AMERICAN SOCIETY FOR MICROBIOLOGY | Molecular and Cellular Biology

# Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse

Gireesh K. Bogu,[a,b,c,d] Pedro Vizán,[b,d] Lawrence W. Stanton,[e,f] Miguel Beato,[b,d] Luciano Di Croce,[b,d,g] Marc A. Marti-Renom[a,b,d,g]

CNAG-CRG, Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain[a]; Gene Regulation, Stem Cells and Cancer Program, Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona, Spain[b]; Bioinformatics and Genomics Program, Center for Genomic Regulation (CRG) , Barcelona Institute of Science and Technology, Barcelona, Spain[c]; Universitat Pompeu Fabra (UPF), Barcelona, Spain[d]; Department of Biological Sciences, National University of Singapore, Singapore, Singapore[e]; Stem Cell and Developmental Biology Group, Genome Institute of Singapore, Singapore, Singapore[f]; Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain[g]

**Discovering and classifying long noncoding RNAs (lncRNAs) across all mammalian tissues and cell lines remains a major challenge. Previously, mouse lncRNAs were identified using transcriptome sequencing (RNA-seq) data from a limited number of tissues or cell lines. Additionally, associating a few hundred lncRNA promoters with chromatin states in a single mouse cell line has identified two classes of chromatin-associated lncRNA. However, the discovery and classification of lncRNAs is still pending in many other tissues in mouse. To address this, we built a comprehensive catalog of lncRNAs by combining known lncRNAs with high-confidence novel lncRNAs 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 lncRNAs with multiple genome-wide chromatin state maps and found two different classes of chromatin state-associated lncRNAs, including promoter-associated (plncRNAs) and enhancer-associated (elncRNAs) lncRNAs, across various tissues. Experimental knockdown of an elncRNA resulted in the downregulation of the neighboring protein-coding *Kdm8* gene, encoding a histone demethylase. Our findings provide 2,803 novel lncRNAs and a comprehensive catalog of chromatin-associated lncRNAs across different tissues in mouse.**

Previous large-scale transcriptome-sequencing (RNA-seq) 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 intergenic long noncoding RNAs (lncRNAs). Comprehensive lncRNA catalogs were recently established for various cell lines and tissues in human, mouse, *Caenorhabditis elegans*, *Drosophila*, and zebrafish (3–8). In addition, we now know the functions of a limited number of the discovered lncRNAs, such as Xist in X chromosome inactivation (9), HOTAIR in cancer metastasis (10), lnc-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 lncRNA regulates neurogenesis by physically interacting with the Sox2 transcription factor (16).

Even though thousands of lncRNAs have been cataloged, it is still unclear how to characterize regulatory lncRNAs. Very recently, regulatory lncRNAs 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 lncRNAs associated with these two chromatin states, since the lncRNA associations were not tested in multiple tissues. In addition, the lncRNA or chromatin state data sets used in the previous study (17) were selected only in a single cell line, which technically limits testing of thousands of lncRNAs. Finally, it is also unknown whether these lncRNAs associate with similar chromatin states across different tissues.

To build a comprehensive chromatin-associated mouse lncRNA data set, we first used billions of mapped RNA-seq reads to identify high-confidence novel lncRNAs and then combined them with thousands of known lncRNAs. Second, we used more than a billion mapped chromatin immunoprecipitation sequencing (ChIP-seq) reads of various histone marks to identify chromatin state maps. Finally, we integrated all these mouse lncRNAs with the 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 lncRNA 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 data set of chromatin state-associated lncRNAs in mouse, and we expect it will be a valuable resource to help researchers select candidate lncRNAs for further experimental studies.

## MATERIALS AND METHODS

**Computational procedures. (i) Data sources.** All data used in the analysis were obtained from public databases. The links through which the data were obtained are listed in Table S7 in the supplemental material. All

Received 19 October 2015 Returned for modification 3 December 2015
Accepted 17 December 2015
Accepted manuscript posted online 28 December 2015
Citation Bogu GK, Vizán P, Stanton LW, Beato M, Di Croce L, Marti-Renom MA. 2016. Chromatin and RNA maps reveal regulatory long noncoding RNAs in mouse. Mol Cell Biol 36:809–819. doi:10.1128/MCB.00955-15.

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

Supplemental material for this article may be found at http://dx.doi.org/10.1128/MCB.00955-15.

Copyright © 2016, American Society for Microbiology. All Rights Reserved.

Downloaded from http://mcb.asm.org/ on February 22, 2016 by Biblioteca de la Universitat Pompeu Fabra

novel lncRNAs identified in this study are listed in Table S2 in the supplemental material, and chromatin state maps can be accessed from https://github.com/gireeshkbogu/chromatin_states_chromHMM_mm9.

**(ii) RNA-seq mapping and transcriptome assembly.** TopHat 2.0.9 (18) was used to map RNA-seq reads against the mouse reference genome (mm9), using default parameters unless otherwise specified (see Table S8 in the supplemental material). Cufflinks (19) was used to assemble 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 otherwise specified (see Table S8 in the supplemental material). Scripture v4 (20) was also used to assemble transcripts, using uniquely mapped reads with default parameters unless otherwise specified (see Table S8 in the supplemental material). Finally, Qualimap v.08 (21) was used with default parameters to count the strand-specific reads overlapping lncRNAs.

**(iii) Identification and genomic annotation of lncRNAs.** We filtered out transcripts from 8 tissues and a primary embryonic stem (ES) cell line pooled by Cuffmerge by using an in-house computational pipeline. Our pipeline relies on previously published software and protocols to identify lncRNAs from transcriptomics data. The pipeline selects transcripts as lncRNAs by their size (≥200 nucleotides [nt]), number of exons (≥2 exons), expression levels (>1 fragment per kilobase of exonic length per million [FPKM] in at least one tissue or cell line that we used), overlap coding regions (no overlap with a known gene set from RefSeq, Ensembl, or UCSC on a similar strand), overlap noncoding regions (no overlap with known snoRNAs, tRNAs, microRNAs [miRNAs], lncRNAs, or pseudogenes), and noncoding potential (<0.44 CPAT [22] and <100 PhyloCSF score). PhyloCSF (23) was used to calculate the coding potential of transcripts. First, we stitched mouse lncRNA exonic sequences into 18 mammals, using mm9-multiz30way alignments from UCSC. Second, we ran PhyloCSF against the stitched sequences, using default parameters unless otherwise specified (see Table S8 in the supplemental material). We then removed the transcripts with open reading frames with a PhyloCSF score greater than 100, as previously suggested (24). The final lncRNA PhyloCSF score is the average deciban 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 lncRNAs.

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

**(iv) Tissue specificity calculations.** To calculate the tissue specificity of lncRNAs, we normalized the raw FPKM expression values, as suggested in previous studies (4, 5). First, we added pseudocount 1 to every raw FPKM value, and second, we applied $\log_2$ normalization to each value to obtain a nonnegative expression vector. Finally, we normalized the expression vector by dividing it by the total expression counts. The resulting matrix of lncRNA-normalized expression levels in each of the replicate experiments per tissue or cell line was clustered by *k* means.

**(v) Transcription factor binding sites, CAGE tags, and DNase I site enrichment analyses.** To identify transcription factor binding sites, we first performed a *de novo* motif analysis of the 2,803 lncRNA 1-kb promoters, using HOMER software with default parameters unless otherwise specified (see Table S8 in the supplemental material). 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 (25). 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 (R package version 3.6.1 1–46, 2014) to perform motif enrichment analysis.

