Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse

Gireesh K. Bogu1,2,3,4,#, Pedro Vizán2,4, Lawrence W. Stanton5,6, Miguel Beato2,4, Luciano Di Croce2,4,7, and Marc A. Marti-Renom1,2,4,7,#

1. CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain
2. Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain
3. Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG) and UPF, Doctor Aiguader, 88, Barcelona 08003, Spain.
4. Universitat Pompeu Fabra (UPF), Barcelona, Spain
5. Department of Biological Sciences, National University of Singapore,
   Singapore.
6. Stem Cell and Developmental Biology Group, Genome Institute of Singapore,
   Singapore.
7. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona. Spain

Keywords: ChromHMM, ChIP-Seq, RNA-seq, lncRNA

#Address correspondence to Gireesh K. Bogu, gireesh.bogu@crg.eu, or Marc A. Marti-Renom, martirenom@cnag.crg.eu
ABSTRACT

Discovering and classifying long noncoding RNAs across all mammalian tissues and cell lines remains a major challenge. Previously, mouse IncRNAs were identified using RNA-seq data from a limited number of tissues or cell lines. Additionally, associating a few hundred IncRNA promoters with chromatin states in a single mouse cell line has identified two classes of chromatin-associated IncRNA. However, the discovery and classification of IncRNAs is still pending in many other tissues in mouse. To address this, we built a comprehensive catalog of IncRNAs by combining known IncRNAs with highly-confident novel IncRNAs identified by mapping and de novo assembling billions of RNA-seq reads from eight tissues and a primary cell line in mouse. Next, we integrated this catalog of IncRNAs with multiple genome-wide chromatin-state maps and found two different classes of chromatin state–associated IncRNAs, including promoter-associated (plncRNAs) and enhancer-associated (elncRNAs) ones across various tissues. Experimental knockdown of an elncRNA resulted in the down-regulation of the neighboring protein-coding gene Kdm8, a histone-demethylase. Our findings provide 2,803 novel IncRNAs and a comprehensive catalog of chromatin-associated IncRNAs across different tissues in mouse.
INTRODUCTION

Previous large-scale transcriptome sequencing studies have confirmed that ~80% of the human genome is transcribed, yet only a minor fraction of it (~3%) codes for protein (1, 2). It is now known that a major fraction of the transcriptome consists of RNAs from intergenic noncoding regions of the genome, which have been termed as intergenic IncRNAs. Comprehensive IncRNA catalogs were recently established for various cell lines and tissues in human, mouse, C. elegans, Drosophila, and zebrafish (3-8). In addition, we now know the functions of a limited number of the discovered IncRNAs, such as Xist in X chromosome inactivation (9), HOTAIR in cancer metastasis (10), Inc-DC in dendritic cell differentiation (11), Braveheart in heart development (12), Megamind and Cyrano in embryonic development (13), Fendrr in cardiac mesoderm differentiation (14), Malat1 in alternative splicing (15), and a few others including one from our previous work showing that RMST IncRNA regulates neurogenesis by physically interacting with Sox2 transcription factor (16).

Even though thousands of IncRNAs have been cataloged, it is still unclear how to characterize regulatory IncRNAs. Very recently, regulatory IncRNAs were shown to associate preferentially with promoter and enhancer chromatin states in a single mouse cell line (17). While this observation is highly interesting, it is not clear whether there were more IncRNAs associate with these two chromatin states since the IncRNA associations were not tested in multiple tissues. In addition, the IncRNA or chromatin state datasets used in the previous study (17)
were selected only in one single cell line, which technically limits testing
thousands of lncRNAs. Finally, it is also unknown whether these lncRNAs
associate with similar chromatin states across different tissues or not.

To build a comprehensive chromatin-associated mouse lncRNA dataset, we first
used billions of mapped RNA-seq reads to identify highly confident novel
lncRNAs and then combined this with thousands of known lncRNAs. Second, we
used more than a billion mapped ChIP-seq reads of various histone marks to
identify chromatin state maps. Finally, we integrated all these mouse lncRNAs
with chromatin state maps, resulting in a comprehensive catalog consisting of
thousands of chromatin state–associated lncRNAs. The analysis across multiple
tissues also revealed a novel set of lncRNAs that are significantly enriched with
promoter and enhancer chromatin states. Interestingly, the majority of the
lncRNAs chromatin states switch from one state to another state across all the
tissues or cell lines we tested. To our knowledge, this is the most comprehensive
dataset of chromatin state–associated lncRNAs in mouse, and we expect this will
be a valuable resource to help researchers select candidate lncRNAs for further
experimental studies.
RESULTS

Transcriptome mapping, assembly and quantification.

About 3 billion raw sequence reads of RNA-seq experiments were downloaded from the ENCODE project (18) and analyzed using a computational pipeline consisting of TopHat (v2.0.9) (19), Cufflinks (v2.1.1) (20), and Scripture (v4) (21) (Fig. 1A). We constructed a map of RNA expression in mouse by first collecting RNA sequencing reads using long (76–108 nucleotides), paired-end, polyadenylated, strand-specific high-throughput RNA sequencing data from 8-week-old adult brain, heart, kidney, small intestine, liver, spleen, testes, thymus and a paired-end embryonic stem (ES) cell line (Table S1). Next, the collected reads were mapped to the reference mouse genome using TopHat, which uniquely mapped 85% (2,631,897,546) of the sequence reads with 2 mismatches allowed. Of the mapped sequences, ~73% aligned with known transcript loci, and the remaining 27% aligned to either intergenic loci or to coding genes in an antisense direction, which suggested that novel transcripts might exist. To test this, we assembled the mapped mouse transcriptome data in a de novo approach using Scripture and Cufflinks to reconstruct transcripts and quantified the expression by masking regions, including those containing snoRNAs, tRNAs, miRNAs and pseudogenes. Transcripts that were significantly covered ($P < 0.01$) were selected to avoid noisy transcripts (Methods). In total, Scripture identified 593,102 multi-exonic transcripts and Cufflinks, 539,775 transcripts, with an overlap of 500,530 transcripts between the two methods. Of those overlapping transcripts, ~86% (429,818) overlapped with known coding transcripts (annotated
in either RefSeq, UCSC, or Ensembl) and 10.2% (51,134) overlapped with known noncoding transcripts (annotated either as snoRNA, tRNA, miRNA, or pseudogenes). This shows the quality of transcripts and their ability to recover known noncoding transcripts. The remaining 3.9% of transcripts (20,018) had no overlap with any known coding or noncoding transcripts.

