Characterization of the genomic and splicing features of long non-coding RNAs using bioinformatics approaches

Abou Alezz M, Celli L, Belotti G, Bione S.
Institute of Molecular Genetics L.L Cavalli-Sforza - National Research Council (IGM-CNR) - Pavia, Italy

INTRODUCTION

Long non-coding RNAs (lncRNAs) are recognized as an important class of genes required for gene expression regulation. They are characterized by their high tissue specificity, low level of expression and low sequence conservation. However, lncRNAs show high conservation at their exon-intron boundaries and their splice sites indicating an importance of their correct splicing. Indeed, the splicing and promoter-proximal splice sites of lncRNAs has been shown to enhance the transcription of their neighboring protein-coding (pc) genes. Initial studies reported that lncRNAs show an overall splicing inefficiency but the mechanisms of this remain poorly understood.

OBJECTIVES

LncRNAs' role in regulating gene expression is well documented but little is known about the transcriptional and post-transcriptional regulation of lncRNAs. While the recognition of lncRNAs intron boundaries and their correct splicing is crucial step for their function, their splicing features are poorly characterized. The objectives of this study are to:

a) **characterize the main genomic features of lncRNA in terms of their gene structure**
b) **Identify the chromatin states and combinatorial histone marks on lncRNAs**
c) **characterize the splicing features of lncRNAs in terms of their splice junctions, intron length, and splice site strength.**

1. LncRNAs show differences in their gene structure

As genomic organization and gene structure may affect gene expression regulation, we characterize genomic features of human lncRNAs in comparison to protein-coding genes. LncRNA genes was significantly shorter than pc ones and significant differences were appreciated in the first and last exons in addition to the first and inner introns of lncRNAs. Interestingly, the reduction in size affect the portions of genes mainly involved in gene regulation. Moreover, unlike what was described for pc genes in which first introns are longer than inner introns, lncRNA first and inner introns appeared similar in length.

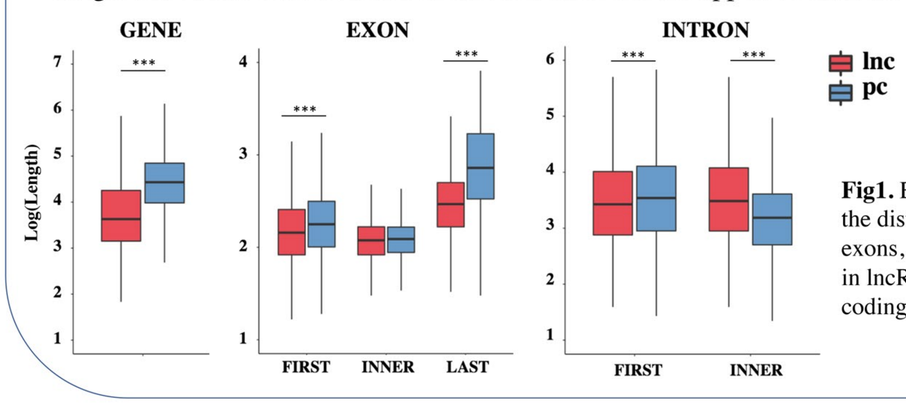

Fig1. Boxplot showing the distribution of gene, exons, and introns length in lncRNA and protein-coding genes

2. LncRNAs show differences in their chromatin states

LncRNAs showed a difference in their epigenetic profiles at the genes, exons and introns level in comparison to protein-coding ones. Using ChromHMM, seven chromatin states has been identified for six histone marks in H1 cell line. LncRNA introns and exons showed an overlapping enrichment with the H2A.Z histone modification described to be associated with promoter regions and required for the splicing of weak introns.