Cap analysis gene expression (CAGE) peak-based annotations for mouse samples were downloaded from the FANTOM5 database (26) and DNase I sites from ENCODE (27). We overlapped these with the 2,803 lncRNA promoters and their corresponding random regions using sitepro from the CEAS program (28) with default parameters. We used the shuffledBed program (29) with default parameters to randomize the coding RNA and lncRNA promoters in the mm9 genome.

**(vi) Discovery of chromatin state maps.** We first collected mapped ChIP-seq reads of H3 lysine 4 monomethylation (H3K4me1), H3 lysine 4 trimethylation (H3K4me3), H3 lysine 36 trimethylation (H3K36me3), H3 lysine 27 trimethylation (H3K27me3), and H3 lysine 27 monoacetylation (H3K27ac), CCCTC-binding factor (CTCF), and RNA polymerase II from ENCODE. These data were originally produced from mouse (strain C57BL/6; embryonic day 14 [E14] or 8 weeks old) brain, heart, kidney, liver, small intestine, spleen, testis, or thymus or from an ES cell line. Second, we used a Poisson-based multivariate hidden Markov model 29 (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 and poised promoter, strong enhancer and poised or weak enhancer, insulator, repressed, transcribed, or heterochromatin state. In total, 3,612,616 regions in the mouse genome were enriched in at least one of the six major chromatin state maps. promoter (active and poised), enhancer (strong and poised/weak), transcribed (transcription transition, elongation, and weak transcription), insulator, repressed, and heterochromatin.

**(vii) Collection of RNA promoters.** We overlapped all 19,873 lncRNAs with protein-coding genes and removed the ones that overlapped by at least 1 nucleotide on either strand. This resulted in 14,147 intergenic lncRNAs. We avoided protein-coding vicinities by removing the lncRNAs that fell 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 expression of more than 1 FPKM in a given tissue. Altogether, 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 testis, 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 fashion similar to that described above (>1 FPKM in a given tissue).

**(viii) Overlapping chromatin state maps with RNA promoters.** We used intersectBed from the BEDtools package (29) 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 (Fisher exact test) in all the tissues we tested. We found both active promoter and strong enhancer chromatin states significantly associated with lncRNA promoters (see Fig. 3B; see Table S4 in the supplemental material). We used CAGE peaks from FANTOM5 and DNase sequencing (DNase-seq) peaks from ENCODE, along with RNA-seq expression, to identify active promoter 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 (see Table S5 in the supplemental material).

**(ix) 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

Downloaded from http://mcb.asm.org/ on February 22, 2016 by Biblioteca de la Universitat Pompeu Fabra

5-kb distance (from both TSS and TES). We then overlapped the lncRNA promoters with active promoter and strong enhancer chromatin states in whole brain. The analysis resulted in 163 enhancer-associated lncRNAs (elncRNAs) and 33 promoter-associated lncRNAs (plncRNAs) in whole brain. We repeated the above-mentioned 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 testis elncRNAs, 258 testis plncRNAs, 82 thymus elncRNAs, and 50 thymus plncRNAs. Finally, we calculated the percentage of transition of chromatin-associated lncRNA from one tissue to another (see Table S6 in the supplemental material).

**(x) Gene ontology analysis.** We ran the GREAT annotation tool (30) on chromatin-associated lncRNA 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. (i) Cell culture.** Wild-type (E14Tg2A) ES cells were cultured feeder free in plates coated with 0.1% 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).

**(ii) Chromatin immunoprecipitation assay.** ES cells were cross-linked in 1% formaldehyde (FA) for 10 min at room temperature. For ChIPs from heart, cross-linking was performed on 1- to 3-$mm^3$ fragments in a conical tube for 10 min with rotation at room temperature in 1.5% FA. Cross-linking was quenched with 0.125 M glycine for 5 min. Pelleted cells and heart fragments were lysed and homogenized. Chromatin extraction and immunoprecipitation were performed as previously described (31), and 300 μg was used for immunoprecipitation. The antibodies used were as follows: 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 quantitative-PCR (qPCR) assays are listed in Table S2 in the supplemental material.

**(iii) Expression and siRNA knockdown analyses.** RNA from organs was extracted with TRIzol (Life Technologies). cDNA was generated from 1 μg of RNA with the First Strand cDNA synthesis kit (Fermentas). The primers used in the quantitative real-time PCR (qRT-PCR) assays are listed in Table S2 in the supplemental material. qRT-PCR was performed in duplicate using the GAPDH (glyceraldehyde-3-phosphate dehydrogenase) gene as a housekeeping gene for normalization. For ES-specific lncRNA knockdowns, 50,000 cells/well in 6-well plates were seeded and then transfected the next day with Lipofectamine RNAiMax reagent and 75 pmol of small interfering RNA (siRNA) duplexes (Invitrogen). The cells were pelleted 24 h posttransfection, and RNA was extracted for qRT-PCR with an RNA extraction kit (Qiagen). cDNA was generated as explained above. The primers used in the qRT-PCR assays and the siRNA duplexes used are listed in Table S9 in the supplemental material. qRT-PCR was performed in triplicate, using the GAPDH gene as a housekeeping gene for normalization.

**(iv) Characterization of mouse lncRNA-*Kdm8* (see below) using RACE.** Total RNA extracted from mouse ES cells (E14) was used to generate rapid amplification of cDNA ends (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) (see Table S9 in the supplemental material). The RACE products were run on a 2% agarose gel, cloned in pRACE (a pUC19-based vector), and sequenced using M13 primers. The recovered fragments were aligned to obtain the different full-length transcripts produced by lncRNA-*Kdm8* (see Table S9 in the supplemental material).

## RESULTS

**Transcriptome mapping, assembly, and quantification.** About 3 billion raw sequence reads from RNA-seq experiments were downloaded from the ENCODE project (32) and analyzed using a computational pipeline consisting of TopHat (v2.0.9) (18), Cufflinks (v2.1.1) (19), and Scripture (v4) (20) (Fig. 1A). We constructed a map of RNA expression in mouse by first collecting RNA sequencing reads using long (76- to 108-nucleotide), paired-end, polyadenylated, strand-specific high-throughput RNA sequencing data from 8-week-old adult brain, heart, kidney, small intestine, liver, spleen, testis, and thymus and a paired-end ES cell line (see Table S1 in the supplemental material). 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 with either intergenic loci or 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 (see Materials and Methods). In total, Scripture identified 593,102 multiexonic transcripts and Cufflinks identified 539,775 transcripts, with an overlap of 500,530 transcripts between the two methods. Of those overlapping transcripts, ~86% (429,818) overlapped known coding transcripts (annotated in either RefSeq, UCSC, or Ensembl) and 10.2% (51,134) overlapped known noncoding transcripts (annotated as either snoRNA, tRNA, miRNA, or pseudogenes). This shows the quality of the transcripts and their ability to recover known noncoding transcripts. The remaining 3.9% of the transcripts (20,018) did not overlap 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 or intronic) (4, 5, 33). We identified 16,185 multiexonic lncRNAs longer than 200 bp and with an expression level of ≥1 FPKM 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 (22) and conservation-dependent PhyloCSF (23) (see Materials and Methods). About 85% of this data set overlapped previously identified lncRNAs (17, 20, 34–38) (see Fig. S1 in the supplemental material), 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 novel lncRNAs in mouse. Further, based on the genomic locations of lncRNAs relative to the nearest protein-coding gene promoters, we annotated 2,174 antisense (i.e., overlapping the protein-coding gene in an antisense direction), 382 intergenic (e.g., located within 10 kb of 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. 1C and Fig. S2 in the supplemental material show examples of a novel lncRNA identified in testes).

