Understanding the transcriptional landscape of non-coding genome (lncRNAs and repetitive elements) in mammals

Gireesh K. Bogu

TESI DOCTORAL UPF 2016

Thesis Directors

Roderic Guigó
Department of Bioinformatics and Genomics
Center for Genome Regulation (CRG), Barcelona

Marc A. Marti-Renom
Centre Nacional d'Anàlisi Genòmica
Department of Gene Regulation and Cancer
Center for Genomic Regulation (CNAG-CRG), Barcelona

Acknowledgments

I would like to thank the following..

Marc Marti-Renome and his group: Marc has shown a great interest in me during my 2012 Ph.D. interviews and eventually I ended up in his group. Marc supported greatly through out my Ph.D. and especially my lncRNA-related work. My first author paper is with him. He always encouraged my ideas and given the freedom that I wanted. I greatly admire his mentorship and also made some very good friends in his group, Fran Martinez.

Roderic Guigo and his group: Roderic is one of the top genomics scientists in the world and adopted me in 3rd year with out any hesitance. He is my co-mentor from the last two years. Though he is incredibly busy, he always allocates time to mentor me through out these years. He made me to become a part of internal projects and encouraged my ideas especially my repetitive elements-related work. His group is full of incredibly smart people who are experts in both experimental and computational methods. I obviously gained so much knowledge from them and group meetings. I would like to thank Roderic and the group, especially, Ferran Reverter, Rory Johnson, Alessandra Breschi, Emilio Palumbo and my GTEx team.

Luciano (and Pedro Vizan): Luciano and his colleague Pedro were some of the co-authors in my lncRNA paper and helped greatly in validating the functional lncRNAs. I thank them so much for their helpful collaboration.

Thesis committee members (Eduarado Eyras, Gian Gaetano Tartaglia)

Rory Johnson: Rory is a great influence on me. He is the one who first introduced me to lncRNA world. I started as an intern student under him at GIS, Singapore. I really thank him for his guidance during my student period.

Lawrence W. Stanton: Larry was my first supervisor after my Masters. I started my real research career under him as an intern and a bioinformatics specialist. With out this, I couldn't have been here. He has great influence on me (both personally and professionally).

Jernej Ule: We worked together on a paper before my Ph.D. He recommended me for both Sangers institute and CRG. Personally and professionally Jernej is a great person.

Shi Yan Ng: We did some great work on lncRNAs together and it really encouraged me to continue my work on it

Bill Keys: Our “elevator talk” really happened and it led to the idea of analyzing repeats.

Juan Valcarcel: One of the main reasons to choose CRG for my Ph.D. He helped greatly during my difficult Ph.D. period. I always thank him for that. He has been my thesis committee member and provided very useful feedback

Family and friends and colleagues: Obviously, with out my dad, mom and my family, I wouldn’t have been here. I always look up to them. I would like to thank my best friend, Sujay Parasa, for encouraging me to apply best places in the world to continue my studies. I would also like to thank my friends who have been supportive of me: Andreas Wilm, Sanjana Nagarajan, Akshay Bhinge, Joao Varela, Sofia Piano, Eduardo Lobenstein, Alessandro Julys, Catalina Rubio, Paula Palomino, Karthik Armugam, Hima Priyanka, Alexandra Santanach, Verónica Lloréns Rico, Lisa Johnsen, Maria Chatzou, Lorenzo Rinaldi and Marcos Perez.

Places

Genome Institute of Singapore (GIS), Singapore
Center for Genomic Regulation CRG, Barcelona
Barcelona (Gracia)

My pre-doctoral fellowship

La Caixa, Barcelona: 4 years fellowship program

Abstract

Widespread transcription in mammals revealed unexpected discovery of non-coding elements like long noncoding RNAs (lncRNAs) and repetitive elements. First, lncRNAs were previously identified in limited number of tissues or cell lines in mouse and the discovery of lncRNAs was still pending in many other tissues in mouse. To address this, we applied a computational pipeline that discovered 2,803 high-confidence novel lncRNAs by mapping and *de novo* assembling billions of RNA-Seq reads in eight tissues and a primary cell line in mouse. Further, we integrated this catalog of lncRNAs with chromatin state maps and found many regulatory lncRNAs (promoter-associated and enhancer-associated lncRNAs). Second, more than half of the human genome contains repetitive elements. However, it is not clear how they are expressed across all mammalian tissues. To address this, as a part of Genotype-Tissue Expression (GTEx) project, we profiled repetitive elements using 8,551 poly-A RNA-Seq datasets from 53 tissues across 550 individuals and found various repeat families transcribed across multiple human tissues in a tissue-specific manner. In summary, to understand the transcriptional landscape of non-coding genome, we mainly analyzed RNA-Seq datasets across many tissues in mammals and show that the non-coding elements like lncRNA and repetitive elements are not only transcribed but also tissue-specific. Together, this thesis work defines a unique collection of non-coding elements that are transcribed and tissue-specific in mammalian tissues.

Resumen

Una gran parte del genoma de mamiefores se expresa en forma de ARNs y se conoce hoy en dia que una gran parte de estos transcritos son no codificantes llamados lncRNAs y que contienen elementos repetitivos. En ratones, estos han sido identificados recientemente en un número limitado de tejidos y líneas celulares. Esta tesis presenta un trabajo exhaustivo de estudio de lnRNAs en ratón en ocho tejidos y una línea celular. En este trabajo se descubrieron 2803 nuevos lncRNAs a los cuáles se les asignó una función reguladora (asociados a promotores o activadores “enhancers”) en el genoma usando datos del estado de la cromatina. Asimismo, más de la mitad del genoma humano contiene elementos repetitivos. Desafortunadamente no se conoce el patrón de expresión de estos elementos repetitivos en los tejidos mamíferos. Como miembros del proyecto GTEx (Genotype-

Tissue Expression), analizamos la expresión de estos elementos repetitivos en 8,551 muestras de polyA RNA-Seq en 53 tejidos de 550 individuos. Encontramos que muchas familias de elementos repetitivos son expresadas en tejidos específicos en varios individuos, y representan una característica peculiar de la identidad de cada tejido en humanos.

Preface

The mammalian genome is mostly (~98%) composed of non-coding regions, that is, regions that do not code for proteins {Kellis:2014gy}. However, a large fraction of these non-coding regions are transcribed. Indeed, thousands of long non-coding RNAs (lncRNAs) transcribed from non-coding regions of the genome have already been cataloged, but their regulatory roles are still unclear {Rinn:2012fh}. Interestingly, a substantial part of the lncRNAs is originated from or contain repetitive elements, which are normally transcriptionally repressed {Kapusta:2013kb}. Despite their general repressed state, specific families of repeats escape repression and their transcripts play a role in gene regulation (Elbarbary et al., 2016). Unfortunately, it is not exactly clear how many of them are expressed across all mammalian tissues and in which tissue they are expressed.

Recently, regulatory lncRNAs have been shown to preferentially associate in the genome to promoter and enhancer chromatin states in a mouse cell line {Marques:2013cm}. While this observation is highly interesting, it was unclear whether the association was universal in different cell types and tissues. Moreover, the fact that the previous studies used only a single cell line statistically limited the testing of thousands of lncRNAs. To address those limitations, we built a comprehensive chromatin-associated mouse lncRNA data set using billions of mapped RNA-Seq reads to identify high-confidence novel lncRNAs (**First Manuscript**) {MouseENCODEConsortium:2012ku}. Next, we used more than a billion mapped chromatin immunoprecipitation sequencing reads of various histone marks to identify chromatin state maps in several cell lines and tissues {MouseENCODEConsortium:2012ku}. Finally, we integrated all these mouse lncRNAs with the chromatin state maps, which resulted in a comprehensive catalog of predicted functional lncRNA transcripts. 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 switched from one state to another state across all the tissues or cell lines we tested. To our knowledge, at the time of the publication, this was the most comprehensive data set of chromatin state-associated lncRNAs in mouse.

To understand the transcriptional landscape of repetitive elements in humans, we used mid-phase RNA sequencing (RNA-Seq) data from the GTEx project (**Second**

Manuscript) (Kellis et al., 2014; The GTEx Consortium et al., 2015). It consisted of 8,551 RNA-Seq samples from 544 human individuals spanning 53 distinct body sites. The samples were extracted from individuals of different age, gender and race and the RNA-Seq data is of 76-base pair (bp), paired-end, unstranded, poly (A)-selected with a median sequencing depth of 60 million reads per sample and with a good RNA quality. The above 53 distinct body sites include 29 solid-organ tissues, 13 brain subregions, two cell lines (EBV-transformed lymphocytes and transformed fibroblasts) and a whole-blood sample. Additionally, around 5 million repetitive genomic instances of various repeat subfamilies, families and classes from the RepeatMasker database were used in the analysis as target sites (Kellis et al., 2014; Tarailo-Graovac and Chen, 2002). We found that about 206,000 repetitive elements expressed in at least once of the 53 distinct human body sites. Cerebellum and testis had the highest number of expressed repetitive elements. To test whether the repetitive elements transcription was tissue-specific, we performed hierarchical clustering on the normalized expression and the clustering successfully recapitulated each tissue type. In total, we found 3,295 tissue-specific repeat instances across the entire human genome. Once published, this work will be a valuable resource to help researchers select candidate regulatory lncRNAs or tissue-specific repetitive elements and conduct further experimental studies.

List of publications during the thesis (2012-2016):

1. **Bogu GK**, Vizán P, Stanton LW, Beato M, Di Croce L, Marti-Renom MA. *Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse*. Mol Cell Biol. 2015 Dec 28;36(5).
2. **Bogu GK**, Ferran Reverter, The GTEx Consortium, Marc A Marti-Renom, Roderic Guigo. *The Transcriptional Landscape of Repetitive Elements in Human Tissues*. (Manuscript in preparation)

Contents

Introduction

Over the past ten years, advances in next-generation sequencing technologies have helped the analysis of transcriptome (total number of transcripts in a cell) of mammalian genomes at unprecedented depth and in an unbiased manner (Goodwin et al., 2016; M.-S. Kim et al., 2014; Rinn and Chang, 2012; Wilhelm et al., 2015). The conventional view of transcription is that mRNAs with defined gene structure are generated from the coding part of the genome. However, recent findings show that almost all of the mammalian genome is transcribed at some level (DGT, 2015; Djebali et al., 2016; Ho et al., 2015; Kapusta et al., 2013; Kellis et al., 2014; Perez-Pinera et al., 2015; The FANTOM Consortium, 2005). It is further supported by other transcriptome studies that detected tens of thousands of transcripts, many of which spliced, coming from non-coding regions of mammalian genomes (Cabili et al., 2011; Guttman et al., 2010; Iyer et al., 2015; Marques et al., 2013; Treangen and Salzberg, 2011). As the occurrence of these transcripts is widespread, this form of transcription has been described as 'pervasive' (Kapranov et al., 2007; Kapusta et al., 2013; Mouse ENCODE Consortium et al., 2012). However, whether pervasive transcription is simply 'biological noise' or whether it actually has distinct functional roles remains a challenging question (Doolittle, 2013; Eddy, 2012; Elbarbary et al., 2016; Kellis et al., 2014; Lander et al., 2001; Mouse ENCODE Consortium et al., 2012; Palazzo, 2016; Treangen and Salzberg, 2011; van Bakel et al., 2011).

Pervasive transcription in human and mouse revealed several findings, including the transcription of noncoding elements like long non-coding RNAs (lncRNAs) and repetitive elements (Cabili et al., 2011; Derrien et al., 2012; Devaux et al., 2015; Djebali et al., 2016; Faulkner et al., 2009; Fort et al., 2014; Guttman et al., 2010; Iyer et al., 2015; McClintock, 1951; Mele et al., 2015). LncRNAs, a group of long RNA transcripts with no apparent protein-coding potential, are found almost in every organism. Interestingly, the complexity of organisms correlate more with the diversity and size of non-coding RNA expression repertoires than with that of protein-coding genes, at least in eukaryotes (Chinwalla et al., 2002; Devaux et al., 2015; Feschotte et al., 2002; Kapusta and Feschotte, 2014; Lander et al., 2001). By recent estimates, the number of human lncRNAs overrun the number of protein-coding genes (Devaux et al., 2015; Djebali et al., 2016; Iyer et al., 2015; Kapusta and Feschotte, 2014). Moreover, the total number of known lncRNAs continues to grow, enhanced by

deeper and advanced RNA, epigenomic sequencing technologies and computational prediction methods.

Repetitive DNA elements are sequences that are similar or identical to sequences elsewhere in the genome (Devaux et al., 2015; Elbarbary et al., 2016; Grote et al., 2013; Treangen and Salzberg, 2011) and their transcription has a key influence upon the transcriptional output of the mammalian genome (Elbarbary et al., 2016; Faulkner et al., 2009; Lorenzen and Thum, 2016). It has been estimated that 6 to 30% of mouse and human RNA transcripts initiate within repetitive elements (Briggs et al., 2015; Faulkner et al., 2009; Smit, 1996). Importantly, it has been shown that repeats can transcribe into RNA and especially into lncRNAs (Kapusta et al., 2013; Lu et al., 2014; Ugarkovic, 2005; Ugarkovic and Plohl, 2002; van de Vondervoort, 2013). Although repetitive elements pervade mammalian genomes and transcribe into lncRNAs, their overall contribution to transcriptional activity is poorly defined.

Analysis of transcription in human and mouse tissues revealed novel lncRNAs and repeat families in various studies (Belgard et al., 2011; Guttman et al., 2010; Knoll et al., 2015; Lander et al., 2001; Luo et al., 2013; Lv et al., 2013; Marques et al., 2013; Morán et al., 2012; Ramos et al., 2013). However, these discoveries were far from complete, as many tissues have not been analyzed in detail. In this thesis work, I discuss my current understanding of non-coding mammalian genome especially by looking at lncRNAs and repetitive elements. In this thesis work, I study the transcriptional landscape of non-coding genome to evaluate whether the non-coding elements show evidence of encoding transcriptional products. In particular, we analyze the transcription of lncRNAs (**First Manuscript**) and repetitive elements (**Second Manuscript**) across various mammalian tissues. We find that specific non-coding elements that result from pervasive transcription are more than 'transcriptional noise' and have important functions in gene regulation.

1. Long noncoding RNAs

1.1. Definition

RNAs that do not encode any protein and are longer than 200 nucleotides in length have been named as lncRNAs (Knoll et al., 2015; Rinn and Chang, 2012; Usdin, 2008). The use of an arbitrary length cut-off contributed to separate them from the so-called small non-coding RNAs (Mercer et al., 2009; Sun et al., 2013; Usdin, 2008). LncRNA transcripts mainly classified as non-coding based on the absence of long open reading frames (ORFs) with a size more than 100 codons or the absence of codon conservation across different species (Knoll et al., 2015; Morris and Mattick, 2014), though there were few exceptions (Banfai et al., 2012; M.-S. Kim et al., 2014; Lorenzen and Thum, 2016; Mele et al., 2015; Wilhelm et al., 2015).

1.2. Classification

According to their relative genomic position with respect to neighboring protein-coding genes, lncRNAs have been classified in five categories (Criscione et al., 2014; Devaux et al., 2015; Lorenzen and Thum, 2016; Mattick and Rinn, 2015):

1. Intergenic lncRNAs (lincRNAs) are located between two protein-coding genes; the majority of lncRNAs belong to this category.
2. Intronic lncRNAs are located within introns of protein-coding genes.
3. Bidirectional promoter lncRNAs are transcribed within 1 kilobases (kb) of promoters in the opposite direction from the protein-coding transcript.
4. Sense lncRNAs are transcribed from the sense strand of protein-coding genes, and can overlap introns and part or all of the exon.
5. Antisense lncRNAs are transcribed from the antisense strand of protein-coding genes, and can overlap an exon of the protein-coding gene in the sense strand, an intron, or both.

This classification has been used mainly to reduce ambiguity and overlap with protein-coding genes. However, there are no evidences of any intrinsic difference between these transcripts, for example, in their interaction with chromatin-activating or chromatin repressive complexes (Göke et al., 2015; Huarte, 2015; Morris and Mattick, 2014; Sahu et al., 2015; Schmitt and Chang, 2016).

1.3. Discovery

Genomic projects over the past decade have used shotgun sequencing and microarray hybridization to obtain evidence for many non-coding transcripts in mammals, However, these approaches have presented limitations including lack of

strand information and low sequence coverage (Elbarbary et al., 2016; Martin and Zhong Wang, 2011; Prensner et al., 2011; Slotkin and Martienssen, 2007). Advances in RNA-Seq have opened the way to unbiased and efficient analysis of the complete transcriptome of mammals (Slotkin and Martienssen, 2007; Tsai et al., 2010; Zhong Wang et al., 2009). Utilizing the RNA-Seq reads, several methods that *de novo* reconstruct the transcriptome have emerged (Bourque, 2009; Elbarbary et al., 2016; Gutschner et al., 2013; Guttman et al., 2010; Pertea et al., 2015; Trapnell et al., 2012; Tripathi et al., 2010) and led to the discovery of thousands of novel multi-exonic lncRNAs across different mammalian cell lines and tissues. Guttman and colleagues took an entirely different approach to discover functional lncRNAs on the basis of exploiting chromatin structure (Guttman et al., 2009; Ichiyanagi, 2013; Prensner et al., 2013). They systematically discover functional lncRNAs by overlapping transcribed regions with RNA polymerase II (Pol II) and marked by trimethylation of lysine 4 of histone H3 (H3K4me3) at their promoter and trimethylation of lysine 36 of histone H3 (H3K36me3). Until now, tens of thousands of lncRNAs have been discovered in mammals including our mouse study (Bogu et al., 2016; Estecio et al., 2012; Huarte, 2015; Mele et al., 2015; Schmitt and Chang, 2016).

1.4. Sequence features

Derrien and colleagues showed that 98% of lncRNAs are spliced, however only 25% of lncRNA genes show evidence of alternative splicing with at least two different transcript isoforms per gene locus (Derrien et al., 2012; Kapusta and Feschotte, 2014; Necsulea and Kaessmann, 2014; Tashiro et al., 2011). Median size of exons (149 bp) and introns (2280 bp) of lncRNA in humans are larger than protein-coding exons (132 bp) and introns (1602 bp). However, overall lncRNA transcripts are shorter than protein-coding (median 592 bp compared with 2453 bp for protein-coding transcript (Derrien et al., 2012; Faulkner et al., 2009; Guttman et al., 2009; Kunarso et al., 2010).

1.5. Localization

According to their cellular localization, lncRNAs can be categorized into nuclear or cytosolic, but some lncRNAs can be found in both compartments (Cabili et al., 2011; Derrien et al., 2012; Djebali et al., 2016; Iyer et al., 2015; Lunyak and Atallah, 2011).

Nuclear lncRNAs such as Xist are likely to exert their functions by modifying chromatin structure, thereby influencing gene transcription, whereas cytosolic lncRNAs regulate target mRNA stability and translational efficiency through RNA-RNA interactions (Cabili et al., 2011; Elbarbary et al., 2016; Knoll et al., 2015). The observation that many nuclear lncRNAs can be depleted in cultured cells by expression of short hairpin RNAs suggests that lncRNAs probably shuttle between the nucleus and cytosol, although this hypothesis has not been tested (Knoll et al., 2015; Mariner et al., 2008).

1.6. Tissue specificity

A striking feature of lncRNAs is that they have greater tissue specificity than protein-coding RNAs, which suggests that lncRNAs might have a crucial role in the formation of multiple, if not all, cell types (Allen et al., 2004; Knoll et al., 2015). Mercer and colleagues first published a study detailing precise and specific expression several lncRNAs in various subcellular compartments of mouse brain (Faulkner et al., 2009; Mercer et al., 2008). Later, transcriptome-wide studies showed that lncRNAs in general exhibit more specific expression profiles than mRNAs (Cabili et al., 2011; Derrien et al., 2012; Kapusta et al., 2013). Furthermore, lncRNA expression patterns are often correlated with mRNA expression patterns both in cis and in trans, suggesting that certain lncRNAs may be co-regulated in expression networks (Kapusta and Feschotte, 2014; Quinn and Chang, 2016; Rinn and Chang, 2012). This tissue or cell specificity has been used as evidence that lncRNA expression is even more tightly regulated than that of protein-coding genes, thereby arguing for the essential role of lncRNAs in determining cell state (Elbarbary et al., 2016; Guttman et al., 2009).

1.7. Evolutionary conservation

Evolution in gene sequence, expression or regulation leads to phenotypic differences between species and therefore the evolution history of lncRNAs can provide their functionality (Lev-Maor et al., 2007; Necsulea and Kaessmann, 2014). A recent study has characterized a major fraction of lncRNA across eight organs and eleven species by discovering 11,000 primate-specific lncRNAs (Necsulea and Kaessmann, 2014; Sorek et al., 2002). Very few lncRNAs (2,500) were highly conserved at both sequence and expression level suggesting that most of the lncRNA sequence and

expression seem to evolve more rapidly compared to protein-coding genes. This rapid turnover impacts the neighboring protein-coding gene expression and contributes biological difference between species (An et al., 2004; Kutter et al., 2012). Even the LncRNAs that were shared between species had lower expression conservation and organ specificity levels than protein-coding counterparts. Another similar study revealed that clustering of lncRNA expression recapitulates organ type with the help of high coverage RNA-Seq data across 9 tissues and 6 species further supporting their functional relevance (Elbarbary et al., 2013; Washietl et al., 2014).

1.8. Molecular Mechanisms of lncRNAs

LncRNA molecular mechanisms, at epigenetic or transcriptional level, range broadly. We can find lncRNA regulating protein-coding gene expression in space and/or time (*signals*), recruiting chromatin-modifier proteins to the promoters of protein-coding genes (*guides*), binding proteins and moving them away from the chromatin (*decoys*), and bringing together multiple proteins to form ribonucleoprotein complexes (*scaffolds*) (Fitzpatrick and Huang, 2014; Guttman and Rinn, 2012; Rinn and Chang, 2012; Kevin C Wang and Chang, 2011).

- Signals are lncRNAs that regulate spatiotemporal gene expression such as activating of genes within specific tissues of an organism at specific times during development. For example, the lncRNA Xist is expressed during female development and binds to the chromosome where it is transcribed. During that process, it helps silencing most of the genes on the inactivated X-chromosome (Tian et al., 2010). Both HOTTIP and HOTAIR lncRNAs are spatially expressed from HOX genomic loci along developmental axes and their expression demarcates broad chromosomal domains of differential histone methylation and RNA polymerase accessibility (Rinn et al., 2007; Kevin C Wang et al., 2011). The expression of eRNAs (enhancer RNAs) transcribed from enhancer regions (enhancers are the non-coding regions that are located far away from the promoter a gene but regulate its transcription) positively correlated with nearby protein-coding genes that regulate neuronal activity in mouse at different time points (T.-K. Kim et al., 2010). Several lncRNAs with enhancer-like functions have also been identified in various human cell lines and shown to regulate several master regulators of cellular differentiation (Qrom et al., 2010).

- Guides are lncRNAs that directly bind to transcription factors or other proteins like chromatin modifiers. These guides recruit the proteins to the promoters of protein-coding genes to either activate or repress them (Kevin C Wang and Chang, 2011). LncRNAs can guide changes in gene expression in cis (nearby genes) or trans (far away genes) manner. For example, eRNAs regulate the gene expression in cis manner (Qrom et al., 2010; Ørom and Shiekhattar, 2011) and in contrast, HOTAIR regulates in trans manner (Meredith et al., 2016). In fact, HOTAIR was the first lncRNA that has been shown to regulate transcription of the genes on the other chromosome. Also it has been shown that knocking down several lncRNAs induce gene expression changes in genes far away in sequence (Guttman et al., 2011). Xist, Air, COLDAIR, CCND1, and HOTTIP are other examples of lncRNAs that act in cis mode and LincRNA-p21, Jpx in trans mode (Kevin C Wang and Chang, 2011).

- Decoys are lncRNAs that transcribe and bind to a protein and then move it away from the target site (Kevin C Wang and Chang, 2011). LncRNA upstream to DHFR protein-coding gene binds to transcription factor IIb and inhibits DHFR transcription by preventing it from binding to the promoter (Martianov et al., 2007). TERRA is a telomeric repeat-lncRNA binds to telomerase and retains it near the telomeric 3' end while inhibiting its action (Kevin C Wang and Chang, 2011). LncRNA PANDA (P21 associated ncRNA DNA damage activated) binds to a transcription factor NF-YA and prevents it from activating the apoptosis (Hung et al., 2011). Several lncRNAs also acts as decoys for miRNAs and splicing factors (Kevin C Wang and Chang, 2011).

- Scaffolds are lncRNAs that contain different structural domains and they could bind to both activator and repressor proteins at the same time to regulate gene expression (Kevin C Wang and Chang, 2011). LncRNA ANRIL serves as a scaffold and recruits multiple sets of polycomb complexes (PRC1 and PRC2) to the target gene for silencing (Hung et al., 2011; Yap et al., 2010). Like ANRIL, HOTAIR also recruits different chromatin proteins (LSD1 and PRC2) and repressed the gene expression (Rinn et al., 2007; Tsai et al., 2010).

Recent evidences suggest that lncRNAs also regulate functions at post-transcriptional level (Göke and Ng, 2016; Mercer et al., 2009). For example, an

antisense lncRNA mask the splice site of *Zeb2* protein-coding gene from the spliceosome by binding to the 5' UTR resulting in intron retention. The translation machinery can then recognize and bind an internal ribosome entry site (IRE) in the retained intron, resulting in efficient *Zeb2* translation and expression (Beltran et al., 2008; Treangen and Salzberg, 2011).

