Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse

Gireesh K. Bogu[1,2,3,4,#], Pedro Vizán[2,4], Lawrence W. Stanton[5,6], Miguel Beato[2,4], Luciano Di Croce[2,4,7], and Marc A. Marti-Renom[1,2,4,7,#]

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 lncRNAs were identified using 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 highly-confident 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) 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 lncRNAs and a comprehensive catalog of chromatin-associated lncRNAs 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 lncRNAs. Comprehensive lncRNA 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 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 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 associate with these two chromatin states since the lncRNA associations were not tested in multiple tissues. In addition, the lncRNA 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 lncRNA identified in testes).

Properties of the 2,803 lncRNAs.

It has been previously shown that lncRNAs 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 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 (**Fig. S3**). To gain more insight, we combined our novel lncRNAs with all the known lncRNAs 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 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. These results further confirmed the genomic features of lncRNA, such as lower expression and conservation levels as compared to protein-coding genes.

153 In mammals, lncRNAs are expressed in a tissue-specific manner (3-5). To
154 assess for any tissue specificity of our dataset of lncRNA, we compared each
155 lncRNA expression in a given tissue to its expression in the remaining 8 tissues
156 (**Fig. 2A; Table S2**). We observed that 62% of our novel intergenic lncRNAs are
157 tissue-specific, which is comparable to known intergenic lncRNAs (68% tissue-
158 specific). Moreover, protein-coding genes resulted in 36.4% tissue specificity
159 across the eight tissues and the ES cell line (**Fig. S4**). Overall, the results clearly
160 show that lncRNAs are highly tissue specific in nature. Next, we selected the
161 tissue-specific lncRNAs from our list as previously defined (e.g., with an entropy
162 > 0.4) (4). To experimentally validate a pair of these selected tissue-specific
163 lncRNAs, we measured the expression levels by RT-PCR of the heart (H-lnc1
164 and H-lnc2), liver (L-lnc1 and L-lnc2), and kidney (K-lnc1 and K-lnc2) lncRNAs
165 with respect to the housekeeping gene GAPDH (**Fig. 2B**), which confirmed their
166 tissue-specificity.
167
168 To assess whether our novel lncRNAs have active TSS and regulatory marks,
169 we overlapped CAGE tags and DNase I tags from the FANTOM and ENCODE
170 projects with the promoters of our lncRNA (30, 31). We observed an enrichment
171 of CAGE tags around our lncRNA promoters, as compared to random lncRNA
172 promoters (**Fig. S5A**). We also observed an enrichment of tissue-specific DNase
173 I tags in lncRNA promoters from the brain, kidney, liver, spleen, and thymus
174 tissues as well as for the ES cell line (**Fig. S5B**). Finally, we performed *de novo*
175 motif analysis using lncRNA promoters to explore whether any transcription

176 factors could be regulating these lncRNAs. Indeed, we found several significant
177 transcription factor binding motifs enriched near lncRNA promoters (**Fig. S5C**).
178 These results show that the 2,803 lncRNA promoters are enriched with various
179 regulatory marks in the mouse genome and could potentially have regulatory
180 roles.
181
182 **Genome-wide identification of chromatin state maps in mouse.**
183 Chromatin marks mapping across different cell lines in mammals have been
184 previously used to detect and annotate novel regulatory regions in the genome,
185 including for putative lncRNAs (5, 17, 32). We hypothesized that integrating
186 chromatin state maps with the promoters of the transcripts identified here using
187 RNA-seq expression could guide us in annotating the potential transcripts and in
188 predicting their mode of regulation. A map of chromatin marks was constructed
189 from ~1.4 billion mapped reads obtained from 72 pooled ENCODE genome-wide
190 ChIP-seq datasets in eight tissues (brain, heart, liver, small intestine, kidney,
191 spleen, testes, and thymus) and the one primary ES cell line. The ChIP-seq
192 datasets used included regulatory histone modifications, such as H3 lysine 4
193 mono-methylation (H3K4me1), H3 lysine 4 tri-methylation (H3K4me3), H3 lysine
194 36 tri-methylation (H3K36me3), H3 lysine 27 tri-methylation (H3K27me3), and H3
195 lysine 27 mono-acetylation (H3K27ac), as well as CCCTC-binding factor (CTCF)
196 marks and RNA polymerase II marks.
197