Downloaded from http://mcb.asm.org/ on February 22, 2016 by Biblioteca de la Universitat Pompeu Fabra

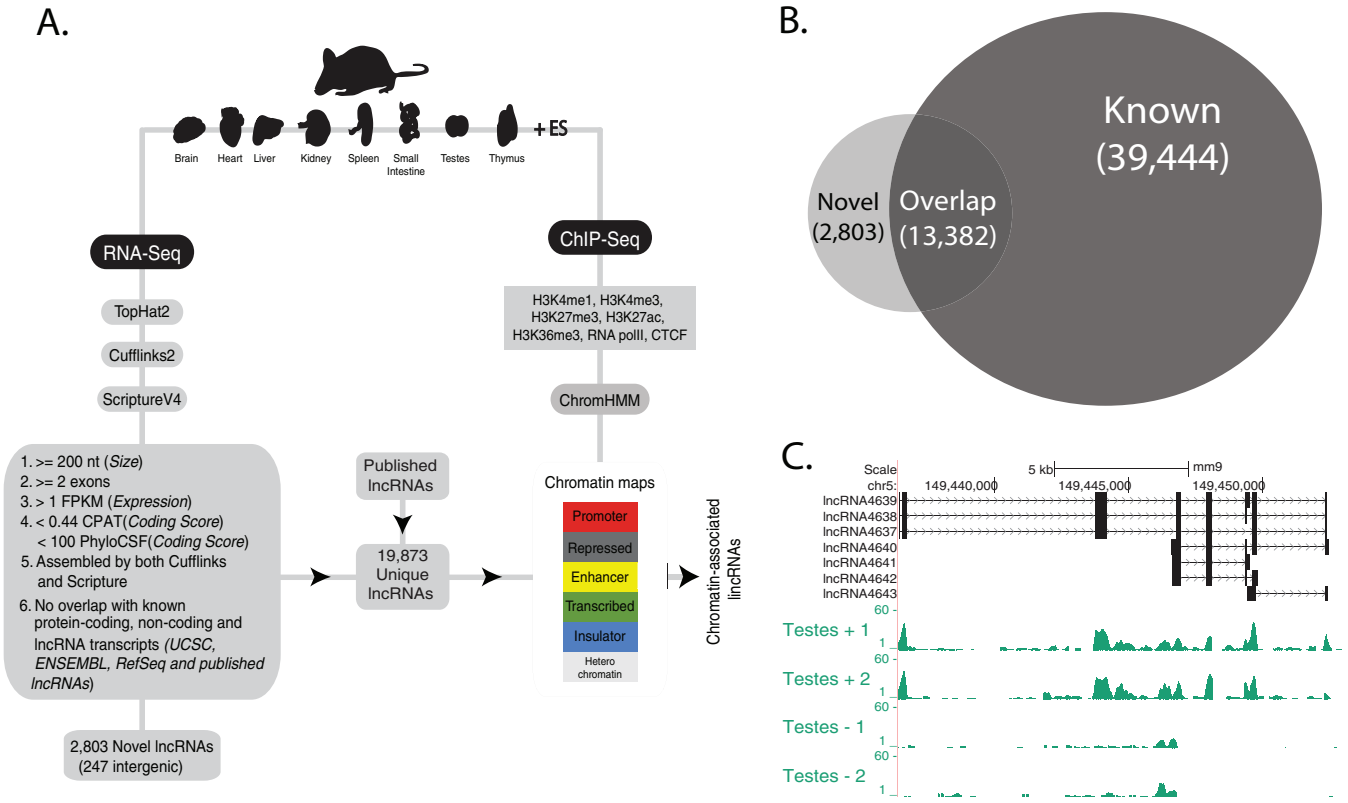


**FIG 1** Overview of the lncRNA discovery and chromatin state map computational pipeline. (A) Overview of the lncRNA discovery and chromatin state map-based classification pipeline that was employed using both RNA-seq and ChIP-seq data from 8 tissues and one primary cell line (ES) in mouse. RNA-seq reads from all the tissues and the cell line were mapped using TopHat 2 against the mouse reference genome (mm9), and transcriptomes were assembled *de novo* using Cufflinks 2 and Scripture v4 assemblers. Common transcripts that were assembled by both Cufflinks 2 and Scripture v4 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 data sets and classified chromatin-associated lncRNAs in mouse. (B) Overlap between lncRNAs identified in this study (small circle) and previously published lncRNAs (large circle; UCSC/Ensembl/RefSeq [5, 17, 20, 34–38]). A total of 2,803 nonannotated 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 1 and 2.

**Properties of the 2,803 lncRNAs.** It has been shown previously that lncRNAs comprise few exons, are shorter, and are expressed at low levels in a highly tissue- or cell-specific manner (3–5). The 2,803 lncRNAs reported here are consistent with these previous studies. On average, our lncRNA 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 (see Fig. S3 in the supplemental material). To gain more insight, we combined our novel lncRNAs with all the known lncRNAs and reanalyzed the genomic features by considering those with an expression level greater than 0.1 FPKM in at least 1 out of 8 tissues and in a cell line and those that are far from protein-coding genes (e.g., 10 kb away from either a TSS or a TES of a protein-coding gene). This resulted in 3,759 lncRNAs. 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 (phylogenetic analysis with space or time conservation). These results further confirmed the genomic features of lncRNA, such as expression and conservation levels lower than those of protein-coding genes.

In mammals, lncRNAs are expressed in a tissue-specific manner (3–5). To assess for any tissue specificity of our data set of lncRNAs, we compared each lncRNA expression level in a given tissue to its expression in the remaining 8 tissues (Fig. 2A; see Table S2 in the supplemental material). 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 (see Fig. S4 in the supplemental material). 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 of >0.4) (4). To experimentally validate a pair of these selected tissue-specific lncRNAs, we measured the expression levels by qRT-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 GAPDH housekeeping gene (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 lncRNAs (26, 27). We observed an enrichment of CAGE tags

Downloaded from http://mcb.asm.org/ on February 22, 2016 by Biblioteca de la Universitat Pompeu Fabra