1.9. Technologies to understand mechanistic functions of lncRNAs

Experimental approaches that validate mechanisms or functions of the genome focused mainly on protein-coding genes or miRNAs but not on lncRNAs. However, recently lncRNA-centric experimental approaches that interrogate structure, genomic organization and protein interactions of lncRNAs are slowly emerging (Chu et al., 2015; Steijger et al., 2013). ChIRP (Chromatin Isolation by RNA purification), developed by Chu Ci and colleagues at Stanford (Cabili et al., 2011; Chu et al., 2011), helps identifying complete genome-wide binding sites of any given lncRNA. Similar methods have been published known as CHART-Seq (Capture Hybridization of RNA Targets Sequencing) and RAP (RNA Antisense Purification) (Engreitz et al., 2013; Simon et al., 2013; Tarailo-Graovac and Chen, 2002) and successfully applied to predict genome-wide lncRNA and chromatin interactions. Recently, CHART-Seq and RAP methods have been applied and revealed detailed mechanisms of Xist on X-chromosome inactivation (Engreitz et al., 2013; Hoen et al., 2016; Simon et al., 2013). Though the above methods capture the chromatin binding sites of lncRNA, not every lncRNA regulate functions by binding DNA. Crosslinking and Immunoprecipitation (CLIP) methods designed by Jernej and colleagues (Ingolia et al., 2011; Ule et al., 2005; Zhen Wang et al., 2010) have been widely used to identify genome-wide interactions between RNA and proteins. Similar methods have been published known as HITS-CLIP (High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation), CLIP-Seq (cross-linked immunoprecipitation followed by next generation sequencing) and PAR-CLIP (Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) have been applied successfully to capture the RNA-protein interactome (Chi et al., 2009; Guttman et al., 2013; Hafner et al., 2010; Yeo et al., 2009). Very recently, CLIP revealed novel proteins that bind to Xist and regulate X-chromosome inactivation (Ingolia et al., 2014; McHugh et al., 2015). LncRNA structure also contribute to its biological function (M.-S. Kim et al., 2014; Mercer and Mattick, 2013; Wilhelm et al., 2015) and recently a method known as RNA-selective 2′-hydroxyl acylation and primer extension (SHAPE-Seq) has been developed to measure the structural reactivity of

all four nucleotides (Cabili et al., 2011; Guttman et al., 2011; Marques et al., 2013; Spitale et al., 2012; Ørom and Shiekhattar, 2011). Very recently, SHAPE-Seq revealed two lncRNAs named roX1 and roX2 contain certain stem loop structures that recruit a set of proteins responsible for activation of X chromosome (Chu et al., 2015). Along with all these methods, direct gene knockout studies using CRISPR (Clustered Regularly Interspaced Short Palindromic Repeat) can reveal more detailed functions of lncRNA (Perez-Pinera et al., 2015). Though most of the methods mainly focused on X-chromosome inactivation, they can be applied in wide-range of biological processes or diseases to gain better understanding of lncRNA functions.

1.10. LncRNAs in diseases

Considering the tissue-specific expression and the wide range of roles that lncRNAs, they could play a major in gene regulation as well as be linked to various diseases (Batista and Chang, 2013). Supporting this notion, many functional evidences have implicated lncRNA in various human diseases in various tissues including heart, brain, kidney, and several endocrine tissues (Satpathy and Chang, 2015). Moreover, recent works have associated lncRNAs with regulating cancer pathways regulated in human (Huarte, 2015; Schmitt and Chang, 2016).

Interestingly, heart is the better-studied organ with respect to finding functions of lncRNAs (Devaux et al., 2015). It has been shown that lncRNAs regulate both the development and ageing of the cardiac cells (Devaux et al., 2015). LncRNAs, MIAT (Myocardial infarction associated transcript), MHRT (myosin heavy chain-associated RNA transcript), CHRF (Cardiac hypertrophy-related factor), CARL (cardiac apoptosis- related lncRNA), and SENCR (smooth muscle and endothelial cell enriched migration/differentiation-associated long non-coding RNA) regulate various heart-related functions by acting as decoys for miRNAs (Devaux et al., 2015). LncRNA FOXF1 adjacent to another noncoding developmental regulatory RNA known as FENDRR acts as a guide lncRNA and by forming complex with PRC2 and inhibit few keys genes involved in cardiac development (Devaux et al., 2015; Grote et al., 2013). Many other lncRNAs involved in cardiac identity, differentiation, heart development, heart failure, cardiac autophagy and myocardial infarction listed in detail by Lorenzen and colleagues in their very recent review (Lorenzen and Thum, 2016).

Genome-wide association studies (GWASs) and comparative transcriptomics studies have associated lncRNAs with neuronal disorders like schizophrenia (lncRNA GOMAFU or MIAT), autism spectrum disorder (Moesin pseudogene-1 anti-sense lncRNA, MSNP1AS) and various brain cancers (Briggs et al., 2015). Van de Vondervoot and colleagues (van de Vondervoort, 2013) listed several lines of published evidences that show lncRNAs regulation in brain-related diseases including: (i) Genomic deletions of several lncRNAs (ST7OT1-3, PTCHD1AS1-3) associate with AD; (ii). SOX2OT lncRNA modulates expression of SOX2 protein-coding gene and cause micropthalmia syndrome-3; (iii) NRON lncRNA regulates nuclear shuttling of NFAT, whose reduced activity leads to down syndrome features. 4. FRM4 lncRNA silenced in Fragile X syndrome patients; knockdown results in alterations in cell cycle regulation and increased apoptotic cell death.

LncRNAs have also been linked to the development and function of endocrine organs like pancreatic β cells, white and brown adipose tissue, and the misregulation of these lncRNAs can lead to hormonal disorders (Knoll et al., 2015). LncRNAs that are expressed in pancreatic islet cells (HI-LINC25, H19, KCNQ1OT1), muscle and adipose tissue (naPINK1) linked to diabetes mellitus (Knoll et al., 2015). Sra1 (Steroid receptor RNA activator 1) lncRNA regulates adipocytes during the development by co-activating a transcription factor PPARγ (Peroxisome proliferator-activated receptor) and it was the first lncRNA that discovered to be associated with the development of adipocytes. FIRRE (functional intergenic repeating RNA element) is another lncRNA that localizes across a 5 megabases domain near its site of transcription and is in close proximity to five distinct transchromosomal loci (four of which were previously described as having regulatory roles in adipogenesis) and helps in their coregulation (Sun et al., 2013). Many other lncRNAs have also been found to be associated with endocrine cancers like breast cancer and pancreatic cancer (Knoll et al., 2015). All the above findings suggest that the lncRNAs have critical roles in the development of endocrine organs.

The study of lncRNAs in kidney (also many other organs) is still in its infancy (Lorenzen and Thum, 2016). Up-regulation of lncRNAs including Xist and NEAT1 associated with membranous nephropathy, MALAT1 with diabetic nephropathy, TapSAKI (TrAnscript Predicting Survival in AKI) with acute kidney injury, HOTAIR, CADM (Cadmium induced) with renal cell carcinoma, *RP11-355P17.15-001* with acute rejection, and down regulation of Mest (Mesoderm specific transcript) and H19 linked to diabetes (Lorenzen and Thum, 2016).

Expression and regulation of several lncRNAs are linked with oncogenic functions and play a major role in cancer pathways (Huarte, 2015; Sahu et al., 2015; Schmitt and Chang, 2016): PCAT1, a breast-cancer associated lncRNA, interacts with PRC2 complex and silences gene expression in trans and also post-transcriptionally activates c-MYC and inhibits BRACA2 genes (Prensner et al., 2011). HOTAIR lncRNA, a tumor inducing lncRNA, is a classic example of epigenetic silencing, where it recruits a polycomb repressive complex (PRC2) and lysine-specific demethylase (LSD1) and silences HOXD gene loci. Increased expression of HOTAIR leads to PRC2 enrichment and ultimately represses several tumor HOX genes (Tsai et al., 2010). LncRNA ANRIL that is antisense to a protein-coding gene that codes an inhibitor of cyclin kinase 4b (INK4b). INK4b encodes three tumor suppressor genes and have been linked to various cancers. MALAT1 is another lncRNA that modulates the distribution of splicing factors nuclear paraspeckles and affects the phosphorylation state of SR RNA binding proteins eventually leading to tumor formation (Gutschner et al., 2013; Tripathi et al., 2010). SCHLAP1 lncRNA interacts with chromatin remodelers (SWIF/SNIF) and antagonizes its functions in prostate cancer (Prensner et al., 2013). For more detailed examples involved in cancer pathways pleaser refer to work of Huarte and Schmitt (Huarte, 2015; Schmitt and Chang, 2016).

In summary, it is evident that many new insights into the functions of lncRNAs and their implication with various diseases have recently emerged. It is expected that these associations will be further analyzed with advances in technological methods as well as the deeper understanding of lncRNAs function.

1.11. Challenges

Identifying lncRNAs and assigning functions in human and also in other species is very challenging. It seems that there is a noticeable difference in the total number of identified lncRNA genes across different species (Kapusta and Feschotte, 2014; Necsulea and Kaessmann, 2014). This could be debatable because of the following reasons: First, this might be because of the difference in the methods used by previous studies. For example, Guttman and colleagues used both chromatin marks and transcription in the same cell lines (Guttman et al., 2009) but others used only transcription to identify transcribed lncRNA genes (Cabili et al., 2011; Derrien et al., 2012; Iyer et al., 2015). Using chromatin marks drastically reduce the number of

lncRNA genes from thousands to hundreds (Cabili et al., 2011). Second, several studies identified lncRNAs in cell types or tissues (Cabili et al., 2011; Guttman et al., 2010; Rinn and Chang, 2012) and many others used whole organism at different stages of development (Kapusta et al., 2013; Necsulea et al., 2015; Ulitsky et al., 2011; Young et al., 2012). A small fraction of lncRNAs actually overlaps between whole-organism based and cell or tissue-based studies (Kapusta and Feschotte, 2014; Marques et al., 2013). Third, different *de novo* assemblers have been used to reconstruct the novel lncRNA transcripts and the differences between organisms could impact in the number of gene predictions (Li et al., 2014; Mouse ENCODE Consortium et al., 2012; Steijger et al., 2013). Fourth, sequencing depth of RNA-Seq can majorly influence the discovery of lncRNAs because of their lower expression levels (Mouse ENCODE Consortium et al., 2012; Rinn and Chang, 2012). Fifth, different methods have been used to assess coding potential using different set of features. For example, many studies predicted coding potential of lncRNAs based on their sequence features, conservation and overlapping with existing protein databases (Derrien et al., 2012; Iyer et al., 2015; Jia et al., 2010) but recently Ingolia and colleagues introduced ribosome footprints to assess more accurate coding potential (Derrien et al., 2012; Guttman et al., 2013; Ingolia et al., 2014; Iyer et al., 2015; Jia et al., 2010; Tarailo-Graovac and Chen, 2002). In addition, recent large-scale human proteome studies based on mass spectrometry revealed hundreds of novel lncRNAs that produce short peptides (Goodwin et al., 2016; Guttman et al., 2013; Ingolia et al., 2014; M.-S. Kim et al., 2014; Tarailo-Graovac and Chen, 2002; Wilhelm et al., 2015). All these could create discrepancies in the total number of putative lncRNAs genes across different species. Ultimately, assigning functions to these thousands of lncRNA is the major challenge in the non-coding RNA field. Though it is a difficult task, now with the advent of CRISPR (clustered regularly interspaced short palindromic repeats) system it is possible to knockout non-coding genes in mammals (DGT, 2015; Djebali et al., 2016; Goodwin et al., 2016; Ho et al., 2015; Kellis et al., 2014; M.-S. Kim et al., 2014; Perez-Pinera et al., 2015; The FANTOM Consortium, 2005; Wilhelm et al., 2015).

2. Repetitive elements

2.1. Definition

Repetitive DNA elements are sequences that occur several times in the genome in identical of similar fashion (Cabili et al., 2011; DGT, 2015; Djebali et al., 2016; Guttman et al., 2010; Ho et al., 2015; Iyer et al., 2015; Kellis et al., 2014; Perez-Pinera et al., 2015; The FANTOM Consortium, 2005; Treangen and Salzberg, 2011). Mammalian genomes are filled with repetitive sequences. For example, nearly half of the human and mouse genome is covered by repeats (Cabili et al., 2011; Guttman et al., 2010; Iyer et al., 2015; Kapranov et al., 2007; Kapusta et al., 2013; Treangen and Salzberg, 2011).

2.2. Classification

Repetitive elements are broadly classified as transposons, satellite repeats and simple repeats (Doolittle, 2013; Eddy, 2012; Elbarbary et al., 2016; Kapranov et al., 2007; Kapusta et al., 2013; Kellis et al., 2014; Lander et al., 2001; Palazzo, 2016; Treangen and Salzberg, 2011; van Bakel et al., 2011). Transposons or Transposable Elements (TEs) are DNA sequences that have the ability to be integrated elsewhere in a genome. TEs are further classified as retrotransposons and DNA transposons. Retrotransposons transpose via an RNA intermediate propagating via copy-and-paste amplification mechanism that has allowed them to accumulate in DNA, giving rise to the bulk of repeats in mammalian genomes. DNA tranpsons transpose directly without an RNA intermediate. The three major retrotransposon classes are long terminal repeat (LTR) retrotransposons, long interspersed elements (LINEs), and short interspersed elements (SINEs). Satellite repeats are further divided into microsatellites constitute a class of repetitive DNA comprising tandem repeats that are ≥2bp in length, minisatellites (10–60 bp in length), and satellites (up to 100 bp in length). Satellites are often associated with centromeric or pericentromeric regions of the genome. Simple sequence repeats, consisting of direct repetitions of relatively short k-mers such as (A)n, (CA)n or (CGG)n.

2.3. Discovery

The first TEs were discovered by Barbara McClintock in Maize more than a half-century ago as the genetic agents that are responsible for the sectors of altered pigmentation on mutant kernels (Doolittle, 2013; Eddy, 2012; Elbarbary et al., 2016; Kellis et al., 2014; Lander et al., 2001; McClintock, 1951; Palazzo, 2016; Treangen

and Salzberg, 2011; van Bakel et al., 2011). This discovery and the ensuing characterization of the genetic properties of TEs led to her being awarded a Nobel Prize in 1983, after TEs had been documented in the genomes of *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *Escherichia coli*, *Caenorhabditis elegans,* mouse and humans (Cabili et al., 2011; Chinwalla et al., 2002; Derrien et al., 2012; Djebali et al., 2016; Faulkner et al., 2009; Feschotte et al., 2002; Fort et al., 2014; Guttman et al., 2010; Iyer et al., 2015; Kapusta and Feschotte, 2014; Lander et al., 2001; McClintock, 1951; Mele et al., 2015). Now it is well established that nearly half of the mouse and human genomes are derived from repetitive elements (Chinwalla et al., 2002; Djebali et al., 2016; Feschotte et al., 2002; Iyer et al., 2015; Kapusta and Feschotte, 2014; Lander et al., 2001).

2.4. Sequence features and repeat families

LINEs are transcribed by RNA polymerase II (Pol II) and around thousand base pairs in length. Their Pol II promoter generates an mRNA-like capped and polyadenylated transcript. LINE-1 (L1), LINE-2 (L2), Chicken Repeat 1 (CR1) and RTE are different families of LINEs and in humans L1 transcript contains two open reading frames (ORFs): The first encodes RNA-binding protein and second encodes a protein with reverse transcriptase. These proteins recognize a specific sequence in the 3′ end of the LINE transcript that encodes them and create two staggered nicks at specific sequences in the genome. Finally by using the genomic sequence as a primer, these protein reverse transcribe the LINE RNA into cDNA that is simultaneously incorporated into the genome (Djebali et al., 2016; Elbarbary et al., 2016; Iyer et al., 2015; Kapusta and Feschotte, 2014; Treangen and Salzberg, 2011).

SINEs are transcribed by RNA polymerase III (Pol III) and range from 85 to 500 bp. Their Pol III promoter generates a 5’ head, a body, and 3’ tail. Arthrobacter luteus (Alu) elements and mammalian-wide interspersed repeats (Mir) are different families of SINEs and they do not encode any protein. In most cases, LINE-encoded proteins recognize SINE RNAs with 3′ sequences that are similar to the 3′ sequence of the LINE RNA from which these proteins were synthesized; subsequently, they generate and integrate a cDNA copy of the SINE RNA into the genome (Elbarbary et al., 2016; Faulkner et al., 2009; Treangen and Salzberg, 2011).

LTR retroposons are characterized by a region of several hundred base pairs called the long terminal repeat, that appears at each end. Some autonomous elements are

cousins of retroviruses (*e.g.,* HIV) but are unable to survive outside of the cell, and are called endogenous retroviruses (ERV1, ERVL, ERVK). And another family called MaLR (mammalian LTR) elements, which arose before the mammalian radiation, seems to be non-autonomous repeats that move via proteins from endogenous retroviruses (Elbarbary et al., 2016; Faulkner et al., 2009; Smit, 1996).

DNA transposons are full-length autonomous elements and encode a protein, called transposase, by which an element can be removed from one position and inserted at another. DNA transposons typically have short inverted repeats at each end. Satellite repeats range from 5 bp to 170 bp in length and mostly centered towards centromere and telomeres of chromosomes (Faulkner et al., 2009; Kapusta et al., 2013; Lu et al., 2014; Smit, 1996; Ugarkovic, 2005; Ugarkovic and Plohl, 2002)**.** Estimates from the human genome sequencing project indicate that simple and low complexity repeats make up ~3% of the sequenced human genome (Belgard et al., 2011; Guttman et al., 2010; Kapusta et al., 2013; Lander et al., 2001; Lu et al., 2014; Luo et al., 2013; Lv et al., 2013; Marques et al., 2013; Morán et al., 2012; Ramos et al., 2013; Ugarkovic, 2005; Ugarkovic and Plohl, 2002). However, the true number and length of such repeats may be much higher than the current database suggests since these sequences are prone to deletion when propagated in bacteria and yeast (Belgard et al., 2011; Guttman et al., 2010; Lander et al., 2001; Luo et al., 2013; Lv et al., 2013; Marques et al., 2013; Morán et al., 2012; Ramos et al., 2013; Rinn and Chang, 2012; Usdin, 2008). Additionally, this repeats are often unstable or hypervariable in mammals as well. Thus, genomes are polymorphic with respect to many of these repeats, with some individuals or families having some tandem repeat tracts that are significantly longer than those seen in the general population (Mercer et al., 2009; Rinn and Chang, 2012; Usdin, 2008).

2.5. Expression

Expression of repetitive elements is usually very low and also very tricky to estimate. Very few studies analyzed the global expression of repetitive elements in human using various methods. For example, Faulkner and Djebali and colleagues used CAGE tags in 8 to 13 human cell lines (Djebali et al., 2016; Faulkner et al., 2009; Mercer et al., 2009; Morris and Mattick, 2014; Usdin, 2008), Mele and colleagues used poly(A)+ RNA-Seq reads in 10 adult human tissues (Banfai et al., 2012; Djebali et al., 2016; Faulkner et al., 2009; M.-S. Kim et al., 2014; Mele et al., 2015; Morris and Mattick, 2014; Wilhelm et al., 2015), Criscone and colleagues used RNA-Seq

reads in normal and cancer lines (Banfai et al., 2012; Criscione et al., 2014; Devaux et al., 2015; M.-S. Kim et al., 2014; Mattick and Rinn, 2015; Mele et al., 2015; Wilhelm et al., 2015) and finally, Goke and colleagues for the first time used single-cell RNA reads across different embryonic tissues (Criscione et al., 2014; Devaux et al., 2015; Göke et al., 2015; Mattick and Rinn, 2015; Morris and Mattick, 2014). All these studies shown that expression of the repetitive elements are positively correlated with near-by protein-coding genes.

2.6. Mechanisms of repetitive elements

Mammalian genomes are subjected to mainly two types of changes. One is direct changes to the DNA sequence and second, changes at epigenetic level (Elbarbary et al., 2016; Göke et al., 2015; Martin and Zhong Wang, 2011; Morris and Mattick, 2014; Slotkin and Martienssen, 2007). Changes to the DNA sequence can result from the errors made during replication or repair or from the insertion of TEs. Depending on its type of insertion, TEs can either disrupt gene expression or create an advantageous modification to gene expression or be of no immediate consequence. Changes can be genetic, epigenetic, or both (Elbarbary et al., 2016; Martin and Zhong Wang, 2011; Slotkin and Martienssen, 2007; Zhong Wang et al., 2009). In this way, TEs restructure genomes and act as evolutionary drivers (Bourque, 2009; Elbarbary et al., 2016; Guttman et al., 2010; Pertea et al., 2015; Slotkin and Martienssen, 2007; Trapnell et al., 2012; Zhong Wang et al., 2009).

Repetitive elements regulate epigenetic events. For example, LINEs and SINEs have high GC content, making them hotspots for DNA methylation, which is used by cells to repress transcription (Bourque, 2009; Elbarbary et al., 2016; Guttman et al., 2010; 2009; Ichiyanagi, 2013; Pertea et al., 2015; Trapnell et al., 2012). The methylation of LINE and SINE in enriched CpG islands has the potential to silence the expression of nearby genes (Bogu et al., 2016; Estecio et al., 2012; Guttman et al., 2009; Ichiyanagi, 2013; Mele et al., 2015). SINEs, and Alu elements in particular, can function as transcriptional enhancers, as shown by two members of the ancient SINE family Amniota SINE1 (AmnSINE1), which act as enhancers for the genes encoding fibroblast growth factor 8 (Fgf8) and special AT-rich sequence-binding protein 2 (Satb2) in the developing brain (**Figure 1**) (Bogu et al., 2016; Derrien et al., 2012; Estecio et al., 2012; Mele et al., 2015; Tashiro et al., 2011). Retrotransposons located immediately upstream of protein-coding genes may function as promoters because putative binding sites for many transcription factors have been identified in

SINEs (**Figure 1**) (Derrien et al., 2012; Faulkner et al., 2009; Kunarso et al., 2010; Tashiro et al., 2011). LINEs and SINEs can demarcate the boundary between heterochromatin and euchromatin. For example, one mouse B2 element functions as a boundary element to prevent cis-residing heterochromatin from silencing developmental expression of the five genes located in the mouse growth hormone locus (**Figure 1**) (Derrien et al., 2012; Djebali et al., 2016; Faulkner et al., 2009; Kunarso et al., 2010; Lunyak and Atallah, 2011).

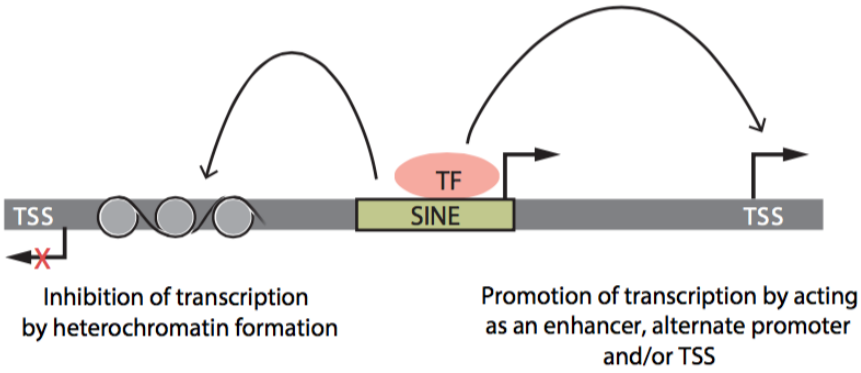


Figure 1. SINEs or LINEs can promote or inhibit the transcription of nearby genes (TSS, transcription start site; TF, transcription factor) (source: (Djebali et al., 2016; Elbarbary et al., 2016; Knoll et al., 2015; Lunyak and Atallah, 2011)***).***

Repetitive elements regulate transcriptional events as well. For example, LINEs and SINEs regulation of transcription not only mediated as DNA elements that recruit transcription factors but also via the RNAs that they encode. Generally, SINEs are transcriptionally repressed in somatic tissues; however, in response to stressors such as heat shock (Elbarbary et al., 2016; Knoll et al., 2015; Mariner et al., 2008), SINE Pol III promoters are activated and SINE RNAs are up-regulated. Stress mediated up-regulation of human Alu and mouse B2 RNAs inhibits the transcription of many genes (**Figure 2**) (Allen et al., 2004; Knoll et al., 2015; Mariner et al., 2008). LINEs and SINEs can also introduce a new transcription start site (TSS). In fact, 6 to 30% of human and mouse 5'-capped transcripts use repetitive sequence-associated TSSs (Allen et al., 2004; Faulkner et al., 2009; Knoll et al., 2015; Mercer et al., 2008). Surprisingly, 75-83% of lncRNAs contain transposable elements (TEs), a considerably higher percentage compared with protein-coding genes (Cabili et al., 2011; Derrien et al., 2012; Faulkner et al., 2009; Kapusta et al., 2013; Mercer et al., 2008). Nineteen percent of lncRNAs consist of more than 50% TE sequence (Cabili

et al., 2011; Derrien et al., 2012; Kapusta et al., 2013; Kapusta and Feschotte, 2014; Quinn and Chang, 2016; Rinn and Chang, 2012), suggesting that exaptation of TEs and evolution of lncRNAs are closely related.

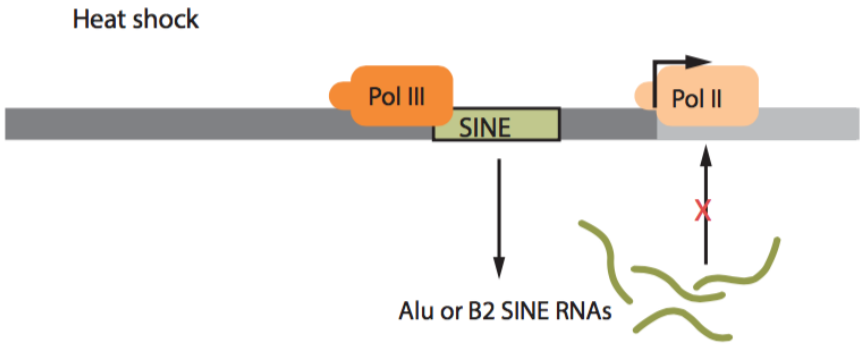


Figure 2. Upon heat shock, RNA polymerase II transcribe RNAs from SINE elements and their increased expression inhibits Pol II, and inhibits the transcription of protein-coding genes (source: (Elbarbary et al., 2016; Guttman et al., 2009; Kapusta and Feschotte, 2014; Quinn and Chang, 2016; Rinn and Chang, 2012)***).***

Repetitive elements regulate co-transcriptional events like pre-mRNA splicing. Alu elements contain cryptic splice sites and usually need few mutations to become functional splice sites and promote exonization (Elbarbary et al., 2016; Guttman et al., 2009; Lev-Maor et al., 2007; Necsulea and Kaessmann, 2014). This leads to inclusion of an intronic sequence within the resulting spliced mRNA and it is estimated that 5% of alternatively spliced exons in humans derive from Alu sequences and that most Alu-containing exons are alternatively spliced (Lev-Maor et al., 2007; Necsulea and Kaessmann, 2014; Sorek et al., 2002).

Repetitive elements also regulate post-transcriptional processes like mRNA stability, localization, translation and exonization of Alu repeats. (1) *mRNA stability:* SINEs have the potential to act in cis and/or in trans to influence mRNA turnover. The poly(T) sequence that exists in antisense Alu elements is the source of ~40% of identified 3' UTR AU-rich elements, which regulate mRNA half-life through the competitive binding of proteins that stabilize or destabilize the transcript (An et al., 2004; Kutter et al., 2012; Necsulea and Kaessmann, 2014; Sorek et al., 2002). (2) *mRNA localization and translation:* Not all mRNAs are efficiently exported from the nucleus to the cytoplasm for translation. A protein known called STAU1 binds to the Alu in the 3' UTR of mRNAs and precludes the binding of another protein known as

p54nrb (a protein component of nuclear paraspeckles) and it eventually leads to their nuclear export (An et al., 2004; Elbarbary et al., 2013; Kutter et al., 2012; Washietl et al., 2014). A subset of 3′UTR Alus mediates translational repression by binding and activating dsRNA-dependent protein kinase (PKR) (Elbarbary et al., 2013; Fitzpatrick and Huang, 2014; Guttman and Rinn, 2012; Rinn and Chang, 2012; Kevin C Wang and Chang, 2011; Washietl et al., 2014). (3) *Exonization of Alu repeats:* Adenosine-to inosine (A-to-I) RNA editing is a tissue-specific posttranscriptional process where adenosine residues located in double stranded RNAs are deaminated to inosines by adenosine deaminase (ADAR) proteins (Athanasiadis et al., 2004; Fitzpatrick and Huang, 2014; Guttman and Rinn, 2012; Rinn and Chang, 2012; Tian et al., 2010; Kevin C Wang and Chang, 2011). Alu repeats acts as binding sites for ADARs and more than 90% of A-to-I editing occurs within Alu elements (Athanasiadis et al., 2004; Rinn et al., 2007; Tian et al., 2010; Kevin C Wang et al., 2011).