198 We applied the ChromHMM program (32) to create a chromatin state model at
199 200 bp resolution, which resulted in six major chromatin state maps (**Fig. 3A**),
200 including promoter (active and poised), enhancer (strong and poised/weak),
201 transcribed (transcription transition, elongation, and weak transcription),
202 insulator, repressed, and heterochromatin states (**Table S3**). In total, we mapped
203 261,175 promoter states (covering ~1% of the mouse genome), 863,677
204 enhancer states (~3%), 1,133,166 transcribed states (~12%), 150,752 repressed
205 states (~1%), 322,521 insulator states (~1%) and 995,562 heterochromatin
206 states (~82%). To validate the accuracy of the predicted chromatin states or
207 maps, we mapped (at ±10 kb) our 206,045 unique non-overlapping active
208 promoter maps to known promoters of 23,431 RefSeq protein-coding genes and
209 3,190 RefSeq noncoding genes from TSSs. Our analysis recalled 82% (19,280)
210 of the protein-coding promoters and 75% (2,401) of the noncoding ones. We
211 repeated the above mapping using the poised promoter map and mapped an
212 additional 709 protein-coding and 92 noncoding gene promoters. All together, we
213 successfully mapped 85% of the known protein-coding and 78% noncoding gene
214 promoters. These results indicate that using combinatorial promoter chromatin
215 states to retrieve promoters results in ~6% higher recall than when using only
216 H3K4me3 as an active promoter chromatin mark (33).
217
218 **Classification of lncRNAs using chromatin state maps.**
219 Previously chromatin state maps at promoters were used to define two distinct
220 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 lncRNAs (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 lncRNAs across wide range of tissues in mouse.

Properties of the chromatin-associated lncRNAs.

To investigate whether the two types of chromatin-associated lncRNAs 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 lncRNAs 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 lncRNAs based on their respective tissue or cell line. In total, ~17% chromatin-associated lncRNAs (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 elncRNA 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

312 (**Fig. 5E**) (38, 39). Together, our results show that chromatin-associated lncRNAs

313 annotated by its chromatin marks could have regulatory roles.

314

315 **DISCUSSION**

316

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

331

332 By using ChromHMM, we further characterized combinatorial chromatin state
333 maps in mouse, using more than 70 ChIP-seq datasets across the same tissues
334 used for lncRNA discovery. In previous studies, promoter, enhancer, and
335 insulator maps were identified using a specific set of ChIP-seq datasets, like
336 H3K4me3 (promoter), H3K4me1 with P300 (enhancer), and CTCF (insulator)
337 (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 lncRNA, 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 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, shorter in length compared to 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 bi-directional lncRNAs 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

Molecular and Cellular Biology

sequencing, we could be missing a large fraction of non-polyadenylated lncRNAs.

In the future, even more comprehensive catalogs of chromatin-associated lncRNAs should be possible to obtain by associating 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. All together, our study provides a novel set of classified lncRNAs, 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 lncRNAs.

Identification and genomic annotation of lncRNAs.

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 lncRNAs from transcriptomics data. The pipeline selects transcripts as lncRNAs 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