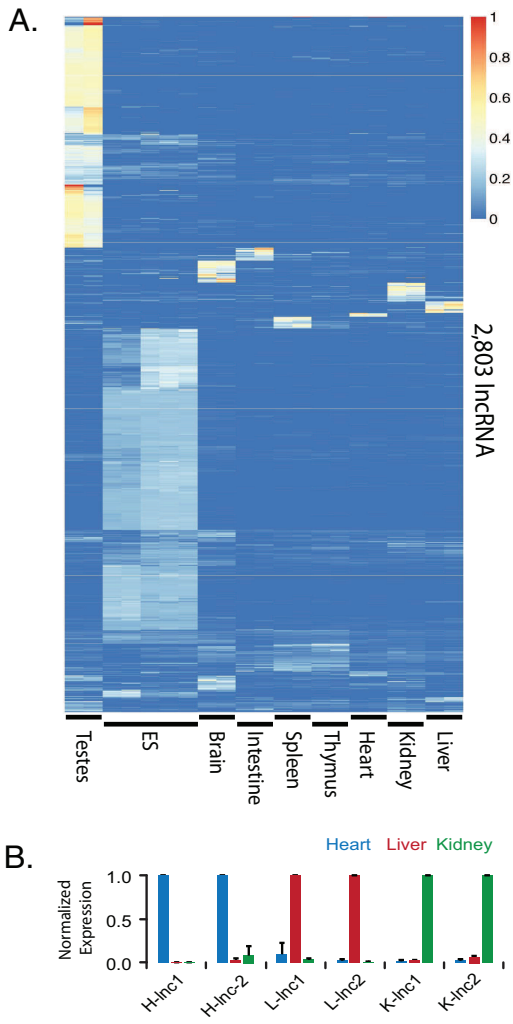


**FIG 2** Tissue- and cell-specific expression of lncRNAs. (A) Heat map representing normalized FPKM expression values of the 2,803 lncRNAs (rows) across eight tissues and a primary cell line (columns). The rows and columns were ordered based on *k* means clustering. The color intensity represents the fractional density across the row of $\log_{10}$-normalized FPKM expression values as estimated by Scripture v4. Each tissue has 2 columns, representing its 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 qRT duplicate normalized (against the GAPDH housekeeping gene) expression levels 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) (see Table S9 in the supplemental material). The error bars indicate standard deviations.

around our lncRNA promoters compared to random lncRNA promoters (see Fig. S5A in the supplemental material). 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 (see Fig. S5B in the supplemental material). 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 (see Fig. S5C in the supplemental material). 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, 39). 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 modes of regulation. A map of chromatin marks was constructed from ~1.4 billion mapped reads obtained from 72 pooled ENCODE genome-wide ChIP-seq data sets in eight tissues (brain, heart, liver, small intestine, kidney, spleen, testis, and thymus) and the one primary ES cell line. The ChIP-seq data sets used included regulatory histone modifications, such as H3K4me1, H3K4me3, H3K36me3, H3K27me3, and H3K27ac, as well as CTCF marks and RNA polymerase II marks.

We applied the ChromHMM program (39) to create a chromatin state model at 200-bp resolution, which resulted in six major chromatin state maps (Fig. 3A), i.e., promoter (active and poised), enhancer (strong and poised/weak), transcribed (transcription transition, elongation, and weak transcription), insulator, repressed, and heterochromatin states (see Table S3 in the supplemental material). 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 nonoverlapping 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 promoters. We repeated the above-described mapping using the poised promoter map and mapped an additional 709 protein-coding and 92 noncoding gene promoters. Altogether, we successfully mapped 85% of the known protein-coding and 78% of the noncoding gene promoters. These results indicate that using combinatorial promoter chromatin states to retrieve promoters results in ~6% higher recall than using only H3K4me3 as an active promoter chromatin mark (40).

**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, elncRNA promoters or TSSs are depleted of H3K4me3 and enriched with H3K4me1, and 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 TSSs and TESs of protein-coding genes). We found a few thousand lncRNAs that passed these 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 testis, and 932 in thymus). Overall, less than 10% (852) of these intergenic lncRNAs significantly overlapped an active promoter

Downloaded from http://mcb.asm.org/ on February 22, 2016 by Biblioteca de la Universitat Pompeu Fabra

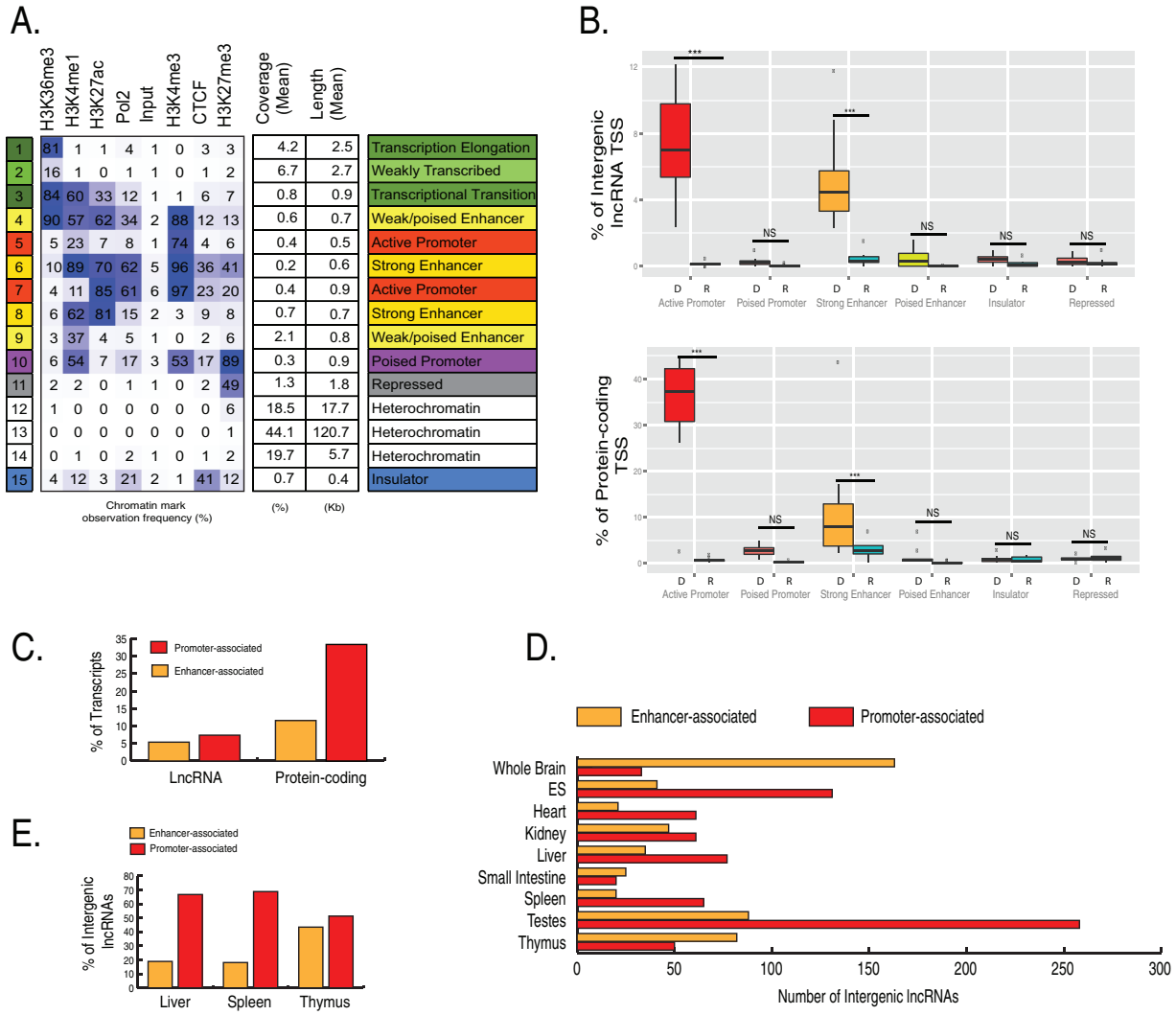


**FIG 3** Discovery of chromatin state maps and their association with lincRNAs. (A) Emission parameters learned *de novo* with ChromHMM on the basis of combinations recurring in chromatin. Each point in the table denotes the frequency with which a given mark is found at genomic positions corresponding to a specific chromatin state. The observation frequencies of various chromatin marks, including H3K36me3, H3K4me1, H3K27ac, Pol II, H3K4me3, CTCF, and H3K27me3, as well as respective inputs showing 6 major chromatin states, including active promoter (red), poised promoter (purple), enhancer (yellow), Polycomb (gray), insulator (blue), and heterochromatin (white), are presented. (B) Percentages of protein-coding TSSs (top) and intergenic lncRNAs (bottom) significantly enriched with both active promoter and strong enhancer (***, $P < 0.001$; NS, not significant; Fisher exact test). D, observed data; R, randomized TSSs. (C) Percentages of lncRNAs and protein-coding genes that are associated with promoter and enhancer chromatin states. (D) Numbers of plncRNAs and elncRNAs across 8 tissues and an ES cell line. (E) Percentages of lncRNAs (overlapping both CAGE peaks and DNase I hypersensitive sites) associated with promoter and enhancer chromatin states.