2.7. Repetitive elements in diseases

The study of repetitive elements associated with diseases is still in its infancy. Most of the previous studies often shown repetitive element genomic insertions causing diseases. Around 100 diseases have been shown to be associated with germline insertions of retrotransposons (Athanasiadis et al., 2004; Belancio et al., 2008; 2010; Hancks and Kazazian, 2012; T.-K. Kim et al., 2010; Rinn et al., 2007; Kevin C Wang et al., 2011). Some of them are hereditary nonpolyposis colorectal cancer, prostate cancer, alport syndrome and breast cancer. LINE or SINE insertions can also affect mRNA stability leading to reduced protein production or altered spatio-temporal expression of protein-coding genes. This phenomenon has been associated with X-linked dilated cardiomyopathy, hemophilia A, hereditary-nonpolyposis colorectal cancer and hyper-immunoglobulin syndrome (Belancio et al., 2010; 2008; Hancks and Kazazian, 2012; Kaer and Speek, 2013; T.-K. Kim et al., 2010; Qrom et al., 2010).

Repetitive elements can disrupt protein-coding sequences and cause aberrant protein production or NMD (Non-sense mRNA Decay) eventually leading to diseases including hemophilia B, breast cancer, colon cancer and neurofibromatosis type 1 (Hancks and Kazazian, 2012; Kaer and Speek, 2013; Qrom et al., 2010; Kevin C Wang and Chang, 2011). Also it has been shown that altered DNA methylation increases expression of LINE and SINE RNA at early stages of various cancers (Belancio et al., 2010; Hancks and Kazazian, 2012; Qrom et al., 2010; Kevin C Wang

and Chang, 2011; Ørom and Shiekhattar, 2011). Further, altered pre-mRNA splicing caused by retrotransposon insertions leading aberrant protein production linked with diseases like fukuyama-type congenital muscular dystrophy, neurofibromatosis type 1, hemophilia A, breast cancer, coffin-lowry syndrome (Belancio et al., 2010; Hancks and Kazazian, 2012; Kaer and Speek, 2013; Meredith et al., 2016; Qrom et al., 2010; Ørom and Shiekhattar, 2011). X-chromosome associated diseases like retinitis pigmentosa and dilated cardiomyopathy have been shown to be caused by retrotransposon mediated premature transcription termination (Guttman et al., 2011; Hancks and Kazazian, 2012; Kaer and Speek, 2013; Meredith et al., 2016). LINEs also participate in X- chromosome inactivation (XCI) via one of two mechanisms: Transcriptionally silent L1 elements contribute to the formation of a silent nuclear compartment during XCI, whereas L1 RNAs that derive from young LINE elements (which are enriched in the X chromosome) participate in inactivating X-chromosome loci that would otherwise escape XCI (Chow et al., 2010; Guttman et al., 2011; Hancks and Kazazian, 2012; Kevin C Wang and Chang, 2011). Very recently retrotransposons have been shown to affect sites of A-to-I editing causing amyotrophic lateral sclerosis (ALS), astrocytoma, metastatic melanoma, aicardi-Goutières syndrome and hepatocellular carcinoma.

Interestingly, increase in copy number of simple repeats by dynamic mutations have been associated with neurodegenerative diseases like mental retardation, kennedy disease, huntington disease, haw-river syndrome, machado joseph disease and nearly thirty hereditary disorders (Chow et al., 2010; Mirkin, 2007; Sutherland and Richards, 1995; Kevin C Wang and Chang, 2011). The repeat expansion diseases are a group of human genetic disorders caused by long and highly polymorphic tandem repeats (Martianov et al., 2007; Mirkin, 2007; Sutherland and Richards, 1995; Usdin, 2008; Kevin C Wang and Chang, 2011). The repeat expansion diseases can be divided into two categories: those like huntington disease or spinobulbar muscular atrophy, where the repeat is located in an exon, and those like myotonic dystrophy or fragile X syndrome, where the repeat is outside of the open reading frame (Martianov et al., 2007; Usdin, 2008; Kevin C Wang and Chang, 2011).

In summary, nearly half of the human genome is derived from repetitive elements and surprisingly very few studies explored their role in human diseases in the past. In the future, with advances in technological methods, it will be possible to understand

expression and functions of these repetitive elements and especially their role in human diseases.

2.8. Challenges

Standard benchmarking and expression quantification are the main computational challenges in the field of repetitive elements. A recent study re-examined the whole human genome using an algorithm that instead relies on relatedness within entire groups of evolutionarily diverged repeats (de Koning et al., 2011; Hung et al., 2011; Usdin, 2008; Kevin C Wang and Chang, 2011). This led them to an increased estimate of 66–69% for the proportion of repeat-derived sequence in the human genome, after correcting for false positives. These results imply that repetitive DNA may have played a larger part in human evolution than was previously assumed. However, there are no standard benchmarks of repetitive elements especially transposable elements annotation and there is a necessity to create standard benchmarks (de Koning et al., 2011; Hoen et al., 2016; Hung et al., 2011; Kevin C Wang and Chang, 2011). Estimating the expression of repetitive elements is also challenging. Challenges include (but are not limited to: short read length. reads that map to more than one genomic location (multi-mapped reads), low sequencing depth, inaccurate alignment methods raise problems in expression quantification (Criscione et al., 2014; Day et al., 2010; Hoen et al., 2016; Treangen and Salzberg, 2011; Kevin C Wang and Chang, 2011). Previously, the inclusion of multi-mapped RNA-Seq reads has questioned the conclusions of a few key studies (Criscione et al., 2014; Day et al., 2010; Hung et al., 2011; Marinov et al., 2015; Royo et al., 2016; Samans et al., 2014; Sienski et al., 2012; Treangen and Salzberg, 2011; Kevin C Wang and Chang, 2011; Yap et al., 2010). Recently, there are few novel methods designed to deal with the multimapping problems, but they are not fast enough to quantify large number of RNA-Seq datasets (Criscione et al., 2014; Day et al., 2010; Hung et al., 2011; Marinov et al., 2015; Rinn et al., 2007; Royo et al., 2016; Samans et al., 2014; Sienski et al., 2012; Tsai et al., 2010; Yap et al., 2010).

Other than benchmarking and quantification, there are other important challenges (Criscione et al., 2014; Day et al., 2010; Göke and Ng, 2016; Mercer et al., 2009; Rinn et al., 2007; Tsai et al., 2010):

1. Among the many repetitive elements that are transcribed or which show regulatory activity, which elements are biologically relevant?

2. What is the function of individual elements and transcripts generated from retrotransposons?
3. Does co-expression or tissue-specific expression of families or subfamilies of repetitive elements indicate common or specific functions?
4. What are the pathways and interaction partners for repeat-derived RNAs?
5. Which repetitive elements are translated, and which repeat-derived RNAs are noncoding?
6. What are the precise sequences of repeat-derived RNAs? Will long read sequencing technology help overcome current limitations?

Objective - 1

Several studies in the past decade have shown tens of thousands of novel lncRNAs across various cell or tissue types in mammals. However, identifying lncRNAs that are most likely to be functional has been a challenging task.

Interestingly, first, Guttman and colleagues revealed functional lncRNAs by identifying K4-K36 chromatin domains that lay outside known protein-coding gene loci (Guttman et al., 2009). These K4-K36 domains marked by H3K4me3 at their promoter and H3K36me3 along the length of the transcribed region. These K4-K36 domains are conserved from yeast to humans and demarcate known Pol II transcribed genes, including protein-coding and miRNA genes. Though this study successfully identified functional lncRNAs it was only in limited to three mouse cell types. Second, Marques and colleagues overlapped lncRNA promoters with enhancer (enriched with H3K4me1 and depleted with H3K4me3) and promoter (enriched with H3K4me3 and depleted with H3K4me1) chromatin states and revealed two distinct classes of lncRNA known as elncRNA (enhancer-associated lncRNA) and plncRNA (promoter-associated lncRNA) (Marques et al., 2013). However, this study also limited to only one mouse cell type. Not only both above mentioned studies limited to few cell types but they also failed to show whether the chromatin status of lncRNAs remains the same in other cell or tissue types.

Functional cataloging or classifying lncRNAs across many mouse tissues have not been explored yet. In addition, large-scale *de novo* discovery of lncRNAs across many mouse tissues is also pending. The main objective (*Objective 1, First manuscript*) here is to address the above challenges by following steps.

1. Assembling transcriptomes from eight different mouse tissues and one cell type using RNA-seq data
2. Classifying them by overlapping with chromatin states defined by various histone marks.
3. Checking how the chromatin state of lncRNA in one tissue switches to another in other tissues.
4. Showing the novel lncRNAs identified by both strategies as functional using experimental methods.

FIRST MANUSCRIPT

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.

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

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).

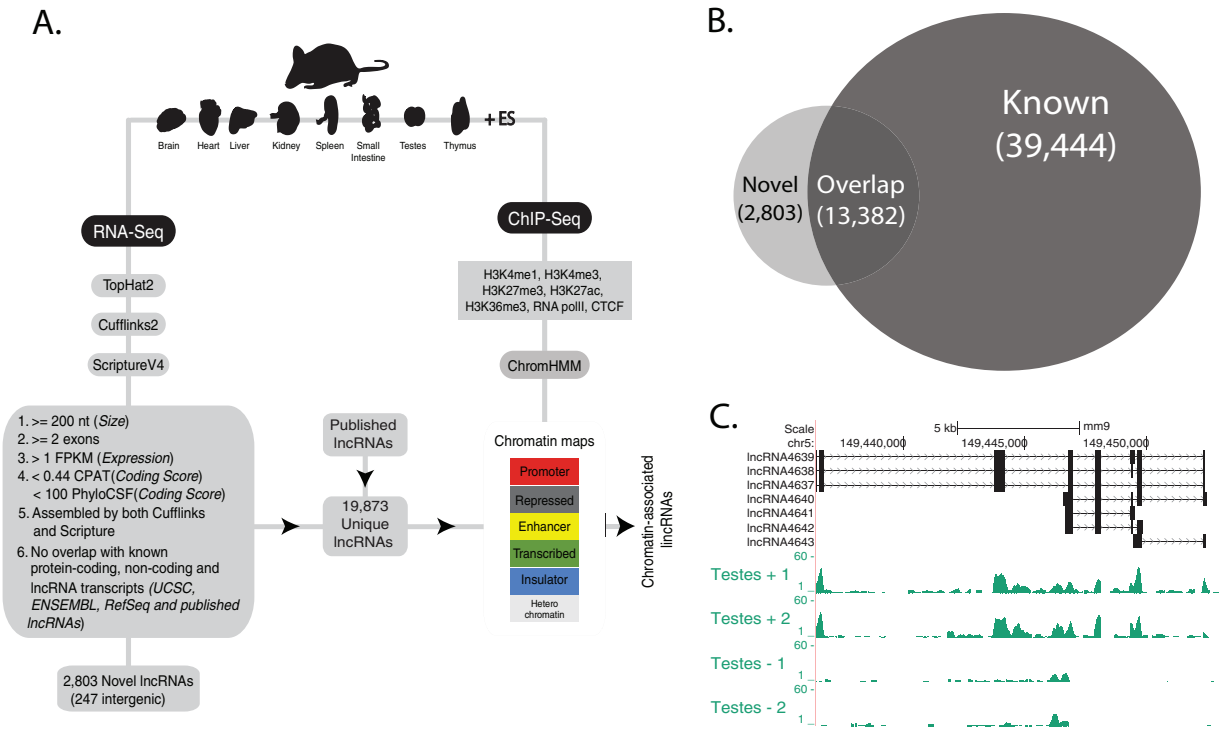


FIG 1 Overview of the lncRNA discovery and chromatin state map computational pipeline. (A) Overview of the lncRNA discovery and chromatin state map-based classification pipeline that was employed using both RNA-seq and ChIP-seq data from 8 tissues and one primary cell line (ES) in mouse. RNA-seq reads 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

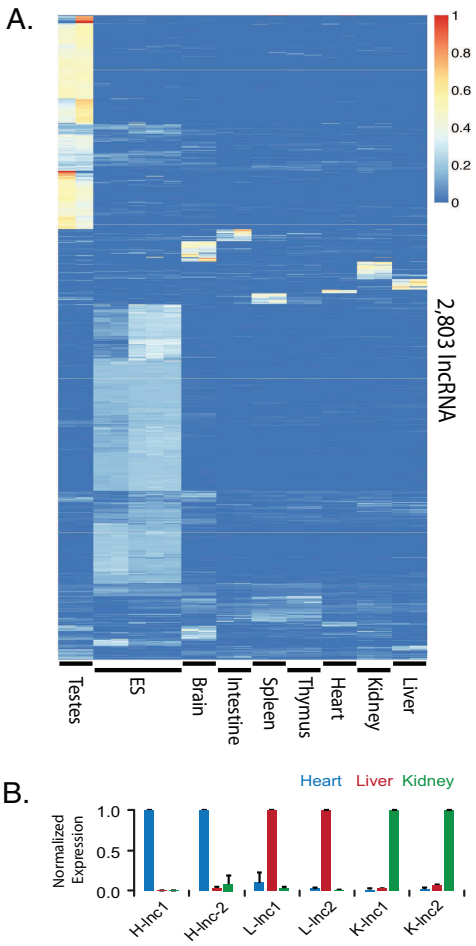


FIG 2 Tissue- and cell-specific expression of lncRNAs. (A) Heat map representing normalized FPKM expression values of the 2,803 lncRNAs (rows) across eight tissues and a primary cell line (columns). The rows and columns were ordered based on *k* means clustering. The color intensity represents the fractional 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

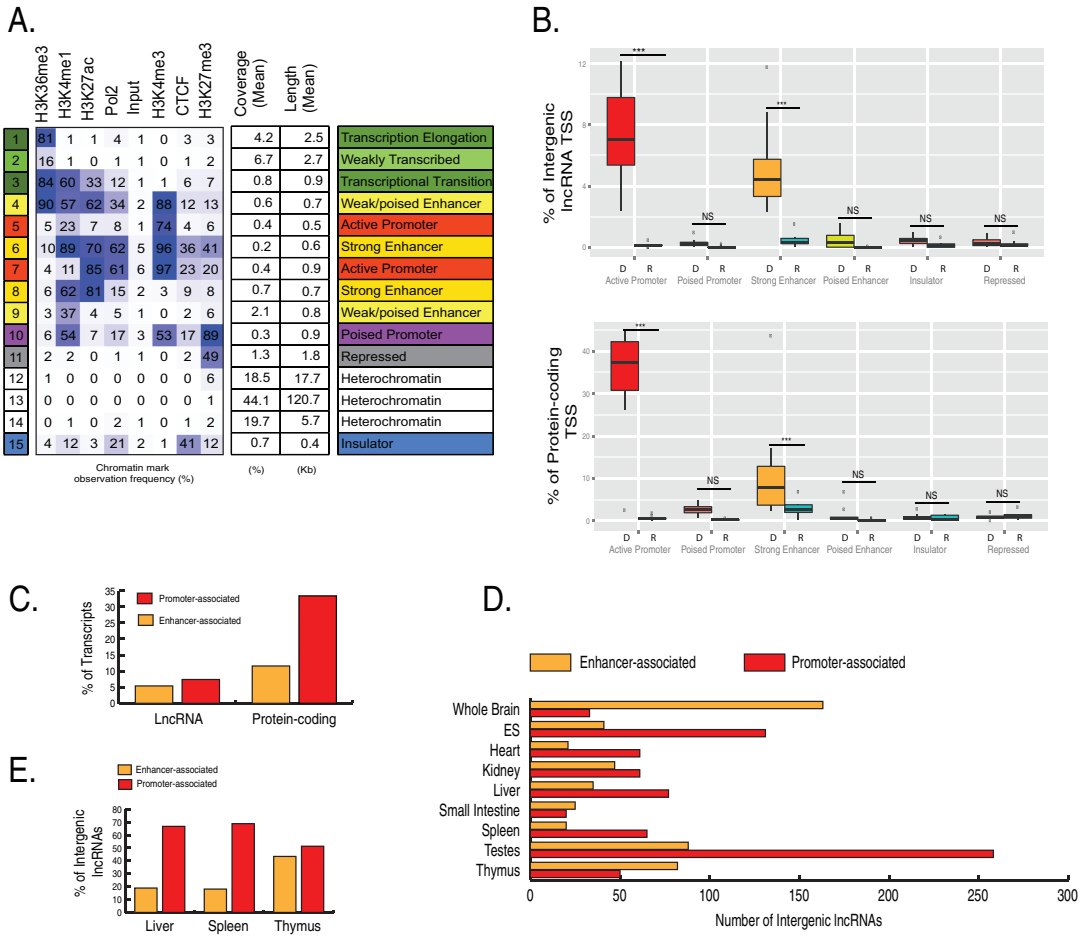


FIG 3 Discovery of chromatin state maps and their association with lincRNAs. (A) Emission parameters learned *de novo* with ChromHMM on the basis of combinations recurring in chromatin. Each point in the table denotes the frequency with which a given mark is found at genomic positions corresponding to 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

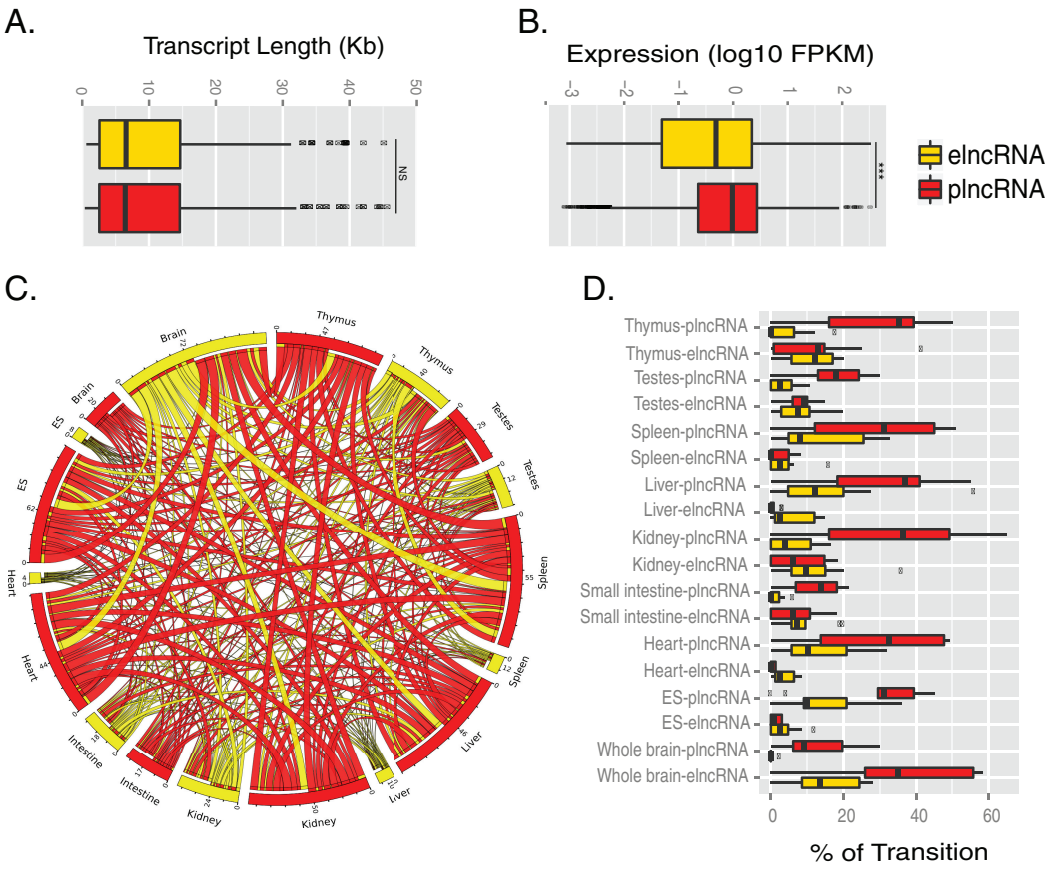


FIG 4 Transcript length, expression, and transition of chromatin-associated lncRNAs in mouse. (A) Transcript lengths of elncRNAs (median = 6,565 nt) and plncRNAs (median = 6,450 nt) across eight tissues and a cell line, showing no difference in length (Mann-Whitney test; NS, not significant; $P = 0.9848$). (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

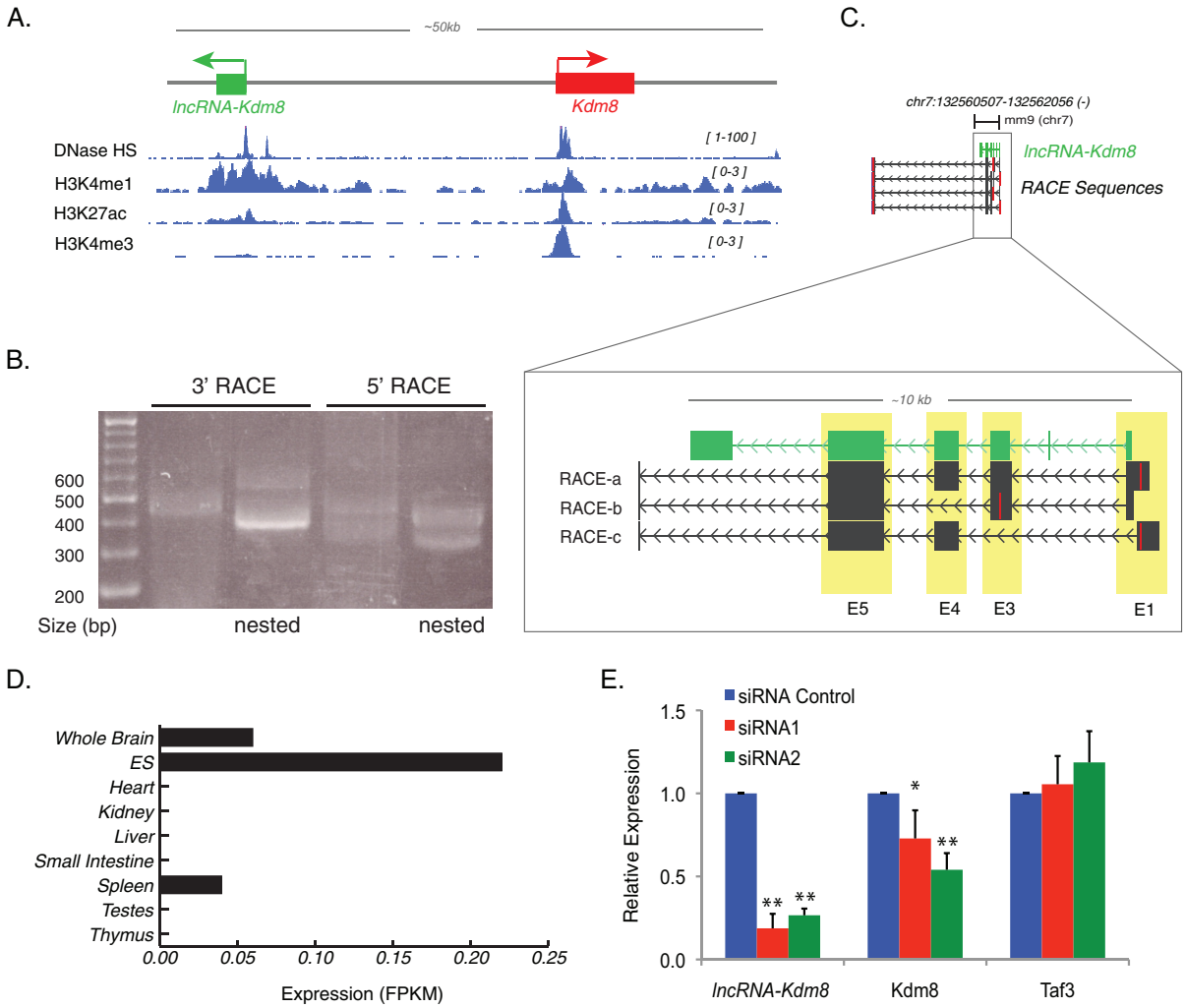


FIG 5 An enhancer-associated lncRNA, lncRNA-*Kdm8*, regulates the expression of a neighboring protein-coding gene, *Kdm8*. (A) The lncRNA-*Kdm8* locus promoter overlaps an enhancer chromatin state and occurs within 20 kb of the TSS of a protein-coding gene, *Kdm8* (e.g., it is an enhancer-associated 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.

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-**

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-

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.