394 strand), overlap with noncoding regions (no overlap with known snoRNAs,
395 tRNAs, miRNAs, lncRNAs, or pseudogenes), and noncoding potential (<0.44
396 CPAT and <100 PhyloCSF score). PhyloCSF (24) was used to calculate the
397 coding potential of transcripts. First, we stitched mouse lncRNA exonic
398 sequences into 18 mammals, using mm9-multiz30way alignments from UCSC.
399 Second, we ran PhyloCSF against stitched sequences using default parameter
400 unless specified (**Table S8**). We then removed the transcripts with open reading
401 frames with a PhyloCSF score greater than 100, as previously suggested (36).
402 The final lncRNA PhyloCSF score is the average decibans score of all its exons
403 based on their strand direction and all possible frames. The transcripts that
404 passed PhyloCSF and CPAT coding potential filters were further selected as
405 potential lncRNAs.
406 LncRNAs that did not overlap with any known protein-coding gene (within a 10 kb
407 window from both TSS and TES) were classified as intergenic lncRNAs or
408 lncRNAs. LncRNAs that overlapped a transcript but on opposite strands were
409 classified as antisense lncRNAs. LncRNAs that were close to a coding gene
410 (within 10 kb from both TSS and TES) were annotated as either convergent
411 (same strand as the nearest coding) or divergent (opposite strand as the nearest
412 coding) lncRNAs.
413 *Tissue specificity calculations.*
414 To calculate tissue specificity of lncRNAs, we normalized the raw FPKM
415 expression values as suggested in previous studies (4, 5). First, we added
416 pseudo-count 1 to every raw FPKM value and second applied log2 normalization

417 to each value, to obtain a non-negative expression vector. Finally, we normalized
418 the expression vector by dividing it by the total expression counts. The resulting
419 matrix of lncRNA normalized expression levels in each of the replica experiments
420 per tissue or cell line was clustered by k-means.

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

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

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

439 *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 with in 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.

485 We selected 200-bp-long promoters of expressed lncRNAs (>1 FPKM) in whole
486 brain and made sure that they did not overlap any protein-coding genes within a
487 5 kb distance (both from TSS and TES). We then overlapped the lncRNA
488 promoters with active promoter and strong enhancer chromatin states in whole
489 brain. The analysis resulted in 163 elncRNAs and 33 plncRNAs in whole brain.
490 We repeated the above steps in other tissues, resulting in hundreds of
491 chromatin-associated lncRNAs. This produced 41 ES-elncRNAs, 131 ES-
492 plncRNAs, 21 heart-elncRNAs, 61 heart-plncRNAs, 47 kidney-elncRNAs, 61
493 kidney-plncRNAs, 35 liver-elncRNAs, 77 liver-plncRNAs, 25 small intestine–
494 elncRNAs, 20 small intestine–plncRNAs, 20 spleen-elncRNAs, 65 spleen-
495 plncRNAs, 88 testes-elncRNAs, 258 testes-plncRNAs, 82 thymus-elncRNAs and
496 50 thymus-plncRNAs. Finally, we calculated the percentage of transition of
497 chromatin-associated lncRNA from one tissue to another (**Table S6**).

498 *Gene ontology analysis.*

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

502 **Experimental Procedures**

503 *Cell culture.*

504 Wild-type (E14Tg2A) ESCs were cultured feeder–free in plates coated with 0.1%
505 of gelatin in Glasgow minimum essential medium (Sigma) supplemented with β-
506 mercaptoethanol, sodium pyruvate, essential amino acids, GlutaMAX, 20% fetal
507 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 lncRNA 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 **Table S9**. RT-PCR was performed in triplicates using GAPDH as a housekeeping gene for normalization.

Characterization of mouse lncRNA-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) (**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 lncRNA-*Kdm8* (**Table S9**).

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 (**Table S7**).

Competing Interests

The authors declare that they have no competing interests.

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.

FIGURES

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 replicas 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 lncRNAs 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 lncRNAs (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 lncRNAs 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* = 1.221e-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 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), testes (29 plncRNAs and 12 elncRNAs) and thymus (47 plncRNAs and 40 elncRNAs). Links inside the bars indicate the number 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. (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 lncRNA, lncRNA-*Kdm8*, regulates the expression of a neighboring protein-coding gene Kdm8. (A) The lncRNA-*Kdm8* locus promoter overlaps with a 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). 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 lncRNA, lncRNA-*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 lncRNA-*Kdm8* are shown in green and red, respectively. Genomic DNA sequence corresponding to the 5′- and 3′-ends of the cloned lncRNA are shown in black at the bottom of the lncRNA-*Kdm8* gene track, defining accurate 5′-end and exon-intron boundaries for exon 1, exon 3, exon 4, and exon 5 of lncRNA-*Kdm8*. (D) Expression levels of lncRNA-*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 lncRNA-*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 lncRNA-*Kdm8* are given in **Table S9**.

