nucleotides of each last exon for their content in the 16 different PAS reported in the paper of Beaudoing et al. (2000). Our results highlighted a higher ratio of lncRNAs lacking any of the 16 PAS considered compared with PCGs in both species (human: lncRNAs PAS = 0 48.0% versus PCGs PAS = 0 29.7%, Chi-square test = 2200.3, 1 df, $p$-value $< 2.2 \times 10^{-16}$; mouse: lncRNAs PAS = 0 42.4% versus PCGs PAS = 0 20.3%, Chi-square test = 2370.9, 1 df, $p$-value $< 2.2 \times 10^{-16}$) (**Supplementary Table S18A**). Considering GC-AG- and GT-AG-containing genes separately, we observed that the higher ratio of PAS = 0 was more evident in GC-AG transcripts but the difference was statistically significant only in human (**Supplementary Table S18B**). Looking at the assortment of different PAS, we observed a preferential usage of non-canonical PAS in lncRNAs with respect to PCGs in both species (human: lncRNAs non-canonical PAS 65.4% versus PCGs non-canonical PAS 57.6%, Chi-square test = 346.1, 1 df, $p$-value $< 2.2 \times 10^{-16}$; mouse: lncRNAs non-canonical PAS 65.8% versus PCGs non-canonical PAS 56.7%, Chi-square test = 309.2, 1 df, $p$-value $< 2.2 \times 10^{-16}$) (**Supplementary Table S18C**). No differences between GC-AG- and GT-AG-containing genes were observed in the usage of different PAS (data not shown).

Our results highlighted differences in alternative splicing and polyadenylation sites and signals between lncRNAs and PCGs which appeared more evident in GC-AG-containing genes thus suggesting that this 5′ss could contribute to gene expression regulation.

## Impact of GC-AG Introns on Gene Expression Level

In order to evaluate a putative effect of the presence of a GC-AG intron on the expression level of the corresponding transcripts, we analyzed a panel of ten different human tissues (i.e., anterior cingulate cortex, amygdala, cerebellum, heart, kidney, liver, lung, skin, spleen, and testis) obtained from the GTEx project. For each tissue, raw RNA-seq data from eight samples were processed using the Salmon tool (Patro et al., 2017) which provide an accurate quantification of transcripts expression. Transcript per million (TPM) of each single transcript, were calculated in each tissue and expressed transcripts were defined based on a threshold of TPM > 0.1 for lncRNAs and of TPM > 0.5 for PCGs to account for highly different level of expression between the two classes of genes. The percentage of expressed transcripts and the mean TPM in each tissue were reported distinguishing between GC-AG- or GT-AG-intron containing transcripts and between transcripts containing a GC-AG intron in the first or inner position (**Supplementary Table S19** and **Supplementary Figure S4**). In addition, we calculated the mean TPM of all tissues combined together in order to provide an overall estimation of expression data.

The mean TPM of GC-AG-containing transcripts appeared always lower with respect to GT-AG containing ones (with the exception of TPM values for lncRNAs in lung) and in the majority of the cases the difference resulted statistically significant. Combining all tissues together, the mean TPM of lncRNAs resulted significantly lower with respect to GT-AG-containing transcripts in both lncRNAs and PCGs (lncRNAs: 1.79 for GC-AG containing transcripts vs. 2.00 for GT-AG containing ones, Wilcoxon test $p$-value = $3.2 \times 10^{-15}$; PCGs: 8.40 for GC-AG containing transcripts vs. 11.10 for GT-AG containing ones, Wilcoxon test $p$-value $< 2.2 \times 10^{-16}$) (**Figure 3**).

The mean TPM of transcripts containing a GC-AG intron in the first position appeared always higher with respect to transcripts having a GC-AG intron in inner positions, both in lncRNAs and PCGs. Considering the combination of all tissues, the mean TPM of GC-first introns lncRNAs resulted significantly higher with respect to GC-inner introns (lncRNAs: GC-first mean TPM 2.00 vs. GC-inner mean TPM 0.58, Wilcoxon test $p$-value $< 2.2 \times 10^{-16}$; PCGs: GC-first mean TPM 7.71 vs. GC-inner mean TPM 6.04, Wilcoxon test $p$-value = $5.4 \times 10^{-11}$)

(**Figure 3**). Interestingly, in some cases the expression levels of transcripts with the GC-AG intron located as the first resulted higher than GT-AG-containing transcripts especially in lncRNAs (i.e., in anterior cingulate cortex, amygdala, lung, skin, and spleen in lncRNAs and in heart for PCGs) (**Supplementary Figure S4** and **Supplementary Table S19**).

These results suggest that the presence of a GC-AG intron may affect transcripts expression by reducing their overall transcription levels, both in lncRNAs and PCGs. Moreover, GC-AG introns may have a different effect on transcript expression levels depending on where they are located as transcripts harboring a GC-AG intron in their first intron showed overall higher expression levels with respect to transcripts with an inner GC-AG intron.

Nevertheless, these data must be taken with caution as: (i) the high variability in expression profiles, that is a common feature of both lncRNAs and PCGs, could affect mean TPM calculation especially for those categories containing a small number of transcripts, and (ii) transcripts containing a GC-AG intron often differ from GT-AG ones for other alternative splicing events which could possibly make differences in expression levels not univocally attributed to the presence of a GC intron.