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 (Fig. 3C and D; see Table S4 in the supplemental material). 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 from 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 reanalysis resulted in more than 40% of the lncRNAs associated with the enhancer chromatin state in thymus (~50% with the promoter chromatin state) and around 20% in liver and spleen. (Fig. 3D; see Table S5 in the supplemental material). Finally, we did not notice any enrichment in the number of elncRNAs over plncRNAs in most of the tissues 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 (41), and

Downloaded from http://mcb.asm.org/ on February 22, 2016 by Biblioteca de la Universitat Pompeu Fabra

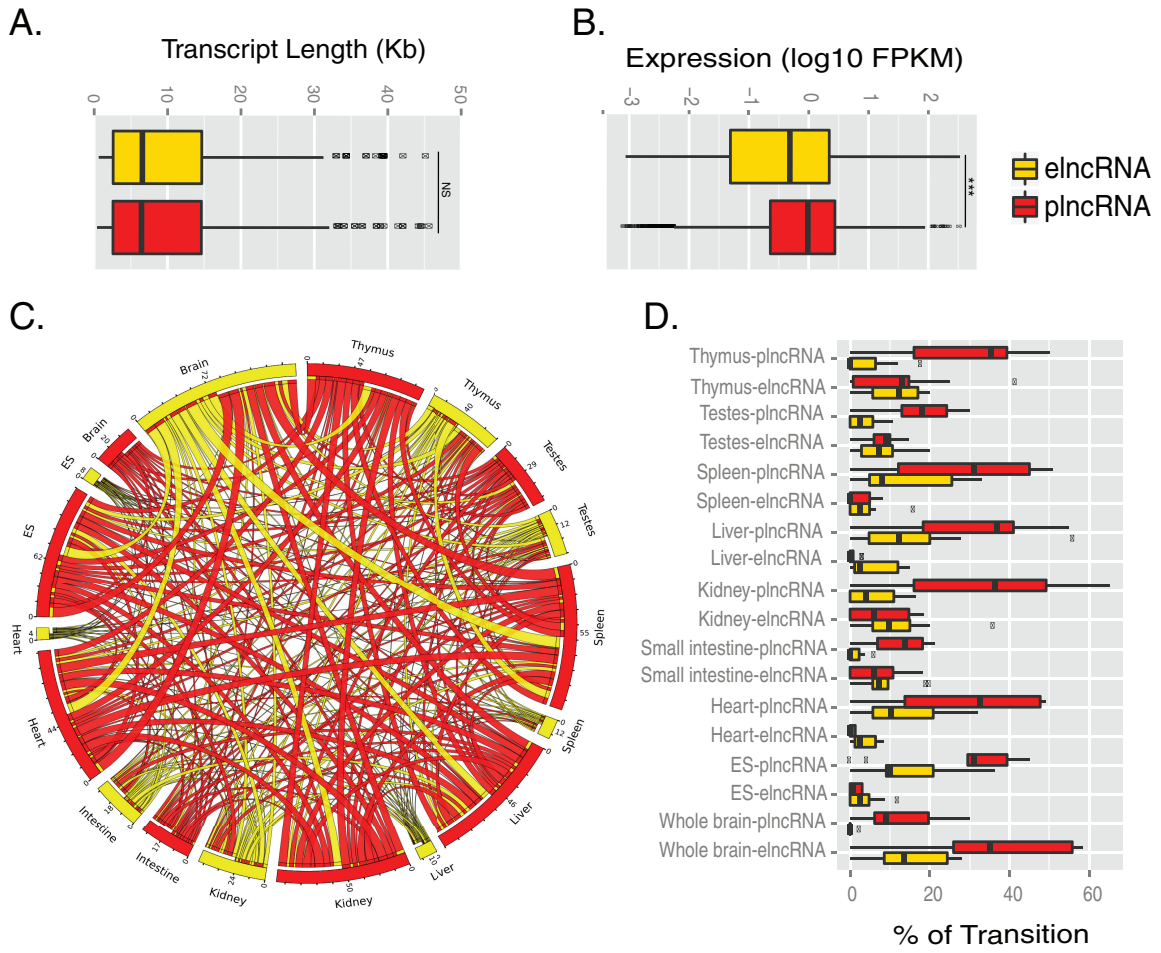


**FIG 4** Transcript length, expression, and transition of chromatin-associated lncRNAs in mouse. (A) Transcript lengths of elncRNAs (median = 6,565 nt) and plncRNAs (median = 6,450 nt) across eight tissues and a cell line, showing no difference in length (Mann-Whitney test; NS, not significant; $P = 0.9848$). (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; ***, $P = 1.221e{-}10$). (C) Circos plot showing the transition of plncRNA to elncRNA, or elncRNA to plncRNA, across eight tissues and an ES cell line. The outer bars indicate the total numbers of chromatin-associated lncRNAs 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), testis (29 plncRNAs and 12 elncRNAs), and thymus (47 plncRNAs and 40 elncRNAs). The links inside the bars indicate the numbers of lncRNAs that switch their chromatin states from one tissue to another (red, plncRNAs; gold, elncRNAs). The lncRNA transition table used to generate the circos plot is shown in Table S6 in the supplemental material. (D) Percentages of chromatin-associated transitions across all the mouse tissues, showing the high percentage of plncRNA-to-plncRNA transitions compared to elncRNA-to-elncRNA transitions.

elncRNAs, such as lincRNA-*Cox2*, lincRNA-*Spasm*, and lincRNA-*Haunt* (42) (see Fig. S6 in the supplemental material). 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) (see Fig. S7 in the supplemental material). 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 lncRNAs (24). 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 in both mouse ES cells and heart cells (see Fig. S8 in the supplemental material), using Klf4 as a negative control and Zic1 as a positive control. Altogether, our study provides a high-confidence list of chromatin-associated lncRNAs across a wide range of tissues in mouse.