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, Parker SCJ, Reddy TE, Rozowsky J, Schlesinger F, Thurman RE, Wang J, Ward LD, Whitfield TW, Wilder SP, Wu W, Xi HS, Yip KY, Zhuang J, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M, Pazin MJ, Lowdon RF, Dillon LAL, Adams LB, Kelly CJ, Zhang J, Wexler JR, Green ED, Good PJ, Feingold EA, Bernstein BE, Birney E, Crawford GE, Dekker J, Elnitski L, Farnham PJ, Gerstein M, Giddings MC, Gingeras TR, Green ED, Guigó R, Hardison RC, Hubbard TJ, Kellis M, Kent WJ, Lieb JD, Margulies EH, Myers RM, Snyder M, Stamatoyannopoulos JA, Tenenbaum SA, Weng Z, White KP, Wold B, Khatun J, Yu Y, Wrobel J, Risk BA, Gunawardena HP, Kuiper HC, Maier CW, Xie L, Chen X, Giddings MC, Bernstein BE, Epstein CB, Shoresh N, Ernst J, Kheradpour P, Mikkelsen TS, Gillespie S, Goren A, Ram O, Zhang X, Wang L, Issner R, Coyne MJ, Durham T, Ku M, Truong T, Ward LD, Altshuler RC, Eaton ML, Kellis M, 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, Batut P, Bell I, Bell K, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena HP, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Li G, Luo OJ, Park E, Preall JB, Presaud K, Ribeca P, Risk BA, Robyr D, Ruan X, Sammeth M, Sandhu KS, Schaeffer L, See L-H, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Hayashizaki Y, Harrow J, Gerstein M, Hubbard TJ, Reymond A, Antonarakis SE, Hannon GJ, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR, Rosenbloom KR, Sloan CA, Learned K, Malladi VS, Wong MC, Barber GP, Cline MS, Dreszer TR, Heitner SG, Karolchik D, Kent WJ, Kirkup VM, Meyer LR, Long JC, Maddren M, Raney BJ, Furey TS, Song L, Grasfeder LL, Giresi PG, Lee B-K, Battenhouse A, Sheffield NC, Simon JM, Showers KA, Safi A, London D, Bhinge AA, Shestak C, Schaner MR, Ki Kim S, Zhang ZZ, Mieczkowski PA, Mieczkowska JO, Liu Z, McDaniell RM, Ni Y, Rashid NU, Kim MJ, Adar S, Zhang Z, Wang T, Winter D, Keefe D, Birney E, Iyer VR, Lieb JD, Crawford GE, Li G, Sandhu KS, Zheng M,**

