SCIENCE CHINA
Life Sciences

• RESEARCH PAPER •

June 2016 Vol.59 No.6: 589–603
doi: 10.1007/s11427-015-4929-x

Identification and analysis of mouse non-coding RNA using transcriptome data

Yuhui Zhao[1,2†], Wanfei Liu[1,3†], Jingyao Zeng[1,2†], Shoucheng Liu[1,2], Xinyu Tan[1], Hasanawad Aljohi[3] & Songnian Hu[1*]

[1]*CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China;*
[2]*University of Chinese Academy of Sciences, Beijing 100049, China;*
[3]*Joint Center for Genomics Research (JCGR), King Abdulaziz City for Science and Technology and Chinese Academy of Sciences, Riyadh 11442, Saudi Arabia*

Received September 23, 2015; accepted October 22, 2015; published online March 3, 2016

Transcripts are expressed spatially and temporally and they are very complicated, precise and specific; however, most studies are focused on protein-coding related genes. Recently, massively parallel cDNA sequencing (RNA-seq) has emerged to be a new and promising tool for transcriptome research, and numbers of non-coding RNAs, especially lincRNAs, have been widely identified and well characterized as important regulators of diverse biological processes. In this study, we used ultra-deep RNA-seq data from 15 mouse tissues to study the diversity and dynamic of non-coding RNAs in mouse. Using our own criteria, we identified totally 16,249 non-coding genes (21,569 non-coding RNAs) in mouse. We annotated these non-coding RNAs by diverse properties and found non-coding RNAs are generally shorter, have fewer exons, express in lower level and are more strikingly tissue-specific compared with protein-coding genes. Moreover, these non-coding RNAs show significant enrichment with transcriptional initiation and elongation signals including histone modifications (H3K4me3, H3K27me3 and H3K36me3), RNAPII binding sites and CAGE tags. The gene set enrichment analysis (GSEA) result revealed several sets of lincRNAs associated with diverse biological processes such as immune effector process, muscle development and sexual reproduction. Taken together, this study provides a more comprehensive annotation of mouse non-coding RNAs and gives an opportunity for future functional and evolutionary study of mouse non-coding RNAs.

non-coding RNA, RNA-seq, transcriptome, lincRNA, mouse

Citation: Zhao, Y., Liu, W., Zeng, J., Liu, S., Tan, X., Aljohi, H.A., and Hu, S. (2016). Identification and analysis of mouse non-coding RNA using transcriptome data. Sci China Life Sci 59, 589−603. doi: 10.1007/s11427-015-4929-x

INTRODUCTION

The previous studies demonstrated that mammalian genomes are pervasively transcribed (Clark et al., 2011; The ENCODE Project Consortium, 2007). For example, more than 80% of human genome are transcribed while ~60% in mouse genome (Carninci et al., 2005; Djebali et al., 2012; Katayama et al., 2005; Okazaki et al., 2002). A recent research further confirmed the active transcription of intron and intergenic regions in mouse genome by analyzing a collection of over 1,000 transcriptome data sets from 123 cell types and primary tissues (Yue et al., 2014). The mammalian genome not only transcribes into mRNAs, but also gives rise to a large amount of non-coding RNAs (Carninci et al., 2005; Guttman et al., 2009; Sati et al., 2012; Zheng et al., 2007). In recent years, thousands of

†Contributed equally to this work
*Corresponding author (email: husn@big.ac.cn)

non-coding RNAs have been identified and they play important roles in various processes including imprinting, X-inactivation, cell cycle and development processes especially in regulation of pluripotency (Brown et al., 1992; Dinger et al., 2008; Guttman et al., 2011; Hawkins and Morris, 2010; Heard and Disteche, 2006; Hu et al., 2012; Pauli et al., 2011, 2012; Yang and Kuroda, 2007).

Recent advances in massively parallel sequencing of RNA (RNA-seq) based on the next-generation sequencing technologies provide an unprecedented method to unbiased identify non-coding RNAs, especially those expressed in a tissue-specific manner with low abundance in mammals (Cloonan et al., 2008; Haas and Zody, 2010; Yassour et al., 2009). RNA-seq has substantially increased the throughput of sequencing which subsequently facilitated the measurement of transcript abundance in practice. Compared with previous sequencing approaches such as serial analysis of gene expression (SAGE) and expressed sequence tag (EST), RNA-seq is capable of capturing almost all transcripts. Moreover, ribo-minus RNA-seq (rmRNA-seq) has been considered to be more accurate and comprehensive in transcriptome profiling than polyadenylated RNA-seq (Cui et al., 2010).

With the rapid growth of RNA-seq data, non-coding RNAs especially lincRNAs are identified rapidly. Thousands of lincRNAs are found in mouse (Guttman et al., 2009, 2010; Liu et al., 2011; Luo et al., 2013; Sigova et al., 2013), human (Cabili et al., 2011; Hangauer et al., 2013; Sigova et al., 2013), chimpanzee (Wetterbom et al., 2010), zebrafish (Pauli et al., 2012; Ulitsky et al., 2011), frog (Tan et al., 2013), nematode (Nam and Bartel, 2012) and *Arabidopsis* (Liu et al., 2012). Some newly identified lincRNAs have well-validated functions. Linc-HOXA1 represses the expression of *Hoxa1* (homeobox A1) by recruiting the protein PURB (purine-rich element binding protein B) as a transcriptional cofactor (Maamar et al., 2013). A class of lincRNAs called ncRNA-activating (ncRNA-a) have enhancer-like function and positively regulate expression of neighboring protein-coding genes (Ørom et al., 2010). *HOTAIR* (HOX transcript antisense RNA) silences transcription across 40 kb of the *HOXD* (homeobox D cluster) locus in trans by inducing a repressive chromatin state, which is proposed to occur by recruitment of the polycomb chromatin remodeling complex PRC2 (polycomb repressive complex 2) by *HOTAIR* (Gupta et al., 2010; Rinn et al., 2007). In summary, lincRNAs perform myriad functions through various mechanisms ranging from the regulation of epigenetic modification and gene expression to acting as scaffolds for protein signaling complexes (Mattick, 2009; Mercer et al., 2009; Sasaki et al., 2009; Wang and Chang, 2011).

As recently reported, most splicing events are highly tissue-specific (Brown et al., 2014) and lncRNAs have restricted expression pattern in specific tissues (Derrien et al., 2012). In order to illustrate the diversity and dynamics of non-coding RNAs in mouse, we developed a pipeline to identify non-coding RNAs in mouse. Totally, 16,249 non-coding genes were found using RNA-seq data of 15 tissues (of which 14 are publicly available mRNA-seq data sets from the encyclopedia of DNA elements (ENCODE), and the other one is sequenced by our own lab using rmRNA-seq method) which covered most of mouse tissue types. These non-coding RNAs have the general features as lower expression, more tissue-specificity and less conservative than protein-coding genes. Our results expand the collection of non-coding RNAs in mouse and provide an important resource for the study of functional elements in mouse. We expect to provide a meaningful new insight into this gene category.

RESULTS

Identification of novel mouse non-coding RNAs from 15 tissues

To comprehensively explore the mouse non-coding RNAs, we developed a pipeline to identify the non-coding RNAs from 15 tissues (Figure 1). First, we downloaded RNA-seq data of 14 mouse tissues (including Adrenal gland, Colon, Heart, Kidney, Large intestine, Liver, Lung, Mammary gland, Ovary, Small intestine, Spleen, Stomach, Testis and Thymus) from NCBI and sequenced mouse cerebrum in our lab by Illumina HiSeq 2000. In total, we obtained 2.28 billion strand-specific pair-end fragments (178.74 Gb, 65.58-fold genome coverage). After pre-processing of the raw data, all the high-quality data were aligned to the mouse genome (mm10) by GSNAP and we got 1.86 billion mapped fragments (142.97 Gb, 52.46-fold genome coverage) (Table 1). Then we constructed transcript models for each tissue by Cufflinks using the mappable fragments. After that, we integrated these assembled transcript models in 15 tissues by Cuffmerge and finally got 75,749 loci and 44,420 loci remained after filtering loci completely overlapping with the genes of RefSeq, UCSC and Ensembl. Those filtered loci accounted for 79.7% of all known genes. Next, 32,862 loci were marked as non-coding while 730 loci were coding due to the coding potential predicted by CPC and CPAT. Moreover, we defined an expressed gene who must have a RPKM value larger than 0.1 in at least one tissue due to the intersection between false positive rate (FPR) and false negative rate (FNR) (Ramskold et al., 2009) ranging from 0.05 to 0.13 in all 15 tissues (Figure S1 in Supporting Information). Subsequently, 31,066 expressed non-coding genes and 722 coding genes were selected. To get credible single-exon non-coding genes, we eliminated those single-exon genes without any support from EST data and Rfam. In the end, we obtained 16,249 expressed non-coding genes (Dataset S1 in Supporting Information) and 722 expressed protein-coding genes respectively. Based

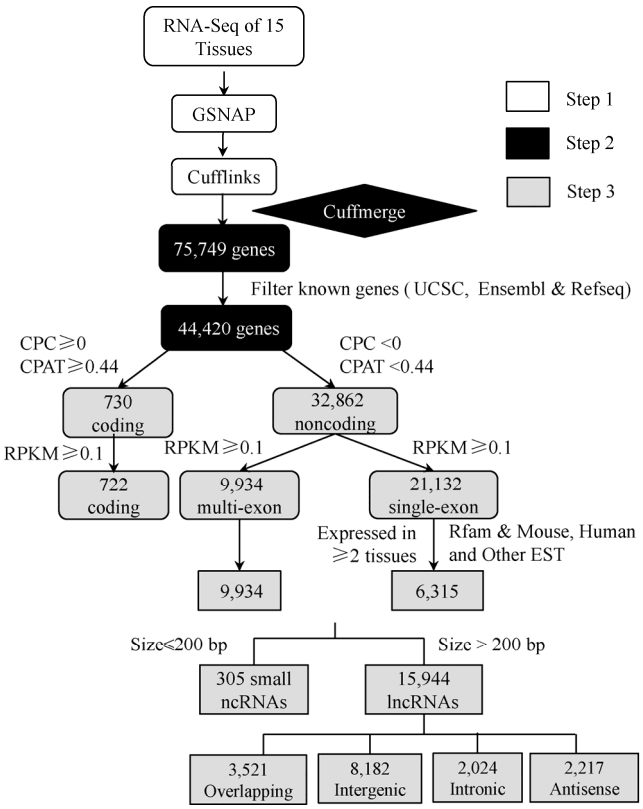

Figure 1 Overview of the computational pipeline for identification of non-coding genes. For identifying non-coding genes across 15 mouse tissues, GSNAP and Cufflinks were used for mapping and transcriptome assembly; CPC and CPAT were used to distinguish non-coding from numerous coding genes. After filtered single-exon non-coding genes unmapped to any data (Rfam and EST data), we finally generated 16,249 non-coding genes.

on their location with respect to known genes' annotation, we further classified these non-coding genes into four sub classes: overlapping with annotation genes, intergenic genes, intronic genes (Nakaya et al., 2007) and antisense genes, which accounted for 23.06%, 50.50%, 12.66% and 13.78% of all genes, respectively (Table 2). We recompiled gene IDs with distinct prefixes: XLOC, INTE, INTR and ANTI for overlapping genes, intergenic genes, intronic genes and antisense genes, respectively.