Supplementary Material to:

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

Table S1. Number of RNA-seq datasets

Tissue/Cell line	Replicate	Type	Dataset	Number of Unique Mapped Reads
Brain	1	Paired End	ENCODE (mm9)	280,831,168
Brain	2	Paired End	ENCODE (mm9)	281,276,946
ES	1	Paired End	Burge lab (mm9)	50,490,483
ES	2	Paired End	Burge lab (mm9)	47,980,155
Heart	1	Paired End	ENCODE (mm9)	130,268,653
Heart	2	Paired End	ENCODE (mm9)	142,341,055
Small Intestine	1	Paired End	ENCODE (mm9)	114,040,101
Small Intestine	2	Paired End	ENCODE (mm9)	113,227,146
Kidney	1	Paired End	ENCODE (mm9)	157,138,280
Kidney	2	Paired End	ENCODE (mm9)	157,138,292
Liver	1	Paired End	ENCODE (mm9)	136,088,014
Liver	2	Paired End	ENCODE (mm9)	143,106,885
Spleen	1	Paired End	ENCODE (mm9)	125,654,511
Spleen	2	Paired End	ENCODE (mm9)	132,310,236
Testes	1	Paired End	ENCODE (mm9)	129,868,370
Testes	2	Paired End	ENCODE (mm9)	137,651,339
Thymus	1	Paired End	ENCODE (mm9)	131,558,427
Thymus	2	Paired End	ENCODE (mm9)	172,523,153

Table S3. Number of chromatin states

Tissue/ Cell line	Active	Polycomb Repressed	Poised	Strong Enhancers	Poised Enhancers	Insulators	Transcribed	Hetero-chromatin
Brain	8,570	12,114	10,856	179,734	2,111	38,885	129,391	144,930
Heart	26,488	20,433	6,216	66,328	24,854	16,650	232,079	90,208
Kidney	24,754	16,708	5,450	112,396	2,712	19,010	121,789	90,534
Liver	25,631	19,749	4,325	74,243	1,903	17,704	114,385	81,515
Spleen	26,578	14,391	5,421	38,546	2,047	8,176	112,847	67,142
Intestine	24,943	11,469	5,690	122,451	2,578	78,888	130,325	145,682
Thymus	14,557	7,164	4,678	98,168	1,307	94,487	106,640	142,589
Testes	57,236	8,516	10,856	104,689	12,556	78,651	145,425	156,962
ES	37,764	43,413	9,831	73,308	4,238	14,086	117,527	99,049

Table S7. URL for data retrieval or access

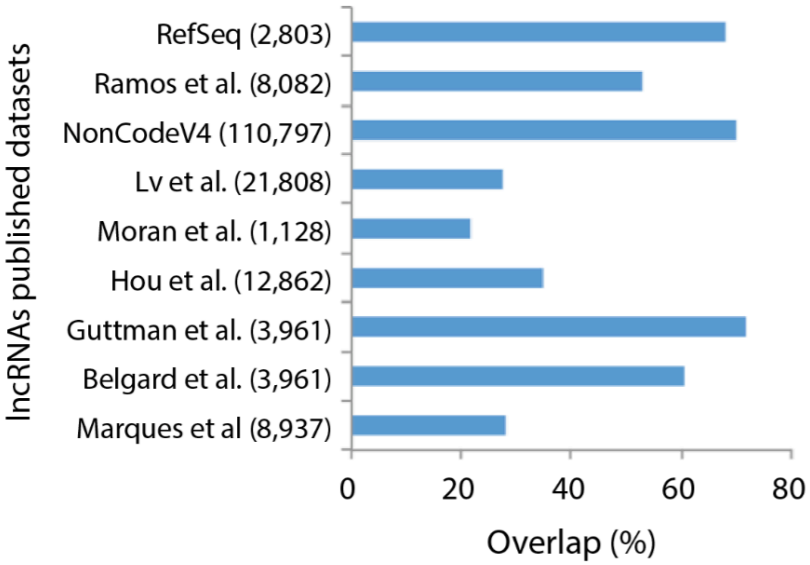
ENCODE RNA-seq	http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeCshlLongRnaSeq/
ENCODE ChIP-seq	http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeLicrHistone/
snoRNA, tRNAs, miRNAs and pseudogenes	UCSC table browser
iGenome	http://cufflinks.cbcb.umd.edu/igenomes.html
CAGE data	http://fantom.gsc.riken.jp/5/data/
PhastCons scores	http://hgdownload.cse.ucsc.edu/goldenPath/mm9/phastCons30way/
DNase I	http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeUwDnase/
18 mammalian genomes	http://hgdownload.soe.ucsc.edu/goldenPath/mm9/multiz30way/

Table S8. Non-default parameters used for the software employed in the study

Mapping
TOPHAT: TopHat-2.0.9 was used to map RNA-seq reads against mouse reference genome (mm9) using parameters –max-multihits (-g) as 1 (to filter multi-mapped reads) and –G (which uses a known gene model annotation called iGenome [a pool of UCSC, NCBI and Ensembl genes] to improve mapping).
Quantification / *De novo* assembly
CUFFLINKS: Cufflinks-2.1.1 was used to assemble transcriptome using uniquely mapped reads with parameters –u/--multi-read-correct (if any multi-mapped reads pass through TopHat by mistake, this parameter accurately weight reads mapping to multiple locations), -M/--mask-file (which masks reads mapped to snoRNA, tRNAs, miRNAs, and pseudogenes) and –b/--frag-bias-correct (which can significantly improve accuracy of transcript abundance estimates). Cuffmerge was used against high-confidence *de novo* transcripts to generate a single transcript annotation file using parameters –min-isoform-fraction 0.1 and –ref-fasta (which filters out transcripts with very low abundance, possibly artifacts).
***De novo* assembly**
SCRIPTURE: Scripture-v4 was also used to assemble transcripts using uniquely mapped reads with parameters including –task reconstruct, -coverage 0.2 and -strand second on each chromosome and pooled them into one. Scripture-Scorer-v4 was also used to quantify expression of transcripts.
Coding Potential
PHYLOCSF
PhyloCSF --bls -f3 --orf=StopStop3 --allScores 18mammals lncRNA.fa
De novo motif analysis
HOMER
findMotifsGenome.pl lncRNA.1kb.promters.bed mm9 output_directory -size 200 -len 8 parameters

Table S9. Primers of experimentally validated lncRNAs

Genomic_corrdinates	Name
chr13:103119662:103121476:+	H-lnc1
chr4:83661876:83662681:+	H-lnc2
chr19:56337904:56341429:-	L-lnc1
chr8:116652366:116655520:-	L-lnc2
chr3:69411867:69414713:+	K-lnc1
chr14:62438986-62440919:+	K-lnc2
Primer sequences for RT-PCR	
Heart-specific lncRNA-1 forward	CATGACGAGCAAGCCAGTAA
Heart-specific lncRNA-1 reverse	CAGCCTTGAAGGAAGTCAGG
Heart-specific lncRNA-2 forward	AGCTCAGCAGTCAGGCTCTC
Heart-specific lncRNA-2 reverse	CTACTTCCAGGCTCCCAGAG
Liver-specific lncRNA-1 forward	CCCACTGAAGGGAACTGGTA
Liver-specific lncRNA-1 reverse	GGGATGACTATGCAGCCACT
Liver-specific lncRNA-2 forward	CATCATGGCAAGAAACATGG
Liver-specific lncRNA-2 reverse	GCCTGCAGAAGACAGTTTCC
Kidney-specific lncRNA-1 forward	AATGGGATCAACTCCCTTCC
Kidney-specific lncRNA-1 reverse	CCATGTTCTCCTGGTGTCCT
Kidney-specific lncRNA-2 forward	GCTCGCTATGTGGTGAGGTT
Kidney-specific lncRNA-2 reverse	GGTCTAGCGAAGGAAACGAA
Kidney-specific lncRNA-3 forward	GGGGACTAGTTTGGGGACAT
Kidney-specific lncRNA-3 reverse	TACCAAACCACCCGCTGTAT
ES-specific lincRNA forward	AACCACATAACTGCCCTGGT
ES-specific lincRNA reverse	AGAACCCCTACCCAAGAGGA
Jmjd5 forward	CCGATTGTATTCGCCGCAAG
Jmjd5 reverse	TTCCACATCAACCTGGCTGG
Primer sequences for ChIP qPCR	
Heart-specific lncRNA forward	ACGACTGCTCTCCTCATGGT
Heart-specific lncRNA reverse	TCCAACCCCACTTACTCGTC
Actb forward	GGATCACTCAGAACGGACACC
Actb reverse	GGCTCATCAAATGCCCACA
ES-specific lincRNA forward	AAACGGCCTGAGCAAGATAA
ES-specific lincRNA reverse	GCCATGGTGGACTCATATCC
K4me1 control forward	AAACGGCCTGAGCAAGATAA
K4me1 control reverse	GCCATGGTGGACTCATATCC
siRNA duplexes	
siRNA_elncRNA	CCAAGAGGACCCUGAGGCUUAUAUU AAUAUAAGCCUCAGGGUCCUCUUGG
siRNA_elncRNA	CACCAGUGUGUGCAUUAGCUGGUAU AUACCAGCUAAUGCACACACUGGUG
siRNA Control	CAUCCUUUCCGCGACUACACGACUU AAGUCGUGUAGUCGCGGAAAGGAUG
RACE Primers	
GSP1 (3′ RACE)	GATTACGCCAAGCTTACTTGTCATCTCACTGCTCCCTTCCATGTCTC
GSP2 (5′ RACE)	GATTACGCCAAGCTTAGCTCGAATCCCTCATCTGCTGGACAAATGA TTGN
NGSP1 (3′ RACE)	GATTACGCCAAGCTTAGAGGCTCTGTAAGCCCACACAGCATGTTGG N
NGSP2 (5′ RACE)	GATTACGCCAAGCTTAGTGTCCTCTGTGAAGAGGGAAGTCGGTCTG
lncRNA-*Kdm8*	chr7 132560507 132562056 lncRNA-*Kdm8* 0 - 132560507 132562056 0,0,0 6 149,194,84,67,3,18, 0,484,857,1054,1259,1531,



Supplemental Figure 1. Overlap of *de novo* assembled transcripts with lncRNA transcripts. Percentage of known lncRNAs (from previous studies) overlapped with lncRNAs identified in our study.

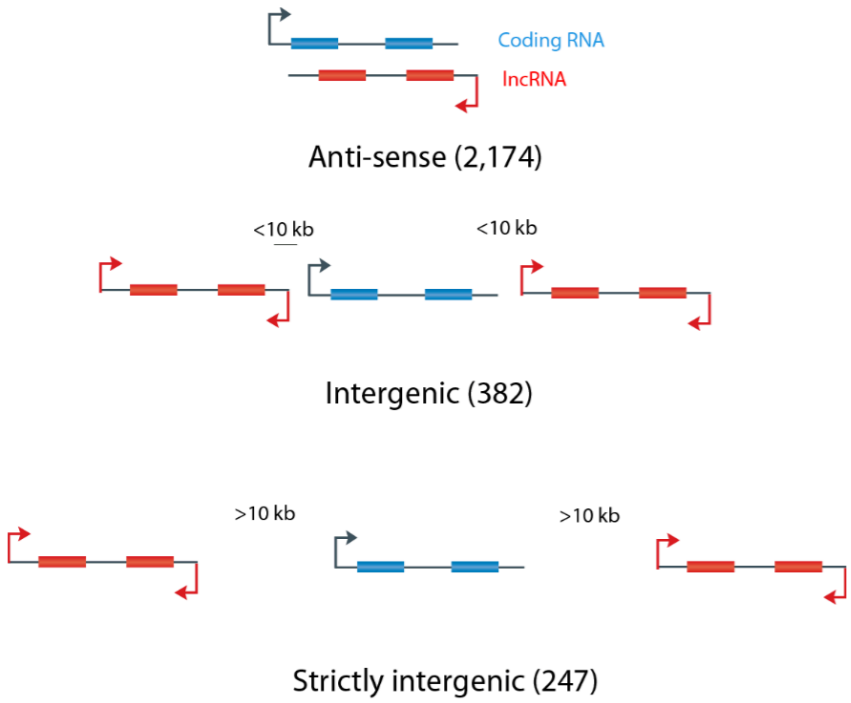


Figure S2: Genomic annotation of lncRNAs. The 2,803 lncRNAs were sub-classified into antisense, intergenic, or strictly intergenic, based on their intersection or proximity to the protein-coding genes.

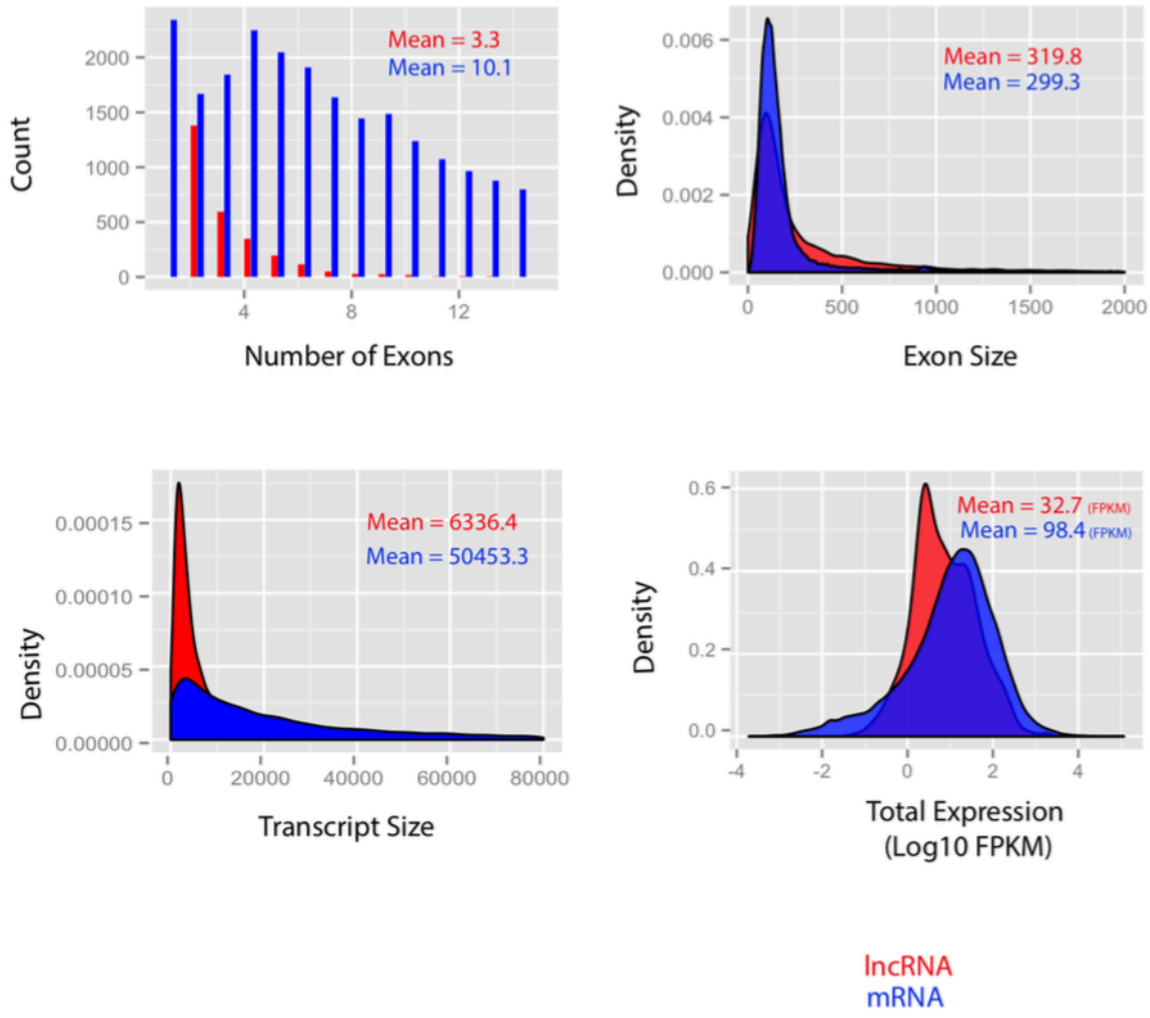


Figure S3: Genomic features of the 2,803 lncRNAs. (A) Number of exons of lncRNAs and mRNAs. (B) Distribution of exon size of lncRNAs and mRNAs. (C) Distribution of transcript size of lncRNAs and mRNAs (D) Distribution of expression levels for lncRNAs and mRNAs. (E) Distribution of conservation levels for lncRNAs and mRNAs. In al panels, lncRNA data are in red, and mRNA data, in blue.

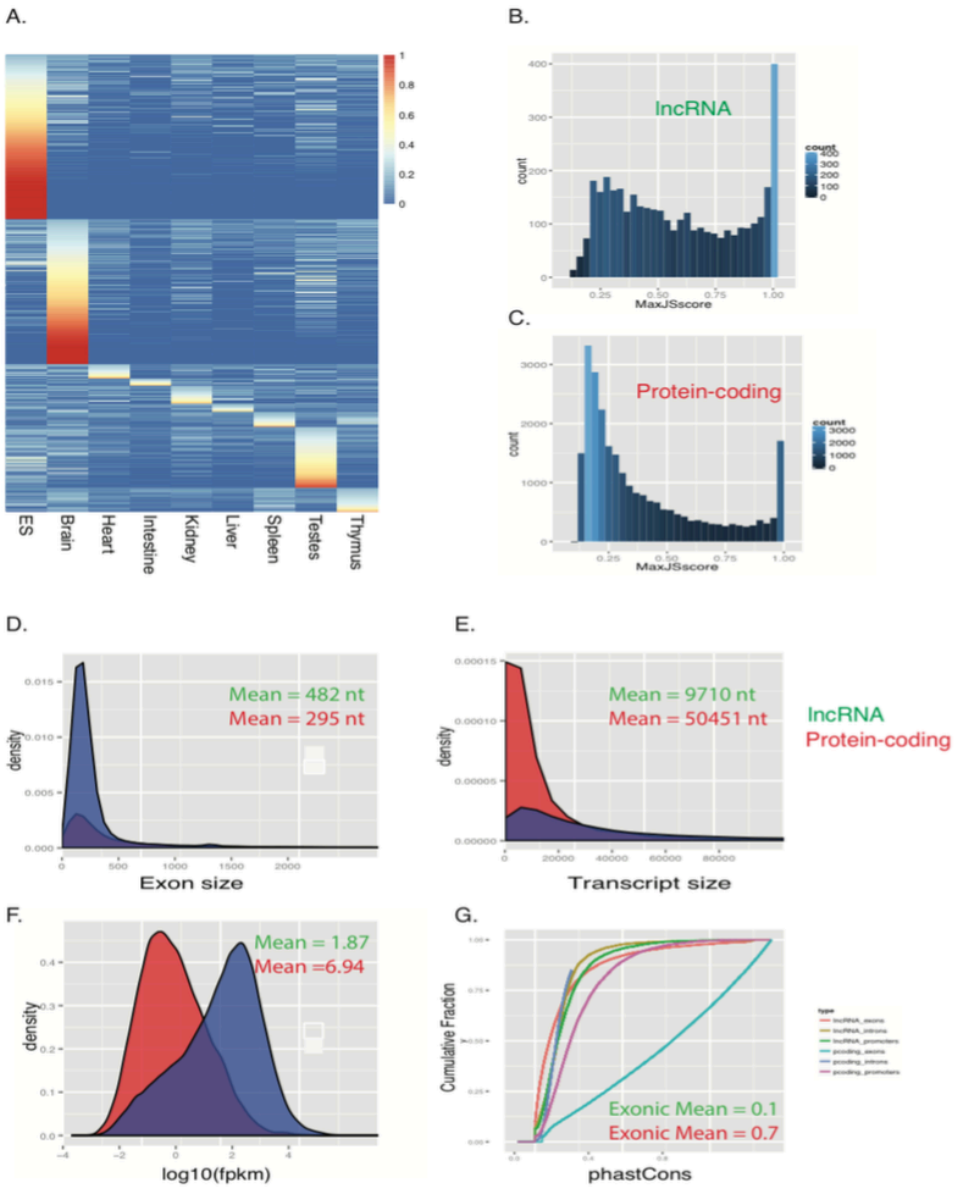


Figure S4: Genomic features of all intergenic lncRNAs. (A) Expression of 3,759 intergenic lncRNAs expressed in at least one out of the 8 tissues or the ES cell line. (B) Higher tissue-specific expression of lncRNAs. (C) Lower tissue-specific expression of protein-coding genes. (D) Exon sizes of lncRNAs (mean = 482 nt) and protein-coding (mean = 295 nt) transcripts. (E) Transcript sizes of lncRNA (mean = 50,451 nt) and protein-coding (mean = 295 nt) transcripts. (F) Expression of lncRNA (mean = 1.87 FPKM) and protein-coding (mean = 6.94 FPKM) transcripts. (G) Conservation scores of lncRNA exons (mean = 0.1 phastCons score) and protein-coding exons (mean = 0.7 phastCons score), along with promoters and introns.

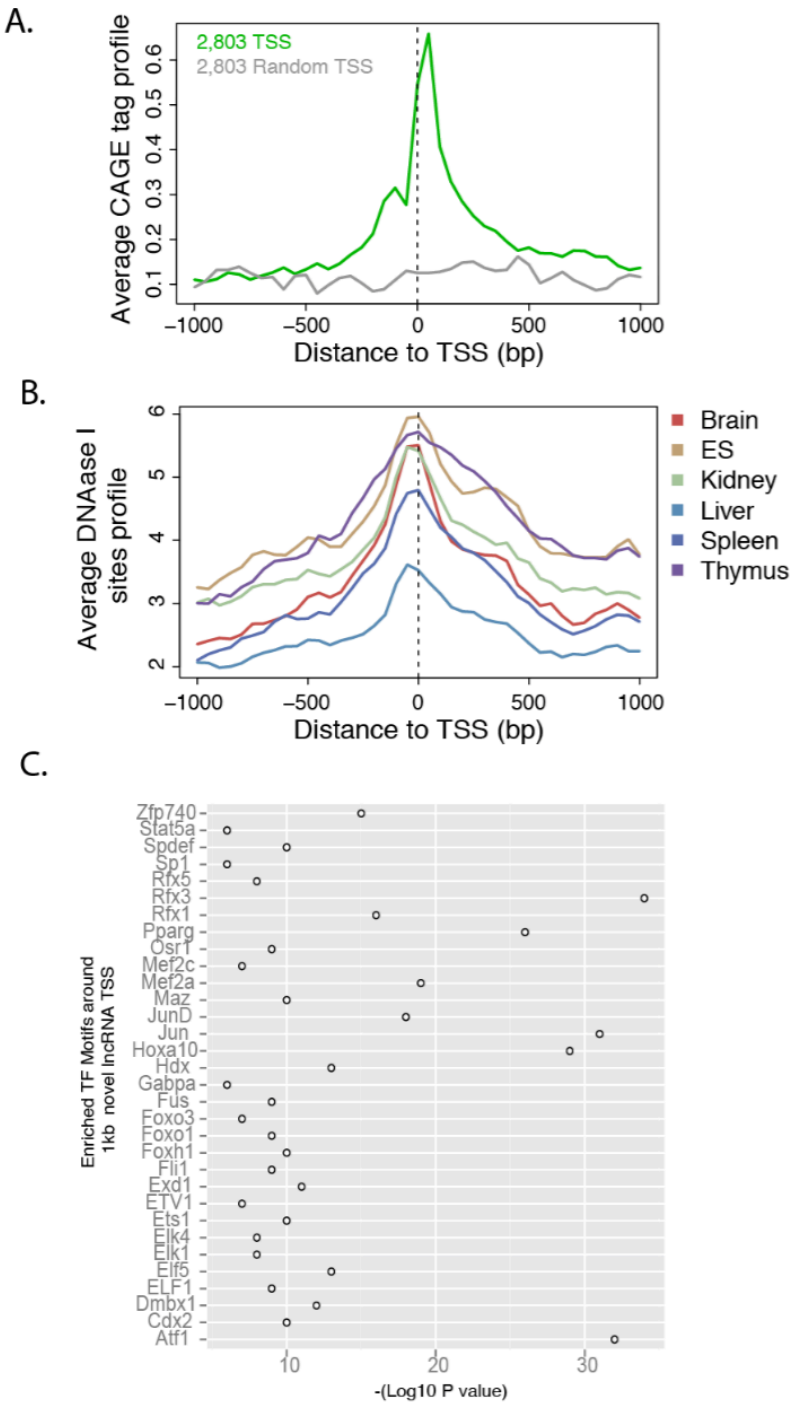


Figure S5: CAGE tags, DNase I, sites and *de novo* motifs of transcription factors enriched at lncRNA promoters. (A) Average CAGE tag profiles around novel lncRNA transcription start sites within 1kb up- or downstream. (B) Average DNase I signals around TSSs of lncRNAs within a distance of 1 kb up- or downstream across several tissues in mouse. (C) Enrichment of *de novo* transcription factor motifs around lncRNA promoters.

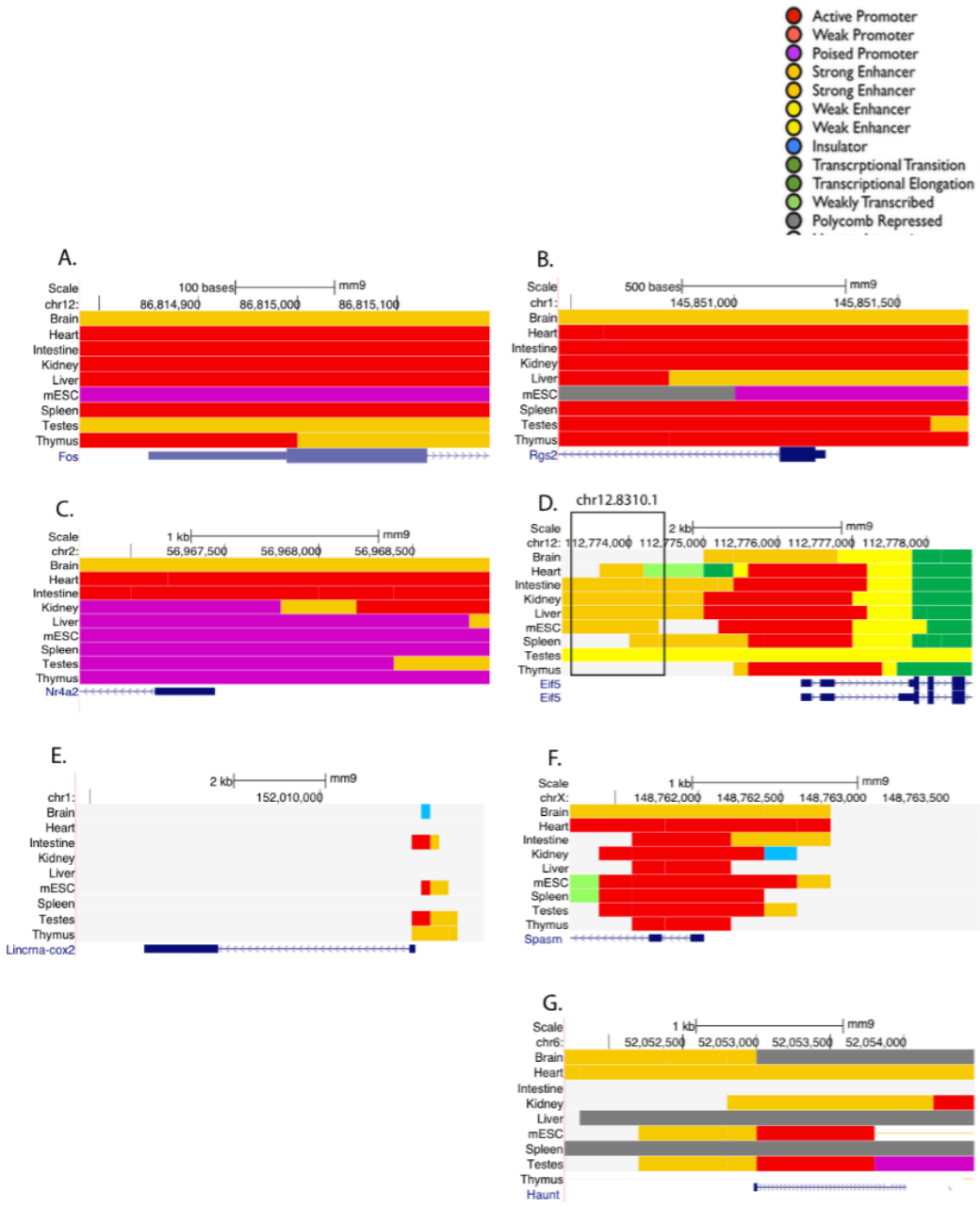


Figure S6: Enhancer-associated RNAs. (A–D) Coding RNA promoters (Fos, Rgs2, Nr4a2, Elf5) associated with an enhancer chromatin map. (F–G) LincRNA promoters (lincRNA-Cox2, Spasm, and Haunt) associated with an enhancer chromatin map.