713 **Wang P**, **Luo OJ**, **Shahab A**, **Fullwood MJ**, **Ruan X**, **Ruan Y**, **Myers RM**,
714 **Pauli F**, **Williams BA**, **Gertz J**, **Marinov GK**, **Reddy TE**, **Vielmetter J**,
715 **Partridge E**, **Trout D**, **Varley KE**, **Gasper C**, **Bansal A**, **Pepke S**, **Jain P**,
716 **Amrhein H**, **Bowling KM**, **Anaya M**, **Cross MK**, **King B**, **Muratet MA**,
717 **Antoshechkin I**, **Newberry KM**, **McCue K**, **Nesmith AS**, **Fisher-Aylor KI**,
718 **Pusey B**, **DeSalvo G**, **Parker SL**, **Balasubramanian S**, **Davis NS**,
719 **Meadows SK**, **Eggleston T**, **Gunter C**, **Newberry JS**, **Levy SE**, **Absher**
720 **DM**, **Mortazavi A**, **Wong WH**, **Wold B**, **Blow MJ**, **Visel A**, **Pennachio LA**,
721 **Elnitski L**, **Margulies EH**, **Parker SCJ**, **Petrykowska HM**, **Abyzov A**,
722 **Aken B**, **Barrell D**, **Barson G**, **Berry A**, **Bignell A**, **Boychenko V**,
723 **Bussotti G**, **Chrast J**, **Davidson C**, **Derrien T**, **Despacio-Reyes G**,
724 **Diekhans M**, **Ezkurdia I**, **Frankish A**, **Gilbert J**, **Gonzalez JM**, **Griffiths**
725 **E**, **Harte R**, **Hendrix DA**, **Howald C**, **Hunt T**, **Jungreis I**, **Kay M**, **Khurana**
726 **E**, **Kokocinski F**, **Leng J**, **Lin MF**, **Loveland J**, **Lu Z**, **Manthravadi D**,
727 **Mariotti M**, **Mudge J**, **Mukherjee G**, **Notredame C**, **Pei B**, **Rodriguez JM**,
728 **Saunders G**, **Sboner A**, **Searle S**, **Sisu C**, **Snow C**, **Steward C**, **Tanzer**
729 **A**, **Tapanari E**, **Tress ML**, **van Baren MJ**, **Walters N**, **Washietl S**,
730 **Wilming L**, **Zadissa A**, **Zhang Z**, **Brent M**, **Haussler D**, **Kellis M**,
731 **Valencia A**, **Gerstein M**, **Reymond A**, **Guigó R**, **Harrow J**, **Hubbard TJ**,
732 **Landt SG**, **Frietze S**, **Abyzov A**, **Addleman N**, **Alexander RP**, **Auerbach**
733 **RK**, **Balasubramanian S**, **Bettinger K**, **Bhardwaj N**, **Boyle AP**, **Cao AR**,
734 **Cayting P**, **Charos A**, **Cheng Y**, **Cheng C**, **Eastman C**, **Euskirchen G**,
735 **Fleming JD**, **Grubert F**, **Habegger L**, **Hariharan M**, **Harmanci A**, **Iyengar**
736 **S**, **Jin VX**, **Karczewski KJ**, **Kasowski M**, **Lacroute P**, **Lam H**, **Lamarre-**
737 **Vincent N**, **Leng J**, **Lian J**, **Lindahl-Allen M**, **Min R**, **Miotto B**, **Monahan**
738 **H**, **Moqtaderi Z**, **Mu XJ**, **O'Geen H**, **Ouyang Z**, **Patacsil D**, **Pei B**, **Raha**
739 **D**, **Ramirez L**, **Reed B**, **Rozowsky J**, **Sboner A**, **Shi M**, **Sisu C**, **Slifer T**,
740 **Witt H**, **Wu L**, **Xu X**, **Yan K-K**, **Yang X**, **Yip KY**, **Zhang Z**, **Struhl K**,
741 **Weissman SM**, **Gerstein M**, **Farnham PJ**, **Snyder M**, **Tenenbaum SA**,
742 **Penalva LO**, **Doyle F**, **Karmakar S**, **Landt SG**, **Bhanvadia RR**,
743 **Choudhury A**, **Domanus M**, **Ma L**, **Moran J**, **Patacsil D**, **Slifer T**,
744 **Victorsen A**, **Yang X**, **Snyder M**, **White KP**, **Auer T**, **Centanin L**,
745 **Eichenlaub M**, **Gruhl F**, **Heermann S**, **Hoeckendorf B**, **Inoue D**, **Kellner**
746 **T**, **Kirchmaier S**, **Mueller C**, **Reinhardt R**, **Schertel L**, **Schneider S**, **Sinn**
747 **R**, **Wittbrodt B**, **Wittbrodt J**, **Weng Z**, **Whitfield TW**, **Wang J**, **Collins PJ**,
748 **Aldred SF**, **Trinklein ND**, **Partridge EC**, **Myers RM**, **Dekker J**, **Jain G**,
749 **Lajoie BR**, **Sanyal A**, **Balasundaram G**, **Bates DL**, **Byron R**, **Canfield**
750 **TK**, **Diegel MJ**, **Dunn D**, **Ebersol AK**, **Frum T**, **Garg K**, **Gist E**, **Hansen**
751 **RS**, **Boatman L**, **Haugen E**, **Humbert R**, **Jain G**, **Johnson AK**, **Johnson**
752 **EM**, **Kutyavin TV**, **Lajoie BR**, **Lee K**, **Lotakis D**, **Maurano MT**, **Neph SJ**,
753 **Neri FV**, **Nguyen ED**, **Qu H**, **Reynolds AP**, **Roach V**, **Rynes E**, **Sabo P**,
754 **Sanchez ME**, **Sandstrom RS**, **Sanyal A**, **Shafer AO**, **Stergachis AB**,
755 **Thomas S**, **Thurman RE**, **Vernot B**, **Vierstra J**, **Vong S**, **Wang H**,
756 **Weaver MA**, **Yan Y**, **Zhang M**, **Akey JM**, **Bender M**, **Dorschner MO**,
757 **Groudine M**, **MacCoss MJ**, **Navas P**, **Stamatoyannopoulos G**, **Kaul R**,
758 **Dekker J**, **Stamatoyannopoulos JA**, **Dunham I**, **Beal K**, **Brazma A**,