**Genome-wide identification and annotation of lncRNAs in mouse.**

We applied a computational pipeline to identify putative intergenic lncRNAs along with other types of lncRNAs (e.g., antisense, intronic) (4, 5, 22). We identified 16,185 multi-exonic lncRNAs longer than 200 bp and with an expression ≥ 1 FPKM (fragments per kilobase of exonic length per million) in at least one given tissue. Importantly, these lncRNAs did not contain transcripts with coding potential as measured by the two independent methods including conservation-independent CPAT (23) and conservation-dependent PhyloCSF (24) (Methods).

About 85% of this dataset overlapped with previously identified lncRNAs (17, 21, 25-29) (Fig. S1), supporting the accuracy of our prediction pipeline with a total of 34% of all known lncRNAs recovered (Fig. 1B). The remaining 2,803 identified lncRNAs were considered as novel lncRNAs in mouse. Further, based on the genomic location of lncRNAs relative to nearest protein-coding gene promoters, we annotated 2,174 antisense (e.g., overlapping the protein-coding gene in an antisense direction), 382 intergenic (e.g., located within 10 kb to the nearest protein-coding gene), and 247 strictly intergenic lncRNAs (e.g., located more
than 10 kb away from the nearest protein-coding gene) (Fig. S2, and Fig. 1C for an example of a novel IncRNA identified in testes).

Properties of the 2,803 IncRNAs.

It has been previously shown that IncRNAs comprise few exons, are shorter in length, and are expressed at low levels in a highly tissue- or cell-specific nature (3-5). The 2,803 IncRNAs reported here are consistent with these previous studies. On average, our IncRNA transcripts have fewer exons (3 exons), are shorter (6,336 nucleotides), and are expressed at lower levels (1.56 FPKM) than the average for the 27,259 RefSeq protein-coding transcripts, which (on average) have 10 exons, a length of 50,453 nucleotides, and expression levels of 4.68 FPKM (Fig. S3). To gain more insight, we combined our novel IncRNAs with all the known IncRNAs and reanalyzed the genomic features by considering the ones with an expression greater than 0.1 FPKM in at least one out of 8 tissues and in a cell line, and the ones that are far from protein-coding genes (e.g., 10 kb away from either a transcription start site [TSS] or a transcriptional end site [TES] of a protein-coding gene). This resulted in 3,759 IncRNAs. On average, these transcripts have an exon size of 482 nucleotides, a transcript size of 9,710 nucleotides, an expression level of 1.87 FPKM, and a conservation score of 0.1 phastCons. These results further confirmed the genomic features of IncRNA, such as lower expression and conservation levels as compared to protein-coding genes.
In mammals, lncRNAs are expressed in a tissue-specific manner (3-5). To assess for any tissue specificity of our dataset of lncRNA, we compared each lncRNA expression in a given tissue to its expression in the remaining 8 tissues (Fig. 2A; Table S2). We observed that 62% of our novel intergenic lncRNAs are tissue-specific, which is comparable to known intergenic lncRNAs (68% tissue-specific). Moreover, protein-coding genes resulted in 36.4% tissue specificity across the eight tissues and the ES cell line (Fig. S4). Overall, the results clearly show that lncRNAs are highly tissue specific in nature. Next, we selected the tissue-specific lncRNAs from our list as previously defined (e.g., with an entropy > 0.4) (4). To experimentally validate a pair of these selected tissue-specific lncRNAs, we measured the expression levels by RT-PCR of the heart (H-lnc1 and H-lnc2), liver (L-lnc1 and L-lnc2), and kidney (K-lnc1 and K-lnc2) lncRNAs with respect to the housekeeping gene GAPDH (Fig. 2B), which confirmed their tissue-specificity.

To assess whether our novel lncRNAs have active TSS and regulatory marks, we overlapped CAGE tags and DNase I tags from the FANTOM and ENCODE projects with the promoters of our lncRNA (30, 31). We observed an enrichment of CAGE tags around our lncRNA promoters, as compared to random lncRNA promoters (Fig. S5A). We also observed an enrichment of tissue-specific DNase I tags in lncRNA promoters from the brain, kidney, liver, spleen, and thymus tissues as well as for the ES cell line (Fig. S5B). Finally, we performed de novo motif analysis using lncRNA promoters to explore whether any transcription
factors could be regulating these lncRNAs. Indeed, we found several significant transcription factor binding motifs enriched near lncRNA promoters (Fig. S5C). These results show that the 2,803 lncRNA promoters are enriched with various regulatory marks in the mouse genome and could potentially have regulatory roles.

Genome-wide identification of chromatin state maps in mouse.

