RESEARCH ARTICLE

Distinct patterns of epigenetic marks and transcription factor binding sites across promoters of sense-intronic long noncoding RNAs

SOURAV GHOSH[1,2], SATISH SATI[1], SHANTANU SENGUPTA[1,2]* and VINOD SCARIA[2,3]*

[1]*Genomics and Molecular Medicine, CSIR Institute of Genomics and Integrative Biology, Mall Road, New Delhi 110 007, India*
[2]*Academy of Scientific and Innovative Research (AcSIR), Mathura Road, New Delhi 110 025, India*
[3]*GN Ramachandran Knowledge Center for Genome Informatics, CSIR Institute of Genomics and Integrative Biology, Mall Road, New Delhi 110 007, India*

Abstract

Long noncoding RNAs (lncRNAs) are a new class of noncoding RNAs that have been extensively studied in the recent past as a regulator of gene expression, including modulation of epigenetic regulation. The lncRNAs class encompasses a number of subclasses, classified based on their genomic loci and relation to protein-coding genes. Functional differences between subclasses have been increasingly studied in the recent years, though the regulation of expression and biogenesis of lncRNAs have been poorly studied. The availability of genome-scale datasets of epigenetic marks has motivated us to understand the patterns and processes of epigenetic regulation of lncRNAs. Here we analysed the occurrence of expressive and repressive histone marks at the transcription start site (TSS) of lncRNAs and their subclasses, and compared these profiles with that of the protein-coding regions. We observe distinct differences in the density of histone marks across the TSS of a few lncRNA subclasses. The sense-intronic lncRNA subclass showed a paucity for mapped histone marks across the TSS which were significantly different than all the lncRNAs and protein-coding genes in most cases. Similar pattern was also observed for the density of transcription factor binding sites (TFBS). These observations were generally consistent across cell and tissue types. The differences in density across the promoter were significantly associated with the expression level of the genes, but the differences between the densities across long noncoding and protein-coding gene promoters were consistent irrespective of the expression levels. Apart from suggesting general differences in epigenetic regulatory marks across long noncoding RNA promoters, our analysis suggests a possible alternative mechanism of regulation and/or biogenesis of sense-intronic lncRNAs.

[Ghosh S., Sati S., Sengupta S. and Scaria V. 2015 Distinct patterns of epigenetic marks and transcription factor binding sites across promoters of sense-intronic long noncoding RNAs. *J. Genet.* **94**, 17–25]

Introduction

A recent large-scale transcriptome analysis has revealed that a large proportion of human genome is transcriptionally active (Bernstein *et al.* 2012). Reports suggest that this is consistent across other vertebrate genomes as well (Kaushik *et al.* 2013). The discovery of novel regions in the genome with transcriptional potential has been primarily brought about by the advancement in sequencing technology to annotate transcriptomes (Morozova *et al.* 2009). Large proportions of these transcripts do not show any potential to encode functional proteins so are generally called noncoding RNA. Theirfunctional roles and regulations are not well understood and are frequently referred to as the dark-matter in the genome (Zhang *et al.* 2007). The noncoding transcripts are categorized and functionally annotated into two classes: small and long, based on the length of the mature/processed RNAs (Bhartiya *et al.* 2012). Small noncoding RNAs are extensively studied and much is understood about their regulatory roles (He and Hannon 2004; Fasanaro *et al.* 2013), e.g. microRNA. LncRNAs, by definition encompass transcripts that are >200 nucleotides in length and do not have an open reading frame (ORF) which encodes peptides having more than 30 amino acids (Liao *et al.* 2011). The lncRNAs further comprise of functionally distinct subclasses, categorized by their loci of origin and relationship with the protein-coding genes (Bhartiya *et al.* 2012). This includes previously annotated classes of regulatory RNAs like antisense, intronic, long intergenic noncoding

*For correspondence. E-mail: Shantanu Sengupta, shantanus@igib.res.in; Vinod Scaria, vinods@igib.in.
Sourav Ghosh and Satish Sati contributed equally to this work.
VS and SSe designed the analysis pipeline and coordinated the study. VS and SSe wrote the manuscript. Bioinformatics analysis was carried out by VS, SG and SSe.

Keywords. epigenetics; lncRNA; histone modifications; sense-intronic.

RNA etc., that are distinct in their genomic locations and possible functional roles in regulation.

Though a large number of lncRNAs do not exhibit any annotated functional role, a small subset of lncRNAs have been functionally studied in detail (Ji *et al*. 2003; Tsai *et al*. 2010; Yap *et al*. 2010; Sati *et al*. 2012a). This includes the involvement in regulation of coding genes both in *cis* and *trans* (Sati *et al.* 2012a). Presently, a number of lncRNAs are known to be associated with various diseases, some of them directly involved in the pathophysiology of disease (Wapinski and Chang 2011; Rajpathak *et al.* 2014). Recent genomewide association studies have also shown that a number of lncRNAs could provide clues to the genetic predisposition of diseases (Pasmant *et al*. 2011; Wapinski and Chang 2011). The lncRNAs mediate their function through biomolecular interactions with other molecules in the cell, like DNA, other transcripts, proteins and protein-DNA complexes (Bhartiya *et al.* 2012; Guttman and Rinn 2012). The specificity, interaction and partners of each lncRNAs vary depending on the core functionality of the lncRNAs (Guttman and Rinn 2012). One of the prominent roles of lncRNAs is with relation to epigenetic regulation of protein-coding genes. A subset of lncRNAs has been extensively studied in relation to the epigenetic regulation of genes by interacting with and recruiting members of the chromatin remodelling complex (Rinn *et al.* 2007; Mercer *et al.* 2009; Ponting *et al.* 2009; Saxena and Carninci 2011). Our group has previously demonstrated the potential of long noncoding RNAs to encode small regulatory RNAs, thus participating in the intricate cellular RNA regulatory network (Sati *et al.* 2012a).