Flicek P, **Herrero J**, **Johnson N**, **Keefe D**, **Lukk M**, **Luscombe NM**, **Sobral D**, **Vaquerizas JM**, **Wilder SP**, **Batzoglou S**, **Sidow A**, **Hussami N**, **Kyriazopoulou-Panagiotopoulou S**, **Libbrecht MW**, **Schaub MA**, **Kundaje A**, **Hardison RC**, **Miller W**, **Giardine B**, **Harris RS**, **Wu W**, **Bickel PJ**, **Banfai B**, **Boley NP**, **Brown JB**, **Huang H**, **Li Q**, **Li JJ**, **Noble WS**, **Bilmes JA**, **Buske OJ**, **Hoffman MM**, **Sahu AD**, **Kharchenko PV**, **Park PJ**, **Baker D**, **Taylor J**, **Weng Z**, **Iyer S**, **Dong X**, **Greven M**, **Lin X**, **Wang J**, **Xi HS**, **Zhuang J**, **Gerstein M**, **Alexander RP**, **Balasubramanian S**, **Cheng C**, **Harmanci A**, **Lochovsky L**, **Min R**, **Mu XJ**, **Rozowsky J**, **Yan K-K**, **Yip KY**, **Birney E**. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature **489**:57–74.

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-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**. 2013. Landscape of transcription in human cells. Nature **488**:101–108.

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 Research **22**:1775–1789.

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 & Development **25**:1915–1927.

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.

6. **Nam J-W**, **Bartel DP**. 2012. Long noncoding RNAs in C. elegans. Genome Research **22**:2529–2540.

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 Biology and Evolution **4**:427–442.

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 Research **22**:577–591.
9. **Panning B**, **Dausman J**, **Jaenisch R**. 1997. X Chromosome Inactivation Is Mediated by Xist RNA Stabilization. Cell **90**:907–916.
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.
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.
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.
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.
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. Developmental Cell **24**:206–214.
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. Molecular Cell **39**:925–938.
16. **Ng S-Y**, **Bogu GK**, **Soh BS**, **Stanton LW**. 2013. The Long Noncoding RNA RMST Interacts with SOX2 to Regulate Neurogenesis. Molecular Cell **51**:349–359.
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.
18. **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**, **oacute RG**, **Gingeras TR**. 1AD. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nat Comms **6**:1–11.

19. **Trapnell C**, **Pachter L**, **Salzberg SL**. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25**:1105–1111.
20. **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.
21. **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.
22. **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.
23. **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 Research.
24. **Lin MF**, **Jungreis I**, **Kellis M**. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics **27**:i275–i282.
25. **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.
26. **Morán I**, **Akerman İ**, **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.
27. **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.
28. **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.
29. **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.
30. **DGT TFCATRPAC**. 2015. A promoter-level mammalian expression atlas. Nature **507**:462–470.
31. **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.