**Properties of chromatin-associated lncRNAs.** To investigate whether the two types of chromatin-associated lncRNAs have different properties, we calculated their sequence lengths 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 an ~6-kb length 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 lncRNAs were enriched in any biological processes by using a nearest-gene approach and whole-genome background with the GREAT software (30). Indeed, they showed enrichment of various biological processes (see Fig. S9 in the supplemental

Downloaded from http://mcb.asm.org/ on February 22, 2016 by Biblioteca de la Universitat Pompeu Fabra

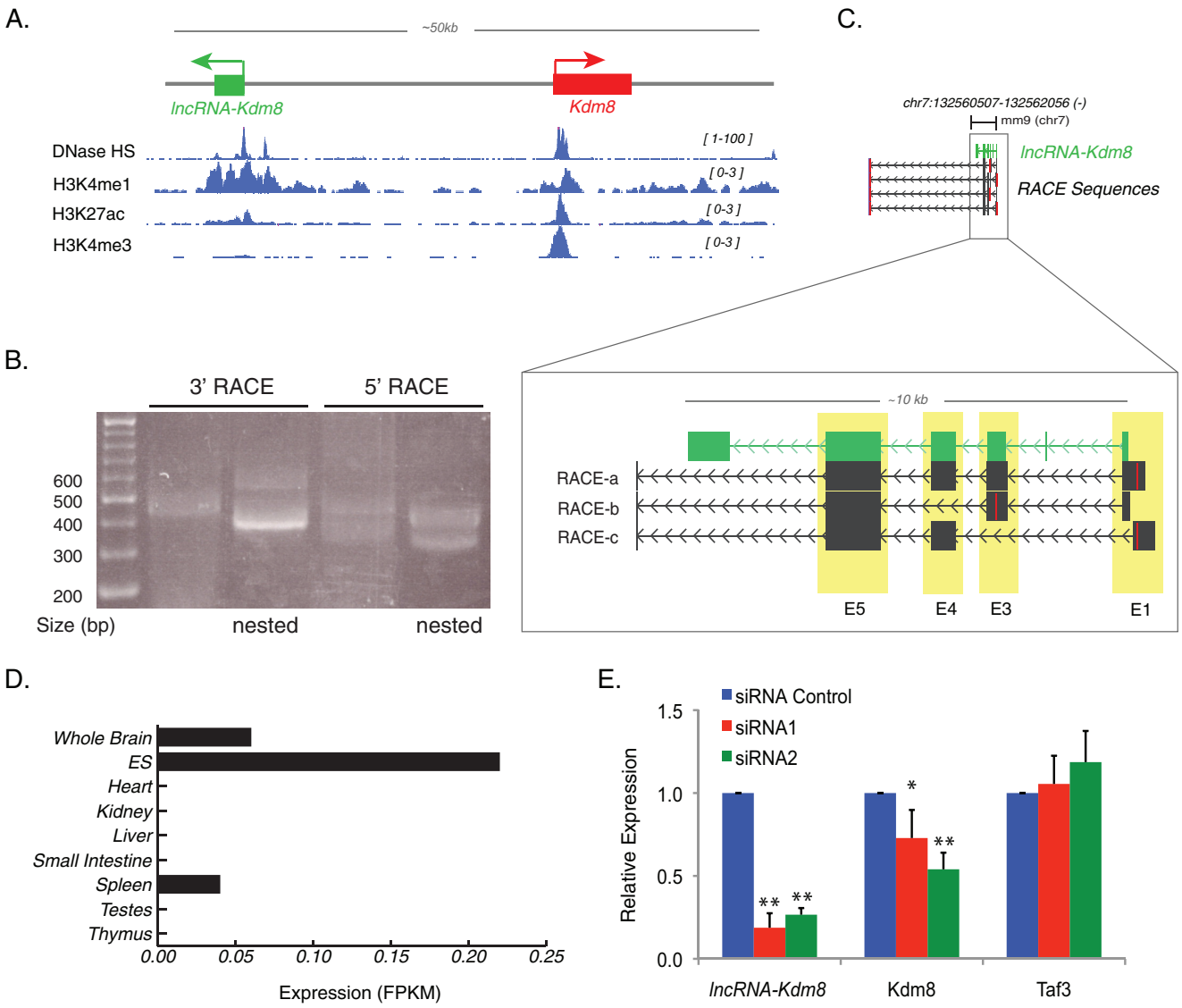


**FIG 5** An enhancer-associated lncRNA, lncRNA-*Kdm8*, regulates the expression of a neighboring protein-coding gene, *Kdm8*. (A) The lncRNA-*Kdm8* locus promoter overlaps an enhancer chromatin state and occurs within 20 kb of the TSS of a protein-coding gene, *Kdm8* (e.g., it is an enhancer-associated lncRNA). The gene tracks represent DNase I hypersensitive sites (HS) and ChIP-seq data for H3K4me1, H3K27ac, and H3K4me3 from ENCODE. The genomic scale is indicated at the top and the scale of both DNase I HS and ChIP-seq data on the upper right. (B and C) The 5′ and 3′ ends and the exon-intron boundaries of the enhancer-associated lncRNA, lncRNA-*Kdm8*, were determined by RACE (see the supplemental material). The black arrows depict TSSs and the directions of transcription for the respective genes. *Kdm8* mRNA and lncRNA-*Kdm8* are shown in green and red, respectively. The genomic DNA sequences corresponding to the 5′ and 3′ ends of the cloned lncRNA are shown in black below the lncRNA-*Kdm8* gene track, defining accurate 5′-end and exon-intron boundaries for exon 1 (E1), exon 3, exon 4, and exon 5 of lncRNA-*Kdm8*. (D) Expression levels of lncRNA-*Kdm8* in mouse ES cells and other tissues, as measured by directional RNA-seq and expressed as FPKM. (E) qRT-PCR expression (triplicates, normalized against the RPO housekeeping gene) after siRNA-based knockdown of lncRNA-*Kdm8* (chr7: 132560406 to 132561472 [−]) resulted in a significant decrease of the neighboring gene, *Kdm8* (*t* test; *, $P \leq 0.05$; **, $P \leq 0.01$), which was not observed for the negative control of the distant coding gene, *Taf3* (chr2: 9836179 to 9970236 [+]). The primers used for siRNA oligonucleotides of lncRNA-*Kdm8* are given in Table S9 in the supplemental material. The error bars indicate standard deviations.

material). Interestingly, we also observed the changes in the status of chromatin-associated lncRNAs based on their respective tissue or cell line. In total, ~17% of chromatin-associated lncRNAs (144 out of 852) tend to switch from one chromatin state to another in multiple tissues (see Table S6 in the supplemental material). plncRNAs are more likely to switch to plncRNAs, and also, the percentage of this type of transition is higher than that of the plncRNA-to-elncRNA or the elncRNA-to-plncRNA transition (Fig. 4C and D; see Table S6 in the supplemental material).

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: (i) associated with a specific chromatin state only in ES cells, (ii) expressed only in ES cells, (iii) associated with DNase I peaks only in ES cell, (iv) associated with pluripotent transcription factors in ES cells, and (v) close to a protein-coding gene associated with pluripotency in ES cells. In total, 12 lncRNAs passed the above-mentioned filters.

Downloaded from http://mcb.asm.org/ on February 22, 2016 by Biblioteca de la Universitat Pompeu Fabra

For validation, we focused on an ES cell-specific predicted regulatory enhancer-associated lncRNA (chromosome 7 [chr7]: 132560406 to 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 D) (30). We named this 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 C). 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 elncRNA knockdown, expression of the *Kdm8* gene significantly decreased compared to that of *Taf3*, which further supported the *cis* mode of enhancer-associated lncRNA gene regulation (Fig. 5E) (43, 44). Together, our results show that chromatin-associated lncRNAs annotated by their 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 high-confidence novel lncRNA transcripts. The transcriptome data used in this study to discover lncRNAs go beyond previous lncRNA studies in terms of depth (32). The tissue-specific nature of these lncRNAs 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 qRT-PCR, thus confirming the quality of the transcripts identified in the study, as well as their expression.