Although, the role of lncRNAs mediated epigenetic regulation has been well studied, the role of epigenetic modifications in the regulation of lncRNA expression is poorly understood. We have recently demonstrated that epigenetic marks including DNA methylation and some histone modifications are differentially distributed across promoters of protein-coding and lncRNA genes, suggesting that epigenetic mechanisms of lncRNAs regulation could be potentially different from protein-coding genes (Sati *et al.* 2012a). Since lncRNAs encompass distinct subtypes in relation to the genomic origin and function, we hypothesized that the epigenetic marks across promoters of lncRNA subclasses could be distinct, as they show potentially distinct biological functions which necessitate finer regulation of expression. In this study, we describe the first comprehensive genome-scale analyses of chromatin marks across promoters of lncRNAs. Analysis of distinct pattern of epigenetic marks across the promoters, signified by paucity of all the marks considered in this analysis suggests a possible alternative mechanism of regulation and/or biogenesis of sense-intronic lncRNAs.

Material and methods

LncRNA annotation datasets

The protein-coding, lncRNA annotation datasets and genomic coordinates of the lncRNAs were derived from the Gencode annotation (Gencode data ver. 9, http://www.gencodegenes.org/releases/9.html). This publicly available genes datasets and transcripts are based on the large-scale annotation of the transcriptome using RNA-sequencing approach from the ENCODE consortium (http://www.genome.gov/encode/). The Gencode ver. 9 dataset is comprised of genomic coordinate of 20,012 protein-coding genes and 11,004 lncRNA genes. The lncRNAs further subclassified into antisense ($N = 3588$), lincRNA ($N = 5890$), processed transcript ($N = 1117$) and sense-intronic ($N = 409$) based on the annotations provided.

Datasets of chromatin modifications

We have used six different histone modifications (H3K4me3, H3K36me3, H3K9me3, H3K27me3, H3K4me1 and H3K9ac) for H1 human embryonic stem cells (H1cell), brain germinal matrix tissue (BrGr), IMR90 embryonic cells (IMR90), CD34 primary cells (CD34), liver control tissue (liver), peripheral blood mononuclear primary cells (PBMC) from NIH Roadmap Epigenomics project (http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/). The read map pings were made available as BED files which were downloaded independently for each corresponding cell/tissue types.

Transcription factor binding sites (TFBS)

Conserved transcription factor binding sites that were predicted on the human genome were downloaded from the University of California Santa Cruz Genome Bioinformatics site (http://genome.ucsc.edu). These datasets comprised of 5,797,267 conserved sites in the human genome. All genomic coordinates are based on the hg19 build.

Analysis of chromatin modification datasets

We used Model-based analysis for Chip-Seq (MACS) (ver. 1.4.0 beta) for peak detection and analysis of immunoprecipitated sequencing data to find the genomic regions that are enriched in a pool of specifically precipitated DNA fragments. Prealigned reads were downloaded from the NIH Epigenomics Roadmap project website as BED files. MACS was used with default parameters on BED formatted files of read alignments of histone modification datasets (H1 cells, BrGr, liver tissue, IMR90 cells, PBMCs, CD34 primary cells) downloaded from the NIH Roadmap Epigenomics project and enriched peaks were generated.

Data integration and analysis

Downstream analysis, data integration and comparison were performed by in-house scripts using Perl. The histone modification summit files generated by MACS were then used for further in-depth analysis. The summit files of histone modification data were used for analysing the differential patterns across 5 kb upstream and downstream of TSS of protein coding, lncRNA and all four sub-lncRNA genes. We used all

six histone modification (H3K4me3, H3K36me3, H3K9me3, H3K27me3, H3K4me1 and H3K9ac) summit files of H1 cell line to determine their occurrence in 100 bins (100 bp each) around 5 kb upstream and downstream of TSS of each gene of protein coding and lncRNA.

Conserved TFBS were downloaded from UCSC Bioinformatics website using the conserved TFBS track. The dataset comprised 5,797,267 conserved TFBS. The conserved TFBS loci were plotted across the TSS of protein coding as well as lncRNA genes. Additionally, we plotted the occurrence of conserved TFBS across TSS of different classes of lncRNAs. We evaluated the conserved TFBS sites and their overlaps in protein coding as long noncoding datasets using custom scripts. For this comparison, the core promoter was defined as the region −5000 to +5000 bases across the TSS of both protein-coding and lncRNAs. This definition was based on finding that the epigenetic marks are enriched for this region across the TSS is both protein-coding as well as noncoding RNA genes. Further, we mapped those conserved TFBS sites on protein coding and lncRNA genes on 20 kb upstream and downstream of TSS, for checking the distal regulatory effects.

Analysis of expression data

The RNAseq reads for H1 cell line available at the Human Epigenome RoadMap Project website were mapped and analysed using TopHat. Two classes of genes were selected as compared to the fragments mapped per kilobase exon per million reads (FPKM) numbers of the H1 cell line RNA-seq data. Briefly, those genes whose corresponding $\log_{10}$(FPKM) number was $>$ or $=$ mean+1SD of all $\log_{10}$(FPKM) were considered as highly expressed genes and those, $<$ or $=$ mean−1SD of all $\log_{10}$(FPKM) were considered as low expressed genes for both protein-coding and lncRNAs genes.