Genomic features of non-coding genes

For studying the genomic features of non-coding genes, we compared non-coding genes with protein-coding genes. First, non-coding genes had fewer exons than protein-coding genes (2.4 exons for non-coding genes and 9.7 exons for protein-coding genes on average). Then, we found that non-coding genes were generally smaller than protein-coding genes (mean length is 11.2 and 42.3 kb for non-coding genes and protein-coding genes). These properties were both consistent with the study by Pauli et al. and Cabili et al. (Pauli et al., 2012; Cabili et al., 2011). Moreover, we were interested in whether the smaller size of non-coding genes was caused by their fewer exons. We thus compared the mean length of genes, exons and introns between non-coding and protein-coding genes which had the equal number of exons, and amazedly found that the non-coding genes had larger gene and intron length, but smaller exon length comparing to protein-coding genes (Figure 2A–C). The result suggested that the relatively smaller size of non-coding genes compared with protein-coding genes was mainly due to their fewer exon numbers and smaller size of exons. Further analysis showed that non-coding genes had more repeats in introns than protein-coding genes (64.7% introns of non-coding genes with repeat vs. 53.8% introns of protein-coding genes; 6.96 repeat elements on average per intron in non-coding genes vs. 3.93 in protein-coding genes). The vast majority of repeats in introns were short interspersed element (SINE), simple repeat, long terminal repeat (LTR) and long interspersed nuclear elements (LINE) (Figure 2D). Although the repeat types were similar, both proportions of number and length

Table 1 The summary of RNA-seq data[a)]

Tissue	sequence read archive (SRA) accession No.	Original fragments	High quality fragments	Percent	Mapping fragments	Total mapping percent
Adrenal gland	SRX135155	148,292,125	139,090,674	93.80%	127,574,606	91.72%
Colon	SRX135165	131,005,753	127,438,807	97.28%	111,576,426	93.50%
Heart	SRX135166	155,581,190	149,271,569	95.94%	119,156,550	94.85%
Kidney	SRX135161	211,079,100	204,777,760	97.01%	190,296,905	92.93%
Large intestine	SRX135156	148,616,147	145,407,725	97.84%	134,480,910	92.49%
Liver	SRX135162	162,688,171	158,919,893	97.68%	150,081,337	94.44%
Lung	SRX135163	133,066,159	128,994,546	96.94%	118,492,917	91.86%
Mammary gland	SRX135151	147,002,618	140,730,296	95.73%	128,854,541	91.56%
Ovary	SRX135150	105,677,057	101,955,077	96.48%	98,032,852	96.16%
Small intestine	SRX135153	138,056,928	134,988,269	97.78%	123,869,804	91.76%
Spleen	SRX135164	152,594,296	148,439,097	97.28%	79,641,990	93.39%
Stomach	SRX135152	168,550,565	158,880,743	94.28%	140,181,599	88.23%
Testis	SRX135160	150,763,638	147,359,308	97.74%	138,020,994	93.66%
Thymus	SRX135159	117,387,927	114,297,797	97.37%	103,907,569	90.91%
Cerebrum*	SRX806806	211,809,646	107,740,712	50.87%	65,484,766	60.78%

a) *, this data was sequenced by our lab.

Table 2 Classification of non-coding genes[a)]

Category	Overlapping	Antisense	Intergenic	Intronic	Total
Non-coding genes	3,747(5,249)	2,239(3,118)	8,206(11,046)	2,057(2,156)	16,249(21,569)
lncRNA genes	3,521(5,023)	2,217(3,096)	8,182(11,022)	2,024(2,123)	15,944(21,264)

a) We classified non-coding genes into four subclasses, which are overlapping genes, intergenic genes, intronic genes and antisense genes. The first number is identified gene number and the second number is number of transcripts derived from the corresponding genes.

for LTR and LINE in non-coding genes were higher than protein-coding genes (Figure 2D and E). When we took the sequence length into consideration, LTR and LINE were longer than other repeat elements (Figure S2 in Supporting Information). To further confirm whether LTR and LINE were responsible for the large intron of non-coding genes, we compared introns with LTR and LINE to introns without LTR and LINE. Then, it was observed that introns with LTR and LINE were obviously longer than introns without LTR and LINE (Figure S3 in Supporting Information). Thus, it indicated that higher proportion of LTR and LINE in non-coding genes contributed to the larger intron, and further led to the longer gene length comparing to protein-coding genes with same exon number. We also compared the gene length among our four subclasses of non-coding genes overall, which indicated that overlapping non-coding genes were the longest, followed by antisense RNA. The most interesting thing was that the length of intergenic non-coding genes seemingly was comparable to intronic non-coding genes (Figure S4A in Supporting Information). To further explore the length difference of them, we did the length comparison with same exon number genes and found that the overlapping non-coding genes and intronic non-coding genes had the shortest and the second shortest length in single exon level, while the intergenic non-coding genes had the shortest length in multiple exon level, which may explain the relatively shorter first peak and longer last peak for overlapping non-coding genes and intronic non-coding genes (Figure S4 in Supporting Information). In addition, we also found that the GC content of non-coding genes was comparable to protein-coding genes.

Characterization of transcriptional activity and conservation of non-coding genes

Previous studies indicated that both protein-coding genes and non-coding genes were associated with diverse chromatin signals (Djebali et al., 2012; Lv et al., 2013). H3K4me3 and H3K27me3 are known to mark promoters for activating and repressing gene expression respectively (Li et al., 2012), while H3K36me3 generally marks the elongation sites of transcribed regions (Lv et al., 2013). So we examined these related transcriptional element of all non-coding RNAs. The results showed a significant enrichment of H3K4me3 and H3K27me3 around TSSs. As for H3K36me3, it was obvious that the gene body was associated with enriched H3K36me3 marks in a higher degree than the other two parts (upstream and downstream) (Figure 3A–C).

Furthermore, we explored the distribution of CAGE tags and RNAPII binding sites around TSS. We observed the obvious peaks at the vicinity of TSSs for both CAGE and RNAPII tags (Figure 3D and E). It suggested that the significant enrichment of these two marks around promoter regions of non-coding genes was consistent with the fact that CAGE was developed to map promoters rather than other genomic regions (Faulkner et al., 2008). A list of non-coding genes supported by at least one of histone methylation, RNAPII and CAGE was shown in Dataset S2.

To assess the evolutionary conservation level of non-coding RNAs, we calculated the conservation scores for the individual exon as well as promoter among non-coding genes, protein-coding genes and random intergenic regions. Consequently, Figure 4 showed that non-coding RNAs' exons had lower conservation scores than protein-coding exons. The random regions had the lowest conservation scores among these three classes, and it has probably reflected a lower degree of constraint on non-coding RNAs structures than on amino-acid codons, as shown in previous study (Guttman et al., 2009). Unlike the conservation levels of exons, promoter conservation of non-coding genes was comparable to protein-coding genes. In addition, we compared the conservation levels between lncRNAs and small RNAs and found lncRNAs were more conserved than small RNAs. Taken together, we concluded that non-coding genes have significant conservation constraint than random regions and these results may provide a meaningful starting point for their further functional studies in mammals.

These series of evidences have illuminated that our non-coding genes have their own transcript indicators, can be transcribed independently and are evolutionarily conserved, which suggested that these non-coding RNAs are functional and not merely transcriptional noise as some previous studies reported (Hüttenhofer and Vogel, 2006).