## Multi-Species Conservation of GC-AG Introns

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

Based on the human-mouse ortholog information provided by the Ensembl project[16], a total of 908 PCGs were conserved between the two species, thus accounting for a considerable fraction of total GC-AG containing genes (47% of human GC-AG containing genes; 54% of mouse GC-AG containing genes). Remarkably, in more than 75% of cases the GC-AG introns also shared the same ordinal position in the homologous genes.

Interestingly, 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 mouse but it extended across evolutionary distant species. For example, the GC-AG splice sites of human *ABI3BP* (ABI family member 3 binding protein) and *NDUFAF6* (NADH:ubiquinone oxidoreductase complex assembly factor 6) genes were shown to be conserved in chimp, macaque, mouse, rat, dog, cow, pig, chicken, fugu, and zebrafish (**Figure 4**). Moreover, the ordinal position of the GC-AG intron was also conserved: in the *ABI3BP* gene, GC-AG introns was always the first intron in all cases and in the *NDUFAF6* gene, the GC-AG intron conserved its position in intron 6 in all species. The GC-AG splice sites of the human genes *BLVRB* (biliverdin reductase B) and *AZI2* (5-azacytidine induced 2) were shown to be conserved in first and inner introns of mammals, respectively, while the canonical GT was found in chicken, fugu and zebrafish (**Supplementary Figure S3**). Despite the assessment of the conservation of lncRNAs was hindered by the lack of annotation in most species, a number of conserved GC-AG splice junctions between human and mouse was determined. Indeed, the *TMEM51-AS1* (TMEM51 antisense RNA 1), the *MALAT1* (metastasis associated lung adenocarcinoma transcript 1) and the *NEAT1* (nuclear paraspeckle assembly transcript 1) genes contained a first GC-AG intron in both species whereas the *JPX* (JPX transcript, XIST activator) gene contained an inner GC-AG intron in both human and mouse.

The high conservation of the GC-AG introns between human and mouse and across multiple species could hint toward their functional importance and suggest their involvement in specific biological processes.

## Biological Processes Enrichment in GC-AG Containing Genes

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 PCGs. By means of the DAVID Functional Annotation Tool (Huang et al., 2009) and the PANTHER Overrepresentation Test (Mi et al., 2019), we selected only those terms that resulted significantly enriched in both species and by both tools (**Figure 5** and **Supplementary Table S20**).

This resulted in the identification of three groups of related terms in the biological process ontology. The first group comprised the GO term "microtubule-based movement" and its ancestors "movement of cell or subcellular component" and "microtubule-based process" and included 221 human and 176 mouse genes. Despite very little is known about the biological processes in which lncRNAs are involved, at least two of the GC-AG-containing lncRNAs were described to have a role in the regulation of the movement of cells or subcellular components: the *MEG3* (maternally expressed 3) gene (Wang et al., 2018; Xu et al., 2019) and the *SOX2-OT* (SOX2 overlapping transcript) gene (Wang et al., 2017). The second group contained the GO term "DNA Repair" and its ancestors "cellular response to DNA damage stimulus" and "cellular response to stress" 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" with its ancestors "neuron development," "generation of neurons," "neurogenesis," and "nervous system development" were included and contained 273 and 220 human and mouse genes. Several lncRNAs with 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-AS1* (GDNF antisense RNA 1) gene and the *MIAT* (myocardial infarction associated transcript) (Clark and Blackshaw, 2014). All the reported GO terms resulted

[16] https://www.ensembl.org/index.html

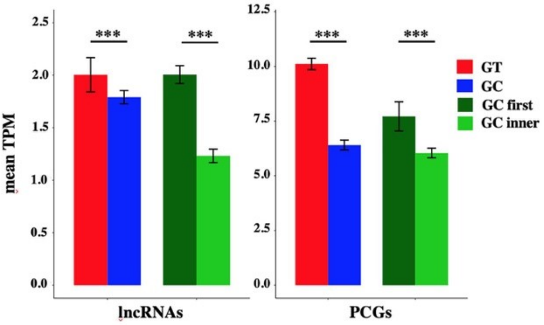
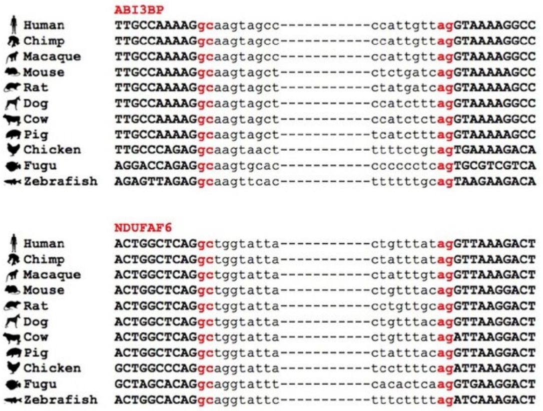