By using ChromHMM, we further characterized combinatorial chromatin state maps in mouse, using more than 70 ChIP-seq data sets 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 data sets, like H3K4me3 (promoter), H3K4me1 with P300 (enhancer), and CTCF (insulator) (40). We built upon that work by further including additional histone marks, allowing us to produce more detailed chromatin state maps. For example, the Fendrr lncRNA, which was previously annotated as enhancer associated, has enhancer histone (p300/H3K4me1) marks (42) 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 lncRNAs.

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

Many of the bidirectional lncRNAs and enhancer-associated RNAs have been shown to be nonpolyadenylated (41, 45). However, recent findings (2, 17), along with our study, suggest the existence of polyadenylated bidirectional transcripts and chromatin-associated RNAs. Still, because of the poly(A)-based RNA sequencing, we could be missing a large fraction of nonpolyadenylated lncRNAs.

In the future, even more comprehensive catalogs of chromatin-associated lncRNAs should be possible to obtain by association of chromatin states and lncRNA promoters across all tissues and cell lines in mammals. In addition, using techniques like CRISPR against regulatory lncRNAs would reveal more valuable information. Altogether, our study provides a novel set of classified lncRNAs, which represents a valuable resource for future genomic experimental studies in mouse.

## ACKNOWLEDGMENTS

We sincerely thank the 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 editing). We also thank the three anonymous reviewers for their critical insights.

We declare that we have no competing interests.

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, RACE, and siRNA experiments. L.W.S., M.B., L.D.C., and M.A.M.-R. contributed ideas and wrote the manuscript.

The project was supported by a grant from la Caixa to G.K.B., by an AIO2014 fellowship from the Spanish Association against Cancer (AECC) to P.V., and by the Spanish MINECO to M.A.M.-R. (BFU2010-19310 and BFU2013-47736-P). We also acknowledge the support of the Spanish Ministry of Economy and Competitiveness, Centro de Excelencia Severo Ochoa 2013-2017 (SEV-2012-0208).

## FUNDING INFORMATION

Spanish MINECO provided funding to Marc A. Marti-Renom under grant numbers BFU2010-19310 and BFU2013-47736-P. Spanish MINECO provided funding to Luciano di Croce under grant number SAF2013-48926-P. Centro de Excelencia Severo Ochoa 2013-2017 provided funding to Luciano Di Croce and Marc A. Marti-Renom under grant number SEV-2012-0208.

## REFERENCES

1. **Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee B-K, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Dunham I, Ernst J, Furey TS, Gerstein M, Giardine B, Greven M, Hardison RC, Harris RS, Herrero J, Hoffman MM, Iyer S, Kellis M, Khatun J, Kheradpour P, Kundaje A, Lassmann T, Li Q, Lin X, Marinov GK, Merkel A, Mortazavi A, et al.** 2012. An integrated encyclopedia of DNA elements in the human genome. Nature **489**:57–74. http://dx.doi.org/10.1038/nature11247.
2. **Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-**

Downloaded from http://mcb.asm.org/ on February 22, 2016 by Biblioteca de la Universitat Pompeu Fabra

Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See L-H, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR. 2012. Landscape of transcription in human cells. Nature **489:**101–108. http://dx.doi.org/10.1038/nature11233.

3. **Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigo R.** 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res **22:**1775–1789. http://dx.doi.org/10.1101/gr.132159.111.
4. **Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL.** 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev **25:**1915–1927. http://dx.doi.org/10.1101/gad.17446611.
5. **Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES.** 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature **458:**223–227. http://dx.doi.org/10.1038/nature07672.
6. **Nam J-W, Bartel DP.** 2012. Long noncoding RNAs in C. elegans. Genome Res **22:**2529–2540. http://dx.doi.org/10.1101/gr.140475.112.
7. **Young RS, Marques AC, Tibbit C, Haerty W, Bassett AR, Liu JL, Ponting CP.** 2012. Identification and properties of 1,119 candidate LincRNA loci in the Drosophila melanogaster genome. Genome Biol Evol **4:**427–442. http://dx.doi.org/10.1093/gbe/evs020.
8. **Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A, Schier AF.** 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res **22:**577–591. http://dx.doi.org/10.1101/gr.133009.111.
9. **Panning B, Dausman J, Jaenisch R.** 1997. X chromosome inactivation is mediated by Xist RNA stabilization. Cell **90:**907–916. http://dx.doi.org/10.1016/S0092-8674(00)80355-4.
10. **Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai M-C, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, van de Vijver MJ, Sukumar S, Chang HY.** 2010. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature **464:**1071–1076. http://dx.doi.org/10.1038/nature08975.
11. **Wang P, Xue Y, Han Y, Lin L, Wu C, Xu S, Jiang Z, Xu J, Liu Q, Cao X.** 2014. The STAT3-binding long noncoding RNA lnc-DC controls human dendritic cell differentiation. Science **344:**310–313. http://dx.doi.org/10.1126/science.1251456.
12. **Klattenhoff CA, Scheuermann JC, Surface LE, Bradley RK, Fields PA, Steinhauser ML, Ding H, Butty VL, Torrey L, Haas S, Abo R, Tabebordbar M, Lee RT, Burge CB, Boyer LA.** 2013. Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. Cell **152:**570–583. http://dx.doi.org/10.1016/j.cell.2013.01.003.
13. **Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP.** 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell **147:**1537–1550. http://dx.doi.org/10.1016/j.cell.2011.11.055.
14. **Grote P, Wittler L, Hendrix D, Koch F, Währisch S, Beisaw A, Macura K, Bläss G, Kellis M, Werber M, Herrmann BG.** 2013. The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Dev Cell **24:**206–214. http://dx.doi.org/10.1016/j.devcel.2012.12.012.
15. **Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA, Blencowe BJ, Prasanth SG, Prasanth KV.** 2010. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol Cell **39:**925–938. http://dx.doi.org/10.1016/j.molcel.2010.08.011.
16. **Ng S-Y, Bogu GK, Soh BS, Stanton LW.** 2013. The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. Mol Cell **51:**349–359. http://dx.doi.org/10.1016/j.molcel.2013.07.017.
17. **Marques AC, Hughes J, Graham B, Kowalczyk MS, Higgs DR, Ponting CP.** 2013. Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. Genome Biol **14:**R131. http://dx.doi.org/10.1186/gb-2013-14-11-r131.
18. **Trapnell C, Pachter L, Salzberg SL.** 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25:**1105–1111. http://dx.doi.org/10.1093/bioinformatics/btp120.
19. **Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L.** 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc **7:**562–578. http://dx.doi.org/10.1038/nprot.2012.016.
20. **Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A.** 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol **28:**503–510. http://dx.doi.org/10.1038/nbt.1633.
21. **Garcia-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Gotz S, Tarazona S, Dopazo J, Meyer TF, Conesa A.** 2012. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics **28:**2678–2679. http://dx.doi.org/10.1093/bioinformatics/bts503.
22. **Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W.** 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res **41:**e74. http://dx.doi.org/10.1093/nar/gkt006.
23. **Lin MF, Jungreis I, Kellis M.** 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics **27:**i275–i282. http://dx.doi.org/10.1093/bioinformatics/btr209.
24. **Alvarez-Dominguez JR, Hu W, Yuan B, Shi J, Park SS, Gromatzky AA, Oudenaarden AV, Lodish HF.** 2014. Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation. Blood **123:**570–581. http://dx.doi.org/10.1182/blood-2013-10-530683.
25. **Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble W.** 2007. Quantifying similarity between motifs. Genome Biol **8:**R24. http://dx.doi.org/10.1186/gb-2007-8-2-r24.
26. **FANTOM Consortium, RIKEN PMI, CLST (DGT).** 2014. A promoter-level mammalian expression atlas. Nature **507:**462–470. http://dx.doi.org/10.1038/nature13182.
27. **Mouse ENCODE Consortium, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, Canfield T, Giste E, Johnson A, Zhang M, Balasundaram G, Byron R, Roach V, Sabo PJ, Sandstrom R, Stehling AS, Thurman RE, Weissman SM, Cayting P, Hariharan M, Lian J, Cheng Y, Landt SG, Ma Z, Wold BJ, Dekker J, Crawford GE, Keller CA, Wu W, Morrissey C, Kumar SA, Mishra T, Jain D, Byrska-Bishop M, Blankenberg D, Lajoie BR, Jain G, Sanyal A, Chen K-B, Denas O, Taylor J, Blobel GA, Weiss MJ, Pimkin M, Deng W, Marinov GK, Williams BA, Fisher-Aylor KI, DeSalvo G, Kiralusha A, Trout D, Amrhein H, Mortazavi A, Edsall L, McCleary D, Kuan S, Shen Y, Yue F, Ye Z, Davis CA, Zaleski C, Jha S, Xue C, Dobin A, Lin W, Fastuca M, Wang H, Guigó R, Djebali S, Lagarde J, Ryba T, Sasaki T, Malladi VS, Cline MS, Kirkup VM, Learned K, Rosenbloom KR, Kent WJ, Feingold EA, Good PJ, Pazin M, Lowdon RF, Adams LB.** 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol **13:**418. http://dx.doi.org/10.1186/gb-2012-13-8-418.
28. **Shin H, Liu T, Manrai AK, Liu XS.** 2009. CEAS: cis-regulatory element annotation system. Bioinformatics **25:**2605–2606. http://dx.doi.org/10.1093/bioinformatics/btp479.
29. **Quinlan AR, Hall IM.** 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26:**841–842. http://dx.doi.org/10.1093/bioinformatics/btq033.
30. **McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G.** 2010. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol **28:**495–501. http://dx.doi.org/10.1038/nbt.1630.
31. **Morey L, Pascual G, Cozzuto L, Roma G, Wutz A, Benitah SA, Di Croce L.** 2012. Nonoverlapping functions of the Polycomb group Cbx family of proteins in embryonic stem cells. Cell Stem Cell **10:**47–62. http://dx.doi.org/10.1016/j.stem.2011.12.006.
32. **Pervouchine DD, Djebali S, Breschi A, Davis CA, Barja PP, Dobin A, Tanzer A, Lagarde J, Zaleski C, See L-H, Fastuca M, Drenkow J, Wang H, Bussotti G, Pei B, Balasubramanian S, Monlong J, Harmanci A, Gerstein M, Beer MA, Notredame C, Guigó R, Gingeras TR.** 2015. Enhanced transcriptome maps from multiple mouse tissues reveal evolu-