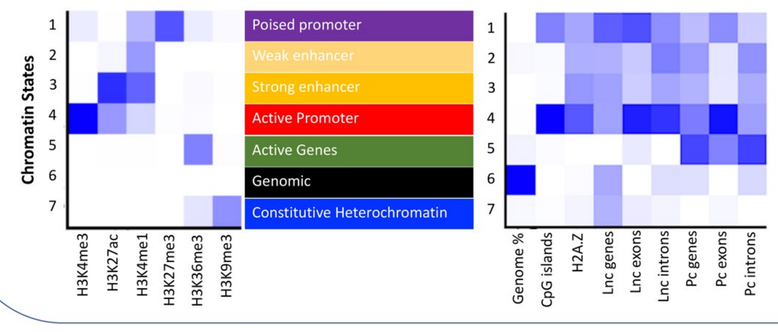

Fig2. Heatmap of 7 chromatin states defined from multiple histone modifications in H1 cells

3. Enrichment of GC-AG introns in lncRNAs

As splicing is a main determinant of post-transcriptional gene expression regulation, we characterized the splicing features of lncRNA introns. The GC-AG splice junctions appeared strongly enriched in lncRNAs representing 3.0% of total splice sites, thus almost four times more than in pc genes (0.8%). Moreover, we notice that GC-AG introns were preferentially located as first introns.

Table 1. Number of different splice sites in lncRNAs and pc genes

	lncRNAs		Protein-coding	
Splice Sites	**Total**	**%**	**Total**	**%**
GT-AG	54667	96.6	517730	98.58
GC-AG	**1683**	**3**	**4351**	**0.8**
AT-AC	9	0.0	583	0.1
Others	223	0.4	2485	0.4
Total	56582		525149	

4. GC-AG introns have a shorter length

The enrichment of GC-AG junctions in lnc genes with their preferential localization in first introns suggested that they could play a particular role in gene expression regulation. GC-AG introns resulted shorter than GT-AG introns both in lncRNA and pc genes and they showed the same trend whether they are first or internal introns.

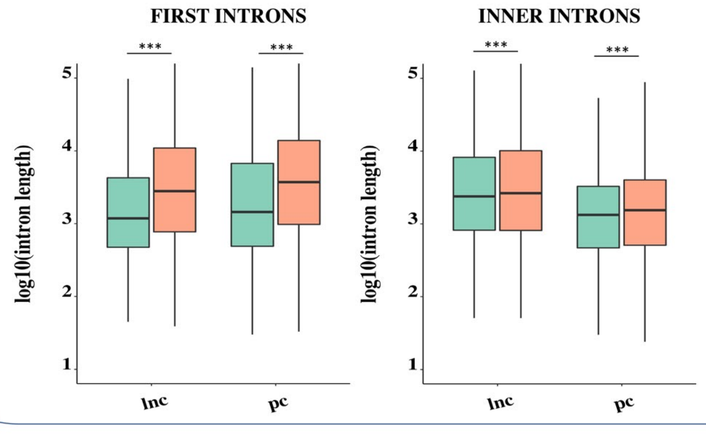

Fig3. Boxplot showing the distribution of GC-AG and GT-AG introns length as first or inner introns in lncRNA and protein-coding genes

5. GC-AG introns show weak donor and acceptor splice site strength in lncRNAs

The strength of 5'and 3'ss appeared lower in lncRNA than in pc genes, presumably one of the causes of the previously reported inefficiency of lncRNA splicing. Although the 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 respect to 5'ss-GC of pc genes in first introns. Despite owning the same consensus sequence, the 3'ss average weight-matrix scores for GC-AG introns appeared overall lower with respect to GT-AG acceptor sites, due to their shorter polypyrimidine tracts. Moreover, the strength of 5'ss and 3'ss was found positively correlated when located in the first intron of lncRNAs ($r = 0.58$, p-value $< 2.2 \times 10^{-16}$).