Chromatin marks mapping across different cell lines in mammals have been previously used to detect and annotate novel regulatory regions in the genome, including for putative lncRNAs (5, 17, 32). We hypothesized that integrating chromatin state maps with the promoters of the transcripts identified here using RNA-seq expression could guide us in annotating the potential transcripts and in predicting their mode of regulation. A map of chromatin marks was constructed from ~1.4 billion mapped reads obtained from 72 pooled ENCODE genome-wide ChIP-seq datasets in eight tissues (brain, heart, liver, small intestine, kidney, spleen, testes, and thymus) and the one primary ES cell line. The ChIP-seq datasets used included regulatory histone modifications, such as H3 lysine 4 mono-methylation (H3K4me1), H3 lysine 4 tri-methylation (H3K4me3), H3 lysine 36 tri-methylation (H3K36me3), H3 lysine 27 tri-methylation (H3K27me3), and H3 lysine 27 mono-acetylation (H3K27ac), as well as CCCTC-binding factor (CTCF) marks and RNA polymerase II marks.
We applied the ChromHMM program (32) to create a chromatin state model at 200 bp resolution, which resulted in six major chromatin state maps (Fig. 3A), including promoter (active and poised), enhancer (strong and poised/weak), transcribed (transcription transition, elongation, and weak transcription), insulator, repressed, and heterochromatin states (Table S3). In total, we mapped 261,175 promoter states (covering ~1% of the mouse genome), 863,677 enhancer states (~3%), 1,133,166 transcribed states (~12%), 150,752 repressed states (~1%), 322,521 insulator states (~1%) and 995,562 heterochromatin states (~82%). To validate the accuracy of the predicted chromatin states or maps, we mapped (at ±10 kb) our 206,045 unique non-overlapping active promoter maps to known promoters of 23,431 RefSeq protein-coding genes and 3,190 RefSeq noncoding genes from TSSs. Our analysis recalled 82% (19,280) of the protein-coding promoters and 75% (2,401) of the noncoding ones. We repeated the above mapping using the poised promoter map and mapped an additional 709 protein-coding and 92 noncoding gene promoters. All together, we successfully mapped 85% of the known protein-coding and 78% noncoding gene promoters. These results indicate that using combinatorial promoter chromatin states to retrieve promoters results in ~6% higher recall than when using only H3K4me3 as an active promoter chromatin mark (33).

Classification of lncRNAs using chromatin state maps.

Previously chromatin state maps at promoters were used to define two distinct classes of lncRNAs (17). For example, enhancer-associated lncRNA (elncRNA)
promoters or transcription start sites (TSSs) are depleted of H3K4me3 and enriched with H3K4me1, and promoter-associated lncRNAs (plncRNAs) are enriched with H3K4me3 and depleted of H3K4me1. Using a similar promoter-overlapping approach for our chromatin state maps, we defined these two classes of chromatin-associated lncRNAs across 8 tissues and an ES cell line.

For this classification, we first listed ~30,000 unique protein-coding promoter loci and ~19,000 intergenic lncRNA promoter loci (200 bp long), which were then passed through an expression filter (requiring >1 FPKM in a given tissue) and an intergenic filter (requiring them to be 5 kb away from both TSS and TES of protein-coding genes). We found a few thousand lncRNAs that passed the above expression and intergenic filters (namely, 1,385 lncRNAs in whole brain, 1,236 in ES cells, 903 in heart, 870 in kidney, 787 in liver, 435 in small intestine, 878 in spleen, 2,083 in testes, and 932 in thymus). Overall, less than 10% (852) of these intergenic lncRNAs significantly overlapped with an active promoter or a strong enhancer chromatin state (P <0.001, Fisher-exact test) (Fig. 3B).

We next focused our analysis on these significant chromatin state-associated lncRNAs. In total, we identified 852 unique intergenic lncRNA transcripts associated with either an active promoter or a strong enhancer chromatin state (Table S4, Fig. 3C and D). This result apparently contradicts a previous study (17), in which 52% of lncRNAs were found to be associated with an enhancer chromatin state, and 48%, with a promoter chromatin state. These differences could arise from several parameters used in the previous study that are distinct
to ours: specifically, the previous study considered single exonic transcripts, used CAGE tags to define 5’ ends, and used DNase-seq peaks to identify active promoters. However, to check the consistency, we also used CAGE peaks from FANTOM5 and DNase-seq peaks from ENCODE, along with RNA-seq expression to identify active promoter lncRNAs in liver, spleen, and thymus. This re-analysis resulted in more than 40% of the lncRNAs associated with enhancer chromatin state in thymus (~50% with promoter chromatin state) and around 20% in liver and spleen. (Table S5 and Figure 3D, Methods). Finally, we did not notice any enrichment in the number of elncRNAs over plncRNAs in most of the tissue we analyzed except brain and thymus. A total of 852 unique intergenic lncRNAs were thus annotated as chromatin-associated, including 514 plncRNAs and 433 elncRNAs.

Our approach successfully identified known enhancer-associated coding RNAs, such as Fos, Rgs2, Nr4a2, and Elf5 (34), and elncRNAs such as lincRNA-Cox2, lincRNA-Spasm, and lincRNA-Haunt (35) (Fig. S6). Moreover, we also found known promoter-associated coding RNAs in our analysis, such as Sox2, Oct4, and Nanog, and plncRNAs, such as linc1405 and linc1428 (5) (Fig. S7). Additionally, by pooling all promoter chromatin state maps into one major promoter chromatin–state map, and enhancers into an enhancer chromatin–state map, we were able to recall 71% of published enhancer-associated IncRNAs (36). Our approach successfully recalled 64% of plncRNAs (74 out of 115) and 56% of elncRNAs (69 out of 124) from another study (17). We also experimentally tested histone modifications around the lncRNA promoters, both
in mouse ES cells and heart cells (Fig. S8) using Klf4 as negative control and Zic1 as positive control. All together, our study provides a confident list of chromatin-associated IncRNAs across wide range of tissues in mouse.

Properties of the chromatin-associated IncRNAs.

To investigate whether the two types of chromatin-associated IncRNAs have different properties, we calculated their sequence length and expression levels (Fig. 4A and B). plncRNAs with a median length of ~6 kb were not significantly different from elncRNAs. However, our finding of a ~6 kb for both elncRNAs and plncRNAs differs from a previous study, which reported them to be ~1 kb long (17). plncRNAs are highly expressed compared to elncRNAs, as previously observed (17). We asked whether these chromatin-associated IncRNAs were enriched in any biological processes by using nearest gene approach and whole-genome background with a GREAT software (37). Indeed, they showed enrichment of various biological processes (Fig. S9). Interestingly, we also observed the changes in the status of chromatin-associated IncRNAs based on their respective tissue or cell line. In total, ~17% chromatin-associated IncRNAs (144 out of 852) tend to switch from one chromatin state to another in multiple tissues (Table S6). plncRNAs are more likely to switch to plncRNAs and also the percentage of this type of transition is higher than the plncRNAs-to-elncRNAs or the elncRNAs-to-plncRNAs transition (Fig. 4C, D and Table S6).
We hypothesized that if a lncRNA is expressed in a specific tissue and also
associated with tissue-specific epigenetic modifications in the same tissue but
not in others, it could be associated with regulatory functions. To test this, we
selected for lncRNAs with the following characteristics: (1) associated with a
specific chromatin state only in ES cells, (2) expressed only in ES cells, (3)
associated with DNase I peaks only in ES cell, (4) associated with pluripotent
transcription factors in ES cells, and (5) close to a protein-coding gene
associated with pluripotency in ES cells. In total, 12 lncRNAs passed the above
filters.