Results

Summary of datasets, mappings and sources

Data corresponding to various histone modifications were analysed to generate the comparative profiles around the TSS of protein-coding and lncRNA genes in different cell and tissue types. Briefly, the appropriate raw sequences were downloaded. The dataset used in this analysis comprised ChIP-Seq data for H3K4me3, H3K36me3, H3K9me3, H3K27me3, H3K4me1 and H3K9ac histone modifications. The data for histone modifications were downloaded for four different cell types (H1, IMR90, CD34 primary cells and peripheral blood mononuclear cells) and two different tissue types (liver and brain) from NIH Roadmap Epigenomics project. We could not access the datasets of brain tissue for H3K4me1 and H3K9ac modifications. Similarly, H3K4me3 and H3K9ac datasets were not available for PBMCs and CD34+ cells, respectively. Thus, these datasets were not considered for further analysis. We downloaded the aligned reads from the NIH RoadMap to Epigenomics website. We used the Human genome build (hg19) from UCSC Genome Bioinformatics site as the reference. The alignments were further processed using a model-based approach implemented in model-based analysis for ChIP Seq (MACS) as previously discussed (Sati *et al*. 2012a, b). Peaks were used independently for each of the ChIP-Seq datasets corresponding to each histone modifications. Table 1, A and B summarizes the datasets used in this analysis and provides a brief outline of the sites/peaks identified. Analysis revealed comparable number of peaks in each of the datasets. Datasets were compiled and further processed for in-depth analysis based on their distribution across genomic loci of protein coding as well and nonprotein coding loci.

Table 1 (A)**.** Summary and number of distinct entities in each of the datasets considered in the analysis.

Name	Count of entries
Protein coding	20012
lncRNA	11004
Antisense	3588
LincRNA	5890
Processed transcript	1117
Sense-intronic	409
HG19 Refseq TFBS conserved track	5797267
HG19 Refseq TFBS unique (name)	259
K562 pol III	953
High-expressed protein-coding genes	3532
High-expressed lncRNA	119
Low-expressed protein-coding genes	1839
Low-expressed lncRNA	2938

Global pattern of histone modification marks across the TSS of protein-coding genes and lncRNAs

The promoters of genes have been extensively studied for the occupancy of nucleosomes. Distinct patterns of nucleosome occupancy have been previously well described and extensively analysed for their relationship with the expression of genes. In addition, specific histone marks have been extensively studied across the promoter of genes and their correlation with expression investigated in detail. Since, histone modifications like H3K4me3, H3K4me1, H3K9ac, H3K27me3 and H3K9me3 are known to be associated with the promoters of protein-coding genes, an attempt has been made to study the promoter architecture of protein coding and lncRNA genes with respect to these histone modifications. Earlier, it was shown that the histone marks like H3K4me3 are associated with the promoters of lncRNAs, in a manner akin to promoters of protein-coding genes. In this study, we analysed the peaks called using MACS across the TSS of both protein-coding as well as lncRNA transcripts for each of the histone modification marks, namely, H3K4me3, H3K4me1, H3K9ac, H3K27me3 and H3K9me3

Table 1 (B). Summary of the number of peaks in each of the histone modification marks analysed in the study.

Name	H3K4me1	H3K4me3	H3K36me3	H3K9ac	H3K9me3	H3K27me3
H1 cell line	63644	21498	25996	13033	43128	4788
CD34	57319	25363	5240	–	27573	3871
IMR90	111281	33148	25488	48665	64601	24929
Liver	50026	32781	22449	15601	54372	4218
PBMC	18171	–	26657	24583	27243	33094
BrGr	–	25417	6810	–	33568	13468

for each of the subclasses of lncRNAs. Further, we classified lncRNA into three more subclasses: known, novel and putative. We also checked the histone modifications of H1 cell line across these three classes. Of the total set, a subset of lncRNAs were closely associated with protein-coding genes, and this includes 409 sense-intronic, 810 processed transcripts, 836 lincRNA and 2825 antisense lncRNAs (table 1 in electronic supplementary material at http://www.ias.ac.in/jgenet). This has also been further classified into sense and antisense classes (table 1 in electronic supplementary material).

We initially analysed patterns of expressive marks H3K4me1, H3K4me3 and H3K9ac by plotting different histone modifications across ± 5 kb of TSS in 100 bp bins (figure 1; figure 1 in electronic supplementary material). The number of genes consisting of each modification, in a single bin was calculated after adjusting for the total number of genes in that category. We found that the expressive histone mark H3K4me1, though had a higher density around the TSS, the marks were found to be scattered across ±5 kb of TSS rather than forming a distinct peak at the TSS as observed in the case of other expressive marks H3K4me3