Expression level and tissue-specificity of non-coding genes

To investigate the expression dynamics of non-coding genes, we calculated the RPKM value of each gene in different tissues. We found that non-coding genes were expressed at lower levels than protein-coding genes (Table 3). First, the mean RPKM value of protein-coding genes is about eight times than non-coding genes (17.18 for protein-coding genes vs. 2.18 for non-coding genes). Then, the

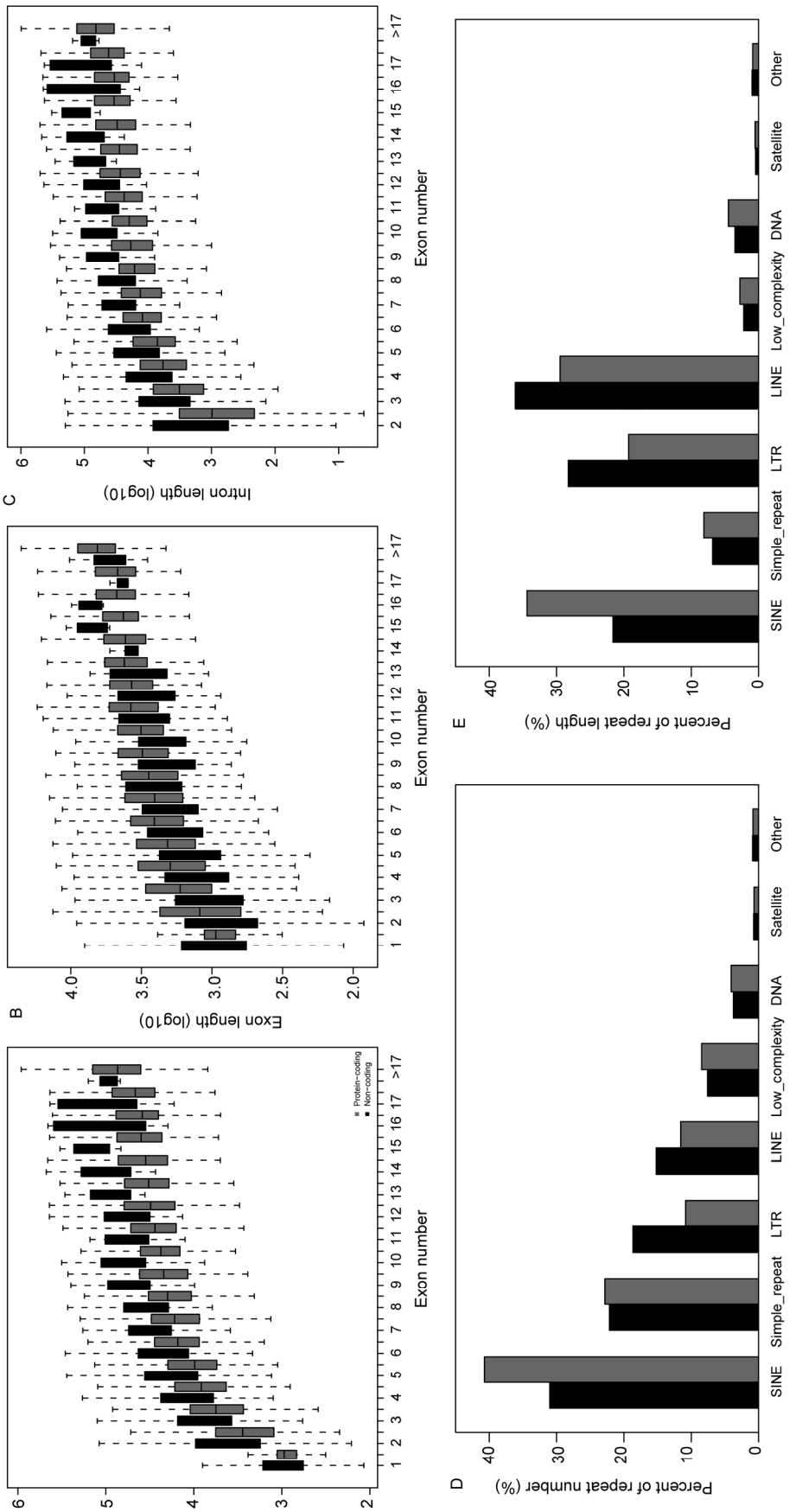

Figure 2 Length and repeat comparison between non-coding genes and protein-coding genes. A–C, Gene length, intron length and exon length are compared between non-coding and protein-coding genes who have the same number of exon. D, Percentage of repeat number in eight types. The main types of repeat elements in introns of non-coding genes were similar to that of protein-coding genes. Introns of non-coding genes had more LTR and LINE than protein-coding genes while there were more SINE in introns of protein-coding genes than non-coding genes. E, Percentage of repeat length in eight types. Similarly, LTR and LINE contributed more to the repeat length in introns of non-coding genes.

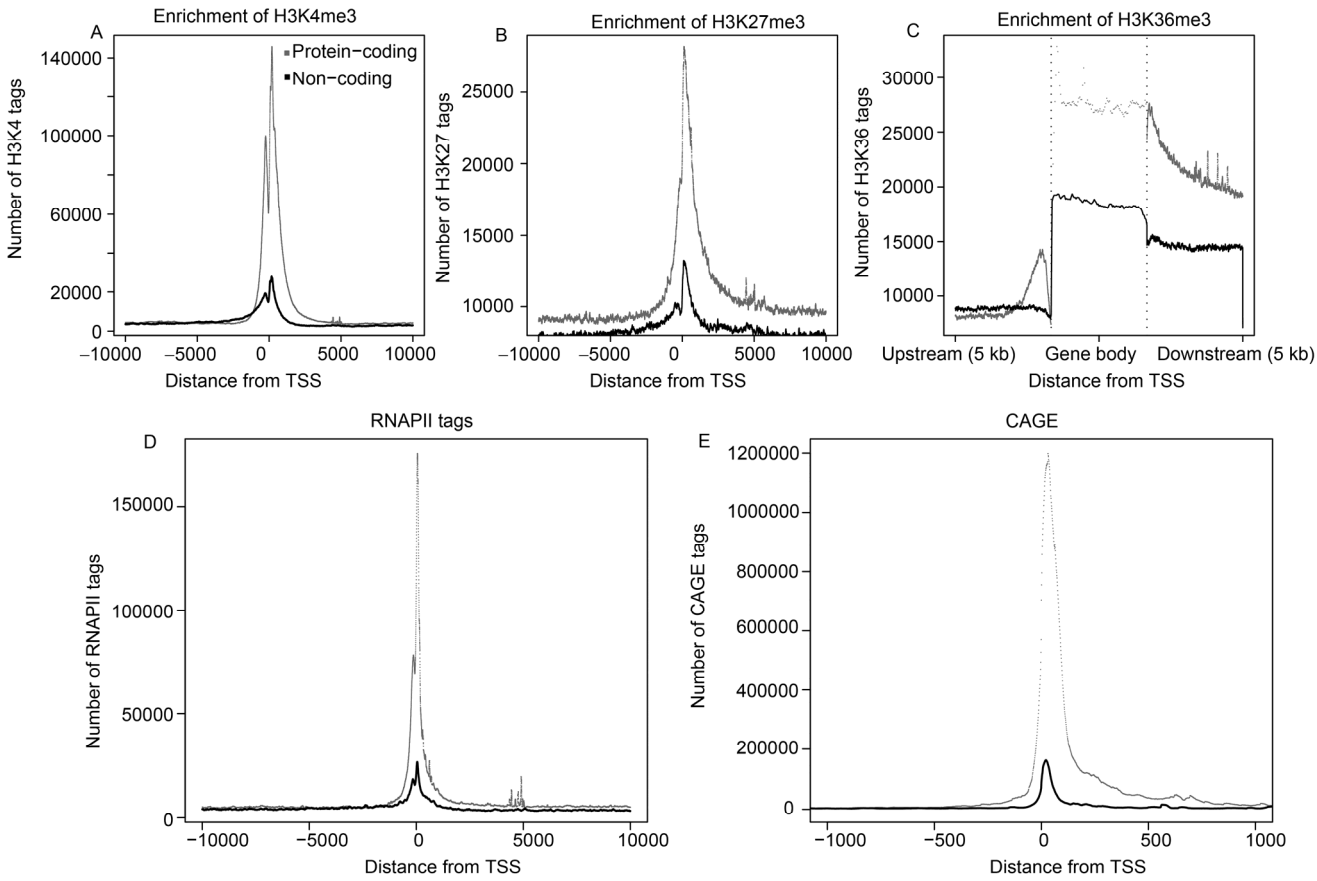

Figure 3 Non-coding genes are enriched with H3K4me3, H3K27me3, RNAPII, CAGE around their TSS regions and H3K36me3 along the gene regions. A and B, The TSS of non-coding genes is significantly enriched with H3K4me3 and H3K27me3 over average level. C, Distribution of H3K36me3 tags are investigated from genes' upstream 5 kb to downstream 5 kb, the bottom row has shown the gene body is enriched with H3K36 tags in a notably higher level than upstream and downstream regions. D and E, Both RNAPII and CAGE tags are enriched around TSS for non-coding and protein-coding genes.

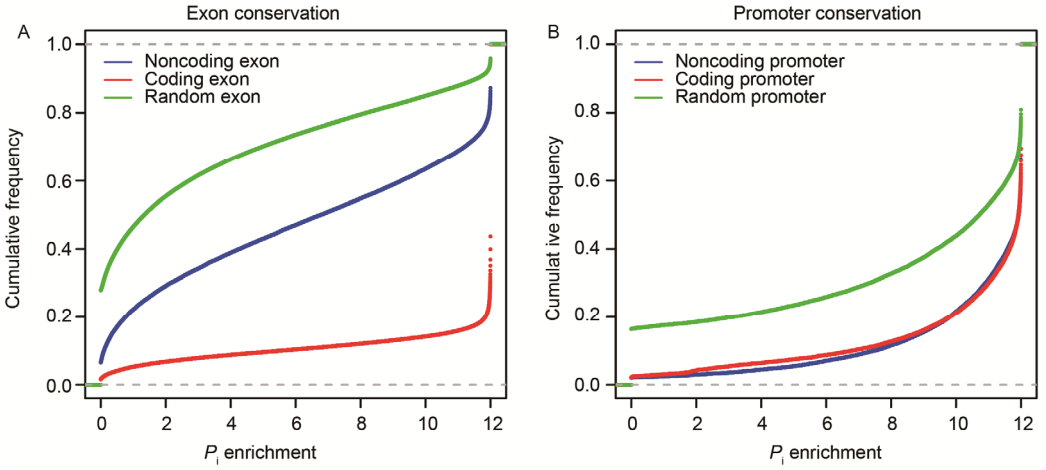

Figure 4 The exon and promoter conservation of non-coding genes in comparison to protein-coding genes. A, Shown is the cumulative distribution of conservation scores across 59 vertebrate species in the exons of non-coding genes, protein-coding genes and random regions. B, Similarly, the conservation of the promoter of non-coding genes, protein-coding genes and random regions is cumulatively plotted. The larger the P_i score, the more highly conserved.