For validation, we focused on a ES cell–specific, predicted regulatory enhancer–
associated lncRNA (chr7:132560406-132561472 (-)) located approximately 20 kb
away from the protein-coding gene Kdm8, which encodes a histone lysine
demethylase and regulates embryonic cell proliferation (Fig. 5A and 5D) (37).
We named this as lncRNA-Kdm8, based on its proximity to the Kdm8 protein-
coding gene. Using the RACE technique, we experimentally characterized the
lncRNA-Kdm8 genomic structure; this revealed at least 3 variants (RACE-a, b
and c) in the 5’ end of lncRNA-Kdm8, and also defined the exon-intron
boundaries (Fig. 5B and 5C). We then knocked-down lncRNA-Kdm8 with two
different siRNAs and checked the expression of the Kdm8 transcript and the
positive control gene Taf3. As predicted, upon lncRNA knockdown, the
expression of the Kdm8 gene significantly decreased as compared to Taf3, which
further supported the cis mode of enhancer-associated lncRNA gene regulation.
Together, our results show that chromatin-associated IncRNAs annotated by its chromatin marks could have regulatory roles.
DISCUSSION

Our study identified novel lncRNAs in mouse by using deep RNA sequencing data from eight tissues and an ES cell line. Public ENCODE large-scale RNA-seq data allowed us to de novo reconstruct confident novel lncRNA transcripts. The transcriptome data used in this study to discover lncRNAs go beyond previous IncRNA studies in terms of depth (18). The tissue-specific nature of these IncRNAs is in agreement with previous findings (3-5). The 2,803 lncRNAs included 2,174 antisense and 629 intergenic transcripts. Antisense lncRNAs have been shown to be key regulators and interestingly, many of the antisense lncRNA transcripts we observed were from ES cells. We used intersection of transcripts assembled by using two different de novo assemblers and also a stringent expression threshold to filter out the spurious transcripts. Further, we validated the expression of the lncRNA transcripts identified in this study by RT-PCR, thus confirming the quality of the transcripts identified in this study as well as their expression.

By using ChromHMM, we further characterized combinatorial chromatin state maps in mouse, using more than 70 ChIP-seq datasets across the same tissues used for lncRNA discovery. In previous studies, promoter, enhancer, and insulator maps were identified using a specific set of ChIP-seq datasets, like H3K4me3 (promoter), H3K4me1 with P300 (enhancer), and CTCF (insulator) (33). We built upon that work by further including additional histone marks.
allowing us to produce more detailed chromatin state maps. For example, the Fendrr IncRNA, which was previously annotated as enhancer-associated, has enhancer histone (p300/H3K4me1) marks (35) at the promoter but is also enriched in H3K27me3 in brain. We conclude that its chromatin status is likely to be poised or to switch to other states rather than to be enhancer-associated, which emphasizes the importance of taking chromatin states into account when classifying chromatin-associated IncRNAs.

By integrating chromatin state maps and promoters of IncRNAs across eight tissues and an ES cell line, we were able to classify IncRNAs into two classes: promoter-associated IncRNAs and enhancer-associated IncRNAs. Our study provides a comprehensive catalog of chromatin-associated IncRNAs across several mouse tissues. We also observed that plncRNAs were highly expressed, shorter in length compared to other chromatin-associated IncRNAs, and retained their embryonic promoter chromatin status in adult tissues. Experimental knockdown of an enhancer-associated IncRNA partially validated the regulatory behavior of chromatin state–associated IncRNAs in mouse.

Many of the bi-directional IncRNAs and enhancer-associated RNAs have been shown to be non-polyadenylated (34, 40). However, recent findings (2, 17), along with our study, suggest the existence of poly-adenylated bi-directional transcripts and chromatin-associated RNAs. Still, because of the polyA-based RNA
sequencing, we could be missing a large fraction of non-polyadenylated IncRNAs.

In the future, even more comprehensive catalogs of chromatin-associated IncRNAs should be possible to obtain by associating of chromatin states and IncRNA promoters across all tissues and cell lines in mammals. In addition, using techniques like CRISPR against regulatory IncRNAs would reveal more valuable information. All together, our study provides a novel set of classified IncRNAs, which presents a valuable resource for future genomic experimental studies in mouse.
MATERIALS AND METHODS

Computational procedures

Data sources.

All data used in the analysis were obtained from public databases. The links from where the data was obtained are listed in Table S7.

RNA-seq mapping and transcriptome assembly.

TopHat-2.0.9 (19) was used to map RNA-seq reads against mouse reference genome (mm9), using default parameters unless specified (Table S8). Cufflinks (20) was used to assembled mapped reads to transcripts de novo, and Cuffmerge was used against high-confidence de novo transcripts to generate a single transcript annotation file, using default parameters unless specified (Table S6). Scripture-v4 (21) was also used to assemble transcripts, using uniquely mapped reads with default parameters unless specified (Table S8). Finally, Qualimap-v.08 (41) was used with default parameters to count the number of strand-specific reads overlapping with IncRNAs.

Identification and genomic annotation of IncRNAs.