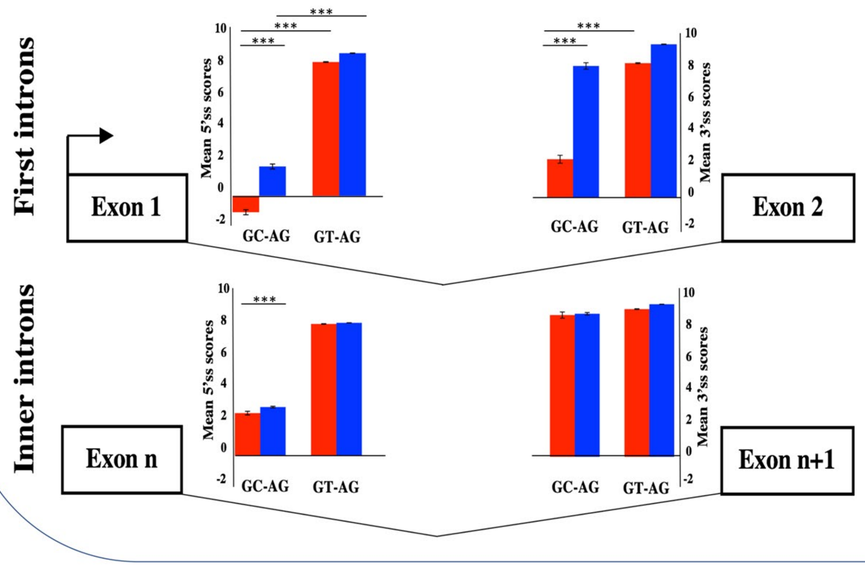

Fig4. Schematic presentation of donor and acceptor splice sites scores of GC-AG and GT-AG introns of lncRNA and pc

6. GC-AG introns show high conservation across multi-species

GC-AG introns together with their relative position inside the gene showed high conservation that extended across evolutionary distant species. For example, the GC-AG splice sites of the first intron of human ABI family member 3 binding protein (ABI3BP) appeared to be conserved in several vertebrates species. The GC-AG splice sites of the inner intron of ceramide kinase like (CERKL) gene appeared conserved in mammals while being GT-AG in chicken, fugu and zebrafish. Indeed, previous studies reported that during evolution there is a trend toward accumulation of GC splice sites and the conversion from GT to a GC splice site is highly tolerated.

Fig5. Alignment of donor and acceptor splice sites consensus sequences of the human ABI3BP and CERKL genes across vertebrate species

7. Biological process enrichment of GC-AG genes

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 pc genes. This resulted in the identification of three groups of linked terms in the biological process ontology: "microtubule-based movement", "DNA-repair" and "neuron projection development".

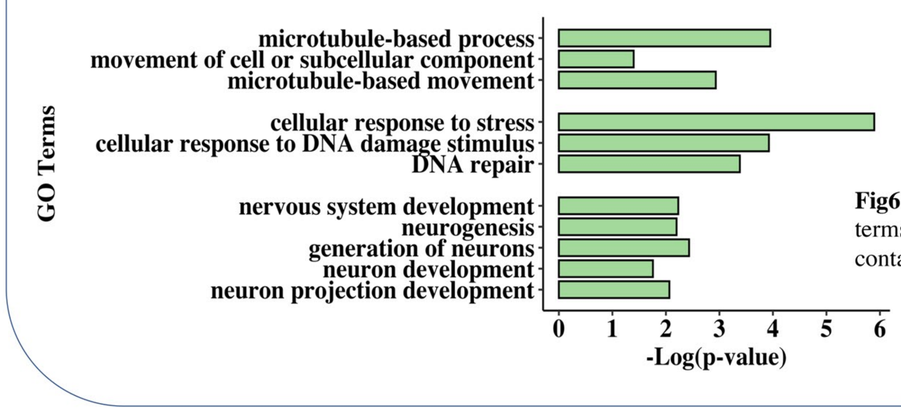

Fig6. Significant GO terms enriched in GC-AG containing genes

CONCLUSION

- LncRNAs show unique features in gene structure, chromatin modifications and splicing
- GC-AG introns may represent a new regulatory motif more abundant in lncRNAs and with important functional aspect
- The elucidation of the mechanisms of action of GC-AG introns would contribute to better understanding of gene expression regulation and comprehension of pathological effects of their mutation
- **GC-AG increased frequency in higher organisms suggest they could contribute to evolution complexity adding a new layer in gene expression regulation.**

Take a picture to download the full paper