expression of protein-coding genes spans seven orders of magnitude (10^{-1}–10^{5}) while non-coding genes only have six orders of magnitude (10^{-1}–10^{4}). We also compared the expression level among four subclasses of lncRNAs (Figure S5A in Supporting Information). The lncRNAs overlapping with annotation (LOWA) were expressed in the highest level, the lincRNAs and antisense non-coding genes were lower expressed in all tissues, and the intronic non-coding

Table 3 Expression level comparison (RPKM)

Tissues	Max-RPKM		Mean RPKM		Non-coding/coding rate
	Non-coding	Coding	Non-coding	Coding	
Cerebrum	27,579.40	2,904.23	9.51	6.56	1.5/1
Ovary	7,815.00	7,094.96	2.17	15.53	1/7
Mammary gland	11,543.73	9,819.95	1.67	15.64	1/9
Stomach	25,140.23	143,063.11	2.78	27.34	1/10
Small intestine	58,749.37	20,770.18	5.28	20.03	1/4
Adrenal gland	23,602.23	26,078.88	4.71	28.08	1/6
Large intestine	12,512.71	17,495.67	1.88	20.96	1/11
Thymus	28,372.71	5,872.64	4.22	13.42	1/3
Testis	6,580.29	9,895.02	2.65	14.93	1/6
Kidney	14,664.59	9,877.03	2.12	16.66	1/8
Liver	18,602.28	21,279.78	2.77	19.50	1/7
Lung	16,647.83	44,579.06	2.67	16.28	1/6
Spleen	52,702.23	134,225.30	7.56	47.05	1/6
Colon	17,686.47	10,574.05	2.68	16.74	1/6
Heart	10,907.67	26,518.70	1.47	16.19	1/11

Table 4 The statistics of tissue-specific non-coding genes[a)]

Types	Gene number	Rate	Overlapping	Intergenic	Intronic	Antisense
Housekeeping	448	2.76%*	210	51	145	42
Total specific	5,535	34.06%*	1,548	3,006	220	761
Cerebrum-specific	413	7.46%**	165	175	41	32
Ovary-specific	115	2.08%**	38	42	18	17
Mammary gland-specific	81	1.46%**	16	45	8	12
Stomach-specific	40	0.72%**	9	14	12	5
Small intestine-specific	23	0.42%**	2	8	11	19
Adrenal gland-specific	97	1.75%**	33	33	20	11
Large intestine-specific	53	0.96%**	13	31	3	6
Thymus-specific	120	2.17%**	35	47	22	16
Testis-specific	3,812	68.87%**	1,054	2,184	21	553
Kidney-specific	177	3.20%**	29	115	6	27
Liver-specific	110	1.99%**	24	55	11	20
Lung-specific	118	2.13%**	24	66	16	12
Spleen-specific	146	2.64%**	56	55	20	15
Colon-specific	61	1.10%**	10	39	5	7
Heart-specific	169	3.05%**	40	97	6	26

a) *, These data were compared with total 16,249 non-coding genes; **, These data were compared with total 5,535 tissue-specific non-coding genes.

genes were expressed fluctuated around the whole long non-coding genes.

On the basis of expression levels of non-coding and protein-coding genes in different tissues, we found an obviously different expression pattern between them. Briefly, non-coding genes revealed a more remarkable tissue-specific manner than protein-coding genes (Table 4, Figure S5B in Supporting Information). Quantificationally, 5,535 (34.06%) non-coding genes expressed in only one tissue while 448 (2.8%) expressed across all 15 tissues. Amongst the former part, 413 (7.46%) genes were cerebrum-specific and 3,812 (68.87%) were testis-specific, and these stringent specific non-coding genes may have their own particular functions in the tissue-specific processes.

Meanwhile, by analyzing the differential expression of all genes in distinct tissues, we got a hierarchical clustering of all tissues based on the similarity relationship between every two of them and found that testis and cerebrum both show distant relationships to other tissues (Figure 5).

Together, the expression dynamic analysis of putative genes highlighted two properties: non-coding genes are generally expressed lower and more likely to be tissue-specific than protein-coding genes. These two properties do not interact as cause and effect substantially.

Gene ontology analysis for neighboring protein-coding genes of lincRNAs

Except for those traditional non-coding RNA including

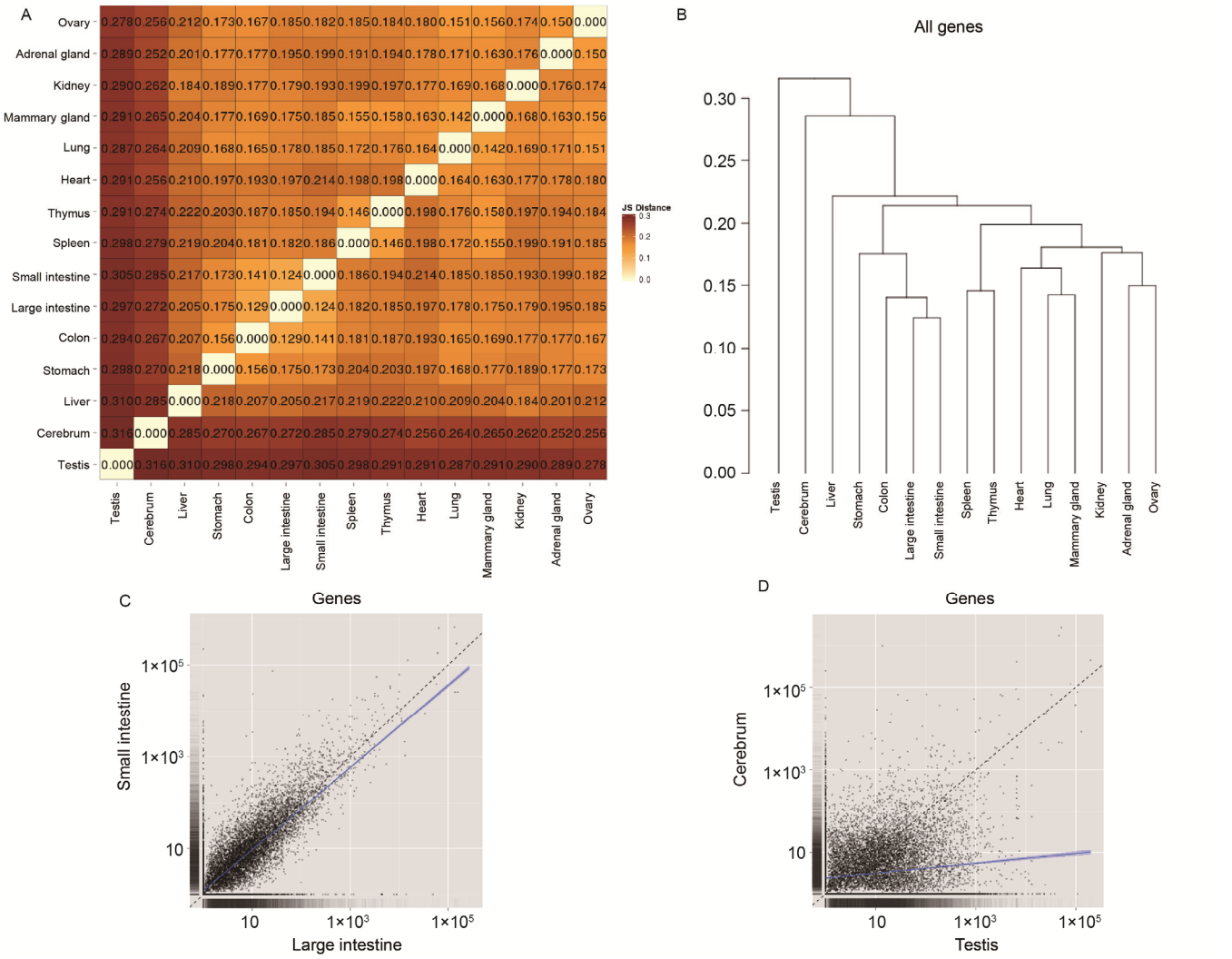

Figure 5 Tissue hierarchical clustering. A, The similarity relationship between every two tissues was measured by Jensen-Shannon (JS) distance. Maximum value represents the lowest correlation and vise verse. Cerebrum and testis both show distant relationships to other tissues. B, The hierarchical clustering was reproduced by a hierarchical tree with different related values marked along *Y*-axis. All tissues are clustered in a way of different physiological systems they belong to (digestive system and immune system for instance). C and D, Shown are the instances for expression correlation between large intestine and small intestine (left panel) and between testis and cerebrum (right panel). Notably, large intestine-small intestine is more correlated than testis-cerebrum pair, and the correlation coefficient of testis-cerebrum pair has far deviated from the axle wire.

tRNA, rRNAand snRNA, the newly identified lincRNAs were considered to possess diverse biological functions in mammals. For instance, a class of lincRNAs called ncRNA-activating (ncRNA-a) have enhancer-like function and positively regulate expression of neighboring protein-coding genes (Ørom et al., 2010). To investigate whether our identified lincRNAs were regulatory functional, we focused on the expression patterns of 1,612 lincRNAs with neighboring protein-coding genes and found that 74.44% of these lincRNA and protein-coding gene pairs exhibited a positive correlation while 25.56% were negative correlative. Furthermore, 20.22% of lincRNA and protein-coding gene pairs were strongly correlated with a Pearson correlation coefficient between 0.8 and 1, and it was remarkably higher than random gene pairs ($P<2.2\times10^{-16}$, *t*-test). When compared lincRNA and protein-coding gene pairs with 3,690 neighboring coding-coding gene pairs, we found the co-expression tendency was similar to a certain degree, and these neighboring coding-coding gene pairs were also more correlative than random gene pairs ($P<2.2\times10^{-16}$, *t*-test) (Figure S6 in Supporting Information). These results are consistent with earlier studies (Cabili et al., 2011; Guttman et al., 2009). Overall, we conclude that lincRNA genes were more or less co-expressed with their neighboring protein-coding genes, and this organization may be important for their specific regulatory functions. We thus used Gostat to cluster the potential functions of those coding genes neighboring lincRNAs and found they were significantly enriched in transcription regulation, intracellular part and metabolic processes.