32. **Ernst J**, **Kellis M**. 2012. ChromHMM: automating chromatin-state discovery and characterization. Nat Meth **9**:215–216.

33. **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.

34. **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.

35. **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–e01749.

36. **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.

37. **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.

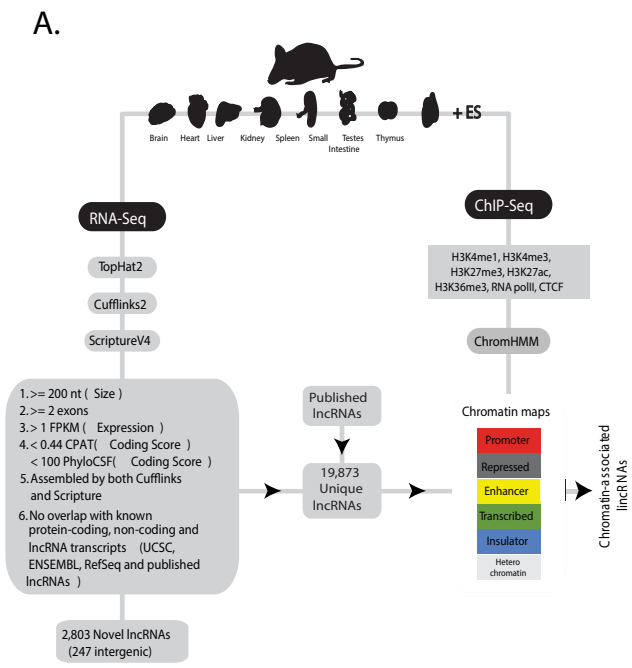
38. **Lai F**, **Orom UA**, **Cesaroni M**, **Beringer M**, **Taatjes DJ**, **Blobel GA**, **Shiekhattar R**. 2013. Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature **494**:497–501.

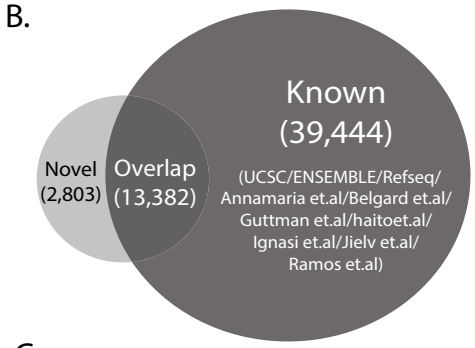
39. **Qrom UA**, **Derrien T**, **Beringer M**, **Gumireddy K**, **Gardini A**, **Bussotti G**. 2010. Long Noncoding RNAs with Enhancer-like Function. Cell **143**:46–58.

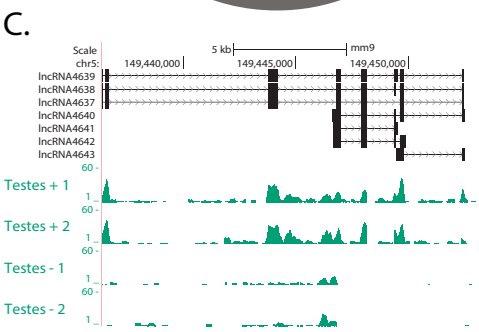
40. **Wu X**, **Sharp PA**. 2013. Perspective. Cell **155**:990–996.