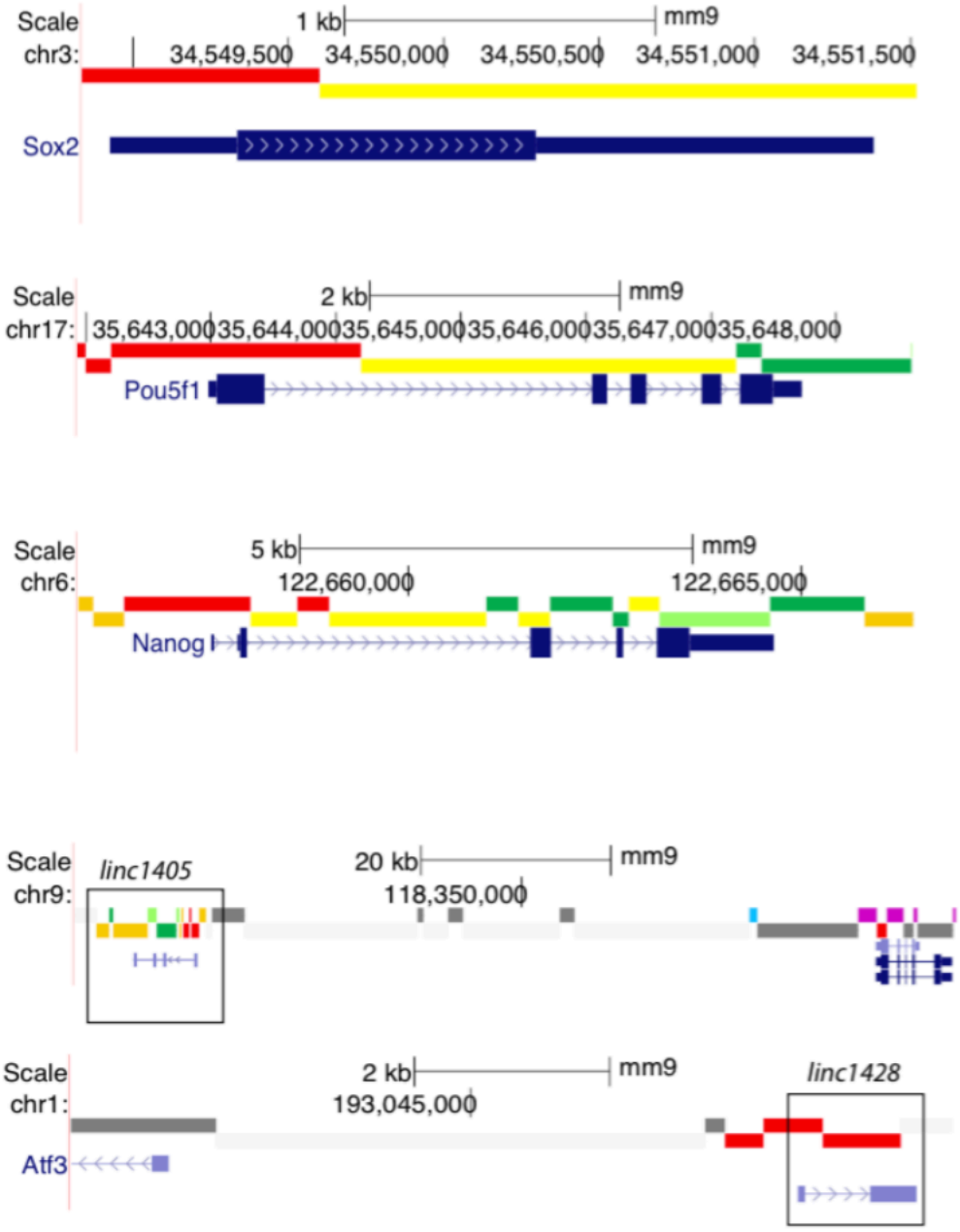


Figure S7: Promoter-associated RNAs in ES cells. (A) Coding RNA promoters (Sox2, Oct4/Pou5f1, and Nanog) associated with a promoter chromatin map. (B) lincRNA promoters (linc1405, linc1428) associated with a promoter chromatin map.

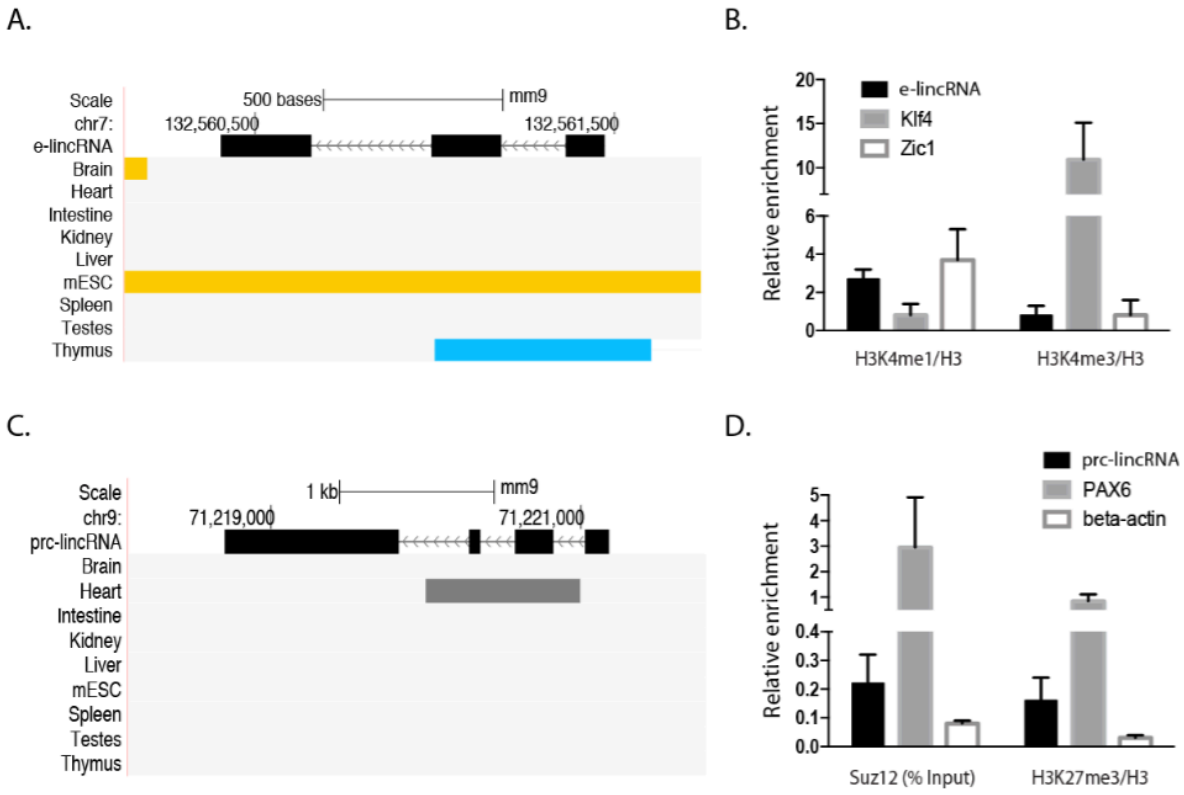


Figure S8: Experimental validation of promoter and enhancer chromatin marks enrichment at lincRNA promoters. (A) e-lincRNA chromatin status across several tissues, with the enhancer chromatin map in ES cells highlighted in yellow. (B) ChIP-qPCR of H3K4me3 and H3K4me1 around e-lincRNA promoter and a positive control (Klf4) along with a negative control (Zic1). (C) prc-lincRNA chromatin status across several tissues highlighting in grey polycomb (repressed) chromatin map in heart. (D) ChIP-qPCR of H3K27me3 and Suz12 around the prc-lincRNA promoter, with a positive control (Pax6) and a negative control (β-actin).

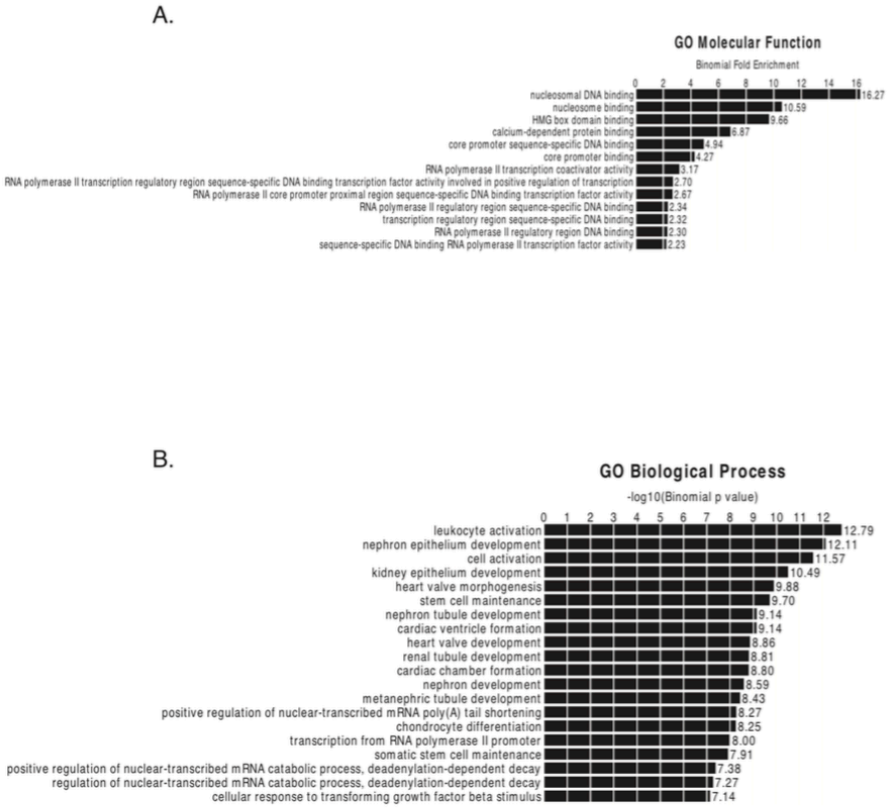


Figure S9: Gene ontology of chromatin-associated lncRNAs, enriched with GO categories such as molecular function and biological processes.

Objective - 2

Not only repetitive elements occupy nearly half of the human genome but also facilitate genome evolution (Bourque, 2009). They also provide alternative promoters, exons and splice junctions to the protein-coding genes (Faulkner et al., 2009). Their genomic insertions and transcription can disrupt gene expression (McClintock, 1951) and cause numerous diseases (Kaer and Speek, 2013). Thus, the insertions and transcription of repetitive elements proximal to protein-coding and lncRNA genes may create new transcriptional landscapes throughout the evolution.

Although repetitive elements pervade mammalian genomes, their overall contribution to transcriptional activity is poorly defined. Previously, using array and CAGE technologies, genome-wide transcription of repetitive elements has been profiled across limited tissues in human and mouse (Djebali et al., 2016; Faulkner et al., 2009; Fort et al., 2014; Nigumann et al., 2002). However, these technologies were limited by cross-hybridization, read length and mapping issues.

Characterization of global transcription of repetitive elements across many human tissues has not been explored yet. The main objective (*Objective 2, Second manuscript*) here is to address the above challenge by utilizing GTEx RNA-Seq data from 53 different human body sites.

SECOND MANUSCRIPT

The Transcriptional Landscape of Repetitive Elements in Human Tissues

Gireesh K. Bogu [1,2,3], Ferran Reverter [1], Marc A. Marti-Renom [2,3,4,5], Roderic Guigo [1,3,6,7]

1. Bioinformatics and Genomics, Center for Genomic Regulation (CRG), Barcelona, Catalonia, Spain.

2. Gene Regulation, Stem Cells and Cancer Program, Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona, Spain.

3. Universitat Pompeu Fabra (UPF), Barcelona, Spain.

4. CNAG-CRG, Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.

5. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

6. Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Barcelona, Catalonia, Spain.

7. Joint CRG-Barcelona Super Computing Center (BSC)–Institut de Recerca Biomedica (IRB) Program in Computational Biology, Barcelona, Catalonia, Spain

Abstract

More than half of the human genome contains repetitive elements and majority of their transcription is repressed. However, it is not clear how many of them are expressed and where they are expressed. To address this, as a part of Genotype-Tissue Expression (GTEx) project, we profiled the transcription of around 5 million repetitive elements using 8,551 poly-A RNA-seq datasets from 53 distinct body sites across 544 individuals. We report 11,502 intergenic repetitive elements originating from various repeat subfamilies that are systematically transcribed across multiple human tissues. Using linear mixed models, we show that, on average, variation in repeat expression is far greater among tissues (~57%) than among individuals (~1%). Further, we found 3,295 tissue-specific repetitive elements and we show that majority of their transcription is not induced by locus-specific effect. We also show that brain and testis consist of higher tissue-specific expression of repetitive elements compared to any other tissue. In summary, we find that repeats are expressed globally and their tissue-specific expression is a hallmark of tissue identity in humans.

Nearly half of the human genome is derived from repetitive elements (Lander et al., 2001). The majority of the repetitive elements belong to a retrotransposon type (a transposon whose sequence shows homology with that of a retrovirus and genetic elements that can amplify themselves via RNA intermediate) and the rest to DNA transposons, satellite repeats and simple or low-complexity repeats (Kazazian, 2004). Retrotransposons are further divided into various classes like SINE (short interspersed nuclear element), LINE (long interspersed nuclear element) and LTR (long terminal repeat) (Elbarbary et al., 2016). Repetitive elements contribute the coding and noncoding transcriptome by acting as enhancers, alternative promoters, repeat-derived RNAs and transcription factor binding sites (Bourque, 2009; Göke and Ng, 2016). Multiple evidences show that insertions of repetitive elements at DNA level linked various diseases (Callinan and Batzer, 2006; Göke and Ng, 2016; Kaer and Speek, 2013).

The overall contribution of repeat expression to the transcriptional activity across human adult tissues is poorly understood. Previously, using array and CAGE (CAP Analysis of Gene Expression) technologies, the transcriptional output of repetitive elements has been analyzed in human adult tissues (Djebali et al., 2016; Faulkner et al., 2009; Fort et al., 2014) but these technologies suffer from various limitations including cross-hybridization, they are expensive and the short tags cannot be uniquely mapped to the reference genome (Wang et al., 2009). Here we investigate the transcription of millions of genomic repetitive instances in the human genome using thousands of RNA-Seq datasets across wide range of tissues and individuals. We show hundreds of thousands of genomic repetitive instances being expressed and various repeat subfamilies varying greatly between tissues rather than individuals. We also identify novel repeat families that are tissue-specific in nature. Our findings highlight the tissue-specific nature of the repetitive elements in the human genome.

We used mid-phase RNA sequencing (RNA-Seq) data from Genotype-Tissue Expression (GTEx) Project, (The GTEx Consortium et al., 2015). It consists of 8,551 RNA-Seq samples from 544 human individuals spanning 53 distinct body sites (**Supplementary Fig. 1**). These individuals are of different age, gender and race and the RNA-Seq data is of 76-base pair (bp), paired-end, unstranded, poly (A)-selected with a median sequencing depth of 60 million reads per sample and with a good RNA

quality (**Supplementary Fig. 1**). The above 53 distinct body sites include 29 solid-organ tissues, 13 brain subregions, two cell lines (EBV-transformed lymphocytes and transformed fibroblasts) and a whole-blood sample. Around 5 million repetitive genomic instances or copies of various repeat subfamilies, families and classes from RepeatMasker database were used in the analysis (**Supplementary Table 1**) (Tarailo-Graovac and Chen, 2002). That majorly includes SINE, LINE, LTR, DNA, satellite, simple, low complexity and unknown repetitive classes. These repetitive classes represent more reliable annotations of 26 families and 1,250 subfamilies covering 46.5% of the human genome (hg19), corresponding to about 1,439 Mb.

To get an overview of repeat transcription, first, we calculated the fraction of transcriptome that originated from all types of repetitive elements across all 53 distinct human body sites (**Fig. 1a**). We found that brain and whole blood show relatively higher fraction of reads mapped to repeats compared to other tissues and skeletal muscle show the least of them. Second, we found 205,779 repetitive elements expressed in at least once of the 53 distinct human body sites **(Supplementary Fig. 2b**). Cerebellum, testis and endocervix seem to have highest number of expressed repetitive elements. 11,502 out of 205,779 repetitive elements are intergenic. Also in this intergenic repeats the highest expression is found in cerebellum and testis (**Fig. 1c, Supplementary Table 2**). Most of the expressed intergenic repeats belong to either LINE or LTR or SINE repeat classes.

To test whether the repetitive elements transcription is tissue-specific, we performed hierarchical clustering on the normalized expression (**Supplementary Fig. 2a**). Remarkably, the clustering largely recapitulates tissue type. We repeated the hierarchical clustering on only intergenic repeats and found that the tissue-specific clustering is not an artifact of overlapping protein-coding or lncRNA genes (**Fig. 2a**). Using linear mixed models, we found that variation in repetitive elements expression is far greater among tissues (57% of total variance in gene expression) than among individuals (1% of total variance, **Supplementary Fig. 3**). During the early stages of development, the genome is largely permissive for transcription; so repetitive elements that reside within transcribed loci are likely to be passively expressed. However, it has been shown that embryonic stages can still be clearly distinguished even after taking the average expression for all repetitive elements that belong to the same family suggesting that expression of ERV (endogenous retrovirus) elements is not a locus-specific effect caused by permissive transcription but instead a property of the families that the transcribed repetitive elements belong to (Göke et al., 2015).

This previous study was only limited to human embryonic tissues and to a specific family of repetitive elements known as ERVs. We extended this analysis to human adult tissues by analyzing all types of repetitive families to see whether expression of repetitive elements is a locus-specific effect or not. To do this, we averaged the expression of all repetitive elements that belong to the same family and performed hierarchical clustering. Interestingly, all adult tissues can still be clearly distinguished confirming the previous study (**Supplemental Fig. 2a**).

To identify which repeat families show tissue-specific expression in adult tissues, we calculated differential expression of all repetitive elements from every family at each tissue level (**Methods**). We found 3,295 tissue-specific repetitive elements in 18 major tissues that showed significant differential expression (**Fig. 2b**). Further, to illustrate the tissue-specific expression; we have shown few examples with RNA-Seq expression across many GTEx tissue samples (**Supplementary Fig. 4**). Interestingly, most of them correspond to either LTRs or LINEs (**Supplementary Table 3**). It is not exactly not known why the other tissues do not have any tissue-specific repetitive elements. Overall, these 18 tissues showed various distributions of tissue-specific repeat classes. Liver, pancreas, thyroid, pituitary, spleen contains almost equal levels of LTR, SINE and LINE repeats. Esophagus mucosa and ovary contains more SINE and LINE repeats than LTR. Lymphocytes contain more LTR repeats than SINE and LINEs. Brain and testis contain more LINE and LTR repeats than SINEs. Skeletal muscle has almost no LTR or LINE repeats. Brain (mostly Cerebellum) and testis showed higher tissue-specific expression of repetitive elements. Interestingly, we found tissue-specific satellite repeats only in testis (REP522, SST1, HSAT5, HSATI, D20S16, GSAT) (**Supplementary Table 4**). However, it is still needed to test whether these different distributions of tissue-specific repeat classes have any impact on tissue development.

Almost every repeat subfamily has similar copies of their sequence spread across many different unique locations in the genome. Even though tissue-specific expression of repeat subfamilies is specific to one tissue, it is possible that their copies could be shared between other tissues. In total, we identified 464 subfamilies that are differentially expressed with at least one copy being expressed in one of the 18 major tissues (**Supplementary Table 4**). 39% of them (181 subfamilies) expressed more than one copy and without any locus-specific effect (**Supplementary Table 4**, *highlighted in green color)*. For example, AluJb is brain-specific repeat subfamily with 29 copies in brain, 18 in testis and few copies in other

tissues. This again confirms the tissue-specificity is not a result of locus-specific effect. However, there are exceptions to this. 14% of repeat subfamilies (64) contain more than one copy of a repeat subfamily but with locus-specific effect (**Supplementary Table 4,** *highlighted in red color*). For example, L1ME5 subfamily contains 3 expressed copies located in 3 different places in the genome. One copy is expressed in pituitary gland, the second one in spleen and the third one in testis. Though these copies are tissue-specific, the overall subfamily average expression is not. Interestingly, 47% (218 subfamilies) of the 464 tissue-specific subfamilies express only one copy (**Supplementary Table 4,** *highlighted in orange color*). For example, LTR39 in brain, HAL1b in liver, MER39 in pancreas and LTR10E in spleen etc. However, 135 subfamilies of this type belong only to testis, 43 to brain and the other 40 to other tissues.

To gain insights into tissue-specific expression, we mapped a global network of all tissue-specific repetitive elements across 18 human tissues (**Supplementary Fig. 5a**). This global network is simply made of nodes and edges, where nodes represent repetitive subfamilies (orange) and tissues (blue), and edges (grey) represent the tissue-origin of repeat subfamilies. The size of the tissue-node represents the number of tissue-specific repetitive elements and the size of the repeat subfamily-nodes represent the number of expressed copies. The distance between tissue-node and repeat subfamily-node along with thickness of the nodes define their strength of association. For example, L2a subfamily is more close to testis than brain (Cerebellum) because it is more associated with brain by having maximum number of expressed copies (62 copies) than testis (33 copies) (**Supplementary Fig. 5**). From the network it is very apparent that all tissue-specific repetitive elements show a clear bias towards brain (**Supplementary Fig. 6**) and testis (**Supplementary Fig. 7**) as expected before. However, it is still to be tested, how this tissue-specific regulatory network of repeat subfamilies across human tissues is controlled.

Repetitive elements have shown to be repressed by various biological processes like DNA methylation or histone modifications. However, few studies suggested that the specific types of repetitive elements (retrotransposons) are transcribed in embryonic tissues at different stages of development and in few adult tissues of human. Overall, in this repeat-analysis, we have taken the advantage of massive RNA-Seq data from human tissues to investigate the genome-wide expression of repetitive elements. Our analysis shows that repetitive elements expression is a common phenomenon in human. To overcome the challenges associated with the analysis, we used only

uniquely mappable reads from 76 bp RNA-Seq data and found that these are sufficient to detect tissue-specific transcription. In conclusion, in this study, we show that all types of repetitive elements can transcribe across 53 distinct human tissue sites and are highly expressed in brain and testis. The transcription of repetitive elements across many tissues is tissue-specific in nature suggests that these elements could play an important role in the tissue's development, function and pathology.

METHODS

Data preparation.

8,551 RNA-Seq samples from 53 distinct body sites across 544 individuals were downloaded from the database of Genotypes and Phenotypes (dbGAP) as a part of GTEx project. We eliminated samples with sex-irregularities and obtained by surgical procedure to avoid heterogeneity.

Mapping

RNA-Seq reads were aligned with TopHat2 (Kim et al., 2013) to the human genome version hg19. Uniquely mapped reads were selected using Samtools (Li et al., 2009) for further analysis. Multi-mapped reads were distributed equally to repeat families using RepEnrich tool ((Criscione et al., 2014)).

Quantification

We used the RepeatMasker annotation (http://www.repeatmasker.org/) to define repeats and we removed those repeats overlapping known genes (Gencode v12, RefSeq, UCSC, ENSEMBL, Caibili et.al lncRNA and Gencode V17 comprehensive annotations). We counted the number of uniquely mapped reads overlapping each annotated repeat instance in the 8,551 samples from the 53 distinct body sites. Read counts were normalized by the length of the repetitive element and by the sum of reads mapping to all repetitive elements and annotated genes (in millions). Normalized read counts were then log transformed after adding a pseudo-count of 0.01. Using this approach, from the 5,285,549 annotated repeats, around 200k repetitive elements were expressed in at least one tissue.

Clustering

We ran hierarchical clustering based on average linkage criterion on the read counts of the 11k intergenic repeats using specific settings (*clustering_distance_rows="euclidean", clustering_distance_cols="correlation",*

clustering_method="average"). The clustering recapitulated tissue classification. This suggesting that the repetitive element tissue-specific expression is not an artifact of overlapping with protein-coding genes.

Tissue-specificity

First the median of the normalized data was calculated for every repetitive element in each tissue. For every repeat family, the distribution of median normalized read count values was then compared to the distribution of median normalized read count values in all other tissues using the Wilcoxon and the t-test. Elements, which had a p-value<0.001 for the t-test and the Wilcoxon test were selected as tissue-specific. *ggnet* was applied to produce networks of tissue-specific repetitive elements (reference).

Variation

To assess the contribution of tissue and individual to gene expression variation, we used a linear mixed model (LMM). Repetitive element expression was modeled as a function of tissue and individual (considered as random factors). The LMM was implemented in the R package lme4 ((Bates et al., 2015)). Repetitive elements not expressed (RPKM > 0) in any of the samples were excluded from the analysis. We used log2-normalized data and pseudo-counts to deal with zero expression values. To obtain the variance components, we divided the restricted maximum likelihood (REML) estimators for the random effects of tissue, individual and residual variance by their sum. We visually examined the scatter plot of the contribution to expression variation of tissue plus individual versus median expression.