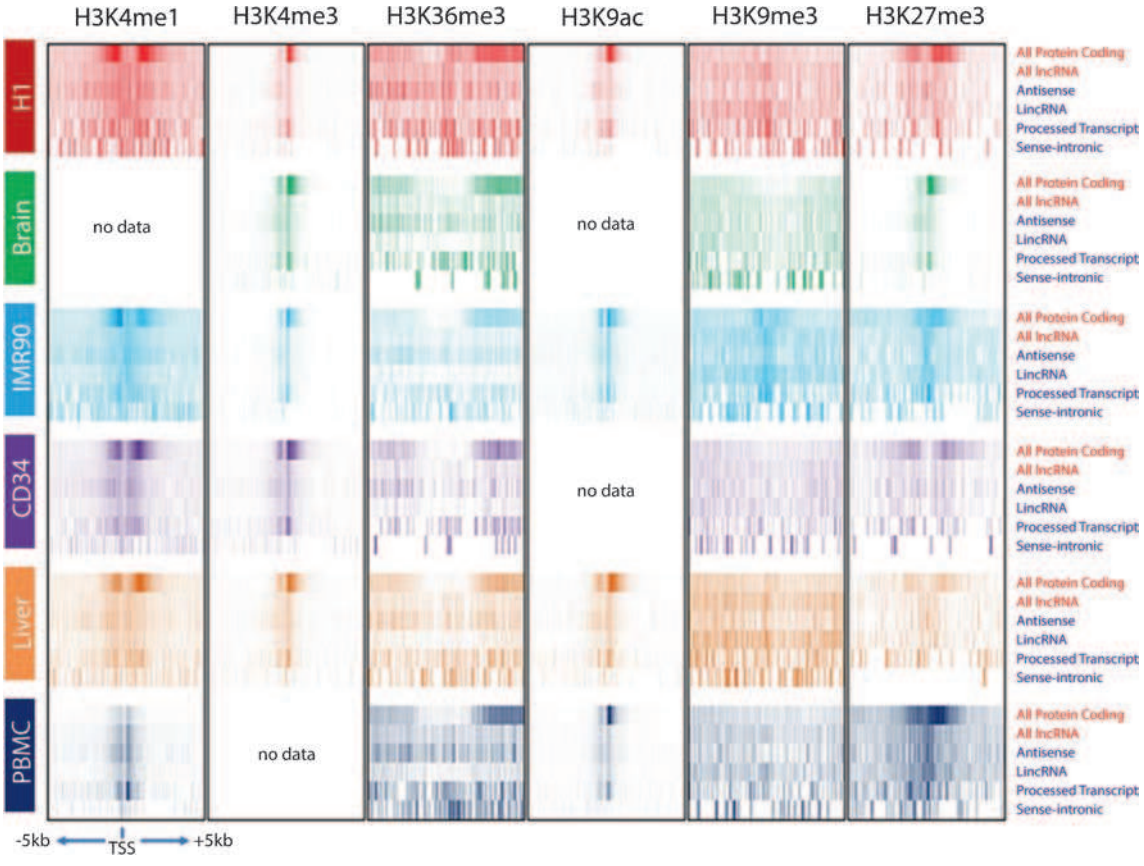

Figure 1. Methylation pattern around TSS. Distribution of methylation peak summit count in 100 bp sliding window, 5 kb upstream and downstream from the start site were calculated for all protein-coding genes, lncRNA genes and lncRNA subclasses in H1 cell, brain tissue, IMR90 cells, CD34 primary cells, liver tissue and PBMCs. The count was normalized by dividing the individual count with the total number of genes in that category. The count was plotted as heat map.

and H3K9ac. This pattern was found consistent across all cell and tissue types analysed in this study. Similar patterns, but with a diminished intensity were found across the TSS of all lncRNAs and its individual subclasses. The other expressive mark, H3K36me3, associated with gene body of actively transcribing genes, was also found to be scattered across ±5 kb of TSS of the coding region for respective subclasses of lncRNAs. This was also consistent with our previous observations of histone modification marks across TSS of lncRNAs and protein-coding genes.

We further analysed patterns of two repressive histone marks H3K9me3 and H3K27me3 across the tissue types and found that both showed a dispersed pattern of localization within ±5 kb of TSS of protein coding, all lncRNAs and lncRNA subclasses. The only exception to this diffused pattern was found in H3K27me3 marks in the brain tissue which exhibited a distinct and accentuated pattern near TSS in all categories of transcripts considered. Further, the enrichment was more pronounced in protein-coding than all lncRNAs taken together or in any of the subclasses (figure 1).

Histone modification marks across TSS of lncRNA subtypes shows distinct absence of patterns for sense-intronic lncRNAs

The patterns of histone methylation marks were generally consistent for all lncRNAs and across tissue types (figure 1). A careful examination of the marks across the TSS of each subtypes across tissues revealed interesting patterns. The global pattern of histone modification marks across the TSS of lncRNAs was in general similar in lincRNAs and antisense transcripts (figure 1). However, in processed transcripts the distribution of histone marks consistently showed a distinct pattern, comparable to protein-coding genes in some cases, with accentuation of the distribution of these marks around TSS as compared to any of the other subtypes of lncRNAs (figure 1).

The distribution of histone modification marks across sense-intronic subclass almost always proved to be distinct and had no consistent pattern across cell and tissue types. In addition, repressive marks H3K9me3/H3k27me3 and expressive marks H3K4me3/H3K9ac showed an attenuated pattern across the TSS while H3K36me3 and H3K4me1 showed an accentuated pattern (figure 1). Further, we performed a nonparametric ANNOVA test with the sense-intronic subclass and found statistically significant (P <0.05 to <0.001) difference from all protein-coding genes and lncRNA genes for most of the cell and tissue types across different histone modifications (figure 2). We believe that this could potentially arise due to a distinctly different regulatory and/or biogenesis mechanism of the sense-intronic lncRNAs.

The distinct pattern of histone mark localized at TSS of sense-intronic lncRNAs might arise due to different mechanisms of their transcription. As few lncRNAs are known to have RNA pol III localization (Dieci *et al*. 2007), we downloaded the pol III localization data for K562 (only data for pol III is available) and pol II and pol III localization data for GM12878 cell line from ENCODE (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19551). On plotting pol III data at the TSS we found no enrichment for both the cell types (K562 and GM12878) (figure 2, a&b in electronic supplementary material). As most of the lncRNAs are known to have RNA pol II localization, we found enrichment of pol II data across the TSS for GM12878 cell type (figure 2c in electronic supplementary material). However, even in this case the pol II marks were less frequent across TSS than the other subclasses.