**FIGURE 3 |** Expression of GC-AG- and GT-AG-containing transcripts. Bar graph representing the expression of lncRNAs and PCGs transcripts containing GC-AG- or GT-AG-introns and between transcripts containing a GC-AG intron in the first or inner position. The expression of transcripts was calculated as mean TPM combining expression data from 10 different tissues together. $^{***}p < 0.001$.



```
            ABI3BP
Human       TTGCCAAAAGgcaagtagcc------------ccattgttagGTAAAAGGCC
Chimp       TTGCCAAAAGgcaagtagcc------------ccattgttagGTAAAAGGCC
Macaque     TTGCCAAAAGgcaagtagcc------------ccattgttagGTAAAAGGCC
Mouse       TTGCCAAAAGgcaagtagct------------ctctgatcagGTAAAAAGCC
Rat         TTGCCAAAAGgcaagtagct------------ctatgatcagGTAAAAAGCC
Dog         TTGCCAAAAGgcaagtagct------------ccatctttagGTAAAAGGCC
Cow         TTGCCAAAAGgcaagtagct------------ccatctctagGTAAAAGGCC
Pig         TTGCCAAAAGgcaagtagct------------tcatctttagGTAAAAAGCC
Chicken     TTGCCCAGAGgcaagtaact------------ttttctgtagTGAAAAGACA
Fugu        AGGACCAGAGgcaagtgcac------------cccccctcagTGCGTCGTCA
Zebrafish   AGAGTTAGAGgcaagttcac------------ttttttgcagTAAGAAGACA

            NDUFAF6
Human       ACTGGCTCAGgctggtatta------------ctgtttatagGTTAAAGACT
Chimp       ACTGGCTCAGgctggtatta------------ctatttatagGTTAAAGACT
Macaque     ACTGGCTCAGgctggtatta------------ctatttgtagGTTAAAGACT
Mouse       ACTGGCTCAGgctggtatta------------ctgtttacagGTTAAGGACT
Rat         ACTGGCTCAGgctggtatta------------cctgttgcagGTTAAGGACT
Dog         ACTGGCTCAGgctggtatta------------ctgtttacagGTTAAGGACT
Cow         ACTGGCTCAGgctggtatta------------ctgtttatagATTAAGGACT
Pig         ACTGGCTCAGgctggtatta------------ctatttacagGTTAAGGACT
Chicken     GCTGGCCCAGgcaggtatta------------tcctttcagATTAAAGACT
Fugu        GCTAGCACAGgcaggtattt------------cacactcaagGTGAAGGACT
Zebrafish   ACTGGCACAGgcaggtattc------------tttcttttagATCAAAGACT
```

**FIGURE 4 |** Conservation of GC-AG introns across multiple species. Multiple sequence alignment of GC-AG splice sites in the first intron of *ABI3BP* gene and the intron 6 of *NDUFAF6* gene across the 11 species indicated.

significantly enriched after correction for multiple testing (**Figure 5** and **Supplementary Table S20**).

## DISCUSSION

In this study, we report a genome-wide comparison of genomic and splicing features of long non-coding and PCGs 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 the genomic features revealed differences in the genetic architecture between long non-coding and PCGs in both human and mouse. We found that lncRNAs were shorter than protein-coding ones in both

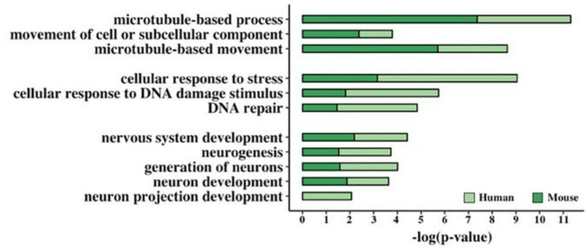
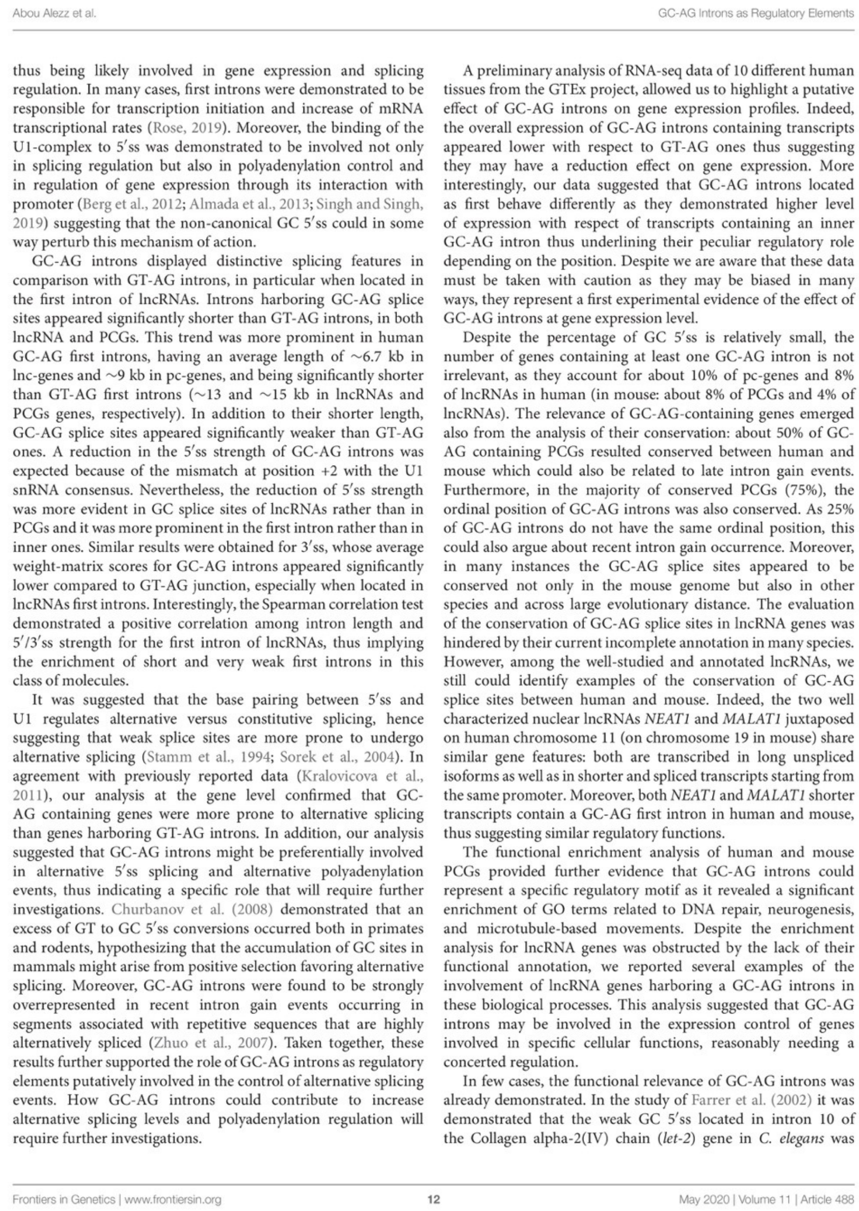