Sense-antisense pair

2,239 antisense were classified into one-to-one type and one-to-multiple type according to the number of their corresponding sense genes. 2,099 one-to-one sense-antisense gene pairs were found on the basis of their genomic location with 449 convergent, 370 divergent and 1,280 non-overlap (Werner, 2013). We evaluated their expression correlation, indicating 1,040 (49.55%) were positively correlative and 905 (43.12%) were negatively correlative, and the correlative degree was obviously higher than random gene pairs ($P<2.2\times10^{-16}$, *t*-test) (Figure S7 in Supporting Information). It is consistent with the regulative function of antisense RNAs. The sense genes were enriched in signal transduction and nervous system development (Figure S8 in Supporting Information). Many further gain-of-function or loss-of-function experiments were required to well-understand their regulatory mechanism.

Functional enrichment of lincRNA loci and tissue-specific non-coding genes

The goal of our study is to infer the biological functions of these numerous lincRNA loci and provide a theoretical foundation for the further experimental validation. To this end, we firstly calculated the Pearson correlation coefficient between each lincRNA and each annotated protein-coding gene on the basis of their expression dynamics across 15 tissues, then we used GSEA (Guttman et al., 2009; Subramanian et al., 2005) to construct a matrix of the association for lincRNAs and protein-coding genes (FDR (false discovery rate)<0.01). By subsequently bi-clustering the matrix into 10 clusters, we identified several groups of lincRNA genes associated with distinct functional categories, including immune effector process, DNA replication initiation, muscle development and sexual reproduction and so on (Figure S9, Table S1 in Supporting Information).

According to the result of GSEA, the lincRNA related genes in each tissue were closely related to the physiological function of the tissue. For example, the lincRNA related genes in testis mainly played an important role in reproductive development including meiosis, development of primary sexual characteristics, sexual reproduction, and gamete generation, and so on. While in cerebrum, lincRNA associated genes mainly involved in brain development, synaptogenesis, axonogenesis and signal transduction.

Validation of lincRNAs by RT-PCR

To validate the existence of lincRNAs identified by our pipeline, we randomly selected 16 house-keeping lincRNAs which expressed in 15 tissues and performed RT-PCR in 12 tissues. 7 out of 16 lincRNAs were found to be expressed in a detectable level in all the 12 tissues while 15 out of 16 lincRNAs fit a loose criterion that each lincRNAs were detected in at least 10 tissues (Figure 6). In addition, 31 out of

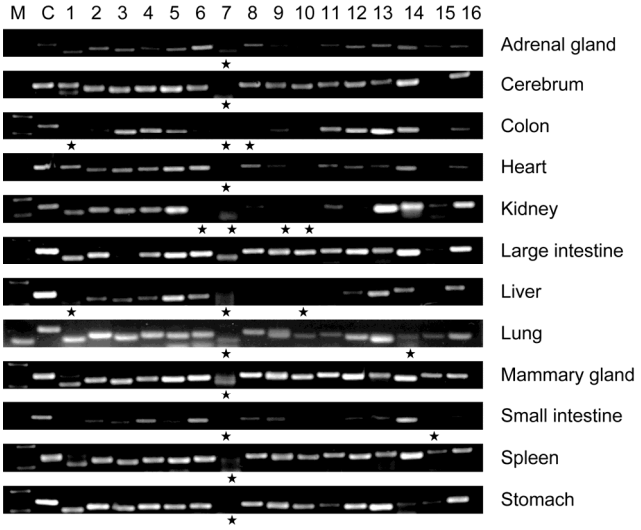

Figure 6 RT-PCR result of housekeeping lincRNA. The meaning of each symbol in first row is as follows: M, marker; C, control and 1–16 represents lincRNAs. The lincRNAs with star labels failed in experiments.

42 tissue-specific lincRNA were detected by RT-PCR (Figure S10 in Supporting Information). Especially, 11 out of 13 cerebrum-specific lincRNAs were found. All these results indicated that our pipeline are believable and can give high quality list of lincRNAs. The details about RT-PCR primers of these lincRNA were shown in Dataset S3.

DISCUSSION

Identification and characterization of non-coding RNA

In this study, we identified non-coding RNAs using ultra-deep RNA-seq data from 15 mouse tissues. These RNA-seq data allowed us to detect those genes that expressed at relatively low levels or in a strictly tissue-specific manner.

We obtained 16,249 genes eventually, of which 6,315 were single-exon non-coding genes. For demonstrating the function of these non-coding genes, we not only explored the genomic features, but also used many supporting evidences such as chromatin modification (H3K4me3, H3K27me3 and H3K36me3), RNAPII, CAGE and evolutional conservation to argue for their functional roles. Comparing with protein-coding genes, these non-coding genes generally have fewer exons, shorter gene and exon length, but larger introns which were viewed as a result of high proportion of LTR and LINE. These non-coding genes were supported by enrichment of histone methylation, RNAPII and CAGE around TSS.

Tissue-specificity of non-coding RNA

From the expression profile of 15 tissues, non-coding genes were found to express in a relatively lower level and more

tissue-specific manner. It is interesting that non-coding genes expressed higher in cerebrum than other tissues (Table 3). We calculated the mean expression rate of non-coding/coding in the downloaded mRNA data of cerebrum to uncover the differences. The non-coding/coding rate in mRNA data of cerebrum is 1∶1.4, which is still higher than that of other 14 tissues. Thus, the high expression rate of non-coding and coding was possibly caused by the cerebrum tissue instead of different library construction methods. To investigate whether the tissue specificity of non-coding genes was caused by their lower expression levels, we contrasted the expression breadth between those non-coding and protein-coding genes with similar expression levels. We found that the tissue specificity was closely related to the expression level (Figure S11 in Supporting Information). However, even if the tissue-specificity decreased along with the increase of expression level, the tissue-specificity of non-coding genes was still more obvious than protein-coding genes. The tissue-specific feature of non-coding genes indicated their biological roles in tissue-specific processes and proved by analyzing their GO enrichment using GSEA.

Non-coding RNA comparison among different studies

Compared with previous annotated lincRNAs from cells (Guttman et al., 2009; Guttman et al., 2010) and based on EST data of multiple tissues (Qu and Adelson, 2012), our lincRNAs overlapped with them less than 10%. In consideration of using tissue data in our paper, we also found another collection of lincRNAs in tissues which were contained in our samples (Luo et al., 2013) and found that 31.16% of those 6,755 lincRNA transcripts (3,965 novel lincRNA genes) overlapped with our lincRNAs. Because the diversity of annotation criteria and cutoffs may one of the reasons causing the minor overlap among lincRNA lists of different studies (Ulitsky and Bartel, 2013), we examined these pipelines in detail and found that our pipeline had many differences with others. First, the tissues or cell types of these researches were differentiated from each other. Second, the transcripts were sequenced by EST, ChIP-seq or RNA-seq in different studies. In addition, these data have different features, like strand specific or not, data depth, and tissue numbers. For example, our data were strand-specific and we identified 2,217 antisense RNAs while some data can't distinguish the reads strand. Third, we totally obtained 2,282 M fragments while the other data size spans from 4,853,460 ESTs to 1,936 M fragments. Fourth, different softwares were used to reconstruct transcripts and predict the coding potential of novel transcripts (Table 5). The specificity and sensitivity varied largely among these softwares. For example, in the Luo et al.'s novel lincRNA dataset, 31% lincRNAs were reconstructed by Scripture only. Fifth, different cutoffs were used to filter lincRNAs in different studies.

Non-coding RNA orthologous

To further validate that our lincRNAs are functional elements, we assessed the evolution origin of these lincRNAs. 85.26% (9,397 of 11,022) of lincRNAs have their orthologous regions in the human genome (Dataset S4 in Supporting Information). Subsequently, we surveyed catalog of mammalian and non-mammalian vertebrate transcripts that were syntenically mapped to the mouse genome by TransMap (Cabili et al., 2011; Kuhn et al., 2009) to estimate the expressed ortholog transcripts in other species. This analysis identified 1,415 lincRNAs syntenically paired with an orthologous transcript from TransMap (Dataset S4, Figure S12 in Supporting Information), accounting for 12.84% of all mouse lincRNAs. The small fraction of homology implied that lincRNAs were less conserved than protein-coding genes.