References:

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. J. Stat. Soft. 67. doi:10.18637/jss.v067.i01

Bourque, G., 2009. Transposable elements in gene regulation and in the evolution of vertebrate genomes. Current Opinion in Genetics & Development 19, 607–612. doi:10.1016/j.gde.2009.10.013

Callinan, P.A., Batzer, M.A., 2006. Retrotransposable elements and human disease. Genome Dyn 1, 104–115. doi:10.1159/000092503

Criscione, S.W., Zhang, Y., Thompson, W., Sedivy, J.M., Neretti, N., 2014. Transcriptional landscape of repetitive elements in normal and cancer human cells. BMC Genomics 15, 583. doi:10.1146/annurev.genet.35.102401.091032

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., 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, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A.M., 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, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., Gingeras, T.R., 2016. Landscape of transcription in human cells. Nature 488, 101–108. doi:10.1038/nature11233

Elbarbary, R.A., Lucas, B.A., Maquat, L.E., 2016. Retrotransposons as regulators of gene expression. Science 351, aac7247–aac7247. doi:10.1126/science.aac7247

Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K., Cloonan, N., Steptoe, A.L., Lassmann, T., Waki, K., Hornig, N., Arakawa, T., Takahashi, H., Kawai, J., Forrest, A.R.R., Suzuki, H., Hayashizaki, Y., Hume, D.A., Orlando, V., Grimmond, S.M., Carninci, P., 2009. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet 41, 563–571. doi:10.1038/ng.368

Fort, A., Hashimoto, K., Yamada, D., Salimullah, M., Keya, C.A., Saxena, A., Bonetti, A., Voineagu, I., Bertin, N., Kratz, A., Noro, Y., Wong, C.-H., de Hoon, M., Andersson, R., Sandelin, A., Suzuki, H., Wei, C.-L., Koseki, H., Hasegawa, Y., Forrest, A.R.R., Carninci, P., 2014. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nature Publishing Group 46, 558–566. doi:10.1038/ng.2965

Göke, J., Lu, X., Chan, Y.-S., Ng, H.-H., Ly, L.-H., Sachs, F., Szczerbinska, I., 2015. Dynamic Transcription of Distinct Classes of Endogenous Retroviral Elements Marks Specific Populations of Early Human Embryonic Cells. Stem Cell 16, 135–141. doi:10.1016/j.stem.2015.01.005

Göke, J., Ng, H.-H., 2016. CTRL+INSERT: retrotransposons and their contribution to regulation and innovation of the transcriptome. EMBO Rep 17, 1131–1144. doi:10.15252/embr.201642743

Kaer, K., Speek, M., 2013. Retroelements in human disease. Gene 518, 231–241. doi:10.1016/j.gene.2013.01.008

Kazazian, H.H., 2004. Mobile elements: drivers of genome evolution. Science 303, 1626–1632. doi:10.1126/science.1089670

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., 2001. Initial sequencing and analysis of the human genome. Nature.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,

Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Tarailo-Graovac, M., Chen, N., 2002. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. John Wiley & Sons, Inc., Hoboken, NJ, USA. doi:10.1002/0471250953.bi0410s25

The GTEx Consortium, Ardlie, K.G., Deluca, D.S., Segre, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., Lek, M., Ward, L.D., Kheradpour, P., Iriarte, B., Meng, Y., Palmer, C.D., Esko, T., Winckler, W., Hirschhorn, J.N., Kellis, M., MacArthur, D.G., Getz, G., Shabalin, A.A., Li, G., Zhou, Y.H., Nobel, A.B., Rusyn, I., Wright, F.A., Lappalainen, T., Ferreira, P.G., Ongen, H., Rivas, M.A., Battle, A., Mostafavi, S., Monlong, J., Sammeth, M., Mele, M., Reverter, F., Goldmann, J.M., Koller, D., Guigo, R., McCarthy, M.I., Dermitzakis, E.T., Gamazon, E.R., Im, H.K., Konkashbaev, A., Nicolae, D.L., Cox, N.J., Flutre, T., Wen, X., Stephens, M., Pritchard, J.K., Tu, Z., Zhang, B., Huang, T., Long, Q., Lin, L., Yang, J., Zhu, J., Liu, J., Brown, A., Mestichelli, B., Tidwell, D., Lo, E., Salvatore, M., Shad, S., Thomas, J.A., Lonsdale, J.T., Moser, M.T., Gillard, B.M., Karasik, E., Ramsey, K., Choi, C., Foster, B.A., Syron, J., Fleming, J., Magazine, H., Hasz, R., Walters, G.D., Bridge, J.P., Miklos, M., Sullivan, S., Barker, L.K., Traino, H.M., Mosavel, M., Siminoff, L.A., Valley, D.R., Rohrer, D.C., Jewell, S.D., Branton, P.A., Sobin, L.H., Barcus, M., Qi, L., McLean, J., Hariharan, P., Um, K.S., Wu, S., Tabor, D., Shive, C., Smith, A.M., Buia, S.A., Undale, A.H., Robinson, K.L., Roche, N., Valentino, K.M., Britton, A., Burges, R., Bradbury, D., Hambright, K.W., Seleski, J., Korzeniewski, G.E., Erickson, K., Marcus, Y., Tejada, J., Taherian, M., Lu, C., Basile, M., Mash, D.C., Volpi, S., Struewing, J.P., Temple, G.F., Boyer, J., Colantuoni, D., Little, R., Koester, S., Carithers, L.J., Moore, H.M., Guan, P., Compton, C., Sawyer, S.J., Demchok, J.P., Vaught, J.B., Rabiner, C.A., Lockhart, N.C., Ardlie, K.G., Getz, G., Wright, F.A., Kellis, M., Volpi, S., Dermitzakis, E.T., 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348, 648–660. doi:10.1126/science.1262110

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10, 57–63. doi:10.1038/nrg2484

Figure 1

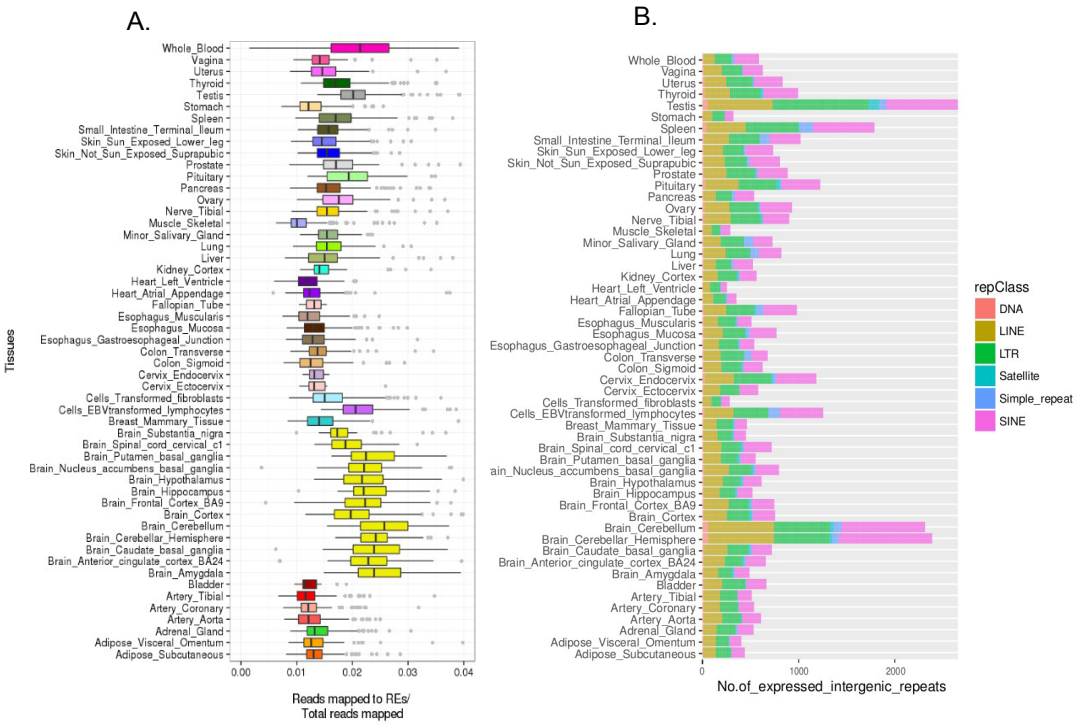
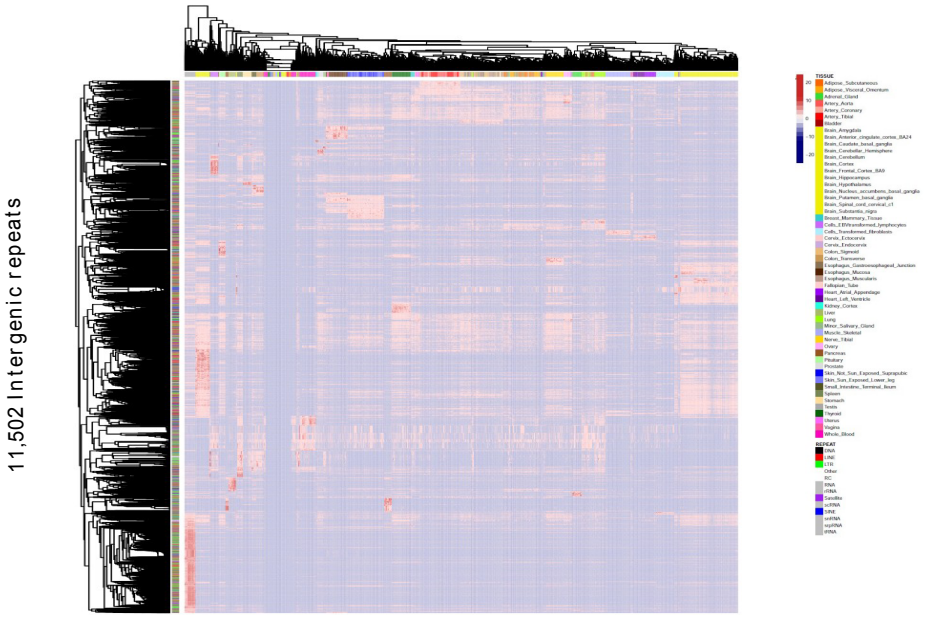


Figure 1. Repetitive elements expression in humans.
(A) Fraction of uniquely-mapped RNA-seq reads that map to repetitive elements. Boxplots show the distribution for all RNA-seq samples from different tissues.
(B) Number of expressed intergenic repetitive elements of various repeat classes by tissue.

Figure 2

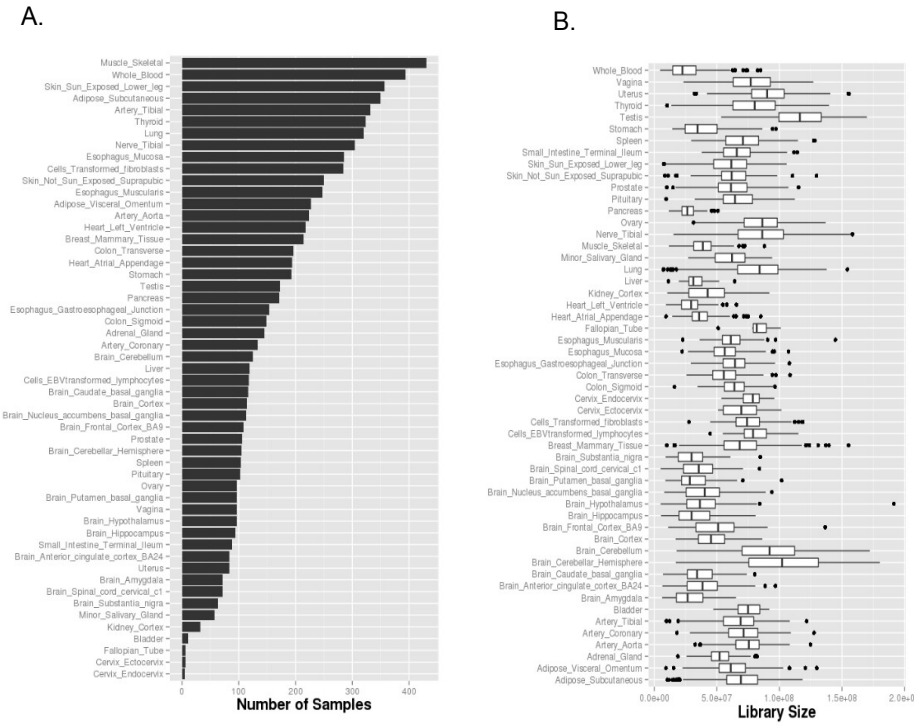
A.



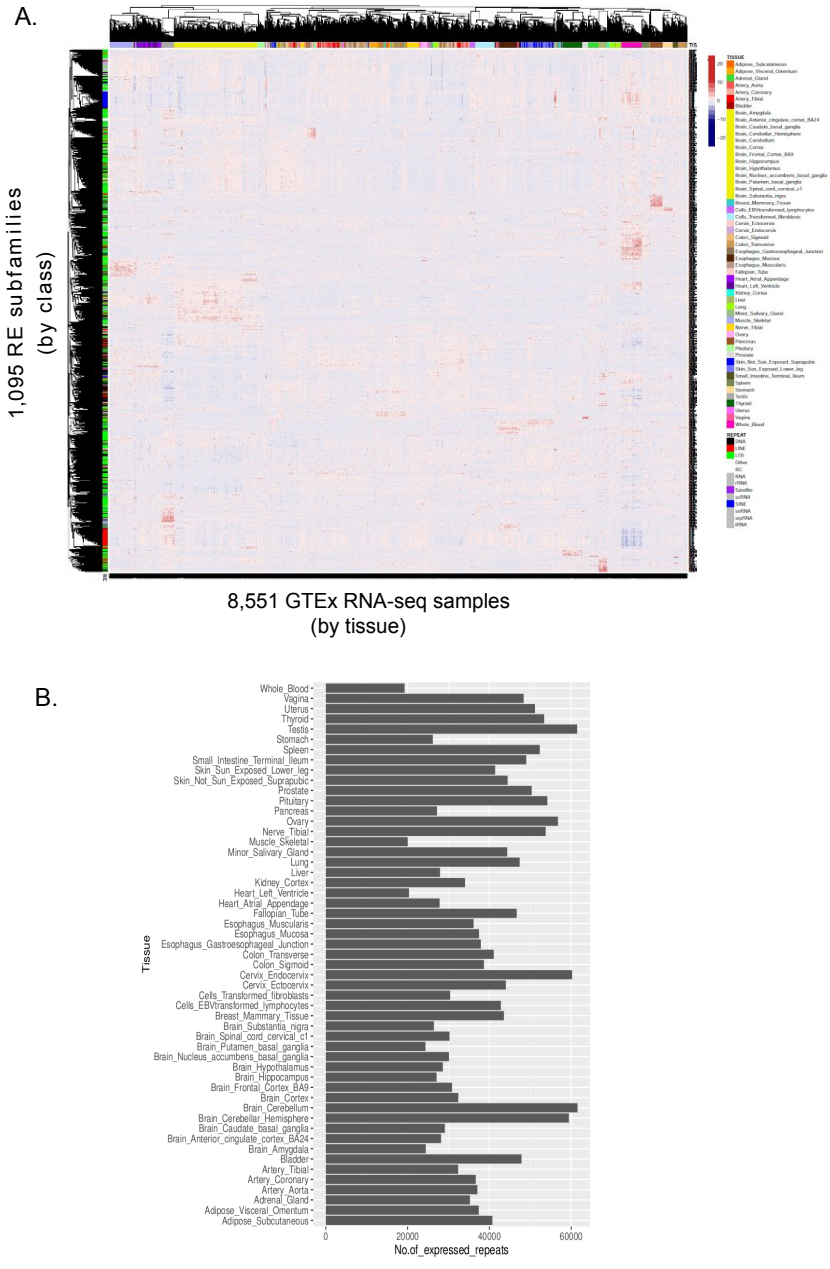
B.

Tissues	Number of tissue-specific repeat instances
Testis	1758
Brain	905
EBV Transformed Lymphocytes	129
Spleen	114
Pituitary	79
Thyroid	76
Pancreas	68
Liver	45
Skin	23
Ovary	18
Esophagus	18
Muscle Skeletal	12
Adrenal Gland	12
Fibrobalsts	9
Kindey Cortex	8
Heart	8
Nerve Tibial	5
Minor Salivary Gland	4
Artery	3
Bladder	1

Figure 2. Tissue-specific expression of repeat families in humans.
(A) Hierarchical clustering of 11,502 intergenic repetitive elements expression across 8,551 RNA-Seq samples spanning all 53 different human body sites. Heatmap is showing clusters of repetitive element expression human by tissue and by repetitive element class. Red indicates overrepresentation, green indicates underrepresentation. Color intensity is based on Log10 (normalised expression + pseudocount, 0.5) scale from -20 to +20.
(B) Number of tissue-specific repeat instances by tissue based on the differential expression from one tissue to another.

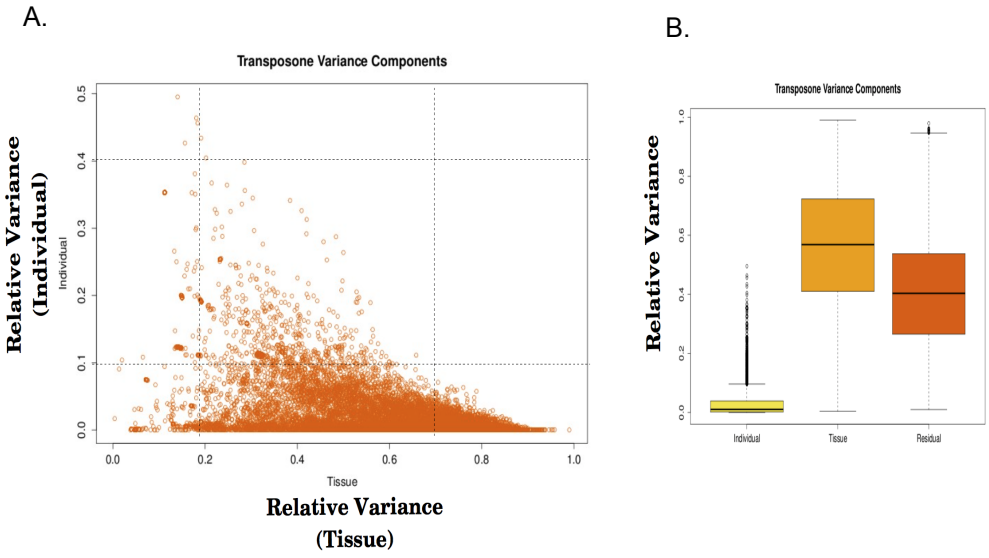


Supplementary Figure 1. GTEx data statistics
(A) Distribution of number of samples per tissue (in total, 8,551).
(B) Distribution of GTEx RNA-Seq samples library size across different tissues.



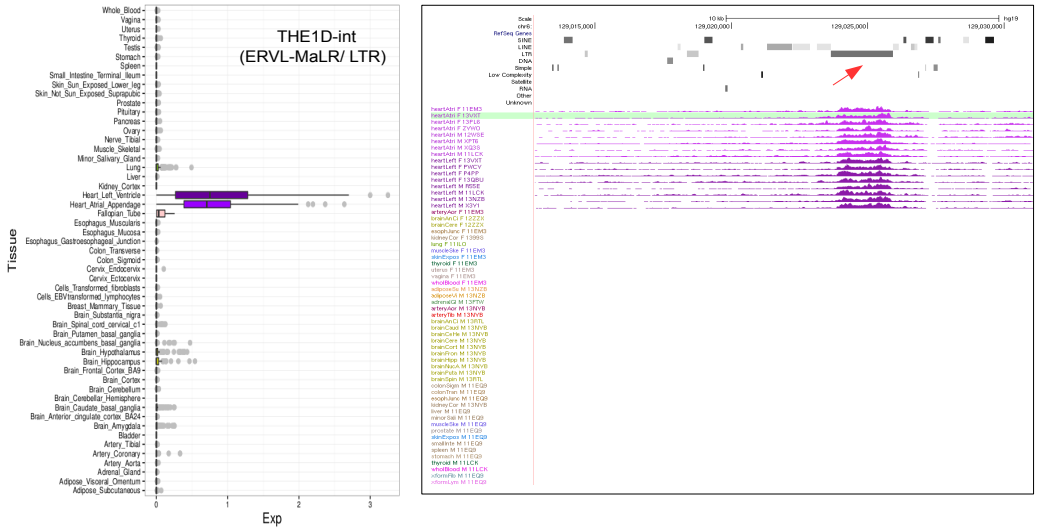
Supplementary Figure 2. Human repetitive elements expression
(A) Heirarchical clustering of repetitive elements expression by family (including the ones that overlap protein-coding genes) recapitualtes tissue type.
(B) Number of expressed repeats (including the ones that overlap protein-coding genes) across different tissues in human,.

Supplementary Figure 3

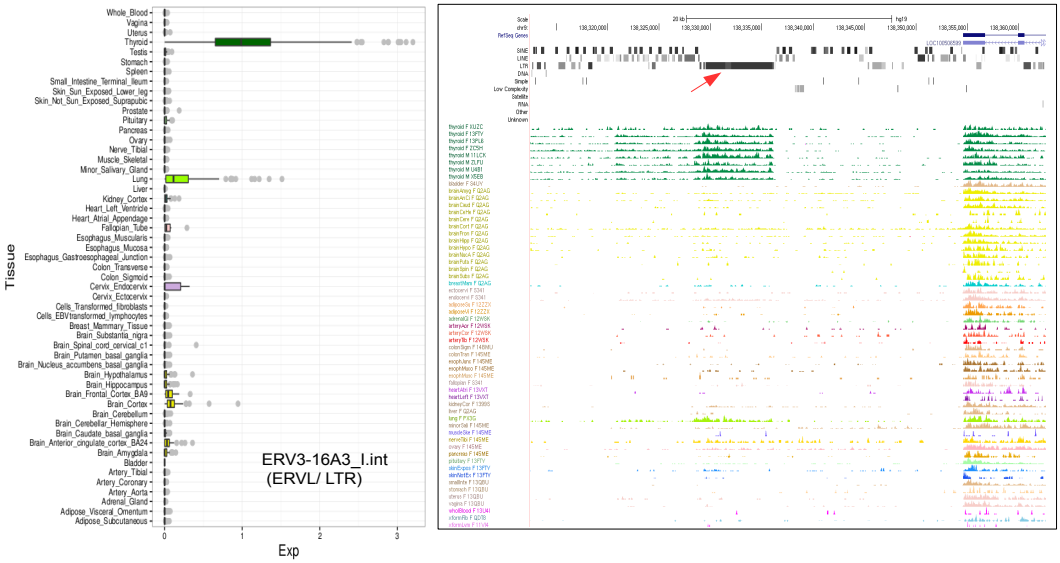


Supplementary Figure 3. Variance in repetitive elements expression across tissues and individuals
(A) Contribution of tissue and individual to gene expression variation of repetitive lements.
Top left: repetitive elements with high individual variation and low tissue variation.
Bottom right: repetitive elements with low individual variation and high tissue variation
(B) Box-plots showing that tissue variance is higher than the individual variation.

A.

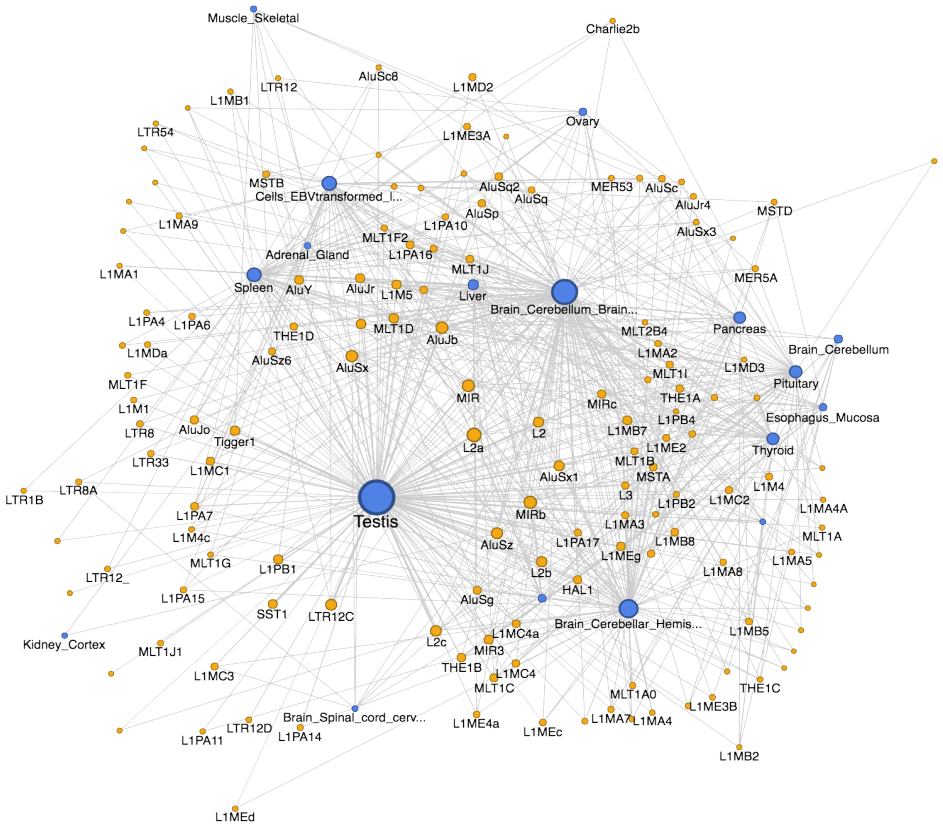


B.



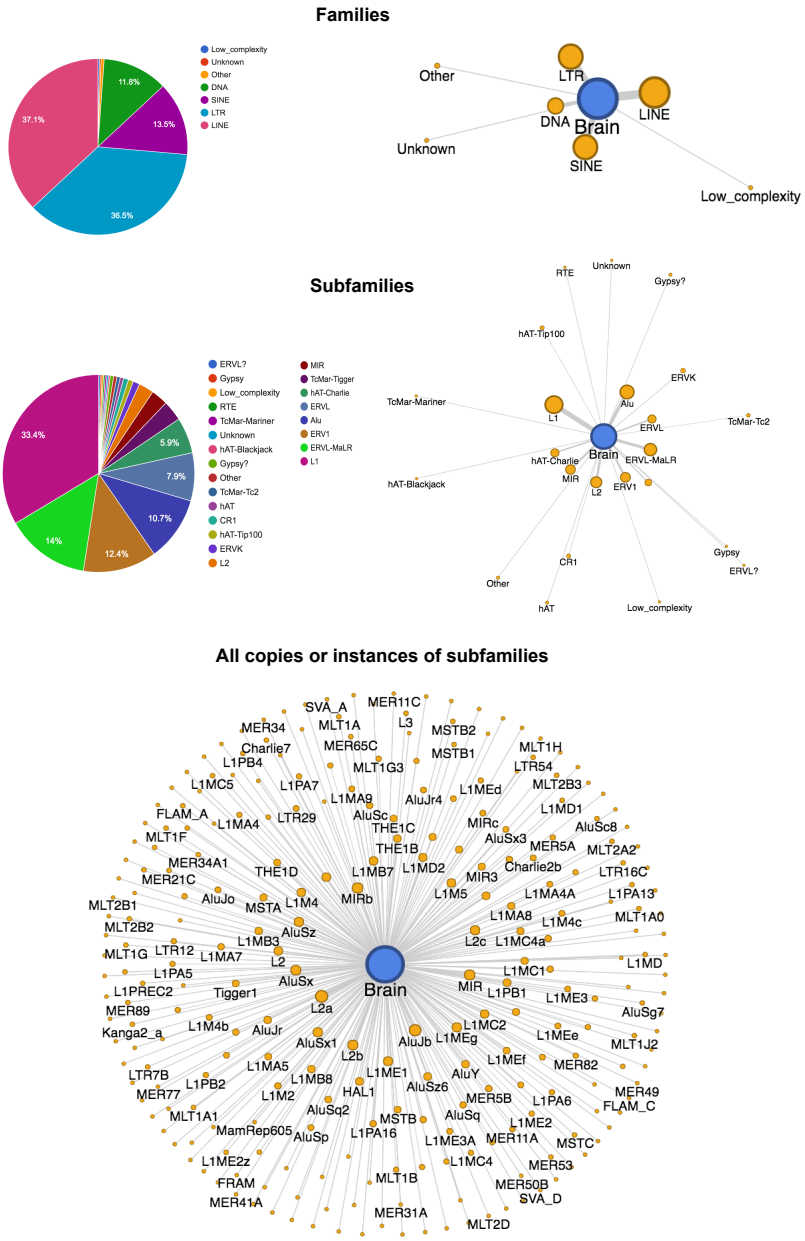
Supplementary Figure 4. Tissue-specific repetitive elements expression at twodifferent loci. Data from heart samples is shown in purple (A) and thyroid in green (B). The box-plots represent the distribution of repetitive elements expression in thousands of RNA-seq samples across various tissues. Arrow marks (red in color) in UCSC browser figures and higher medians in box-plots highlight the stage-specific ERV elements (THE1D-int, ERV3-16A3_I.int).