**FIGURE 5 |** Functional enrichment analysis of GC-AG-containing genes. Bar graph representing the GO terms found significantly enriched in GC-AG containing PCGs. The GO term name is indicated on the *Y*-axis, and the (–)log$_{10}$ of the *p*-values is indicated on the *X*-axis.

species in agreement with previous studies (Ravasi et al., 2005; Cabili et al., 2011; Derrien et al., 2012); however, this was not only due to the lower number of exons but also to the shorter length of exons in lncRNAs. The shorter length and the limited number of exons in lncRNA genes might be attributed to their incomplete annotation as their low expression level and high tissue specificity hampers the complete characterization, as suggested by the studies of Lagarde et al. (2016, 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) and they were confirmed in six more lncRNAs annotation datasets (FANTOM5, NONCODEv5, BIGTranscriptome, MiTranscriptome, LNCipedia, and LncBook). In particular, the FANTOM CAT robust lncRNA annotations specifically providing accurate annotations of transcripts' TSS and 5′ ends through the Cap Analyses of Gene Expression (CAGE) method, and the BIGTranscriptome dataset employing both CAGE and poly(A)-position profiling by sequencing (3P-seq) to assess 5′ and 3′ end completeness, indicated that our results are not subjected to the bias of incompleteness. It is nevertheless interesting to note that the reduction in size in lncRNAs affect those portions of the gene mainly involved in gene expression regulation. The length of the first exon was described to be related to transcription efficiency and it was reported that short first exons could promote transcriptional accuracy as they exhibit a more concentrated assembly of transcription factors near transcription start sites (Bieberstein et al., 2012). Moreover, last exons tend to be longer than first and inner exons due to the presence of 3′UTR sequences, essential for the regulation of multiple aspects including nuclear export, cytoplasmic localization, stability, and translational efficiency (Kalari et al., 2006; Andreassi and Riccio, 2009). Thus, our results hints toward a difference in the regulatory potential contained in the first and last exons of lncRNAs. Taken together, our data suggested that the difference in gene architecture between lncRNAs and PCGs could imply their involvement in different mechanisms of genomic control and gene expression 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% in human and mouse genomes, respectively (Sheth et al., 2006; Parada et al., 2014). In agreement with what was previously reported, we assessed the same frequency of GC-AG introns in both species when considering only PCGs (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, GC-AG first introns corresponded to 4.2 and 1.2% of total first introns of lncRNAs and PCGs, respectively. The same trend was observed in mouse in which a higher ratio of GC-AG splice junctions were found in the first intron in both lncRNAs (2.4%) and PCGs (1.2%). The enrichment of GC-AG introns in lncRNAs and their preferential position in the first intron did not appear to be driven by a mis-annotation bias as the same trend was also observed in the FANTOM5 dataset. The same enrichment was also assessed in the lower organisms *D. melanogaster* and *C. elegans*, despite this analysis could not be conclusive due to incomplete annotations and limited number. The significant increase of GC-AG introns in lncRNAs, 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 expression 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 (Bieberstein et al., 2012; Park et al., 2014),