In addition, we analysed the antisense lncRNA subclass and found 787 intron associated lncRNAs. Further, these antisense intronic RNA genes were used to check the distribution of different histone marks of H1 cell line at TSS and found a similar pattern as sense-intronic lncRNA (figure 3 in electronic supplementary material).

Cell type specific patterns show consistency across TSS of lncRNAs and protein-coding genes

We further investigated the relative distribution and profile of each of the histone modification marks across the TSS of two different cell types, i.e. a stem cell (H1) and a differentiated cell (IMR90), of both protein-coding and lncRNA genes, in an attempt to understand the patterns of profiles and to compare differential patterns across the TSS (figure 3, a&b). Briefly the number of genes with a distinct peak for a histone mark were plotted across each of the 100 nucleotide

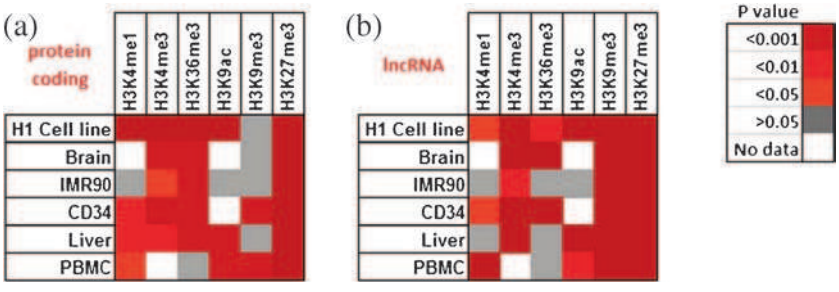

Figure 2. Heat map predicting the significant difference level (P value) of different histone modification between (a) sense-intronic lncRNA class and protein-coding class and (b) sense-intronic and lncRNA class throughout H1cell, brain tissue, IMR90 cells, CD34 primary cells, liver tissue and PBMCs.

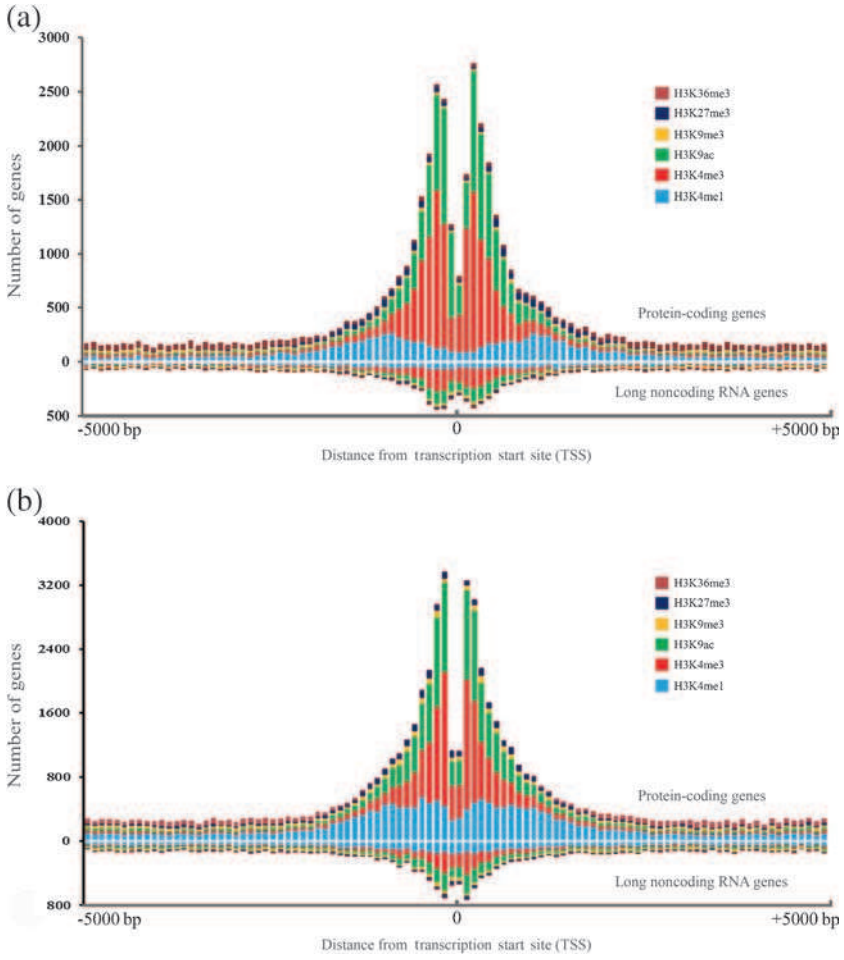

Figure 3. Histone modification density around TSS of protein coding and lncRNA genes in H1 cell line (a) and IMR90 cells (b). Peak summit count in 100 bp sliding window was calculated for each histone modification. The count for each 100 bp bin for different histone modifications were added and plotted on Y-axis, while the X-axis represents the individual bin.

bins as described previously. The number of genes having the histone marks were plotted across the TSS for each of the bins. The profile shows a distinct pattern of interplay between the histone modifications across the TSS. We observed that the proximal promoter region (±2 kb of TSS) in protein-coding genes was highly enriched for expressive marks like H3K4me3, H3K4me1 and H3K9ac (figure 3, a&b). The repressive histone marks H3K9me3 and H3K27me3 were noted by significant attenuation across the TSS of protein-coding genes. These global patterns were found to be conserved between the two different cell types in both protein-coding and lncRNA genes, showing a conserved attenuation at the TSS (figure 3, a&b). In protein-coding genes and lncRNAs coding genes, H3K36me3 marks were better represented in the distal promoter region (2–5 kb upstream of TSS) and gene body. Along with H3K36me3 marks, lncRNA distal promoter and gene body regions also have higher localization of H3K4me1 marks. The sense-intronic lncRNAs consistently showed a distinct absence of any pattern shared between other lncRNA subclasses or with protein-coding genes (figure 3, a&b).