Supplementary Figure 5.



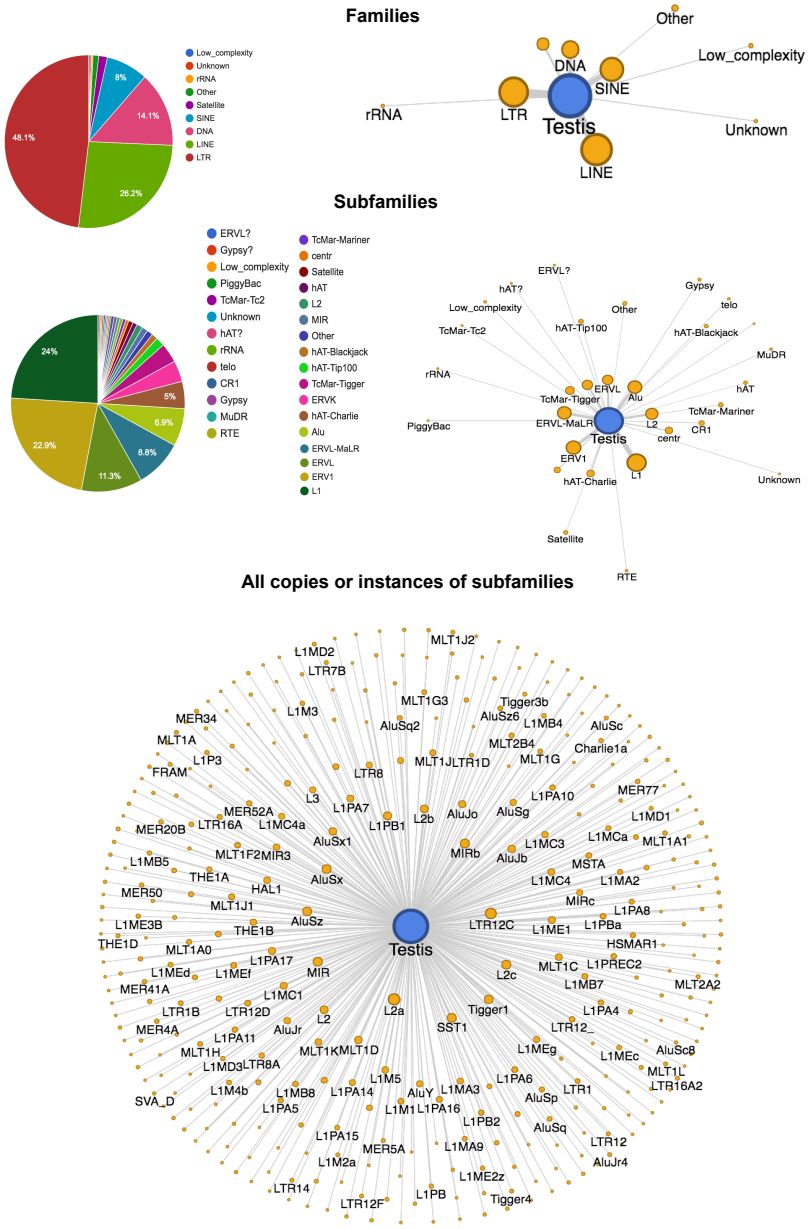
Supplementary Figure 5. Tissue-specific repetitive elements networks in humans.
A complete network of tissue-specific repetitive elements in human. Blue color represents tissues and orange repeat subfamilies. Node size represents the total number of tissue-specific repetitive instances.

Supplementary Figure 6.



Supplementary Figure 6. Brain tissue-specific repetitive elements network. Compisition of repetitive elements in brain by family, subfamily (Top) and complete tissue-specific repeat netwrok of brain (Bottom).

Supplementary Figure 7.



Supplementary Figure 7. Testis tissue-specific repetitive elements network. Compisition of repetitive elements in testisby family, subfamily (Top) and complete tissue-specific repeat netwrok of testis (Bottom).

Discussion

This thesis work is a survey of the lncRNA transcriptome in mouse and repetitive elements transcriptome in human. Characterizing the deepest part of the transcriptome for each specific tissue or sub-cellular compartment is essential to an understanding of the complexity of the non-coding genome in both human and mouse. The non-coding genome has a plethora of potential regulatory features, such as the chromatin-associated lncRNAs and tissue-specific repetitive elements described in this study.

This work relies on the capacity of RNA-Seq can accurately detect transcription of both lncRNAs and repetitive elements. More than 90% of the RNA-Seq reads successfully mapped to the mammalian genomes using advanced mappers (Trapnell et al., 2009; 2012b). Also to the rise of consortiums, especially ENCODE and GTEx, which make publicly available high quality RNA-Seq data with greater sequencing depth and longer read length and make it availability to the public (Mouse ENCODE Consortium et al., 2012; The GTEx Consortium et al., 2015). Such high-quality RNA-Seq data enables single-nucleotide resolution and therefore enables the discovery of shorter length repetitive elements and obviously the longer non-coding RNAs. Additionally, we confirmed the expression of lncRNAs by qRT-PCR in multiple tissues.

Our work also integrated ChIP-Seq to support RNA-Seq based discoveries. By using advanced machine learning method, 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) (Shen et al., 2012). 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 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, 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 was the first one to provide a comprehensive catalog of chromatin-associated lncRNAs across several mouse tissues.

We discovered 2,803 lncRNAs by mapping reads from 19 RNA-Seq datasets, assembling them *de novo* in mouse brain, heart, kidney, small intestine, liver, spleen, testis, and thymus and a paired-end ES cell line, Further, we identified 11,502 expressed intergenic repetitive elements by mapping reads from 8,551 RNA-Seq datasets in human tissues including, adipose, adrenal gland, artery, bladder, brain, breast, lymphocytes, whole blood, fibroblasts, cervix, colon, esophagus, fallopian tube, heart, kidney, liver, lung, salivary gland, skeletal muscle, tibial nerve, ovary, pancreas, pituitary, prostate, skin, small intestine, stomach, testis, thyroid, uterus and vagina. Despite being accurately detected, both lncRNAs and repetitive elements were less expressed on average than protein-coding genes. This level of low expression is likely to result even more underestimated because lncRNA transcripts are preferentially localized in the nucleus, which may be obscured by the much larger number of cytoplasmic RNA transcripts.

Because signal from repetitive elements in many cases is likely to be weaker than from genes, as a consequence of their low level of activity, previous studies favored a strategy that assigned reads to repetitive element subfamilies as opposed to individual instances (Criscione et al., 2014; Djebali et al., 2016; Faulkner et al., 2009b; Fort et al., 2014b). However, by taking advantage of the sequencing depth of GTEx data, we have quantified all ~5 million instances of repetitive elements in the human genome.

Another limitation on the analysis of transcripts from repetitive regions is the nature of multi-mapping sequencing reads (read mapping in several localizations in the genome) (Göke and Ng, 2016; Mercer et al., 2009; Treangen and Salzberg, 2011). These reads are routinely omitted from further analysis, leading to experimental bias and reduced coverage. RNA-Seq reads that mapped to multiple genomic locations were mostly excluded from the main analysis of lncRNAs and repetitive elements. However, we did analyze the data assigning multi-mapped reads in a probabilistic or “fractional-count” manner to made sure that results were not biased by this problem.

Many of the bidirectional lncRNAs and enhancer-associated RNAs have been shown to be non-polyadenylated (T.-K. Kim et al., 2010). However, recent findings, along with our study (Chapters 2 and 3), suggest the existence of polyadenylated bidirectional transcripts and chromatin-associated RNAs (Djebali et al., 2016; Marques et al., 2013). In our thesis work, we used poly(A)-based RNA-Seq data in both mouse and human and because of this, we could be missing a large fraction of non-polyadenylated lncRNAs and also some repetitive elements.

Different methods have been used to reconstruct the transcripts of protein-coding genes and lncRNAs. However, the comparison of these methods revealed larger differences between the low abundance transcripts especially lncRNA transcripts (Beltran et al., 2008; Steijger et al., 2013; Treangen and Salzberg, 2011). This could be due to the differences in the algorithms that assemble low abundance transcripts (Cabili et al., 2011; Chu et al., 2015; Steijger et al., 2013). To avoid this problem, we applied both cufflinks and scripture methods on RNA-Seq data and considered the lncRNA transcripts that were built by both methods. However, interpretation of unannotated lncRNA transcripts assembled from RNA-Seq data should be cautious and should be subject to experimental validation. In our study, we experimentally validated the exact start and end exon splice junctions, and number of isoforms, using rapid amplification of cDNA ends (RACE). This transcriptome reconstruction could be improved by increase in sequence depth, computational methods validated by experimental methods.

Several others and we used around 5 million repeat instances from RepeatMasker where it provides exact genomic location and names of millions of repetitive elements (Cabili et al., 2011; Chu et al., 2011; Tarailo-Graovac and Chen, 2002). However, still there are still repeats of unknown families exist in human genome. Recently, several scientists who work on repetitive elements called for a standard benchmarking of repetitive elements across all the species. In future, this work would help accurate annotations of repetitive elements (Engreitz et al., 2013; Hoen et al., 2016; Simon et al., 2013; Tarailo-Graovac and Chen, 2002).

For long time it has been believed that the lncRNAs do not code any protein. However, a landmark study from Ingolia and colleagues in 2011 shown that lncRNAs have higher ribosome occupancy than 3' UTRs and raised many questions in the non-coding RNA field (Engreitz et al., 2013; Hoen et al., 2016; Ingolia et al., 2011; Simon et al., 2013). However, in the same study they have shown XIST, a classic

lncRNA that have been experimentally annotated as non-coding showed higher ribosome occupancy doubting the quality of their ribosome sequencing data. Supporting this, in 2013, Guttman and colleagues published another follow-up study showing that ribosome occupancy alone is not sufficient to predict the coding potential of a transcript and supporting the claim that lncRNAs are indeed non-coding (Guttman et al., 2013; Ingolia et al., 2011; Ule et al., 2005; Zhen Wang et al., 2010). Again in 2014, Ingolia and colleagues published their experimental findings indicating coding potential of several lncRNAs (Chi et al., 2009; Guttman et al., 2013; Hafner et al., 2010; Ingolia et al., 2014; Yeo et al., 2009). To make things even more complicated two different studies that capture the proteome of wide-range of human tissues catalogued hundreds of peptide signatures from lncRNAs (Ingolia et al., 2014; M.-S. Kim et al., 2014; McHugh et al., 2015; Wilhelm et al., 2015). However, it is now apparent that there are thousands of lncRNAs exist in mammals and the above ribosome profiling or proteomic studies however revealed very small fraction of them as coding. Most of the lncRNA-discovery based studies including our study only used computational filters that assess coding potential of the lncRNAs and these above findings raise the awareness of experimentally validating the true coding potential. In addition, proteome of repetitive elements is completely unexplored area so far. It would be interesting to check which repetitive families are translated, and which repeat family derived RNAs are noncoding.

LncRNAs are highly tissue-specific compared to protein-coding genes. This was previously shown by several studies based on differential expression between tissues and cell lines and also proposed that these tissue-specific lncRNAs could act as enhancers, promoting the transcription of neighboring genes (Cabili et al., 2011; Guttman et al., 2011; M.-S. Kim et al., 2014; Marques et al., 2013; Mercer and Mattick, 2013; Wilhelm et al., 2015; Ørom and Shiekhattar, 2011). In fact, in our mouse study, we validate the functional capacity of an enhancer lncRNA (*lncRNA-KDM8*) by using siRNA knockdown strategy and show that it positively regulates its neighboring a protein-coding gene (KDM8). Thousands of specific repetitive elements in human are also tissue-specific in nature suggesting the possibility of their functions. Interestingly, similar to lncRNAs, most of the tissue-specific repetitive elements belong to testis and brain.

Although a model of transcribed non-coding elements acting as functional elements is attractive, it is far from proven. Testing this hypothesis for lncRNAs and repetitive elements function will require more systematic studies. Taken together, this thesis

work along with recent reports has probably just begun to unravel the set of unexpected functions of non-coding elements. We anticipate extensive biological analyses as a consequence of this work, such as phenotypic effects of large-scale knockdown of lncRNA and repetitive elements expression using methods like CRISPR. Ultimately, lncRNAs and repetitive elements are a pervasive source of transcription and transcriptional regulation and therefore important to be considered in future genomic research.

Conclusions

By developing and applying computational pipelines to analyze RNA-Seq and ChIP-Seq data, we have been able to further characterize lncRNA in mammals. Our results suggest that lncRNAs and in particular transcripts from repetitive elements despite having low levels of expression are not negligible. We also demonstrated that a large proportion of such transcripts are tissue specific indicating a potential functional role in the cell. Overall, our work has resulted in a series of conclusions:

- We discovered a set of nove2,803 lncRNAs in mouse using published RNA-Seq data (mostly ENCODE) from 8-week-old adult brain, heart, kidney, small intestine, liver, spleen, testis, and thymus and an ES cell line.

- We reported that, 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 protein-coding transcripts, which (on average) have 10 exons, a length of 50,453 nucleotides, and expression levels of 4.68 FPKM.

- 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.

- We defined genome-wide chromatin states in mouse by using 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 one 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.

- In total, we identified 852 unique intergenic lncRNA transcripts associated with either an active promoter or a strong enhancer chromatin state, and named them as promoter-associated (plncRNA) and enhancer-associated lncRNA (elncRNA). We also showed that these lncRNAs switch their chromatin state from one tissue to another.

- To assess the regulatory potential of lncRNAs, we focused on an ES cell-specific enhancer-associated lncRNA located approximately 20 kb away from the protein-coding gene Kdm8 *(lncRNA-Kdm8*), which encodes a histone lysine demethylase and regulates embryonic cell proliferation. Using the RACE technique, we characterized the genomic structure of this lncRNAs and then knocked it down using two different siRNAs. As predicted, upon this lncRNA knockdown, expression of the Kdm8 gene significantly decreased.

- We reported a set of 205,779 repetitive elements as expressed by analyzing 8,551 poly-A RNA-Seq datasets from 53 distinct body sites across 544 human individuals. We also showed that 11,502 of them are intergenic in nature. Most of the expressed repetitive elements belong to retrotransposons (LTR, SINE and LINEs).

- Clustering of intergenic repeat expression recapitulates tissue-type. Using linear mixed model, we found that on average variation in repeat expression is far greater among tissues (~57%) than among individuals (~1%).

- We identified 3,295 unique repeat instances that are tissue-specific. We found 18 tissues with at least one location tissue-specific repeat instance and most of the tissue-specific repeat instances belong to cerebellum and testis tissues.

References:

Allen, T.A., Kaenel, Von, S., Goodrich, J.A., Kugel, J.F., 2004. The SINE-encoded mouse B2 RNA represses mRNA transcription in response to heat shock. Nat Struct Mol Biol 11, 816–821. doi:10.1038/nsmb813

An, H.J., Lee, D., Lee, K.H., Bhak, J., 2004. The association of Alu repeats with the generation of potential AU-rich elements (ARE) at 3' untranslated regions. BMC Genomics 5, 97. doi:10.1186/1471-2164-5-97

Athanasiadis, A., Rich, A., Maas, S., 2004. Widespread A-to-I RNA Editing of Alu-Containing mRNAs in the Human Transcriptome. Plos Biol 2, e391. doi:10.1371/journal.pbio.0020391.st001

Banfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W.E., Kundaje, A., Gunawardena, H.P., Yu, Y., Xie, L., Krajewski, K., Strahl, B.D., Chen, X., Bickel, P., Giddings, M.C., Brown, J.B., Lipovich, L., 2012. Long noncoding RNAs are rarely translated in two human cell lines. Genome Research 22, 1646–1657. doi:10.1101/gr.134767.111

Batista, P.J., Chang, H.Y., 2013. Long Noncoding RNAs: Cellular Address Codes in Development and Disease. Cell 152, 1298–1307. doi:10.1016/j.cell.2013.02.012

Belancio, V.P., Hedges, D.J., Deininger, P., 2008. Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health. Genome Research 18, 343–358. doi:10.1101/gr.5558208

Belancio, V.P., Roy-Engel, A.M., Deininger, P.L., 2010. All y“all need to know ”bout retroelements in cancer. Semin. Cancer Biol. 20, 200–210. doi:10.1016/j.semcancer.2010.06.001

Belgard, T.G., Marques, A.C., Oliver, P.L., Abaan, H.O., Sirey, T.M., Hoerder-Suabedissen, A., García-Moreno, F., Molnár, Z., Margulies, E.H., Ponting, C.P., 2011. NeuroResource. Neuron 71, 605–616. doi:10.1016/j.neuron.2011.06.039

Beltran, M., Puig, I., Pena, C., Garcia, J.M., Alvarez, A.B., Pena, R., Bonilla, F., de Herreros, A.G., 2008. A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. Genes & Development 22, 756–769. doi:10.1101/gad.455708

Bogu, G.K., Vizán, P., Stanton, L.W., Beato, M., Di Croce, L., Marti-Renom, M.A., 2016. Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse. Mol. Cell. Biol. 36, 809–819. doi:10.1128/MCB.00955-15

Bourque, G., 2009. Transposable elements in gene regulation and in the evolution of vertebrate genomes. Current Opinion in Genetics & Development 19, 607–612. doi:10.1016/j.gde.2009.10.013

Briggs, J.A., Wolvetang, E.J., Mattick, J.S., Rinn, J.L., Barry, G., 2015. Mechanisms of Long Non-coding RNAsin Mammalian Nervous System Development, Plasticity, Disease, and Evolution. Neuron 88, 861–877. doi:10.1016/j.neuron.2015.09.045

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., Rinn, J.L., 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & Development 25, 1915–1927. doi:10.1101/gad.17446611

Chi, S.W., Zang, J.B., Mele, A., Darnell, R.B., 2009. Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. Nature. doi:10.1038/nature08170

Chinwalla, A.T., Cook, L.L., Delehaunty, K.D., Fewell, G.A., Fulton, L.A., Fulton, R.S., Graves, T.A., Hillier, L.D.W., Mardis, E.R., McPherson, J.D., 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.

Chow, J.C., Ciaudo, C., Fazzari, M.J., Mise, N., Servant, N., Glass, J.L., Attreed, M., Avner, P., Wutz, A., Barillot, E., Greally, J.M., Voinnet, O., Heard, E., 2010.

LINE-1 Activity in Facultative Heterochromatin Formation during X Chromosome Inactivation. Cell 141, 956–969. doi:10.1016/j.cell.2010.04.042
Chu, C., Qu, K., Zhong, F.L., Artandi, S.E., Chang, H.Y., 2011. Genomic Maps of Long Noncoding RNA Occupancy Reveal Principles of RNA-Chromatin Interactions. Molecular Cell 44, 667–678. doi:10.1016/j.molcel.2011.08.027
Chu, C., Spitale, R.C., Chang, H.Y., 2015. Technologies to probe functions and mechanisms of long noncoding RNAs. Nat Struct Mol Biol 22, 29–35. doi:10.1038/nsmb.2921
Criscione, S.W., Zhang, Y., Thompson, W., Sedivy, J.M., Neretti, N., 2014. Transcriptional landscape of repetitive elements in normal and cancer human cells. BMC Genomics 15, 583. doi:10.1146/annurev.genet.35.102401.091032
Day, D.S., Luquette, L.J., Park, P.J., Kharchenko, P.V., 2010. Estimating enrichment of repetitive elements from high-throughput sequence data. Genome Biol 11, R69. doi:10.1186/gb-2010-11-6-r69
de Koning, A.P.J., Gu, W., Castoe, T.A., Batzer, M.A., Pollock, D.D., 2011. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. PLoS Genet. 7, e1002384. doi:10.1371/journal.pgen.1002384.s013
Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J.B., Lipovich, L., Gonzalez, J.M., Thomas, M., Davis, C.A., Shiekhattar, R., Gingeras, T.R., Hubbard, T.J., 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. doi:10.1101/gr.132159.111
Devaux, Y., Zangrando, J., Schroen, B., Creemers, E.E., Pedrazzini, T., Chang, C.-P., Dorn, G.W., Thum, T., Heymans, S., 2015. Long noncoding RNAs in cardiac developmentand ageing. Nature Publishing Group 1–11. doi:10.1038/nrcardio.2015.55
DGT, T.F.C.A.T.R.P.A.C., 2015. A promoter-level mammalian expression atlas. Nature 507, 462–470. doi:10.1038/nature13182
Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., 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, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A.M., 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, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., Gingeras, T.R., 2016. Landscape of transcription in human cells. Nature 488, 101–108. doi:10.1038/nature11233
Doolittle, W.F., 2013. Is junk DNA bunk? A critique of ENCODE. Proc. Natl. Acad. Sci. U.S.A. 110, 5294–5300. doi:10.1073/pnas.1221376110
Eddy, S.R., 2012. The C-value paradox, junk DNA and ENCODE. CURBIO 22, R898–R899. doi:10.1016/j.cub.2012.10.002
Elbarbary, R.A., Li, W., Tian, B., Maquat, L.E., 2013. STAU1 binding 3' UTR IRAlus complements nuclear retention to protect cells from PKR-mediated translational shutdown. Genes & Development 27, 1495–1510. doi:10.1101/gad.220962.113
Elbarbary, R.A., Lucas, B.A., Maquat, L.E., 2016. Retrotransposons as regulators of gene expression. Science 351, aac7247–aac7247. doi:10.1126/science.aac7247
Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C.,

Kadri, S., Xing, J., Goren, A., Lander, E.S., Plath, K., Guttman, M., 2013. The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. Science 341, 1237973–1237973. doi:10.1126/science.1237973

Estecio, M.R.H., Gallegos, J., Dekmezian, M., Lu, Y., Liang, S., Issa, J.P.J., 2012. SINE Retrotransposons Cause Epigenetic Reprogramming of Adjacent Gene Promoters. Molecular Cancer Research 10, 1332–1342. doi:10.1158/1541-7786.MCR-12-0351

Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K., Cloonan, N., Steptoe, A.L., Lassmann, T., Waki, K., Hornig, N., Arakawa, T., Takahashi, H., Kawai, J., Forrest, A.R.R., Suzuki, H., Hayashizaki, Y., Hume, D.A., Orlando, V., Grimmond, S.M., Carninci, P., 2009. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet 41, 563–571. doi:10.1038/ng.368

Feschotte, C., Jiang, N., Wessler, S.R., 2002. PLANT TRANSPOSABLE ELEMENTS: WHERE GENETICS MEETS GENOMICS. Nat Rev Genet 3, 329–341. doi:10.1038/nrg793

Fitzpatrick, T., Huang, S., 2014. 3'-UTR-located inverted Alu repeats facilitate mRNA translational repression and stress granule accumulation. Nucleus 3, 359–369. doi:10.4161/nucl.20827

Fort, A., Hashimoto, K., Yamada, D., Salimullah, M., Keya, C.A., Saxena, A., Bonetti, A., Voineagu, I., Bertin, N., Kratz, A., Noro, Y., Wong, C.-H., de Hoon, M., Andersson, R., Sandelin, A., Suzuki, H., Wei, C.-L., Koseki, H., Hasegawa, Y., Forrest, A.R.R., Carninci, P., 2014. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nat Genet 46, 558–566. doi:10.1038/ng.2965

Goodwin, S., McPherson, J.D., McCombie, W.R., 2016. Coming of age: ten years of next- generation sequencing technologies. Nat Rev Genet 17, 333–351. doi:10.1038/nrg.2016.49

Göke, J., Lu, X., Chan, Y.-S., Ng, H.-H., Ly, L.-H., Sachs, F., Szczerbinska, I., 2015. Dynamic Transcription of Distinct Classes of Endogenous Retroviral Elements Marks Specific Populations of Early Human Embryonic Cells. Stem Cell 16, 135–141. doi:10.1016/j.stem.2015.01.005

Göke, J., Ng, H.-H., 2016. CTRL+INSERT: retrotransposons and their contribution to regulation and innovation of the transcriptome. EMBO Rep 17, 1131–1144. doi:10.15252/embr.201642743

Grote, P., Wittler, L., Hendrix, D., Koch, F., Währisch, S., Beisaw, A., Macura, K., Bläss, G., Kellis, M., Werber, M., Herrmann, B.G., 2013. The Tissue-Specific lncRNA Fendrr Is an Essential Regulator of Heart and Body Wall Development in the Mouse. Developmental Cell 24, 206–214. doi:10.1016/j.devcel.2012.12.012

Gutschner, T., Hämmerle, M., Diederichs, S., 2013. MALAT1 — a paradigm for long noncoding RNA function in cancer. J Mol Med 91, 791–801. doi:10.1007/s00109-013-1028-y

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., Cabili, M.N., Jaenisch, R., Mikkelsen, T.S., Jacks, T., Hacohen, N., Bernstein, B.E., Kellis, M., Regev, A., Rinn, J.L., Lander, E.S., 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458, 223–227. doi:10.1038/nature07672

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., Yang, X., Amit, I., Meissner, A., Regev, A., Rinn, J.L., Root, D.E., Lander, E.S., 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477, 295–300. doi:10.1038/nature10398