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 (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 distinctive splicing features in comparison with GT-AG introns, in particular when located in the first intron of lncRNAs. Introns harboring GC-AG splice sites appeared significantly shorter than GT-AG introns, in both lncRNA and PCGs. 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 being significantly shorter than GT-AG first introns (∼13 and ∼15 kb in lncRNAs and PCGs 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 lncRNAs rather than in PCGs and it was more prominent in the first intron rather than in inner ones. Similar results were obtained for 3′ss, whose average weight-matrix scores for GC-AG introns appeared significantly lower compared to GT-AG junction, especially when located in lncRNAs first introns. Interestingly, the Spearman correlation test demonstrated a positive correlation among intron length and 5′/3′ss strength for the first intron of lncRNAs, 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 previously reported data (Kralovicova et al., 2011), our analysis at the gene level confirmed that GC-AG containing genes were more prone to alternative splicing than genes harboring GT-AG introns. In addition, our analysis suggested that GC-AG introns might be preferentially involved in alternative 5′ss splicing and alternative polyadenylation events, thus indicating a specific role that will require further investigations. 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 favoring alternative splicing. Moreover, GC-AG introns were found to be strongly overrepresented in recent intron gain events occurring in segments associated with repetitive sequences that are highly alternatively spliced (Zhuo et al., 2007). 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 polyadenylation regulation will require further investigations.

A preliminary analysis of RNA-seq data of 10 different human tissues from the GTEx project, allowed us to highlight a putative effect of GC-AG introns on gene expression profiles. Indeed, the overall expression of GC-AG introns containing transcripts appeared lower with respect to GT-AG ones thus suggesting they may have a reduction effect on gene expression. More interestingly, our data suggested that GC-AG introns located as first behave differently as they demonstrated higher level of expression with respect of transcripts containing an inner GC-AG intron thus underlining their peculiar regulatory role depending on the position. Despite we are aware that these data must be taken with caution as they may be biased in many ways, they represent a first experimental evidence of the effect of GC-AG introns at gene expression level.

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 lncRNAs in human (in mouse: about 8% of PCGs and 4% of lncRNAs). The relevance of GC-AG-containing genes emerged also from the analysis of their conservation: about 50% of GC-AG containing PCGs resulted conserved between human and mouse which could also be related to late intron gain events. Furthermore, in the majority of conserved PCGs (75%), the ordinal position of GC-AG introns was also conserved. As 25% of GC-AG introns do not have the same ordinal position, this could also argue about recent intron gain occurrence. 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 still could 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 PCGs 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 harboring 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, reasonably needing a concerted regulation.

In few cases, the functional relevance of GC-AG introns was already demonstrated. In the study of 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 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. An 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 PCGs 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 and Krause, 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 supporting the idea 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 number of alternative splicing events occurring at weak GC-AG introns (Abril et al., 2005; Churbanov et al., 2008).

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 further investigated despite preliminary evidence suggested that they could favor alternative splicing. The elucidation of the mechanisms of action of GC-AG introns could contribute to a deeper and 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.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

## AUTHOR CONTRIBUTIONS

SB conceived the project, set up the study design, and coordinated the work. MA contributed to conceive the study and performed almost all the analyses. LC performed the analysis of splice site strength of mouse introns. GB performed the analysis of splice junctions multi-species conservation. AL gave support for the statistical analysis and for the computational resources. SB and MA wrote the manuscript. LC, GB, and AL revised the manuscript.

## ACKNOWLEDGMENTS

We thank Prof. C. Calvio, Dr. S. Sabbioneda, Dr. R. Alfieri, and Dr. C. Mondello for critical reading of the manuscript. MA and LC are enrolled in the Ph.D. program in "Genetics, Molecular and Cellular Biology" of the University of Pavia, Italy.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00488/full#supplementary-material

## REFERENCES

Abril, J. F., Castello, R., and Guigò, R. (2005). Comparison of splice sites in mammals and chicken. *Genome Res.* 15, 111–119. doi: 10.1101/gr.3108805

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

Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B., and Sharp, P. A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* 499, 360–363. doi: 10.1038/nature12349

Anderson, K. M., Anderson, D. M., McAnally, J. R., Shelton, J. M., Bassel-Duby, R., and Olson, E. N. (2016). Transcription of the non-coding RNA upperhand controls Hand2 expression and heart development. *Nature* 539, 433–436. doi: 10.1038/nature20128

Andreassi, C., and Riccio, A. (2009). To localize or not to localize: mRNA fate is in 3'UTR ends. *Trends Cell Biol.* 19, 465–474. doi: 10.1016/j.tcb.2009.06.001

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Barry, G., Briggs, J. A., Hwang, D. W., Nayler, S. P., Fortuna, P. R. J., Jonkhout, N., et al. (2017). The long non-coding RNA NEAT1 is responsive to neuronal

activity and is associated with hyperexcitability states. *Sci. Rep.* 7:40127. doi: 10.1038/srep40127

Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M., and Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* 10, 1001–1010. doi: 10.1101/gr.10.7.1001

Berg, M. G., Singh, L. N., Younis, I., Liu, Q., Pinto, A. M., Kaida, D., et al. (2012). U1 snRNP determines mRNA length and regulates isoform expression. *Cell* 150, 53–64. doi: 10.1016/j.cell.2012.05.029

Bieberstein, N. I., Carrillo Oesterreich, F., Straube, K., and Neugebauer, K. M. (2012). First exon length controls active chromatin signatures and transcription. *Cell Rep.* 2, 62–68. doi: 10.1016/j.celrep.2012.05.019