We filtered out transcripts from 8 tissues and a primary ES cell line pooled by Cuffmerge by using an in-house computational pipeline. Our pipeline relies in previously published software and protocols for identifying IncRNAs from transcriptomics data. The pipeline selects transcripts as IncRNAs by their size (≥200 nucleotides), number of exons (≥2 exons), expression levels (>1 FPKM in at least one tissue or cell line that we used), overlap with coding regions (no overlap with a known gene set from RefSeq, Ensembl, or UCSC on a similar
strand), overlap with noncoding regions (no overlap with known snoRNAs, tRNAs, miRNAs, IncRNAs, or pseudogenes), and noncoding potential (<0.44 CPAT and <100 PhyloCSF score). PhyloCSF (24) was used to calculate the coding potential of transcripts. First, we stitched mouse IncRNA exonic sequences into 18 mammals, using mm9-multiz30way alignments from UCSC. Second, we ran PhyloCSF against stitched sequences using default parameter unless specified (Table S8). We then removed the transcripts with open reading frames with a PhyloCSF score greater than 100, as previously suggested (36). The final IncRNA PhyloCSF score is the average decibans score of all its exons based on their strand direction and all possible frames. The transcripts that passed PhyloCSF and CPAT coding potential filters were further selected as potential IncRNAs.

LncRNAs that did not overlap with any known protein-coding gene (within a 10 kb window from both TSS and TES) were classified as intergenic IncRNAs or IncRNAs. LncRNAs that overlapped a transcript but on opposite strands were classified as antisense IncRNAs. LncRNAs that were close to a coding gene (within 10 kb from both TSS and TES) were annotated as either convergent (same strand as the nearest coding) or divergent (opposite strand as the nearest coding) IncRNAs.

Tissue specificity calculations.

To calculate tissue specificity of IncRNAs, we normalized the raw FPKM expression values as suggested in previous studies (4, 5). First, we added pseudo-count 1 to every raw FPKM value and second applied log2 normalization.
to each value, to obtain a non-negative expression vector. Finally, we normalized the expression vector by dividing it by the total expression counts. The resulting matrix of IncRNA normalized expression levels in each of the replica experiments per tissue or cell line was clustered by k-means.

**TFBS, CAGE tags, and DNase I sites enrichment analyses.**

To identify transcription factor binding sites, we first performed a *de novo* motif analysis on the 2,803 IncRNA 1 kb promoters, using the HOMER software with default parameters unless specified (Table S8). Second, the significant (*P* < 1e-5) *de novo* motifs from HOMER were used as input to the TOMTOM program to search against the JASPAR CORE and UNIPROBE databases (42). Next, we combined all identified motifs from both searches into a final list of transcription factor motifs. We then checked the expression of genes in the master list and required that the candidate transcription factor be expressed in the tissue. Finally, we used the PWMEnrich program (43) to perform motif enrichment analysis.

CAGE peak based annotations for mouse samples were downloaded from the FANTOM5 database (30), and DNase I sites from ENCODE (31). We overlapped these with the 2,803 IncRNA promoters and their corresponding random regions using *sitepro* from the CEAS program (44) with default parameters. We used the shuffledBed program ([https://code.google.com/p/bedtools/](https://code.google.com/p/bedtools/)) (45) with default parameters to randomize the coding RNA and IncRNA promoters in the mm9 genome.

**Discovery of chromatin state maps.**
We first collected mapped ChIP-seq reads of H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K27ac, CTCF, and RNA polymerase II from ENCODE. This data was originally produced from mouse (C57BL/6-strain, E14, or 8-week-old) brain, heart, kidney, liver, small intestine, spleen, testes, or thymus, or from an embryonic stem (ES) cell line. Second, we used a Poisson-based multivariate hidden Markov model29 (ChromHMM, http://compbio.mit.edu/ChromHMM/) to identify regions enriched in specific combinations of histone modifications as previously described but without extending the read lengths. We ran the ChromHMM software to produce classified maps containing from 2 to 50 states. The 15-state model was rich enough and, at the same time, allowed us to interpret the chromatin frequency observed across various tissues and cell lines. Next, we classified the 15-state model into the final six major chromatin state maps of active promoter, poised promoter, strong enhancer, poised or weak enhancer, insulator, repressed, transcribed, or heterochromatin states. In total, 3,612,616 regions in the mouse genome were enriched in at least one of the six major chromatin state maps.

Collection of RNA promoters.

We overlapped all 19,873 lncRNAs with protein-coding genes and removed the ones that overlapped by at least one nucleotide on either strand. This resulted in 14,147 intergenic lncRNAs. We avoided protein-coding vicinities by removing the lncRNAs that fall within 1 kb from either the TSS or the TES of any known protein-coding gene. This resulted in 12,129 strictly intergenic lncRNAs. Further, we selected lncRNAs with an expressed of more than 1 FPKM in a given tissue.
All together, the filters resulted in 1,385 lncRNAs in whole brain, 1,236 in ES cells, 903 in heart, 870 in kidney, 787 in liver, 435 in small intestine, 878 in spleen, 2,083 in testes and 932 in thymus. We created 200 bp promoters of these expressed lncRNAs by extending the TSS 100 bp upstream and downstream. We created random promoters by shuffling across intergenic space and then overlapped these promoters with chromatin states in each tissue separately. Next, we used ~30,000 RefSeq protein-coding gene promoters and overlapped them with chromatin states in a similar fashion as above (>1 FPKM in a given tissue).

*Overlapping chromatin state maps with RNA promoters.*

We used intersectBed from BEDtools package (45) to overlap RNA promoters with chromatin state maps in each tissue or cell line. We considered the chromatin association to be significant if the *P* value was less than 0.001 (Fischer-exact test) in all the tissues we tested. We found both active promoter and strong enhancer chromatin states significantly associated with lncRNA promoters (*Table S4* and *Fig. 3B*). We used CAGE peaks from FANTOM5 and DNase-seq peaks from ENCODE, along with RNA-seq expression, to identify active promoters lncRNA in liver, spleen, and thymus. We could not find both CAGE and DNase-seq data for other tissues. We used the same 200 bp promoter size for CAGE peaks (more than 1 tag) and overlapping DNase-seq peaks (*Table S5*).