Poly(A)– non-coding RNA

Currently, most of transcriptome studies are focused on poly(A)+ transcripts, therefore, the expression of poly(A)– transcripts need to be explored. In human, about 20% transcripts are poly(A)– or bimorphic transcripts and some im-

Table 5 Summary of lincRNA studies

Dataset	Sample	Data type &data size	Software for transcript reconstruction	Coding potential prediction	Reference
~1,600 multi-exonic lincRNAs	four mouse cell types (ESC, MEF, MLF and NPC)	ChIP-seq	Authors' own program by a sliding window approach	Codon Substitution Frequency (CSF) method	Guttman et al. (Guttman et al., 2009)
1,140 multi-exonic lincRNAs	three mouse cell types (ESC, NPC and MLF)	RNA-Seq (493 M)	Scripture	CSF and ORF	Guttman et al. (Guttman et al., 2010)
9,490 lincRNAs	Multiple tissue/cell types	all publically available mouse EST(4,853,460)	TGICL	Blastx and EMBOSS	Qu and Adelson (Qu and Adelson, 2012)
6,755 novel lincRNAs	six tissues	RNA-Seq (1,936 M)	Cufflinks and Scripture	CPC and CNCI	Luo et al. (Luo et al., 2013)
11022 lincRNAs	fifteentissues	RNA-Seq (2,282 M)	Cufflinks	CPC and CPAT	Our research

portant lncRNAs are poly(A)− transcripts such as well characterized lincRNAs *MALAT1* and *NEAT1* (Djebali et al., 2012; Yang et al., 2011). To detect poly(A)− transcripts, we sequenced cerebrum in our laboratory using ribo-minus method which can capture both poly(A)+ and poly(A)− transcripts by Hiseq2000. For identification of the poly(A)− lincRNAs, first, we found 255 lincRNAs that only expressed in our ribo-minus RNA-seq data of cerebrum. Second, to remove the possible poly(A)+ transcripts, we used mRNA-seq data of cerebrum (SRX191149) and found 8 expressed lincRNAs in the 255 lincRNAs. It implied that the remaining 247 lincRNAs (18.70% of all expressed lincRNAs in cerebrum) were possible poly(A)− transcripts. As for the 247 possible poly(A)− lincRNAs in cerebrum, we also found there are 82 (1.43%), 57 (2.27%) and 92 (4.20%) lincRNAs are expressed in the ribo-minus RNA-seq data of testis (Liu et al., 2011), mammary (Zhou et al., 2014) and ovary (Pan et al., 2014) which were sequenced by SOLiD in our laboratory. It indicated that about 1.43%~18.70% lincRNAs are possible poly(A)− lincRNAs. Furthermore, we found that poly(A)− lincRNAs didn't have significant enrichment of RNAPII tags and had less GC content than poly(A)+ lincRNAs.

CONCLUSION

Our work provides a way to identify and characterize non-coding RNAs especially lincRNAs. More importantly, we expand the lincRNome of mouse and give the opportunity to investigate the physiological function of lincRNAs. These non-coding RNAs provide an important source for functional experiment. The specific function for each lincRNA needs to be further validated by experiments, such as the lincRNAs which are required for life and brain development (Sauvageau et al., 2013). Hopefully, the physiological and pathological role of lincRNAs can be applied to human diseases.

MATERIALS AND METHODS

Datasets

The mRNA-seq data of 14 mouse tissues produced by ENCODE were downloaded from National Center of Biotechnology Information (NCBI) (GSM900188, GSM900198, GSM900199, GSM900194, GSM900189, GSM900195, GSM900196, GSM900184, GSM900183, GSM900186, GSM900197, GSM900185, GSM900193, and GSM900192) (ENCODE Project Consortium et al., 2012; Mouse ENCODE Consortium et al., 2012). These data are strand-specific and each tissue has 147 million pair-end fragments on average with 76 bp read length (Table 1). We also retrieved ChIP-seq (chromatin immunoprecipitation followed by sequencing) data of RNA polymerase II (RNAPII), H3K4me3, H3K27me3 and H3K36me3 from mouse heart (Karolchik et al., 2014; Pervouchine et al., 2015). A mRNA-seq data of cerebrum (SRX191149) was downloaded from NCBI to confirm some results (Nuno L. Barbosa-Morais, 2012). The data processing method of 15 tissues was applied to this mRNA data.

RNA sequencing

Cerebrum was obtained from an 8 week old adult male C57BL/6J mouse. We used Trizol (Invitrogen, USA) to isolate total RNA and then used the Ribo-minus Eukaryote kit (cat.10837-08, Invitrogen, USA) to deplete rRNA. We constructed the transcriptomic library using RNA-Seq Library Preparation Kit for Whole Transcriptome Discovery-Illumina Compatible Kit (Wujiang Huijie Biotech, Suzhou), with a starting material of 500 ng rRNA-depleted RNA. The cDNA library was amplified, cleaned using the AMPure XP beads and then sequenced on the GA-analyzer by the Illumina HiSeq 2000 (Illumina, USA). Finally, we obtained ~211 million strand-specific pair-end fragments with 101 bp read length (SRX806806).

Reads mapping and transcripts assembly

As an initial step, we removed fragments mapping to ribosomal RNA. We accomplished this by directly aligning each fragments against the mouse DNA sequence of ribosomal RNA. We removed low-quality reads with an in-house Perl script. For obtaining all known transcripts as a comprehensive annotation, we integrated the mouse transcripts from UCSC (2013.5.6), NCBI (Refseq v58) and Ensembl (v72). Then RNA-seq data of 15 mouse tissues (14 downloaded tissues and one new sequenced tissue) were mapped to mouse genome (mm10) by GSNAP version 2013.7.16 with the options "-N 1, -force-xs-dir and -s splice sites of integrated mouse transcripts from UCSC, Ensembl and NCBI" (all the other options are default). The aligned reads were assembled into transcripts for each tissue separately by Cufflinks v2.1.1.

Non-coding RNAs identification pipeline

In order to accurately and completely identify the transcripts, all transcripts of 15 tissues were merged into an unique set of transcripts using Cuffmerge. Then, loci completely overlapping with the genes of RefSeq, UCSC and Ensembl were filtered out. To precisely distinguish coding and non-coding locus, we used Coding Potential Calculator (CPC) and Coding-Potential Assessment Tool (CPAT-1.2.1) to predict the coding potential for each transcript locus. The locus was classified into non-coding/coding locus only when two softwares gave the same result. Next, locus with RPKM (reads per kilobase of exon model per million

mapped reads)<0.1 was filtered out. To credibly identify single-exon non-coding genes, we performed BLASTN to align these single-exon non-coding genes to EST data (mouse, human and other species) and scanned them with Rfam covariance models, and then we only kept those genes which mapped to at least one type of data. In addition, the single exon genes must be expressed in at least two tissues. For these non-coding loci, they were classified into overlapping, antisense, intronic and intergenic non-coding genes according to their genomic locations.

Histone methylation and RNAPII binding sites analysis

Mapping results and peak files of these ChIP-seq data including H3K4me3, H3K27me3, H3K36me3 and RNAPII were downloaded. The promoter was confined to 5,000 bp upstream and downstream of transcription start site (TSS). Chromatin states of non-coding genes were determined based on the overlapping between promoters of non-codinggenes and enrichment peaks of H3K4me3, H3K27me3 and RNAPII, or between the transcription region of non-coding genes and enrichment peaks of H3K36me3.

TSS and CAGE evidences

129,466 mouse TSS-like classified CAGE peaks published by the Fantom5 project were used to provide the TSS support for non-coding genes (FANTOM Consortium and the RIKEN PMI and CLST et al., 2014) (http://fantom.gsc. riken. jp/5/tet/). We first converted the coordinate of non-coding genes in mm10 to mm9 by LiftOver of UCSC. Then, we defined the TSS regions for non-coding genes from the upstream 500 bp to downstream 500 bp of TSSs. The non-coding genes were considered supported by CAGE only when the TSS-like CAGE peaks overlapped with the TSS regions.

Mapping result of 5′ CAGE tags from mouse testis were downloaded from DDBJ (ftp://ftp.ddbj.nig.ac.jp/ddbj_ database/dra/fastq/DRA000/DRA000991/DRZ001400/). We calculated the coverage per nucleotide position covered by CAGE tags and then accumulated the coverage of non-coding genes from upstream 1 kb to downstream 1 kb of TSSs.

Conservation

To evaluate the conservation of non-coding genes, PhastConsElements60way data for mouse genome (mm10) were downloaded from UCSC (Meyer et al., 2013). PhastConsElements60way was genomic conservation profiles generated by the phyloP (phylogenetic *P*-values) and PhastCons algorithms (http://compgen.bscb.cornell. edu/ phast/) for multiple alignments of 59 vertebrate genomes to the mouse genome. We calculated the conservative scores in a 12 bp window and a step length of 1 bp for every region and selected the maximal score as its conservation value. The higher the conservation score the more conservative the region is.

Gene set enrichment analysis

Gene set enrichment analysis (GSEA) was performed as previous studies (Guttman et al., 2009; Pauli et al., 2012). The Pearson correlations between each lincRNA and all protein-coding genes were calculated on the basis of their expression dynamics across 15 tissues. Then, protein-coding genes were ranked according to their correlation value. For each lincRNA locus, a list of correlation-based ranked protein-coding genes was submitted to GSEA (Subramanian et al., 2005). Finally, an association matrix between lincRNA locus and Gene Ontology (GO) terms, which was obtained from the result of GSEA with *P* value below 0.01, was used to cluster. To screen the most enrichment GO terms of each cluster among all positively associated GO terms, we ranked the positively associated enrichment GO terms based on binominal test.

Co-expression profiling of lincRNAs and their neighboring protein-coding genes

On the basis of the location of each lincRNA gene and its neighboring protein-coding genes, we extracted lincRNAs and protein-coding gene pairs within 10 kb distance. We also produced the neighboring protein-coding gene pairs using similar method. The random gene pairs were obtained by randomly selected two genes. We then used R to calculate the Pearson correlation coefficient for them and analyzed the distribution of correlation scores. For the neighboring protein-coding genes of lincRNAs, GOstat (http://gostat.wehi.edu.au/cgi-bin/goStat.pl) was used to analyze their functional enrichment.

RT-PCR validation

RNA was extracted from mouse tissues using Trizol reagent (Invitrogen) according to the manufacturer's protocol and then treated with DNaseI (1 U μL^{-1}; Invitrogen) to remove possible contaminating genomic DNA. DNA-free RNAs were reverse transcribed to cDNAs using Superscript III (Invitrogen) and then cDNAs were diluted 20-fold for further PCR analysis. PCR was performed on 10 μL volumes containing 10×PCR buffer 1 μL, dNTPmix (2.5 mmol L^{-1}) 1.5 μL, primers (10 μmol L^{-1} each) 0.5 μL, Taq DNA polymerase (5 U μL^{-1} TaKaRa, Dalian) 0.1 μL, 20-fold diluting cDNA 1 μL, and 5.9 μL nuclease-free water by Applied Biosystems 9700 PCR System (Life Technologies, USA). PCR conditions were 94°C for 3 min, followed by 35 cycles of 98°C for 10s, 60°C for 30s, 72°C for 30s and at last, 5 min at 72°C. Finally, PCR products were separated on 2.0%

agrose gels. For amplifying the lincRNA transcripts, primer sets were designed by web-based tool GenScript (https://www.genscript.com/ssl-bin/app/primer) and their specificity was checked with mouse RefSeq mRNA.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *The handling of mouse and experimental procedures were guided and approved by 2013A017. This work was supported by grants from Natural Science Foundation of China (31271385), and Knowledge Innovation Program of the Chinese Academy of Sciences (KSCX2-EW-R-01-04).*

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafrenière, R.G., Xing, Y., Lawrence, J., and Willard, H.F. (1992). The human *XIST* gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell 71, 527–542.