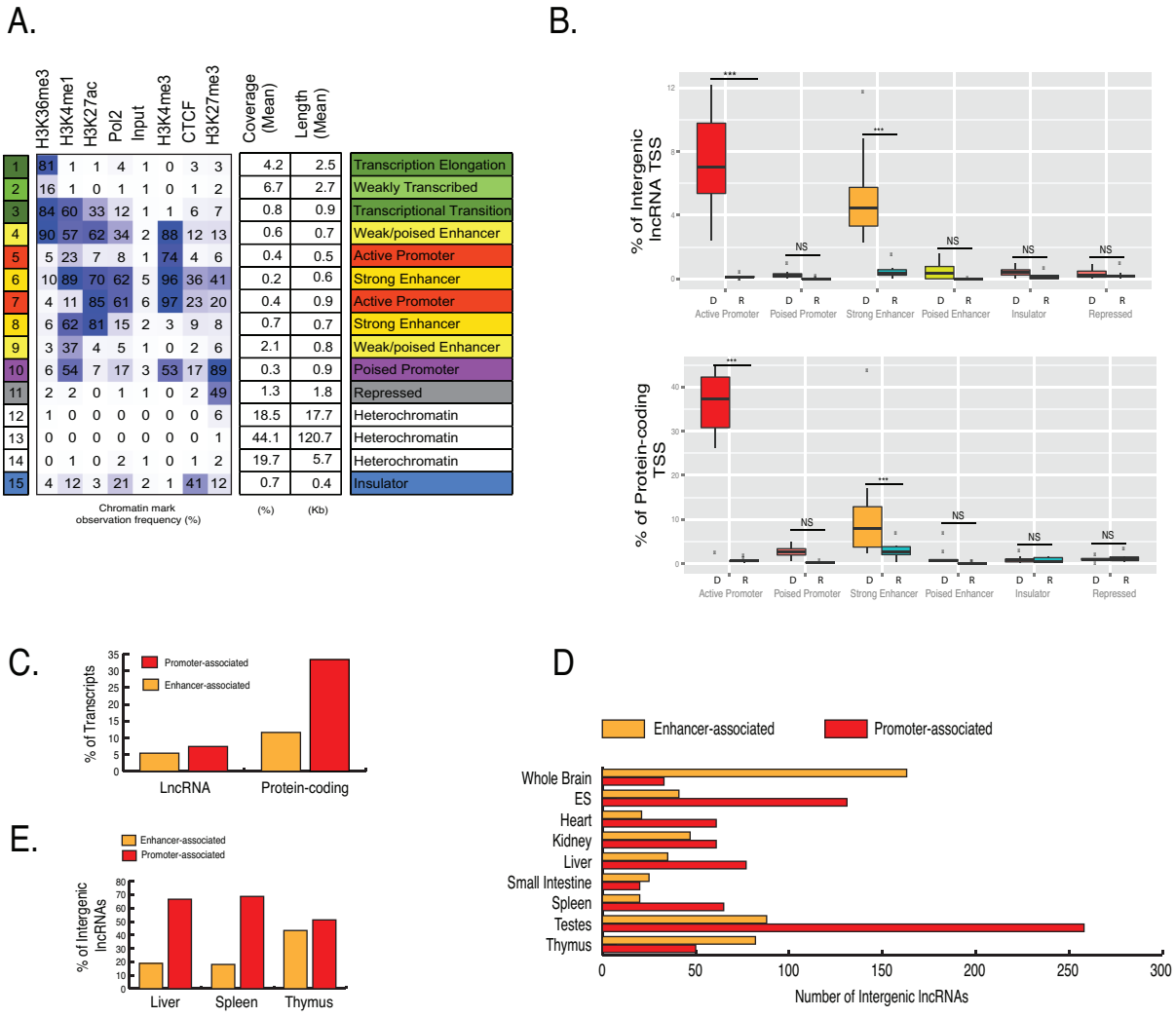
41. **Garcia-Alcalde F**, **Okonechnikov K**, **Carbonell J**, **Cruz LM**, **Gotz S**,

Tarazona S, **Dopazo J**, **Meyer TF**, **Conesa A**. 2012. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics **28**:2678–2679.
42. **Gupta S**, **Stamatoyannopoulos JA**, **Bailey TL**, **Noble W**. 2007. Quantifying similarity between motifs. Genome Biol **8**:R24.
43. **Stojnic R**. 2014. PWMEnrich: PWM enrichment analysis. R package version 3.6.1 1–46.
44. **Shin H**, **Liu T**, **Manrai AK**, **Liu XS**. 2009. CEAS: cis-regulatory element annotation system. Bioinformatics **25**:2605–2606.
45. **Quinlan AR**, **Hall IM**. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**:841–842.









Molecular and Cellular Biology