*Transition of chromatin-associated lncRNAs.*
We selected 200-bp-long promoters of expressed lncRNAs (>1 FPKM) in whole brain and made sure that they did not overlap any protein-coding genes within a 5 kb distance (both from TSS and TES). We then overlapped the IncRNA promoters with active promoter and strong enhancer chromatin states in whole brain. The analysis resulted in 163 elncRNAs and 33 plncRNAs in whole brain. We repeated the above steps in other tissues, resulting in hundreds of chromatin-associated lncRNAs. This produced 41 ES-elncRNAs, 131 ES-plncRNAs, 21 heart-elncRNAs, 61 heart-plncRNAs, 47 kidney-elncRNAs, 61 kidney-plncRNAs, 35 liver-elncRNAs, 77 liver-plncRNAs, 25 small intestine–elncRNAs, 20 small intestine–plncRNAs, 20 spleen-elncRNAs, 65 spleen-plncRNAs, 88 testes-elncRNAs, 258 testes-plncRNAs, 82 thymus-elncRNAs and 50 thymus-plncRNAs. Finally, we calculated the percentage of transition of chromatin-associated IncRNA from one tissue to another (Table S6).

Gene ontology analysis.

We ran GREAT annotation tool on chromatin-associated IncRNA genomic locations by taking the two nearest genes, using a default of a 1,000 kb distance window. A whole-genome background was selected as a control.

Experimental Procedures

Cell culture.

Wild-type (E14Tg2A) ESCs were cultured feeder–free in plates coated with 0.1% of gelatin in Glasgow minimum essential medium (Sigma) supplemented with β-mercaptoethanol, sodium pyruvate, essential amino acids, GlutaMAX, 20% fetal bovine serum (Hyclone), and leukemia inhibitory factor (LIF). Heart, liver, and
kidneys were isolated from 8-week-old C57BL/6J mice and snap-frozen before RNA extraction for chromatin immunoprecipitation assays (only heart).

**Chromatin immunoprecipitation assay.**

ESCs were cross-linked in 1% formaldehyde (FA) for 10 min at room temperature (RT). For ChIPs from heart, crosslinking was performed on 1- to 3-mm³ fragments in a conical tube for 10 min rotating at RT in 1.5% FA. Crosslinking was quenched with 0.125 M glycine for 5 min. Pelleted cells and heart fragments were lysed and homogenized. Chromatin extraction and immunoprecipitation was performed as previously described (Morey et al, 2012).

300 μg were used for immunoprecipitation. Antibodies used were: Suz12, Abcam ab12073; histone H3, Abcam ab1791; histone H3K4me1, Abcam ab8895; histone H3K27me3, Active-Motif 39155; and histone H3K27ac, Millipore 07-360. The primers used in the qPCR assays are listed in Table S2.

**Expression and siRNA knockdown analyses.**

RNA from organs was extracted with Trizol (Life Technologies). cDNA was generated from 1 mg of RNA with the First Strand cDNA Synthesis Kit (Fermentas). The primers used in the RT-qPCR assays are listed in Table S2. RT-PCR was performed in duplicates using GAPDH as a housekeeping gene for normalization. For ES-specific IncRNA knock-downs, 50,000 cells/well in 6-well plates were seeded and then transfected the next day with Lipofectamine RNAiMAX Reagent and 75 pmol of siRNA duplexes (Invitrogen). Cells were pelleted 24 h post-transfection, and RNA was extracted for RT-qPCR with the RNA extraction kit (QIAGEN). cDNA was generated as explained above. The
primers used in the RT-qPCR assays and the siRNA duplexes used are listed in
\textbf{Table S9}. RT-PCR was performed in triplicates using GAPDH as a
housekeeping gene for normalization.

\textit{Characterization of mouse IncRNA-Kdm8 using RACE.}

Total RNA extracted from mouse ES cells (E14) was used to generate RACE-ready
3'- and 5'-cDNA using the SMARTer RACE cDNA Amplification Kit
(Clontech) following the manufacturer's protocol. cDNA ends were amplified with
universal primer mix and gene-specific primers (GSP), followed by a 'nested'
PCR with the nested universal primer and the nested gene-specific primers
(NGSP) (\textbf{Table S9}). RACE products were run on a 2% agarose gel, cloned in
pRACE (pUC19-based vector), and sequenced using M13 primers. Recovered
fragments were aligned to obtain the different full-length transcripts produced by
the IncRNA-Kdm8 (\textbf{Table S9}).

\textit{Data access.}

All lncRNAs and chromatin state maps identified in this work for mouse (mm9)
are listed in the additional files lncRNAs.xlsx and ChromatinMaps.zip (\textbf{Table S7}).

\textit{Competing Interests}

The authors declare that they have no competing interests.

\textit{Authors’ Contributions}

G.K.B. conceived the study, collected the data, analyzed the data, interpreted the
data, and wrote the manuscript. P.V. conducted qPCR and ChIP-PCR
experiments. L.W.S., M.B., L.D.C. and M.A.M.-R. contributed ideas and wrote the manuscript.

ACKNOWLEDGMENTS

We sincerely thank ENCODE consortium for publicly providing rich data. We are thankful for the many productive discussions, especially with Rory Johnson (lncRNAs), Jason Ernst and Guillaume Fillion (chromatin state maps), Irwin Jungreis (PhyloCSF), Jochen Hecht (RACE), Sabah Kadri (Scripture), and Veronica Raker (manuscript edition). We also thank the three anonymous reviewers for their critical insights. The project was supported by a grant from la Caixa to G.K.B. and by the Spanish MINECO to M.A.M-R. (BFU2010-19310 and BFU2013-47736-P). We also acknowledge support of the Spanish Ministry of Economy and Competitiveness, ‘Centro de Excelencia Severo Ochoa 2013-2017’, SEV-2012-0208.
Figure 1. Overview of the lncRNA discovery and chromatin state map computational pipeline. (A) Overview of the lncRNA discovery and chromatin state map based classification pipeline that was employed using both RNA-seq and ChIP-seq data from 8 tissues and one primary cell line (ES) in mouse. RNA-seq reads from all the tissues and the cell line were mapped using TopHat2 against mouse reference genome (mm9), and transcriptomes were assembled in de novo using Cufflink2 and ScriptureV4 assemblers. Common transcripts that were assembled by both Cufflinks2 and ScriptureV4 were scanned for lncRNA features like size, length, exon number, expression and coding score. A library of intergenic lncRNAs was constructed by pooling lncRNAs identified in this study and previous studies. In total, 10,728 unique lncRNAs were overlapped with chromatin state maps discovered by using ChromHMM by pooling various ChIP-Seq datasets and classified chromatin-associated lncRNAs in mouse. (B) Overlap between lncRNAs identified in this study (light grey, left) and previously published lncRNAs (dark grey, right). 2,803 non-annotated lncRNAs were identified, and 34% (13,382) of the known lncRNAs were recovered in this study. (C) RNA-seq coverage tracks showing the expression of a novel lncRNA identified in this study (black). Transcription in testes is shown. “+” and “−” indicate sense and antisense directions, respectively, and experimental replicates are numbered as “1” and “2”.