Promoters defined by conserved TFBS show a similar architecture of organization in TSS of protein-coding and lncRNA subclasses except sense-intronic lncRNAs

Also, we futher explored the distribution of conserved TFBS across TSS of protein-coding and lncRNAs. Approximately we downloaded 5,797,267 predicted conserved binding sequence motifs for conserved TFBS in humans from the UCSC Bioinformatics site (figure 4 in electronic supplementary material). Analysis of the TFBS distribution across the TSS of lncRNAs showed similar patterns as the protein-coding genes albeit with an attenuation (figure 4). Further indepth analysis of the distribution across the subclasses of lncRNA TSS revealed antisense and lincRNAs showing lower enrichment of TFBS at their TSS. The profile across TSS of processed pseudogenes revealed a pattern

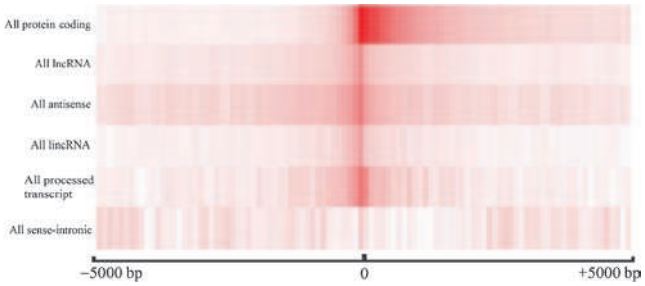

Figure 4. Distribution of TFBS around TSS. Distribution of TFBS count in 100 bp sliding window, 5 kb upstream and downstream from the start site were calculated for all protein-coding genes, lncRNA genes and lncRNA subclasses. Count was normalized by dividing the individual count with the total number of genes in that category. The count was plotted as heat map.

distinctly similar to those of protein-coding genes and distinctly accentuated compared with other lncRNA subclasses. The only exception to this enrichment at TSS was in the sense-intronic class (figure 4). Since transcription binding factors at distal-binding sites are known to influence the transcriptional regulation, we checked TFBS site distribution, upto 20 kb upstream and downstream of the TSS (figure 5 in electronic supplementary material). We found that unlike the proximal promoter site, the distal promoter sites showed a similar pattern between sense-intronic and antisense genes.

Difference in the density of H3K4me3 marks in the promoter of lncRNAs and protein-coding genes are consistent irrespective of the expression level differences between them

The differences in density of histone marks (H3K4me3) across the TSS of protein-coding genes and lncRNA genes could potentially be attributed to the general expression differences between the genes. If this was true, the density of marks across the TSS of both sets should show consistent patterns in genes which are expressed high or low in a particular cell or tissue type. We therefore analysed the density of marks across the TSS of high and low expressed genes in H1 cell cline (figure 6 in electronic supplementary material). The difference in density of H3K4me3 was significantly different in the promoters of high and low expressed genes, but the differences in density between the lncRNA and protein-coding sets were consistently maintained.

Discussion and conclusions

The advent of breakthrough technologies to understand genomewide epigenetic profiles of cells has contributed a significant wealth of understanding of the epigenetic state of a cell and how the organization of epigenetic marks modulate the expression of genes (Hurd and Nelson 2009). Many of these transcribed loci in the genome do not encode for proteins, but rather encompass a fast-growing set of transcripts which, otherwise, are termed noncoding RNA. One of the major classes of noncoding RNAs are long noncoding RNAs which are increasingly being shown to be involved in a variety of biological processes in the cell (Mattick *et al*. 2009; Clark and Mattick 2011; Nagano and Fraser 2011; Wang and Chang 2011). Many lncRNAs appear to be spatiotemporally regulated as they are expressed in cell type and stage specific manner (Dinger *et al.* 2008; Mercer *et al.* 2008). Though in recent years there have been a number of landmark studies on the regulation of protein-coding genes, little is known on the regulation of lncRNA. Recently, we have shown that there are some similarities and differences in the distribution of epigenetic marks around the TSS of protein-coding and lncRNA genes (Sati *et al.* 2012a). Considering the DNA methylation and histone modification data (H3K4me3, H3K9me3, H3K27me3 and H3K36me3) from various cell lines and tissue types we have shown that expressive histone marks (H3K4me3 and H3K36me3) and repressive histone marks (H3K27me3) have similar distribution at the TSS of both protein coding and lncRNA genes. However, there were differences in the distribution of DNA methylation and H3K9me3 marks at the TSS of these classes. We also analysed an empirical list of 10-well-established lncRNA loci and analysed the histone modification marks in the promoters of these lncRNAs. Our analysis revealed consistent patterns of histone modification marks across the promoters (figure 4 in electronic supplementary material). The lncRNA subset included distinct subtypes of lncRNAs which was categorized into four major classes: lincRNAs (5890 genes), antisense (3588 genes), processed transcripts (1117 genes) and sense-intronic transcripts (409 genes). Since earlier we suggested a global difference in the profiles, we hypothesised those subclasses which had a distinct genomic loci of origin and a distinct spatial association with protein-coding genes could have different profiles for epigenetic marks and different contributions to the global profile as previously reported. In this study, we performed a comprehensive analysis of histone modification marks in the lncRNA subclasses across six distinct cell types.