Bush, S. J., Chen, L., Tovar-Corona, J. M., and Urrutia, A. O. (2017). Alternative splicing and the evolution of phenotypic novelty. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 372, 20150474. doi: 10.1098/rstb.2015.0474

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., et al. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–1927. doi: 10.1101/gad.17446611

Chen, C., Gao, S., Sun, Q., Tang, Y., Han, Y., Zhang, J., et al. (2017). Induced splice site mutation generates alternative intron splicing in starch synthase II (SSII) gene in rice. *Biotechnol. Biotechnol. Equip.* 31, 1093–1099. doi: 10.1080/13102818.2017.1370984

Chernikova, D., Managadze, D., Glazko, G., Makalowski, W., and Rogozin, I. (2016). Conservation of the exon-intron structure of long intergenic non-coding RNA genes in eutherian mammals. *Life* 6:27. doi: 10.3390/life6030027

Churbanov, A., Winters-Hilt, S., Koonin, E. V., and Rogozin, I. B. (2008). Accumulation of GC donor splice signals in mammals. *Biol. Direct.* 3:30. doi: 10.1186/1745-6150-3-30

Clark, B. S., and Blackshaw, S. (2014). Long non-coding RNA-dependent transcriptional regulation in neuronal development and disease. *Front. Genet.* 5:164. doi: 10.3389/fgene.2014.00164

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22, 1775–1789. doi: 10.1101/gr.132159.111

Deveson, I. W., Hardwick, S. A., Mercer, T. R., and Mattick, J. S. (2017). The dimensions, dynamics, and relevance of the mammalian noncoding transcriptome. *Trends Genet.* 33, 464–478. doi: 10.1016/j.tig.2017.04.004

Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., et al. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* 539, 452–455. doi: 10.1038/nature20149

Fang, S., Zhang, L., Guo, J., Niu, Y., Wu, Y., Li, H., et al. (2018). NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.* 46, 308–314. doi: 10.1093/nar/gkx1107

Farrer, T., Roller, A. B., Kent, W. J., and Zahler, A. M. (2002). Analysis of the role of *Caenorhabditis elegans* GC-AG introns in regulated splicing. *Nucleic Acids Res.* 30, 3360–3367. doi: 10.1093/nar/gkf465

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* 15, 7–21. doi: 10.1038/nrg3606

Fong, Y. W., and Zhou, Q. (2001). Stimulatory effect of splicing factors on transcriptional elongation. *Nature* 414, 929–933. doi: 10.1038/414929a

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 47, D766–D773. doi: 10.1093/nar/gky955

Fumasoni, I., Meani, N., Rambaldi, D., Scafetta, G., Alcalay, M., and Ciccarelli, F. D. (2007). Family expansion and gene rearrangements contributed to the functional specialization of PRDM genes in vertebrates. *BMC Evol. Biol.* 7:187. doi: 10.1186/1471-2148-7-187

GTEx Consortium. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P., and Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* 11, 1110–1122. doi: 10.1016/j.celrep.2015.04.023

Hon, C.-C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J. L., Gough, J., et al. (2017). An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* 543, 199–204. doi: 10.1038/nature21374

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

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* 47, 199–208. doi: 10.1038/ng.3192

Jandura, A., and Krause, H. M. (2017). The New RNA World: growing evidence for long noncoding RNA functionality. *Trends Genet.* 33, 665–676. doi: 10.1016/j.tig.2017.08.002

Kalari, K. R., Casavant, M., Bair, T. B., Keen, H. L., Comeron, J. M., Casavant, T. L., et al. (2006). First exons and introns – A survey of GC content and gene structure in the human genome. *In Silico Biol.* 6, 237–242.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., et al. (2004). The UCSC table browser data retrieval tool. *Nucleic Acids Res.* 32, 493D–496D. doi: 10.1093/nar/gkh103

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., et al. (2002). The human genome browser at UCSC. *Genome Res.* 12, 996–1006. doi: 10.1101/gr.229102

Kralovicova, J., Hwang, G., Asplund, A. C., Churbanov, A., Smith, C. I. E., and Vorechovsky, I. (2011). Compensatory signals associated with the activation of human GC 5' splice sites. *Nucleic Acids Res.* 39, 7077–7091. doi: 10.1093/nar/gkr306

Krchòáková, Z., Thakur, P. K., Krausová, M., Bieberstein, N., Haberman, N., Müller-McNicoll, M., et al. (2019). Splicing of long non-coding RNAs primarily depends on polypyrimidine tract and 5' splice-site sequences due to weak interactions with SR proteins. *Nucleic Acids Res.* 47, 911–928. doi: 10.1093/nar/gky1147

Lagarde, J., Uszczynska-Ratajczak, B., Carbonell, S., Pérez-Lluch, S., Abad, A., Davis, C., et al. (2017). High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat. Genet.* 49, 1731–1740. doi: 10.1038/ng.3988

Lagarde, J., Uszczynska-Ratajczak, B., Santoyo-Lopez, J., Gonzalez, J. M., Tapanari, E., Mudge, J. M., et al. (2016). Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput sequencing (RACE-Seq). *Nat. Commun.* 7:12339. doi: 10.1038/ncomms12339