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., Rinn, J.L., Lander, E.S., 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. doi:10.1038/nbt.1633
Guttman, M., Rinn, J.L., 2012. Modular regulatory principles of large non-coding RNAs. Nature 482, 339–346. doi:10.1038/nature10887
Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S., Lander, E.S., 2013. Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. Cell 154, 240–251. doi:10.1016/j.cell.2013.06.009
Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr, Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G.S., Dewell, S., Zavolan, M., Tuschl, T., 2010. Transcriptome-wide Identificationof RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. Cell 141, 129–141. doi:10.1016/j.cell.2010.03.009
Hancks, D.C., Kazazian, H.H., Jr, 2012. Active human retrotransposons: variation and disease. Current Opinion in Genetics & Development 22, 191–203. doi:10.1016/j.gde.2012.02.006
Ho, T.T., Zhou, N., Huang, J., Koirala, P., Xu, M., Fung, R., Wu, F., Mo, Y.Y., 2015. Targeting non-coding RNAs with the CRISPR/Cas9 system in human cell lines. Nucleic Acids Research 43, e17–e17. doi:10.1093/nar/gku1198
Hoen, D.R., Hickey, G., Bourque, G., Casacuberta, J., Cordaux, R., Feschotte, C., Fiston-Lavier, A.-S., Hua-Van, A., Hubley, R., Kapusta, A., Lerat, E., Maumus, F., Pollock, D.D., Quesneville, H., Smit, A., Wheeler, T.J., Bureau, T.E., Blanchette, M., 2016. A call for benchmarking transposable element annotation methods. Mobile DNA 1–9. doi:10.1186/s13100-015-0044-6
Huarte, M., 2015. The emerging role of lncRNAs in cancer. Nat Med 21, 1253–1261. doi:10.1038/nm.3981
Hung, T., Wang, Y., Lin, M.F., Koegel, A.K., Kotake, Y., Grant, G.D., Horlings, H.M., Shah, N., Umbricht, C., Wang, P., Wang, Y., Kong, B., Langerød, A., Børresen-Dale, A.-L., Kim, S.K., van de Vijver, M., Sukumar, S., Whitfield, M.L., Kellis, M., Xiong, Y., Wong, D.J., Chang, H.Y., 2011. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nat Genet 43, 621–629. doi:10.1038/ng.848
Ichiyanagi, K., 2013. Epigenetic regulation of transcription and possible functions of mammalian short interspersed elements, SINEs. Genes Genet. Syst. 88, 19–29.
Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., Talhouarne, G.J.S., Jackson, S.E., Wills, M.R., Weissman, J.S., 2014. Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. CellReports 8, 1365–1379. doi:10.1016/j.celrep.2014.07.045
Ingolia, N.T., Lareau, L.F., Weissman, J.S., 2011. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. Cell 147, 789–802. doi:10.1016/j.cell.2011.10.002
Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S.M., Wu, Y.-M., Robinson, D.R., Beer, D.G., Feng, F.Y., Iyer, H.K., Chinnaiyan, A.M., 2015. The landscape of long noncoding RNAs in the human transcriptome. Nature Publishing Group 1–13. doi:10.1038/ng.3192
Jia, H., Osak, M., Bogu, G.K., Stanton, L.W., Johnson, R., Lipovich, L., 2010. Genome-wide computational identification and manual annotation of human long noncoding RNA genes. RNA 16, 1478–1487. doi:10.1261/rna.1951310
Kaer, K., Speek, M., 2013. Retroelements in human disease. Gene 518, 231–241. doi:10.1016/j.gene.2013.01.008
Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermuller, J., Hofacker, I.L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., Sementchenko, V., Tammana, H., Gingeras, T.R., 2007. RNA Maps Reveal

New RNA Classes and a Possible Function for Pervasive Transcription. Science 316, 1484–1488. doi:10.1126/science.1138341
Kapusta, A., Feschotte, C., 2014. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. Trends Genet. 30, 439–452. doi:10.1016/j.tig.2014.08.004
Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., Feschotte, C., 2013. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. PLoS Genet. 9, e1003470. doi:10.1371/journal.pgen.1003470.s016
Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E., Dekker, J., Dunham, I., Elnitski, L.L., Farnham, P.J., Feingold, E.A., Gerstein, M., Giddings, M.C., Gilbert, D.M., Gingeras, T.R., Green, E.D., Guigo, R., Hubbard, T., Kent, J., Lieb, J.D., Myers, R.M., Pazin, M.J., Ren, B., Stamatoyannopoulos, J.A., Weng, Z., White, K.P., Hardison, R.C., 2014. Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. U.S.A. 111, 6131–6138. doi:10.1073/pnas.1318948111
Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., Thomas, J.K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N.A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L.D.N., Patil, A.H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S.K., Marimuthu, A., Sathe, G.J., Chavan, S., Datta, K.K., Subbannayya, Y., Sahu, A., Yelamanchi, S.D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K.R., Syed, N., Goel, R., Khan, A.A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T.-C., Zhong, J., Wu, X., Shaw, P.G., Freed, D., Zahari, M.S., Mukherjee, K.K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C.J., Shankar, S.K., Satishchandra, P., Schroeder, J.T., Sirdeshmukh, R., Maitra, A., Leach, S.D., Drake, C.G., Halushka, M.K., Prasad, T.S.K., Hruban, R.H., Kerr, C.L., Bader, G.D., Iacobuzio-Donahue, C.A., Gowda, H., Pandey, A., 2014. A draft map of the human proteome. Nature 509, 575–581. doi:10.1038/nature13302
Kim, T.-K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P.F., Kreiman, G., Greenberg, M.E., 2010. Widespread transcription at neuronal activity-regulated enhancers. Nature 465, 182–187. doi:10.1038/nature09033
Knoll, M., Lodish, H.F., Sun, L., 2015. Long non-coding RNAs as regulators of the endocrine system. Nature Publishing Group 1–10. doi:10.1038/nrendo.2014.229
Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., Bourque, G., 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet 42, 631–634. doi:10.1038/ng.600
Kutter, C., Watt, S., Stefflova, K., Wilson, M.D., Goncalves, A., Ponting, C.P., Odom, D.T., Marques, A.C., 2012. Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression. PLoS Genet. 8, e1002841. doi:10.1371/journal.pgen.1002841.s027
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., 2001. Initial sequencing and analysis of the human genome. Nature.
Lev-Maor, G., Sorek, R., Levanon, E.Y., Paz, N., Eisenberg, E., Ast, G., 2007. RNA-editing-mediated exon evolution. Genome Biol 8, R29. doi:10.1186/gb-2007-8-2-r29
Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J.A., Stewart, R., Dewey, C.N., 2014. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol 15, 663. doi:10.1186/s13059-014-0553-5
Lorenzen, J.M., Thum, T., 2016. Long noncoding RNAs in kidney and cardiovascular

diseases. Nature Publishing Group 1–14. doi:10.1038/nrneph.2016.51
Lu, X., Sachs, F., Ramsay, L., Jacques, P.-É., Göke, J., Bourque, G., Ng, H.-H., 2014. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat Struct Mol Biol 21, 423–425. doi:10.1038/nsmb.2799
Lunyak, V.V., Atallah, M., 2011. Genomic relationship between SINE retrotransposons, Pol III–Pol II transcription, and chromatin organization: the journey from junk to jewel. Biochem. Cell Biol. 89, 495–504. doi:10.1139/o11-046
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. doi:10.1371/journal.pone.0070835.s015
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. doi:10.1371/journal.pone.0071152.s012
Mariner, P.D., Walters, R.D., Espinoza, C.A., Drullinger, L.F., Wagner, S.D., Kugel, J.F., Goodrich, J.A., 2008. Human Alu RNA Is a Modular Transacting Repressor of mRNA Transcription during Heat Shock. Molecular Cell 29, 499–509. doi:10.1016/j.molcel.2007.12.013
Marinov, G.K., Wang, J., Handler, D., Wold, B.J., Weng, Z., Hannon, G.J., Aravin, A.A., Zamore, P.D., Brennecke, J., Toth, K.F., 2015. Pitfalls of Mapping High-Throughput Sequencing Data to Repetitive Sequences: Piwi's Genomic Targets Still Not Identified. Developmental Cell 32, 765–771. doi:10.1016/j.devcel.2015.01.013
Marques, A.C., Hughes, J., Graham, B., Kowalczyk, M.S., Higgs, D.R., Ponting, C.P., 2013. Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. Genome Biol 14, R131. doi:10.1186/gb-2013-14-11-r131
Martianov, I., Ramadass, A., Serra Barros, A., Chow, N., Akoulitchev, A., 2007. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. Nature 445, 666–670. doi:10.1038/nature05519
Martin, J.A., Wang, Zhong, 2011. Next-generation transcriptomeassembly. Nat Rev Genet 12, 671–682. doi:10.1038/nrg3068
Mattick, J.S., Rinn, J.L., 2015. Discovery and annotation oflong noncoding RNAs. Nat Struct Mol Biol 22, 5–7. doi:10.1038/nsmb.2942
McClintock, B., 1951. CHROMOSOME ORGANIZATION AND GENIC EXPRESSION. Cold Spring Harbor Symposia on Quantitative Biology 16, 13–47. doi:10.1101/SQB.1951.016.01.004
McHugh, C.A., Chen, C.-K., Chow, A., Surka, C.F., Tran, C., McDonel, P., Pandya-Jones, A., Blanco, M., Burghard, C., Moradian, A., Sweredoski, M.J., Shishkin, A.A., Su, J., Lander, E.S., Hess, S., Plath, K., Guttman, M., 2015. The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. Nature 521, 232–236. doi:10.1038/nature14443
Mele, M., Ferreira, P.G., Reverter, F., Deluca, D.S., 2015. The human transcriptome across tissues and individuals. …. doi:10.1126/science.1262110
Mercer, T.R., Dinger, M.E., Mattick, J.S., 2009. Long non-coding RNAs: insights into functions. Nat Rev Genet 10, 155–159. doi:10.1038/nrg2521
Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F., Mattick, J.S., 2008. Specific expression of long noncoding RNAs in the mouse brain. Proc. Natl. Acad. Sci. U.S.A. 105, 716–721. doi:10.1073/pnas.0706729105
Mercer, T.R., Mattick, J.S., 2013. Structure and function of long noncoding RNAs in epigenetic regulation. Nat Struct Mol Biol 20, 300–307. doi:10.1038/nsmb.2480
Meredith, E.K., Balas, M.M., Sindy, K., Haislop, K., Johnson, A.M., 2016. An RNA

matchmaker protein regulates the activity of the long noncoding RNA HOTAIR. RNA 22, 995–1010. doi:10.1261/rna.055830.115
Mirkin, S.M., 2007. Expandable DNA repeats and human disease. Nature 447, 932–940. doi:10.1038/nature05977
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, P.R., Berry, A., Lee, C., Harkins, T., Gmyr, V., Pattou, F., Kerr-Conte, J., Piemonti, L., Berney, T., Hanley, N., Gloyn, A.L., Sussel, L., Langman, L., Brayman, K.L., Sander, M., McCarthy, M.I., 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. doi:10.1016/j.cmet.2012.08.010
Morris, K.V., Mattick, J.S., 2014. PERSPECTIVES. Nat Rev Genet 15, 423–437. doi:10.1038/nrg3722
Mouse ENCODE Consortium, Stamatoyannopoulos, J.A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D.M., Groudine, M., Bender, M., Kaul, R., Canfield, T., Giste, E., Johnson, A., Zhang, M., Balasundaram, G., Byron, R., Roach, V., Sabo, P.J., Sandstrom, R., Stehling, A.S., Thurman, R.E., Weissman, S.M., Cayting, P., Hariharan, M., Lian, J., Cheng, Y., Landt, S.G., Ma, Z., Wold, B.J., Dekker, J., Crawford, G.E., Keller, C.A., Wu, W., Morrissey, C., Kumar, S.A., Mishra, T., Jain, D., Byrska-Bishop, M., Blankenberg, D., Lajoie, B.R., Jain, G., Sanyal, A., Chen, K.-B., Denas, O., Taylor, J., Blobel, G.A., Weiss, M.J., Pimkin, M., Deng, W., Marinov, G.K., Williams, B.A., Fisher-Aylor, K.I., DeSalvo, G., Kiralusha, A., Trout, D., Amrhein, H., Mortazavi, A., Edsall, L., McCleary, D., Kuan, S., Shen, Y., Yue, F., Ye, Z., Davis, C.A., 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, V.S., Cline, M.S., Kirkup, V.M., Learned, K., Rosenbloom, K.R., Kent, W.J., Feingold, E.A., Good, P.J., Pazin, M., Lowdon, R.F., Adams, L.B., 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol 13, 418. doi:10.1186/gb-2012-13-8-418
Necsulea, A., Kaessmann, H., 2014. Evolutionary dynamics of coding and non-coding transcriptomes. Nat Rev Genet 1–15. doi:10.1038/nrg3802
Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J.C., Grützner, F., Kaessmann, H., 2015. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505, 635–640. doi:10.1038/nature12943
Nigumann, P., Redik, K., Mätlik, K., Speek, M., 2002. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics 79, 628–634. doi:10.1006/geno.2002.6758
Palazzo, A.F., 2016. Non-coding RNA: what is functional and what is junk? 1–11. doi:10.3389/fgene.2015.00002/abstract
Perez-Pinera, P., Jones, M.F., Lal, A., Lu, T.K., 2015. Technology Preview. Molecular Cell 59, 146–148. doi:10.1016/j.molcel.2015.07.002
Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., Salzberg, S.L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33, 290–295. doi:10.1038/nbt.3122
Prensner, J.R., Iyer, M.K., Balbin, O.A., Dhanasekaran, S.M., Cao, Q., Brenner, J.C., Laxman, B., Asangani, I.A., Grasso, C.S., Kominsky, H.D., Cao, X., Jing, X., Wang, X., Siddiqui, J., Wei, J.T., Robinson, D., Iyer, H.K., Palanisamy, N., Maher, C.A., Chinnaiyan, A.M., 2011. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol 29, 742–749. doi:10.1038/nbt.1914
Prensner, J.R., Iyer, M.K., Sahu, A., Asangani, I.A., Cao, Q., Patel, L., Vergara, I.A., Davicioni, E., Erho, N., Ghadessi, M., Jenkins, R.B., Triche, T.J., Malik, R.,

Bedenis, R., McGregor, N., Ma, T., Chen, W., Han, S., Jing, X., Cao, X., Wang, X., Chandler, B., Yan, W., Siddiqui, J., Kunju, L.P., Dhanasekaran, S.M., Pienta, K.J., Feng, F.Y., Chinnaiyan, A.M., 2013. The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. Nat Genet 45, 1392–1398. doi:10.1038/ng.2771

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

Quinn, J.J., Chang, H.Y., 2016. Unique features of long non-coding RNA biogenesis and function. Nat Rev Genet 17, 47–62. doi:10.1038/nrg.2015.10

Ramos, A.D., Diaz, A., Nellore, A., Delgado, R.N., Park, K.-Y., Gonzales-Roybal, G., Oldham, M.C., Song, J.S., Lim, D.A., 2013. Integration of Genome-wide Approaches Identifies lncRNAs of Adult Neural Stem Cells and Their Progeny In Vivo. Cell Stem Cell 12, 616–628. doi:10.1016/j.stem.2013.03.003

Rinn, J.L., Chang, H.Y., 2012. Genome Regulation by Long Noncoding RNAs. Annu. Rev. Biochem. 81, 145–166. doi:10.1146/annurev-biochem-051410-092902

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., Chang, H.Y., 2007. Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. Cell 129, 1311–1323. doi:10.1016/j.cell.2007.05.022

Royo, H., Stadler, M.B., Peters, A.H.F.M., 2016. Alternative Computational Analysis Shows No Evidence for Nucleosome Enrichment at Repetitive Sequences in Mammalian Spermatozoa. Developmental Cell 37, 98–104. doi:10.1016/j.devcel.2016.03.010

Sahu, A., Singhal, U., Chinnaiyan, A.M., 2015. Long Noncoding RNAs inCancer: From Function toTranslation. TRENDS in CANCER 1–17. doi:10.1016/j.trecan.2015.08.010

Samans, B., Yang, Y., Krebs, S., Sarode, G.V., Blum, H., Reichenbach, M., Wolf, E., Steger, K., Dansranjavin, T., Schagdarsurengin, U., 2014. Uniformity of NucleosomePreservation Pattern in Mammalian Spermand Its Connection to Repetitive DNA Elements. Developmental Cell 30, 23–35. doi:10.1016/j.devcel.2014.05.023

Satpathy, A.T., Chang, H.Y., 2015. Long Noncoding RNA in Hematopoiesis and Immunity. Immunity 42, 792–804. doi:10.1016/j.immuni.2015.05.004

Schmitt, A.M., Chang, H.Y., 2016. Perspective. Cancer Cell 29, 452–463. doi:10.1016/j.ccell.2016.03.010

Sienski, G., Dönertas, D., Brennecke, J., 2012. Transcriptional Silencing of Transposons by Piwi and Maelstrom and Its Impact on Chromatin State and Gene Expression. Cell 151, 964–980. doi:10.1016/j.cell.2012.10.040

Simon, M.D., Pinter, S.F., Fang, R., Sarma, K., Rutenberg-Schoenberg, M., Bowman, S.K., Kesner, B.A., Maier, V.K., Kingston, R.E., Lee, J.T., 2013. High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. Nature 504, 465–469. doi:10.1038/nature12719

Slotkin, R.K., Martienssen, R., 2007. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8, 272–285. doi:10.1038/nrg2072

Smit, A., 1996. The origin of interspersed repeats in the human genome. Current Opinion in Genetics & Development.

Sorek, R., Ast, G., Graur, D., 2002. Alu-containing exons are alternatively spliced. Genome Research 12, 1060–1067. doi:10.1101/gr.229302

Spitale, R.C., Crisalli, P., Flynn, R.A., Torre, E.A., Kool, E.T., Chang, H.Y., 2012. RNA SHAPE analysis in living cells. Nat Chem Biol 9, 18–20. doi:10.1038/nchembio.1131

Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., Abril, J.F., Akerman, M., Alioto, T., Ambrosini, G., Antonarakis, S.E., Behr, J., Bertone, P., Bohnert, R., Bucher, P., Cloonan, N., Derrien, T., Djebali, S., Du, J., Dudoit, S., Engström, P.G., Gerstein, M., Gingeras, T.R., Gonzalez, D., Grimmond, S.M., Guigó, R.,

Habegger, L., Harrow, J., Hubbard, T.J., Iseli, C., Jean, G., Kahles, A., Kokocinski, F., Lagarde, J., Leng, J., Lefebvre, G., Lewis, S., Mortazavi, A., Niermann, P., Rätsch, G., Reymond, A., Ribeca, P., Richard, H., Rougemont, J., Rozowsky, J., Sammeth, M., Sboner, A., Schulz, M.H., Searle, S.M.J., Solorzano, N.D., Solovyev, V., Stanke, M., Steijger, T., Stevenson, B.J., Stockinger, H., Valsesia, A., Weese, D., White, S., Wold, B.J., Wu, J., Wu, T.D., Zeller, G., Zerbino, D., Zhang, M.Q., Hubbard, T.J., Guigó, R., Harrow, J., Bertone, P., 2013. Assessment of transcript reconstruction methods for RNA-seq. Nat Meth 10, 1177–1184. doi:10.1038/nmeth.2714

Sun, L., Goff, L.A., Trapnell, C., Alexander, R., Lo, K.A., Hacisuleyman, E., Sauvageau, M., Tazon-Vega, B., Kelley, D.R., Hendrickson, D.G., Yuan, B., Kellis, M., Lodish, H.F., Rinn, J.L., 2013. Long noncoding RNAs regulate adipogenesis. Proc. Natl. Acad. Sci. U.S.A. 110, 3387–3392. doi:10.1073/pnas.1222643110

Sutherland, G.R., Richards, R.I., 1995. Simple tandem DNA repeats and human genetic disease., in:. Presented at the Proceedings of the National ….

Tarailo-Graovac, M., Chen, N., 2002. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. John Wiley & Sons, Inc., Hoboken, NJ, USA. doi:10.1002/0471250953.bi0410s25

Tashiro, K., Teissier, A., Kobayashi, N., Nakanishi, A., Sasaki, T., Yan, K., Tarabykin, V., Vigier, L., Sumiyama, K., Hirakawa, M., Nishihara, H., Pierani, A., Okada, N., 2011. A Mammalian Conserved Element Derived from SINE Displays Enhancer Properties Recapitulating Satb2 Expression in Early-Born Callosal Projection Neurons. PLoS ONE 6, e28497. doi:10.1371/journal.pone.0028497.s007

The FANTOM Consortium, 2005. The Transcriptional Landscape of the Mammalian Genome. Science 309, 1559–1563. doi:10.1126/science.1112014

The GTEx Consortium, Ardlie, K.G., Deluca, D.S., Segre, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., Lek, M., Ward, L.D., Kheradpour, P., Iriarte, B., Meng, Y., Palmer, C.D., Esko, T., Winckler, W., Hirschhorn, J.N., Kellis, M., MacArthur, D.G., Getz, G., Shabalin, A.A., Li, G., Zhou, Y.H., Nobel, A.B., Rusyn, I., Wright, F.A., Lappalainen, T., Ferreira, P.G., Ongen, H., Rivas, M.A., Battle, A., Mostafavi, S., Monlong, J., Sammeth, M., Mele, M., Reverter, F., Goldmann, J.M., Koller, D., Guigo, R., McCarthy, M.I., Dermitzakis, E.T., Gamazon, E.R., Im, H.K., Konkashbaev, A., Nicolae, D.L., Cox, N.J., Flutre, T., Wen, X., Stephens, M., Pritchard, J.K., Tu, Z., Zhang, B., Huang, T., Long, Q., Lin, L., Yang, J., Zhu, J., Liu, J., Brown, A., Mestichelli, B., Tidwell, D., Lo, E., Salvatore, M., Shad, S., Thomas, J.A., Lonsdale, J.T., Moser, M.T., Gillard, B.M., Karasik, E., Ramsey, K., Choi, C., Foster, B.A., Syron, J., Fleming, J., Magazine, H., Hasz, R., Walters, G.D., Bridge, J.P., Miklos, M., Sullivan, S., Barker, L.K., Traino, H.M., Mosavel, M., Siminoff, L.A., Valley, D.R., Rohrer, D.C., Jewell, S.D., Branton, P.A., Sobin, L.H., Barcus, M., Qi, L., McLean, J., Hariharan, P., Um, K.S., Wu, S., Tabor, D., Shive, C., Smith, A.M., Buia, S.A., Undale, A.H., Robinson, K.L., Roche, N., Valentino, K.M., Britton, A., Burges, R., Bradbury, D., Hambright, K.W., Seleski, J., Korzeniewski, G.E., Erickson, K., Marcus, Y., Tejada, J., Taherian, M., Lu, C., Basile, M., Mash, D.C., Volpi, S., Struewing, J.P., Temple, G.F., Boyer, J., Colantuoni, D., Little, R., Koester, S., Carithers, L.J., Moore, H.M., Guan, P., Compton, C., Sawyer, S.J., Demchok, J.P., Vaught, J.B., Rabiner, C.A., Lockhart, N.C., Ardlie, K.G., Getz, G., Wright, F.A., Kellis, M., Volpi, S., Dermitzakis, E.T., 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348, 648–660. doi:10.1126/science.1262110

Tian, D., Sun, S., Lee, J.T., 2010. The Long Noncoding RNA, Jpx, Is a Molecular Switch for X Chromosome Inactivation. Cell 143, 390–403. doi:10.1016/j.cell.2010.09.049

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H.,

Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578. doi:10.1038/nprot.2012.016
Treangen, T.J., Salzberg, S.L., 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. doi:10.1038/nrg3117
Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y., Pan, Q., Watt, A.T., Freier, S.M., Bennett, C.F., Sharma, A., Bubulya, P.A., Blencowe, B.J., Prasanth, S.G., Prasanth, K.V., 2010. The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation. Molecular Cell 39, 925–938. doi:10.1016/j.molcel.2010.08.011
Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., Chang, H.Y., 2010. Long noncoding RNA as modular scaffold of histone modification complexes. Science 329, 689–693. doi:10.1126/science.1192002
Ugarkovic, D., 2005. Functional elements residing within satellite DNAs. EMBO Rep 6, 1035–1039. doi:10.1038/sj.embor.7400558
Ugarkovic, D., Plohl, M., 2002. Variation in satellite DNA profiles--causes and effects. EMBO J 21, 5955–5959.
Ule, J., Jensen, K., Mele, A., Darnell, R.B., 2005. CLIP: A method for identifying protein–RNA interaction sites in living cells. Methods 37, 376–386. doi:10.1016/j.ymeth.2005.07.018
Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., Bartel, D.P., 2011. Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution. Cell 147, 1537–1550. doi:10.1016/j.cell.2011.11.055
Usdin, K., 2008. The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases. Genome Research 18, 1011–1019. doi:10.1101/gr.070409.107
van Bakel, H., Nislow, C., Blencowe, B.J., Hughes, T.R., 2011. Response to “The Reality of Pervasive Transcription.” Plos Biol 9, e1001102. doi:10.1371/journal.pbio.1001102.g001
van de Vondervoort, I.I.G.M., 2013. Long non-coding RNAs in neurodevelopmental disorders 1–9. doi:10.3389/fnmol.2013.00053/abstract
Wang, Kevin C, Chang, H.Y., 2011. Molecular Mechanisms of Long Noncoding RNAs. Molecular Cell 43, 904–914. doi:10.1016/j.molcel.2011.08.018
Wang, Kevin C, Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., Wysocka, J., Lei, M., Dekker, J., Helms, J.A., Chang, H.Y., 2011. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472, 120–124. doi:10.1038/nature09819
Wang, Zhen, Kayikci, M., Briese, M., Zarnack, K., Luscombe, N.M., Rot, G., Zupan, B., Curk, T., Ule, J., 2010. iCLIP Predicts the Dual Splicing Effects of TIA-RNA Interactions. Plos Biol 8, e1000530. doi:10.1371/journal.pbio.1000530.s015
Wang, Zhong, Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10, 57–63. doi:10.1038/nrg2484
Washietl, S., Kellis, M., Garber, M., 2014. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. Genome Research 24, 616–628. doi:10.1101/gr.165035.113
Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J.-H., Bantscheff, M., Gerstmair, A., Faerber, F., Kuster, B., 2015. Mass-spectrometry-based draft of the human proteome. Nature 509, 582–587. doi:10.1038/nature13319
Yap, K.L., Li, S., Muñoz-Cabello, A.M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J.,

Walsh, M.J., Zhou, M.-M., 2010. Molecular Interplay of the Noncoding RNA ANRIL and Methylated Histone H3 Lysine 27 by Polycomb CBX7 in Transcriptional Silencing of INK4a. Molecular Cell 38, 662–674. doi:10.1016/j.molcel.2010.03.021

Yeo, G.W., Coufal, N.G., Liang, T.Y., Peng, G.E., Fu, X.-D., Gage, F.H., 2009. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nat Struct Mol Biol 16, 130–137. doi:10.1038/nsmb.1545

Young, R.S., Marques, A.C., Tibbit, C., Haerty, W., Bassett, A.R., Liu, J.L., Ponting, C.P., 2012. Identification and Properties of 1,119 Candidate LincRNA Loci in the Drosophila melanogaster Genome. Genome Biology and Evolution 4, 427–442. doi:10.1093/gbe/evs020

Ørom, U.A., Shiekhattar, R., 2011. Noncoding RNAs and enhancers: complications of a long-distance relationship. Trends in Genetics 27, 433–439. doi:10.1016/j.tig.2011.06.009