Brown, J.B., Boley, N., Eisman, R., May, G.E., Stoiber, M.H., Duff, M.O., Booth, B.W., Weng, J., Park, S., and Suzuki, A.M. (2014). Diversity and dynamics of the *Drosophila* transcriptome. Nature 512, 393–399.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25, 1915–1927.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., and Wells, C. (2005). The transcriptional landscape of the mammalian genome. Science 309, 1559–1563.

Clark, M.B., Amaral, P.P., Schlesinger, F.J., Dinger, M.E., Taft, R.J., Rinn, J.L., Ponting, C.P., Stadler, P.F., Morris, K.V., and Morillon, A. (2011). The reality of pervasive transcription. PLoS Biol 9, e1000625.

Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., and Bethel, G. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5, 613–619.

Cui, P., Lin, Q., Ding, F., Xin, C., Gong, W., Zhang, L., Geng, J., Zhang, B., Yu, X., and Yang, J. (2010). A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. Genomics 96, 259–265.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., and Knowles, D.G. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 22, 1775–1789.

Dinger, M.E., Amaral, P.P., Mercer, T.R., Pang, K.C., Bruce, S.J., Gardiner, B.B., Askarian-Amiri, M.E., Ru, K., Soldà, G., and Simons, C. (2008). Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. Genome Res 18, 1433–1445.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., and Schlesinger, F. (2012). Landscape of transcription in human cells. Nature 489, 101–108.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

FANTOM Consortium and the RIKEN PMI and CLST(DGT), Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., Itoh, M., Andersson, R., Mungall, C.J., Meehan, T.F., Schmeier, S., Bertin, N., Jorgensen, M., Dimont, E., Arner, E., Schmidl, C., Schaefer, U., Medvedeva, Y.A., Plessy, C., Vitezic, M., Severin, J., Semple, C., Ishizu, Y., Young, R.S., Francescatto, M., Alam, I., Albanese, D., Altschuler, G.M., Arakawa, T., Archer, J.A., Arner, P., Babina, M., Rennie, S., Balwierz, P.J., Beckhouse, A.G., Pradhan-Bhatt, S., Blake, J.A., Blumenthal, A., Bodega, B., Bonetti, A., Briggs, J., Brombacher, F., Burroughs, A.M., Califano, A., Cannistraci, C.V., Carbajo, D., Chen, Y., Chierici, M., Ciani, Y., Clevers, H.C., Dalla, E., Davis, C.A., Detmar, M., Diehl, A.D., Dohi, T., Drablos, F., Edge, A.S., Edinger, M., Ekwall, K., Endoh, M., Enomoto, H., Fagiolini, M., Fairbairn, L., Fang, H., Farach-Carson, M.C., Faulkner, G.J., Favorov, A.V., Fisher, M.E., Frith, M.C., Fujita, R., Fukuda, S., Furlanello, C., Furino, M., Furusawa, J., Geijtenbeek, T.B., Gibson, A.P., Gingeras, T., Goldowitz, D., Gough, J., Guhl, S., Guler, R., Gustincich, S., Ha, T.J., Hamaguchi, M., Hara, M., Harbers, M., Harshbarger, J., Hasegawa, A., Hasegawa, Y., Hashimoto, T., Herlyn, M., Hitchens, K.J., Ho Sui, S.J., Hofmann, O.M., Hoof, I., Hori, F., Huminiecki, L., Iida, K., Ikawa, T., Jankovic, B.R., Jia, H., Joshi, A., Jurman, G., Kaczkowski, B., Kai, C., Kaida, K., Kaiho, A., Kajiyama, K., Kanamori-Katayama, M., Kasianov, A.S., Kasukawa, T., Katayama, S., Kato, S., Kawaguchi, S., Kawamoto, H., Kawamura, Y.I., Kawashima, T., Kempfle, J.S., Kenna, T.J., Kere, J., Khachigian, L.M., Kitamura, T., Klinken, S.P., Knox, A.J., Kojima, M., Kojima, S., Kondo, N., Koseki, H., Koyasu, S., Krampitz, S., Kubosaki, A., Kwon, A.T., Laros, J.F., Lee, W., Lennartsson, A., Li, K., Lilje, B., Lipovich, L., Mackay-Sim, A., Manabe, R., Mar, J.C., Marchand, B., Mathelier, A., Mejhert, N., Meynert, A., Mizuno, Y., de Lima Morais, D.A., Morikawa, H., Morimoto, M., Moro, K., Motakis, E., Motohashi, H., Mummery, C.L., Murata, M., Nagao-Sato, S., Nakachi, Y., Nakahara, F., Nakamura, T., Nakamura, Y., Nakazato, K., van Nimwegen, E., Ninomiya, N., Nishiyori, H., Noma, S., Noma, S., Noazaki, T., Ogishima, S., Ohkura, N., Ohimiya, H., Ohno, H., Ohshima, M., Okada-Hatakeyama, M., Okazaki, Y., Orlando, V., Ovchinnikov, D.A., Pain, A., Passier, R., Patrikakis, M., Persson, H., Piazza, S., Prendergast, J.G., Rackham, O.J., Ramilowski, J.A., Rashid, M., Ravasi, T., Rizzu, P., Roncador, M., Roy, S., Rye, M.B., Saijyo, E., Sajantila, A., Saka, A., Sakaguchi, S., Sakai, M., Sato, H., Savvi, S., Saxena, A., Schneider, C., Schultes, E.A., Schulze-Tanzil, G.G., Schwegmann, A., Sengstag, T., Sheng, G., Shimoji, H., Shimoni, Y., Shin, J.W., Simon, C., Sugiyama, D., Sugiyama, T., Suzuki, M., Suzuki, N., Swoboda, R.K., t Hoen, P.A., Tagami, M., Takahashi, N., Takai, J., Tanaka, H., Tatsukawa, H., Tatum, Z., Thompson, M., Toyodo, H., Toyoda, T., Valen, E., van de Wetering, M., van den Berg, L.M., Verado, R., Vijayan, D., Vorontsov, I.E., Wasserman, W.W., Watanabe, S., Wells, C.A., Winteringham, L.N., Wolvetang, E., Wood, E.J., Yamaguchi, Y., Yamamoto, M., Yoneda, M., Yonekura, Y., Yoshida, S., Zabierowski, S.E., Zhang, P.G., Zhao, X., Zucchelli, S., Summers, K.M., Suzuki, H., Daub, C.O., Kawai, J., Heutink, P., Hide, W., Freeman, T.C., Lenhard, B., Bajic, V.B., Taylor, M.S., Makeev, V.J., Sandelin, A., Hume, D.A., Carninci, P., and Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. Nature 507, 462–470.

Faulkner, G.J., Forrest, A.R., Chalk, A.M., Schroder, K., Hayashizaki, Y., Carninci, P., Hume, D.A., and Grimmond, S.M. (2008). A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. Genomics 91, 281–288.

Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.C., Hung, T., Argani, P., and Rinn, J.L. (2010). Long noncoding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464, 1071–1076.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M., Feldser, D., Huarte, M., Zuk, O., Carey, B., and Cassady, J. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458, 223–227.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., and Bruhn, L. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477, 295–300.

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., and 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.

Hüttenhofer, A., and Vogel, J. (2006). Experimental approaches to identify non-coding RNAs. Nucleic Acids Res 34, 635–646.

Haas, B.J., and Zody, M.C. (2010). Advancing RNA-seq analysis. Nat Biotechnol 28, 421.

Hangauer, M.J., Vaughn, I.W., and McManus, M.T. (2013). Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genetics 9, e1003569.

Hawkins, P.G., and Morris, K.V. (2010). Transcriptional regulation of Oct4 by a long non-coding RNA antisense to Oct4-pseudogene 5. Transcription 1, 165–175.

Heard, E., and Disteche, C.M. (2006). Dosage compensation in mammals: fine-tuning the expression of the X chromosome. Genes Dev 20, 1848–1867.

Hu, W., Alvarez‐Dominguez, J.R., and Lodish, H.F. (2012). Regulation of mammalian cell differentiation by long non-coding RNAs. EMBO Rep 13, 971–983.

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., Harte, R.A., Heitner, S., Hinrichs, A.S., Learned, K., Lee, B.T., Li, C.H., Raney, B.J., Rhead, B., Rosenbloom, K.R., Sloan, C.A., Speir, M.L., Zweig, A.S., Haussler, D., Kuhn, R.M., and Kent, W.J. (2014). The UCSC Genome Browser database: 2014 update. Nucleic Acids Res 42, D764–D770.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nashida, H., Yap, C., Suzuki, M., and Kawai, J. (2005). Antisense transcription in the mammalian transcriptome. J Biol Chem 309, 1564–1566.

Kuhn, R.M., Karolchik, D., Zweig, A.S., Wang, T., Smith, K.E., Rosenbloom, K.R., Rhead, B., Raney, B.J., Pohl, A., and Pheasant, M. (2009). The UCSC genome browser database: update 2009. Nucleic acids Res 37, D755–D761.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., and Zhang, J. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148, 84–98.

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., Arenas-Huertero, C., and Chua, N.-H. (2012). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. Plant Cell 24, 4333–4345.

Liu, W., Zhao, Y., Cui, P., Lin, Q., Ding, F., Xin, C., Tan, X., Song, S., Yu, J., and Hu, S. (2011). Thousands of novel transcripts identified in mouse cerebrum, testis, and ES cells based on ribo-minus RNA sequencing. Front Genet 2, 93.