Latos, P. A., Pauler, F. M., Koerner, M. V., Senergin, H. B., Hudson, Q. J., Stocsits, R. R., et al. (2012). Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science* 338, 1469–1472. doi: 10.1126/science.1228110

Liu, G., Mattick, J., and Taft, R. J. (2013). A meta-analysis of the genomic and transcriptomic composition of complex life. *Cell Cycle* 12, 2061–2072. doi: 10.4161/cc.25134

Long, Y., Wang, X., Youmans, D. T., and Cech, T. R. (2017). How do lncRNAs regulate transcription? *Sci. Adv.* 3:eaao2110. doi: 10.1126/sciadv.aao2110

Luo, S., Lu, J. Y., Liu, L., Yin, Y., Chen, C., Han, X., et al. (2016). Divergent lncRNAs regulate gene expression and lineage differentiation in pluripotent cells. *Cell Stem Cell* 18, 637–652. doi: 10.1016/j.stem.2016.01.024

Ma, L., Cao, J., Liu, L., Du, Q., Li, Z., Zou, D., et al. (2019). LncBook: a curated knowledgebase of human long non-coding RNAs. *Nucleic Acids Res.* 47, 128–134.

Mattick, J. S. (2018). The state of long non-coding RNA biology. *Non-Coding RNA* 4:17. doi: 10.3390/ncrna4030017

Melé, M., Mattioli, K., Mallard, W., Shechner, D. M., Gerhardinger, C., and Rinn, J. L. (2017). Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res.* 27, 27–37. doi: 10.1101/gr.214205.116

Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P. D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* 47, D419–D426. doi: 10.1093/nar/gky1038

Nitsche, A., Rose, D., Fasold, M., Reiche, K., and Stadler, P. F. (2015). Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved. *RNA* 21, 801–812. doi: 10.1261/rna.046342.114

Nitsche, A., and Stadler, P. F. (2017). Evolutionary clues in lncRNAs: evolutionary clues in lncRNAs. *Wiley Interdiscip. Rev. RNA*. 8:e1376. doi: 10.1002/wrna.1376

Palaniswamy, R., Teglund, S., Lauth, M., Zaphiropoulos, P. G., and Shimokawa, T. (2010). Genetic variations regulate alternative splicing in the 5' untranslated regions of the mouse glioma-associated oncogene 1, Gli1. *BMC Mol. Biol.* 11:32. doi: 10.1186/1471-2199-11-32

Papasaikas, P., and Valcárcel, J. (2016). The spliceosome: the ultimate RNA chaperone and sculptor. *Trends Biochem. Sci.* 41, 33–45. doi: 10.1016/j.tibs.2015.11.003

Parada, G. E., Munita, R., Cerda, C. A., and Gysling, K. (2014). A comprehensive survey of non-canonical splice sites in the human transcriptome. *Nucleic Acids Res.* 42, 10564–10578. doi: 10.1093/nar/gku744

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

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419.

Ponjavic, J., Ponting, C. P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* 17, 556–565. doi: 10.1101/gr.6036807

Ravasi, T., Suzuki, H., Pang, K. C., Katayama, S., Furuno, M., Okunishi, R., et al. (2005). Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.* 16, 11–19. doi: 10.1101/gr.4200206

Rose, A. B. (2019). Introns as gene regulators: a brick on the accelerator. *Front. Genet.* 9:72. doi: 10.3389/fgene.2018.00672

Schaefke, B., Sun, W., Li, Y.-S., Fang, L., and Chen, W. (2018). The evolution of posttranscriptional regulation. *Wiley Interdiscip. Rev. RNA* 9:e1485. doi: 10.1002/wrna.1485

Schüler, A., Ghanbarian, A. T., and Hurst, L. D. (2014). Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs. *Mol. Biol. Evol.* 31, 3164–3183. doi: 10.1093/molbev/msu249

Sheth, N., Roca, X., Hastings, M. L., Roeder, T., Krainer, A. R., and Sachidanandam, R. (2006). Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.* 34, 3955–3967. doi: 10.1093/nar/gkl556

Signal, B., Gloss, B. S., Dinger, M. E., and Mercer, T. R. (2018). Machine learning annotation of human branchpoints. *Bioinformatics* 34, 920–927. doi: 10.1093/bioinformatics/btx688

Sigova, A. A., Mullen, A. C., Molinie, B., Gupta, S., Orlando, D. A., Guenther, M. G., et al. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.* 110, 2876–2881. doi: 10.1073/pnas.1221904110

Singh, R. N., and Singh, N. N. (2019). A novel role of U1 snRNP: splice site selection from a distance. *Biochim. Biophys. Acta Gene Regul. Mech.* 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. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* 43, W589–W598. doi: 10.1093/nar/gkv350

Sorek, R., Lev-Maor, G., Reznik, M., Dagan, T., Belinky, F., Graur, D., et al. (2004). Minimal conditions for exonization of intronic sequences: 5'. Splice site formation in alu exons. *Mol. Cell* 14, 221–231. doi: 10.1016/s1097-2765(04)00181-9