Of the different subclasses analysed in this study, sense-intronic transcripts shows low density of expressive marks H3K4me3 and H3K36me3 at their TSS, the reason being that these lncRNAs originate from introns of protein-coding genes and are thus intragenic. It is well known that H3K4me3 modification marks the TSS of protein-coding genes and is generally not present in intragenic regions (Cedar and Bergman 2009; Hahn *et al.* 2011). Similarly, H3K36me3 modifications are associated with exons of protein coding genes and thus as expected are absent from the introns (Kolasinska-Zwierz *et al.* 2009; Dhami *et al.* 2010). Sense-intronic lncRNAs are sequences present within the introns of coding genes on the sense strand. It seems that this category should be treated separately whenever analysing lncRNA class. As for the other classes, there needs to be specific transcription initiation modules restricted to solely their expression, thus their promoter chromatin architecture shows similarities with that of protein-coding gene promoters. In contrast, the sense-intronic lncRNAs can be either *de novo*

transcribed or originate from the spliced out fragment of the gene. The histone modification profiles will reflect both the mentioned scenarios, yielding a mixed profile. A special condition will arise whenever there is a *de novo* transcription initiation of these RNAs, as the nearby periphery needs to be as euchromatin to allow access to RNA pol II, keeping the parent protein-coding gene in the repressed state. Further, we found no enrichment of RNA pol III at the TSS of sense-intronic lncRNAs, which suggests a potentially novel mechanism of biogenesis.

It is known that transcription factor binding is influenced by sequence elements and local chromatin architecture (Cedar and Bergman 2009). The histone modifications like H3K4me3 which are markers of open chromatin are associated with transcription factor binding sites. As in case of protein-coding genes, lncRNA coding genes enriched in H3K4me3 marks had higher enrichment of TF binding sites, with the exception of sense-intronic lncRNAs. Further, sense-intronic lncRNA seems to be expressed via noncanonical transcription factor binding and might harbour different histone modification at the TSS than other subclasses. The rest of the subclasses showed similar distribution patterns for these histone modifications. When all of these repressive and expressive histone modification marks were co-analysed in H1 and IMR90 cells, it was found that most of the protein-coding and lncRNA genes have expressive marks like H3K4me3, H3H4me1 and H3K9ac at their TSS. This observation suggests that most of the genes, either protein-coding or lncRNA coding, are poised for expression and very few of them reside in the repressed chromatin. The presence of very low amount of H3K9me3 and H3K27me3 marks at the promoters suggest that there might be additional factors that are involved in gene repression, like DNA methylation. H3K4me3 distribution at TSS of protein-coding genes has been previously demonstrated as the characteristic twin peaks of distribution on either side of TSS. We have also observed similar twin peak distribution character for H3K4me3 at TSS of protein and lncRNA coding genes. Interestingly, such twin peak distribution patterns was also discernible for the expressive marks like H3K4me1 and H3K9ac. We hypothesise that this distinct pattern of histone modification marks across the TSS could potentially be contributed by the eviction of nucleosomes around the TSS and distinctly at the +1 position of actively transcribing genes during gene expression. Overall, this study suggests similar patterns of histone distribution in lncRNA and protein-coding genes. It seems that all RNAs either protein-coding or lncRNA follow a universal epigenetic code except for the sense-intronic class of lncRNAs, which we believe should be treated as a separate class of ncRNAs. Although it might be over simplistic as our analysis relies on the ChIP data produced from few cell and tissue types. Further validation is required in more number of cell and tissue types and compared with expression profiles that will highlight the functional implication of these patterns. Our observations show that the promoter architecture of protein-coding and lncRNA genes share similar distribution of histone marks. Another major finding was that, the sense-intronic lncRNA have different histone modifications and TFBS localization patterns at their TSS. It was observed that their expression occurs via noncanonical transcription factor binding and they might harbour different histone modification at the TSS than other subclasses, leading to a view that this subclass should be treated differently when analysing lncRNAs.

Acknowledgements

The authors acknowledge the discussions with Dr. Sheetal Gandotra which considerably improved thought process and content of the manuscript. SSa and SG acknowledges a Senior Research Fellowship from the Council of Scientific and Industrial Research, India. This work was funded by the Council of Scientific and Industrial Research, India through project GENCODE-C (BSC0123). We acknowledge the NIH Roadmap to Epigenomics Consortium members for producing the data used in our analysis procedures. We are thankful to NIH Roadmap, Epigenomics Consortium for maintaining an open access to the datasets deposited in NIH Roadmap Epigenomics Project Data Listings at http://nihroadmap.nih.gov/epigenomics/.

References

Bernstein B. E., Birney E., Dunham I., Green E. D., Gunter C. and Snyder M. 2012 An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74.

Bhartiya D., Kapoor S., Jalali S., Sati S., Kaushik K. *et al.* 2012 Conceptual approaches for lncRNA drug discovery and future strategies. *Expert. Opin. Drug Discov.* **7**, 503–513.

Cedar H. and Bergman Y. 2009 Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.* **10**, 295–304.

Clark M. B. and Mattick J. S. 2011 Long non-coding RNAs in cell biology. *Semin. Cell. Dev. Biol.* **22**, 366–376.