Luo, H., Sun, S., Li, P., Bu, D., Cao, H., and Zhao, Y. (2013). Comprehensive characterization of 10,571 mouse large intergenic noncoding RNAs from whole transcriptome sequencing. PLoS One 8, e70835.

Lv, J., Liu, H., Huang, Z., Su, J., He, H., Xiu, Y., Zhang, Y., and Wu, Q. (2013). Long non-coding RNA identification over mouse brain development by integrative modeling of chromatin and genomic features. Nucleic Acids Res, 41, 10044–10061.

Maamar, H., Cabili, M.N., Rinn, J., and Raj, A. (2013). linc-HOXA1 is a noncoding RNA that represses Hoxa1 transcription *in cis*. Genes Dev 27, 1260–1271.

Mattick, J.S. (2009). The genetic signatures of noncoding RNAs. PLoS Genet 5, e1000459.

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

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., and Rhead, B. (2013). The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res 41, D64–D69.

Mouse ENCODE Consortium,Stamatoyannopoulos, J.A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D.M., Groudine, M., Bender, M., Kaul, R., and Canfield, T. (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol 13, 418.

Nakaya, H.I., Amaral, P.P., Louro, R., Lopes, A., Fachel, A.A., Moreira, Y.B., El-Jundi, T.A., da Silva, A.M., Reis, E.M., and Verjovski-Almeida, S. (2007). Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. Genome Biol 8, R43.

Nam, J.W., and Bartel, D. (2012). Long non-coding RNAs in *C. elegans*. Genome Res 22, 2529–2540.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., Kim, T., Misquitta-Ali, C.M., Wilson, M.D., Kim, P.M., Odom, D.T., Frey, B.J., and Blencowe, B.J. (2012). The evolutionary landscape of alternative splicing in verterbrate species. Science 338, 1587–1593.

Okazaki, Y., Furuno, M., Kasukawa, T., and Adachi, J. (2002). Analysis of the mouse transcriptome based on functional annotation of 60770 full-length cDNAs. Nature 420, 563–573.

Ørom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., and Huang, Q. (2010). Long noncoding RNAs with enhancer-like function in human cells. Cell 143, 46–58.

Pan, L., Gong, W., Zhou, Y., Li, X., Yu, J., and Hu, S. (2014). A comprehensive transcriptomic analysis of infant and adult mouse ovary. Genomics Proteomics Bioinformatics 12, 239–248.

Pauli, A., Rinn, J.L., and Schier, A.F. (2011). Non-coding RNAs as regulators of embryogenesis. Nat Rev Genet 12, 136–149.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., and Regev, A. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res 22, 577–591.

Pervouchine, D.D., Djebali, S., Breschi, A., Davis, C.A., Barja, P.P., 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, M.A., Notredame, C., Guigo, R., and Gingeras, T.R. (2015). Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nat Commun 6, 5903.

Qu, Z., and Adelson, D.L. (2012). Identification and comparative analysis of ncRNAs in human, mouse and zebrafish indicate a conserved role in regulation of genes expressed in brain. PLoS One 7, e52275.

Ramsk ld, D., Wang, E., Burge, C., and Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput Biol 5, e1000598.

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

Sasaki, Y.T., Ideue, T., Sano, M., Mituyama, T., and Hirose, T. (2009). MENε/β noncoding RNAs are essential for structural integrity of nuclear paraspeckles. Proc Natl Acad Sci USA 106, 2525–2530.

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

Sauvageau, M., Goff, L.A., Lodato, S., Bonev, B., Groff, A.F., Gerhardinger, C., Sanchez-Gomez, D.B., Hacisuleyman, E., Li, E., and Spence, M. (2013). Multiple knockout mouse models reveal lincRNAs are required for life and brain development. ELife 2, e01749.

Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., and Giallourakis, C.C. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc Natl Acad Sci USA 110, 2876–2881.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., and Lander, E.S. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102, 15545–15550.

The ENCODE Project Consortium. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799–816.

Tan, M.H., Au, K.F., Yablonovitch, A.L., Wills, A.E., Chuang, J., Baker, J.C., Wong, W.H., and Li, J.B. (2013). RNA sequencing reveals a diverse and dynamic repertoire of the *Xenopus tropicalis* transcriptome over development. Genome Res 23, 201–216.

Ulitsky, I., and Bartel, D.P. (2013). lincRNAs: genomics, evolution, and mechanisms. Cell 154, 26–46.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell 147, 1537–1550.

Wang, K.C., and Chang, H.Y. (2011). Molecular mechanisms of long noncoding RNAs. Mol Cell 43, 904–914.

Werner, A. (2013). Biological functions of natural antisense transcripts. BMC Biol 11, 31.

Wetterbom, A., Ameur, A., Feuk, L., Gyllensten, U., and Cavelier, L. (2010). Identification of novel exons and transcribed regions by

chimpanzee transcriptome sequencing. Genome Biol 11, R78.
Yang, L., Duff, M.O., Graveley, B.R., Carmichael, G.G., and Chen, L.-L. (2011). Genomewide characterization of non-polyadenylated RNAs. Genome Biol 12, R16.
Yang, P.K., and Kuroda, M.I. (2007). Noncoding RNAs and intranuclear positioning in monoallelic gene expression. Cell 128, 777–786.
Yassour, M., Kaplan, T., Fraser, H.B., Levin, J.Z., Pfiffner, J., Adiconis, X., Schroth, G., Luo, S., Khrebtukova, I., and Gnirke, A. (2009). *Ab initio* construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc Natl Acad Sci USA 106, 3264–3269.
Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D., Shen, Y., Pervouchine, D.D., Djebali, S., Thurman, R.E., Kaul, R., Rynes, E., Kirilusha, A., Marinov, G.K., Williams, B.A., Trout, D., Amrhein, H., Fisher-Aylor, K., Antoshechkin, I., DeSalvo, G., See, L.H., Fastuca, M., Drenkow, J., Zaleski, C., Dobin, A., Prieto, P., Lagarde, J., Bussotti, G., Tanzer, A., Denas, O., Li, K., Bender, M.A., Zhang, M., Byron, R., Groudine, M.T., McCleary, D., Pham, L., Ye, Z., Kuan, S., Edsall, L., Wu, Y.C., Rasmussen, M.D., Bansal, M.S., Kellis, M., Keller, C.A., Morrissey, C.S., Mishra, T., Jain, D., Dogan, N., Harris, R.S., Cayting, P., Kawli, T., Boyle, A.P., Euskirchen, G., Kundaje, A., Lin, S., Lin, Y., Jansen, C., Malladi, V.S., Cline, M.S., Erickson, D.T., Kirkup, V.M., Learned, K., Sloan, C.A., Rosenbloom, K.R., Lacerda de Sousa, B., Beal, K., Pignatelli, M., Flicek, P., Lian, J., Kahveci, T., Lee, D., James Kent, W., Ramalho Santos, M., Herrero, J., Notredame, C., Johnson, A., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Canfield, T., Sabo, P.J., Wilken, M.S., Reh, T.A., Giste, E., Shafer, A., Kutyavin, T., Haugen, E., Dunn, D., Reynolds, A.P., Neph, S., Humbert, R., Scott Hansen, R., De Bruijn, M., Selleri, L., Rudensky, A., Josefowicz, S., Samstein, R., Eichler, E.E., Orkin, S.H., Levasseur, D., Papayannopoulou, T., Chang, K.H., Skoultchi, A., Gosh, S., Disteche, C., Treuting, P., Wang, Y., Weiss, M.J., Blobel, G.A., Cao, X., Zhong, S., Wang, T., Good, P.J., Lowdon, R.F., Adams, L.B., Zhou, X.Q., Pazin, M.J., Feingold, E.A., Wold, B., Taylor, J., Mortazavi, A., Weissman, S.M., Stamatoyannopoulos, J.A., Snyder, M.P., Guigo, R., Gingeras, T.R., Gilbert, D.M., Hardison, R.C., Beer, M.A., Ren, B., and The Mouse, E.C. (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature 515, 355–364.
Zheng, D., Frankish, A., Baertsch, R., Kapranov, P., Reymond, A., Choo, S.W., Lu, Y., Denoeud, F., Antonarakis, S.E., and Snyder, M. (2007). Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res 17, 839–851.
Zhou, Y., Gong, W., Xiao, J., Wu, J., Pan, L., Li, X., Wang, X., Wang, W., Hu, S., and Yu, J. (2014). Transcriptomic analysis reveals key regulators of mammogenesis and the pregnancy-lactation cycle. Sci China Life Sci 57, 340–355.

SUPPORTING INFORMATION

Figure S1 Cutoff for expressed gene.

Figure S2 Length of eight repeat types.

Figure S3 Intron length comparison among introns with or without LTR and LINE.

Figure S4 Gene length comparison among four subclass of non-coding genes.

Figure S5 Expression patterns of non-coding genes in comparison to protein-coding genes.

Figure S6 Co-expression between neighboring lincRNA and protein-coding gene.

Figure S7 Correlation of sense-antisense gene pairs.

Figure S8 Function of sense genes.

Figure S9 Expression-based association matrix of lincRNA loci (rows) and functional gene sets (columns) resulted from GSEA.

Figure S10 RT-PCR result of tissue-specific lincRNA.

Figure S11 Expression comparison between protein-coding and non-coding genes with similar RPKM.

Figure S12 An example of mouse novel lincRNA with orthologous transcripts.

Table S1 Top 10 GO functional annotation for all the clusters of lincRNA

Dataset S1 The catalog of 16,249 non-coding genes (21,569 non-coding RNAs) identified from 15 mouse tissues

Dataset S2 Non-coding genes supported by ChIP-seq data and CAGE

Dataset S3 RT-PCR primers of lincRNAs

Dataset S4 LincRNAs with orthologous regions in human (hg19) and the Transmap orthologous

The supporting information is available online at life.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.