Figure 2. Tissue- and cell-specific expression of lncRNAs. (A) Heatmap representing normalized FPKM expression values of the 2,803 lncRNAs (rows) across eight tissues and a primary cell line (columns). Rows and columns were ordered based on k-means clustering. Legend color intensity represents the fractional density across the row of log10-normalized FPKM expression values as estimated by ScriptureV4. Each tissue has 2 columns, representing their replicates, and the ES cell line has 5 columns. (B) Experimentally validated examples of lncRNAs with tissue-specific expression across heart, liver, and kidney. Shown are RT-PCR duplicates normalized (against housekeeping gene GAPDH) expression of heart-specific lncRNAs (H-lnc1 and H-lnc2), liver-specific lncRNAs (L-lnc1 and L-lnc2), and kidney-specific lncRNAs (K-lnc1 and K-lnc2) (Table S9).

Figure 3. Discovery of chromatin state maps and their association with lincRNAs. (A) Emission parameters learned de novo with ChromHMM on the basis of combinations recurring in chromatin. Each point in the table denotes the frequency with which a given mark is found at genomic positions corresponding to a specific chromatin state. The observation frequency of various chromatin marks, including H3K36me3, H3K4me1, H3K27ac, Pol II, H3K4me3, CTCF, and H3K27me3, as well as respective input showing 6 major chromatin states, including active promoter (red), poised promoter (purple), enhancer (yellow), Polycomb (grey), insulator (blue), and heterochromatin (white). (B) Percentage of protein-coding TSS (top) and intergenic lncRNAs (bottom) significantly enriched...
with both active promoter and strong enhancer (***, \( P < 0.001 \), Fisher-exact test). “D” and “R” labels correspond to the observed data and randomized TSSs, respectively. (C) Percentage of IncRNAs and protein-coding genes that are associated with promoter and enhancer chromatin states. (D) The number of plncRNAs and elncRNAs across 8 tissues and an ES cell line. (E) Percentage of IncRNAs (overlapped with both CAGE peaks and DNase I hypersensitive sites) associated with promoter and enhancer chromatin states.

Figure 4. Transcript length, expression, and transition of chromatin-associated IncRNAs in mouse. (A) Transcript length of elncRNAs (median = 6565 nt) and plncRNAs (median = 6450 nt) across eight tissues and a cell line, showing no difference in length (Mann-Whitney test; NS, not significant; \( P = 0.9848 \)). (B) Log-normalized expression (FPKM) of elncRNAs (median = 0.08 FPKM) and plncRNAs (median = 0.33 FPKM) across eight tissues and an ES cell line, showing a significant difference between them (Mann-Whitney test, ***, \( P = 1.221 \times 10^{-10} \)). (C) Circos plot showing the transition of plncRNA to elncRNA, or elncRNA to plncRNA, across eight tissues and an ES cell line. Top bars indicate the total number of chromatin-associated IncRNAs that undergo a transition per tissue or cell line, which included whole brain (20 plncRNAs and 72 elncRNAs), ES cells (62 plncRNAs and 8 elncRNAs), heart (44 plncRNAs and 4 elncRNAs), small intestine (17 plncRNAs and 18 elncRNAs), kidney (50 plncRNAs and 24 elncRNAs), liver (46 plncRNAs and 10 elncRNAs), spleen (55 plncRNAs and 12 elncRNAs), testes (29 plncRNAs and 12 elncRNAs) and thymus (47 plncRNAs and 40 elncRNAs). Links inside the bars indicate the number of IncRNAs that switch their chromatin states from one tissue to another (red, plncRNAs; gold, elncRNAs). The IncRNA transition table used to generate the circos plot is shown in Table S6. (D) Percentage of chromatin-associated transitions across all the mouse tissues, showing the high percentage of plncRNA-to-plncRNA transitions as compared to elncRNA-to-elncRNA transitions.

Figure 5. An enhancer-associated IncRNA, IncRNA-Kdm8, regulates the expression of a neighboring protein-coding gene Kdm8. (A) The IncRNA-Kdm8 locus promoter overlaps with an enhancer chromatin state and occurs within 20 kb of the TSS of a protein-coding gene, Kdm8 (e.g., it is an enhancer-associated IncRNA). Gene tracks represent DNase I hypersensitive sites (HS) and ChIP-seq data for H3K4me1, H3K27ac, and H3K4me3 from ENCODE. The genomic scale is indicated on the top, and the scale of both DNase I HS and ChIP-seq data on the top right. (B and C) The 5’- and 3’-ends and the exon-intron boundaries of the enhancer-associated IncRNA, IncRNA-Kdm8, were determined by RACE (see Supplemental Materials and Methods). Black arrows depict TSSs and direction of transcription for respective genes. Kdm8 mRNA and the IncRNA-Kdm8 are shown in green and red, respectively. Genomic DNA sequence corresponding to the 5’- and 3’-ends of the cloned IncRNA are shown in black at the bottom of the IncRNA-Kdm8 gene track, defining accurate 5’-end and exon-intron boundaries for exon 1, exon 3, exon 4, and exon 5 of IncRNA-Kdm8. (D) Expression levels of IncRNA-Kdm8 in mES cells and other tissues, as
measured by directional RNA-seq and expressed as fragments per kilobase of
exonic length per million (FPKM). (E) RT-PCR expression (triplicates, normalized
against housekeeping gene RPO) after siRNA-based knockdown of IncRNA-
Kdm8 (chr7:132560406-132561472, −) resulted in a significant decrease of
neighboring gene Kdm8 (t-test *P ≤0.05, **P ≤0.01), which was not observed for
the negative control of the distant coding gene Taf3 (chr2:9836179-9970236,
+). Primers used for siRNA oligonucleotides of the IncRNA-Kdm8 are given
in Table S9.
REFERENCES