Downloaded from http://mcb.asm.org/ on February 22, 2016 by Biblioteca de la Universitat Pompeu Fabra

tionary constraint in gene expression. Nat Commun **6:**5903. http://dx.doi.org/10.1038/ncomms6903.

33. **Jia H, Osak M, Bogu GK, Stanton LW, Johnson R, Lipovich L.** 2010. Genome-wide computational identification and manual annotation of human long noncoding RNA genes. RNA **16:**1478–1487. http://dx.doi.org/10.1261/rna.1951310.
34. **Luo H, Sun S, Li P, Bu D, Cao H, Zhao Y.** 2013. Comprehensive characterization of 10,571 mouse large intergenic noncoding RNAs from whole transcriptome sequencing. PLoS One **8:**e70835. http://dx.doi.org/10.1371/journal.pone.0070835.
35. **Morán I, Akerman ˙ I, van de Bunt M, Xie R, Benazra M, Nammo T, Arnes L, Nakić N, García-Hurtado J, Rodríguez-Seguí S, Pasquali L, Sauty-Colace C, Beucher A, Scharfmann R, van Arensbergen J, Johnson PR, Berry A, Lee C, Harkins T, Gmyr V, Pattou F, Kerr-Conte J, Piemonti L, Berney T, Hanley N, Gloyn AL, Sussel L, Langman L, Brayman KL, Sander M, McCarthy MI, Ravassard P, Ferrer J.** 2012. Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. Cell Metabolism **16:**435–448. http://dx.doi.org/10.1016/j.cmet.2012.08.010.
36. **Lv J, Cui W, Liu H, He H, Xiu Y, Guo J, Liu H, Liu Q, Zeng T, Chen Y, Zhang Y, Wu Q.** 2013. Identification and characterization of long non-coding RNAs related to mouse embryonic brain development from available transcriptomic data. PLoS One **8:**e71152. http://dx.doi.org/10.1371/journal.pone.0071152.
37. **Ramos AD, Diaz A, Nellore A, Delgado RN, Park K-Y, Gonzales-Roybal G, Oldham MC, Song JS, Lim DA.** 2013. Integration of genome-wide approaches identifies lncRNAs of adult neural stem cells and their progeny in vivo. Cell Stem Cell **12:**616–628. http://dx.doi.org/10.1016/j.stem.2013.03.003.
38. **Belgard TG, Marques AC, Oliver PL, Abaan HO, Sirey TM, Hoerder-Suabedissen A, García-Moreno F, Molnár Z, Margulies EH, Ponting CP.** 2011. NeuroResource. Neuron **71:**605–616. http://dx.doi.org/10.1016/j.neuron.2011.06.039.
39. **Ernst J, Kellis M.** 2012. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods **9:**215–216. http://dx.doi.org/10.1038/nmeth.1906.
40. **Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B.** 2012. A map of the cis-regulatory sequences in the mouse genome. Nature **488:**116–120. http://dx.doi.org/10.1038/nature11243.
41. **Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME.** 2010. Widespread transcription at neuronal activity-regulated enhancers. Nature **465:**182–187. http://dx.doi.org/10.1038/nature09033.
42. **Sauvageau M, Goff LA, Lodato S, Bonev B, Groff AF, Gerhardinger C, Sanchez-Gomez DB, Hacisuleyman E, Li E, Spence M, Liapis SC, Mallard W, Morse M, Swerdel MR, D'Ecclessis MF, Moore JC, Lai V, Gong G, Yancopoulos GD, Frendewey D, Kellis M, Hart RP, Valenzuela DM, Arlotta P, Rinn JL.** 2013. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. ELife **2:**e01749. http://dx.doi.org/10.7554/eLife.01749.
43. **Lai F, Orom UA, Cesaroni M, Beringer M, Taatjes DJ, Blobel GA, Shiekhattar R.** 2013. Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature **494:**497–501. http://dx.doi.org/10.1038/nature11884.
44. **Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G.** 2010. Long noncoding RNAs with Enhancer-like function. Cell **143:**46–58. http://dx.doi.org/10.1016/j.cell.2010.09.001.
45. **Wu X, Sharp PA.** 2013. Perspective. Cell **155:**990–996. http://dx.doi.org/10.1016/j.cell.2013.10.048.

Downloaded from http://mcb.asm.org/ on February 22, 2016 by Biblioteca de la Universitat Pompeu Fabra