Stamm, S., Zhang, M. Q., Marr, T. G., and Helfman, D. M. (1994). A sequence compilation and comparison of exons that are alternatively spliced in neurons. *Nucleic Acids Res.* 22, 1515–1526. doi: 10.1093/nar/22.9.1515

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

The Gene and Ontology Consortium. (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47, D330–D338. doi: 10.1093/nar/gky1055

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

Trincado, J. L., Entizne, J. C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D. J., et al. (2018). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* 19:40. doi: 10.1186/s13059-018-1417-1

Tsai, K., Tseng, H., and Lin, W. (2008). Two wobble-splicing events affect ING4 protein subnuclear localization and degradation. *Exp. Cell Res.* 314, 3130–3141. doi: 10.1016/j.yexcr.2008.08.002

Uesaka, M., Nishimura, O., Go, Y., Nakashima, K., Agata, K., and Imamura, T. (2014). Bidirectional promoters are the major source of gene activation-associated non-coding RNAs in mammals. *BMC Genomics* 15:35. doi: 10.1186/1471-2164-15-35

Volders, P. J., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., et al. (2013). LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.* 41, 246–251. doi: 10.1093/nar/gks915

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

Wang, Z., Tan, M., Chen, G., Li, Z., and Lu, X. (2017). LncRNA SOX2-OT is a novel prognostic biomarker for osteosarcoma patients and regulates osteosarcoma cells proliferation and motility through modulating SOX2: the role of LNC RNA SOX2-OT in osteosarcoma. *IUBMB Life* 69, 867–876. doi: 10.1002/iub.1681

Xu, D., Chi, G., Zhao, C., and Li, D. (2019). 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.* 120, 7516–7526. doi: 10.1002/jcb.28026

Yeo, G., and Burge, C. B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* 11, 377–394. doi: 10.1089/1066527041410418

You, B. H., Yoon, S. H., and Nam, J. W. (2017). High-confidence coding and noncoding transcriptome maps. *Genome Res.* 27, 1050–1062.

You, D., and You, H. (2019). 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.* 111, 1447–1457. doi: 10.1016/j.biopha.2018.12.067

Zhuo, D., Madden, R., Elela, S. A., and Chabot, B. (2007). Modern origin of numerous alternatively spliced human introns from tandem arrays. *Proc. Natl. Acad. Sci. U.S.A.* 104, 882–886. doi: 10.1073/pnas.0604777104

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# *Acknowledgements*

*" If I have seen further, it is by standing on the shoulders of giants."*
Sir Isaac Newton (1643 – 1727)

"Goodness can never be defied and good human beings can never be denied." Undertaking this PhD has been a truly life-changing experience. A doctoral thesis is often described as a solitary endeavor; however it would not have been possible to do without the support and guidance from several people. This thesis represents not only my work at the keyboard; it is a milestone in 3 years of work at CNR and specifically within the Computational Biology Laboratory. Although it is just my name on the cover, many people have contributed to the research in their own particular way and for that I want to give them special thankfulness.

First and foremost, my deepest gratitude goes to my supervisor, Prof. Silvia Bione. I would like first to thank her for introducing me to the exciting field of Bioinformatics during my Masters studies in which not only she taught me the state-of-art of Bioinformatics but also she ignited the spark of curiosity in this research field upon which I decided to build on my career. I am deeply grateful to her, as following my Masters graduation, she believed enough in what she saw in a young graduate to accept becoming my supervisor. She has supported me not only by providing a research assistantship over almost 3 years, but also by creating the invaluable space for me to do this research and develop myself as a researcher in the best possible way. Throughout my PhD, her enthusiasm, integral view on research and her mission for providing high-quality work, has made a deep impression on me. I have learnt extensively from her, including how to raise new possibilities, how to regard an old question from a new perspective, how to approach a problem by systematic thinking, data-driven decision making and exploiting serendipity. I owe her lots of gratitude for her dedicated help, advice, inspiration, encouragement and continuous support.

I greatly appreciate the support received from our staff and team members Antonella, Ornella, Daniela and Roberta for their kind support throughout my project and who were always ready to give their timely help whenever required. I gratefully acknowledge their moral support, constant cooperation, personal helps and friendly nature that made my working environment feel like home.

While supervision, discussion and inspiration have all been important for the materialization of this thesis, some people have directly contributed to its contents, namely my students who worked on the empirical projects. I am grateful to my first student, colleague and friend Ludovica for her hard work, enthusiasm, creativity and joyful spirit. Many thanks also to my students Giulia Nicole, Lorenzo, Giulia, Sofia and Arianna for your excellent work, precious help, friendliness, enthusiasm, for all the laughs, chats, coffee breaks and meals together.

I would also like to say a heartfelt thank you to my lab colleagues and friends Leonida and Umberto for their friendly support, help, and joyful times throughout these years. A special mention of thanks to all my friends in Pavia too.

As always, it is impossible to mention everybody who had an impact to this work; however there are those whose spiritual support is even more important. My acknowledgement will never be complete without the special mention of my deep sense of gratitude for my mother, father, Remy and Morad who formed part of my vision and taught me good things that really matter in life. Their infallible love and support has always been my strength. Their patience and sacrifice will remain my inspiration throughout my life.

*Monah Abou Alezz*

*Pavia 28/09/2020*