31. Mouse ENCODE Consortium, Stamatoyannopoulos JA, Snyder M,
897  Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M,  
898  Kaul R, Canfield T, Giste E, Johnson A, Zhang M, Balasundaram G,  
899  Byron R, Roach V, Sabo PJ, Sandstrom R, Stehling AS, Thurman RE,  
901  Z, Wold BJ, Dekker J, Crawford GE, Keller CA, Wu W, Morrissey C,  
902  Kumar SA, Mishra T, Jain D, Byrsk-Bishop M, Blankenberg D, Lajoie  
904  MJ, Pimkin M, Deng W, Marinov GK, Williams BA, Fisher-Aylor KI,  
905  DeSalvo G, Kiralusha A, Trout D, Amrhein H, Mortazavi A, Edsall L,  
906  McCleary D, Kuan S, Shen Y, Yue F, Ye Z, Davis CA, Zaleski C, Jha S,  
908  Lagarde J, Ryba T, Sasaki T, Malladi VS, Cline MS, Kirkup VM,  
909  Learned K, Rosenbloom KR, Kent WJ, Feingold EA, Good PJ, Pazin  
910  M, Lowdon RF, Adams LB. 2012. An encyclopedia of mouse DNA  
914  33. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon  
915  J, Lee L, Lobanenkov VV, Ren B. 2012. A map of the cis-regulatory  
917  34. Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA,  
918  Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E,  
921  35. Sauvageau M, Goff LA, Lodato S, Bonev B, Groff AF, Gerhardinger C,  
922  Sanchez-Gomez DB, Hacisuleyman E, Li E, Spence M, Liapis SC,  
923  Mallard W, Morse M, Sverdel MR, D’Eccelessis MF, Moore JC, Lai V,  
924  Gong G, Yancopoulos GD, Frenedewy D, Kellis M, Hart RP,  
925  Valenzuela DM, Arlotta P, Rinn JL. 2013. Multiple knockout mouse  
926  models reveal lincRNAs are required for life and brain development. eLife  
927  2:e01749–e01749.  
928  36. Alvarez-Dominguez JR, Hu W, Yuan B, Shi J, Park SS, Gromatzky AA,  
929  Oudenaarden AV, Lodish HF. 2014. Global discovery of erythroid long  
930  noncoding RNAs reveals novel regulators of red cell maturation. Blood  
931  123:570–581.  
932  37. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB,  
933  Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of  
935  38. Lai F, Orom UA, Cesaroni M, Beringer M, Taatjes DJ, Blobel GA,  
936  Shiekhattar R. 2013. Activating RNAs associate with Mediator to enhance  
941  41. Garcia-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Gotz S,


A. Novel lncRNAs (2,803) and their overlap (13,382) with known lncRNAs.

1. >= 200 nt (Size)
2. >= 2 exons
3. >= 1 FPKM (Expression)
4. < 0.44 CPAT (Coding Score)
5. Assembled by both Cufflinks and Scripture
6. No overlap with known protein-coding, non-coding and lncRNA transcripts (UCSC, ENSEMBL, RefSeq and published lncRNAs)

B. Known lncRNAs (39,444) and novel lncRNAs (2,803) overlapped with known lncRNAs.

C. Chromatin maps showing expression patterns in various tissues (e.g., Brain, Heart, Liver, Kidney, Spleen, Small Intestine, Testes, Thymus).
A.

B.

Normalized Expression

Testes  ES  Brain  Intestine  Spleen  Thymus  Heart  Kidney  Liver

H-lnc1  H-lnc2  L-lnc1  L-lnc2  K-lnc1  K-lnc2

Heart  Liver  Kidney

2,803 IncRNA

0.0  0.5  1.0
A. Table showing the number of intergenic lncRNAs and their transcriptional states.

<table>
<thead>
<tr>
<th>Chromatin mark</th>
<th>Observation frequency (%)</th>
<th>Coverage (Mean)</th>
<th>Length (Mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription Elongation</td>
<td>4.2</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>Weakly Transcribed</td>
<td>6.7</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td>Transcriptional Transition</td>
<td>0.8</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>Weak/poised Enhancer</td>
<td>0.6</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>Active Promoter</td>
<td>0.4</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Strong Enhancer</td>
<td>0.2</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>Weak/poised Enhancer</td>
<td>0.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>Active Promoter</td>
<td>0.7</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>Strong Enhancer</td>
<td>0.3</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>Weak/poised Enhancer</td>
<td>0.3</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>Active Promoter</td>
<td>0.7</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Strong Enhancer</td>
<td>1.3</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>Repressed</td>
<td>18.5</td>
<td>17.7</td>
<td></td>
</tr>
<tr>
<td>Heterochromatin</td>
<td>44.1</td>
<td>120.7</td>
<td></td>
</tr>
<tr>
<td>Heterochromatin</td>
<td>19.7</td>
<td>5.7</td>
<td></td>
</tr>
<tr>
<td>Insulator</td>
<td>19.7</td>
<td>19.7</td>
<td></td>
</tr>
</tbody>
</table>

B. Graph showing the percentage of intergenic lncRNAs associated with promoters and enhancers.

C. Bar graph comparing the percentage of lincRNA and protein-coding transcripts associated with promoters and enhancers.

D. Bar graph showing the number of intergenic lncRNAs associated with enhancers and promoters in different tissues.

E. Bar graph showing the percentage of intergenic lncRNAs associated with promoters in different tissues.