Dhami P., Saffrey P., Bruce A. W., Dillon S. C., Chiang K. *et al.* 2010 Complex exon-intron marking by histone modifications is not determined solely by nucleosome distribution. *PLoS One* **5**, e12339.

Dieci G., Fiorino G., Castelnuovo M., Teichmann M. and Pagano A. 2007 The expanding RNA polymerase III transcriptome. *Trends Genet.* **23**, 614–622.

Dinger M. E., Amaral P. P., Mercer T. R., Pang K. C., Bruce S. J. *et al.* 2008 Long non-coding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res.* **18**, 1433–1445.

Fasanaro P., D'Alessandra Y., Magenta A., Pompilio G. and Capogrossi M. C. 2013 MicroRNAs: promising biomarkers and therapeutic targets of acute myocardial ischemia. *Curr. Vas. Pharmacol.* (in press).

Guttman M. and Rinn J. L. 2012 Modular regulatory principles of large non-coding RNAs. *Nature* **482**, 339–346.

Hahn M. A., Wu X., Li A. X., Hahn T. and Pfeifer G. P. 2011 Relationship between gene body DNA methylation and intragenic H3K9me3 and H3K36me3 chromatin marks. *PLoS One* **6**, e18844.

He L. and Hannon G. J. 2004 MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.* **5**, 522–531.

Hurd P. J. and Nelson C. J. 2009 Advantages of next-generation sequencing versus the microarray in epigenetic research. *Brief Funct. Genomic Proteomic* **8**, 174–183.

Ji P., Diederichs S., Wang W., Boing S., Metzger R. *et al.* 2003 MALAT-1, a novel non-coding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22**, 8031–8041.

Kaushik K., Leonard V. E., Kv S., Lalwani M. K., Jalali S. *et al.* 2013 Dynamic expression of long non-coding RNAs (lncRNAs) in adult zebrafish. *PLoS One* **8**, e83616.

Kolasinska-Zwierz P., Down T., Latorre I., Liu T., Liu X. S. and Ahringer J. 2009 Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat. Genet.* **41**, 376–381.

Liao Q., Liu C., Yuan X., Kang S., Miao R. *et al.* 2011 Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Res.* **39**, 3864–3878.

Mattick J. S., Amaral P. P., Dinger M. E., Mercer T. R. and Mehler M. F. 2009 RNA regulation of epigenetic processes. *BioEssays* **31**, 51–59.

Mercer T. R., Dinger M. E. and Mattick J. S. 2009 Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* **10**, 155–159.

Mercer T. R., Dinger M. E., Sunkin S. M., Mehler M. F. and Mattick J. S. 2008 Specific expression of long non-coding RNAs in the mouse brain. *Proc. Natl. Acad. Sci. USA* **105**, 716–721.

Morozova O., Hirst M. and Marra M. A. 2009 Applications of new sequencing technologies for transcriptome analysis. *Annu. Rev. Genomics Hum. Genet.* **10**, 135–151.

Nagano T. and Fraser P. 2011 No-nonsense functions for long non-coding RNAs. *Cell* **145**, 178–181.

Pasmant E., Sabbagh A., Vidaud M. and Bieche I. 2011 ANRIL, a long, non-coding RNA, is an unexpected major hotspot in GWAS. *FASEB J.* **25**, 444–448.

Ponting C. P., Oliver P. L. and Reik W. 2009 Evolution and functions of long non-coding RNAs. *Cell* **136**, 629–641.

Rajpathak S. N., Vellarikkal S. K., Patowary A., Scaria V., Sivasubbu S. and Deobagkar D. D. 2014 Human 45,X fibroblast transcriptome reveals distinct differentially expressed genes including long non-coding RNAs potentially associated with the pathophysiology of turner syndrome. *PLoS One* **9**, e100076.

Rinn J. L., Kertesz M., Wang J. K., Squazzo S. L., Xu X. *et al.* 2007 Functional demarcation of active and silent chromatin domains in human HOX loci by non-coding RNAs. *Cell* **129**, 1311–1323.

Sati S., Ghosh S., Jain V., Scaria V. and Sengupta S. 2012a Genome-wide analysis reveals distinct patterns of epigenetic features in long non-coding RNA loci. *Nucleic Acids Res.* **40**, 10018–10031.

Sati S., Tanwar V. S., Kumar K. A., Patowary A., Jain V. *et al.* 2012b High resolution methylome map of rat indicates role of intragenic DNA methylation in identification of coding region. *PLoS One* **7**, e31621.

Saxena A. and Carninci P. 2011 Long non-coding RNA modifies chromatin: epigenetic silencing by long non-coding RNAs. *BioEssays* **33**, 830–839.

Tsai M. C., Manor O., Wan Y., Mosammaparast N., Wang J. K. *et al.* 2010 Long non-coding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689–693.

Wang K. C. and Chang H. Y. 2011 Molecular mechanisms of long non-coding RNAs. *Mol. Cell.* **43**, 904–914.

Wapinski O. and Chang H. Y. 2011 Long non-coding RNAs and human disease. *Trends Cell Biol.* **21**, 354–361.

Yap K. L., Li S., Munoz-Cabello A. M., Raguz S., Zeng L. *et al.* 2010 Molecular interplay of the non-coding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell* **38**, 662–674.

Zhang Z., Pang A. W. and Gerstein M. 2007 Comparative analysis of genome tiling array data reveals many novel primate-specific functional RNAs in human. *BMC Evol. Biol.* **7**, suppl 1, S14.

Received 14 February 2014, in revised form 10 September 2014; accepted 16 September 2014
Unedited version published online: 1 October 2014
Final version published online: 17 March 2015