RNA

A PUBLICATION OF THE RNA SOCIETY

Conserved expression of lincRNA during human and macaque prefrontal cortex development and maturation

Zhisong He, Hindrike Bammann, Dingding Han, et al.

RNA published online May 20, 2014

Conserved expression of lincRNA during human and macaque prefrontal cortex development and maturation

ZHISONG HE,[1,2,4] HINDRIKE BAMMANN,[1,3,4] DINGDING HAN,[1,4] GANGCAI XIE,[1,2] and PHILIPP KHAITOVICH[1,3,5]

[1]CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai 200031, China
[2]Graduate School of Chinese Academy of Sciences, 100039 Beijing, China
[3]Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany

ABSTRACT

The current annotation of the human genome includes more than 12,000 long intergenic noncoding RNAs (lincRNA). While a handful of lincRNA have been shown to play important regulatory roles, the functionality of most remains unclear. Here, we examined the expression conservation and putative functionality of lincRNA in human and macaque prefrontal cortex (PFC) development and maturation. We analyzed transcriptome sequence (RNA-seq) data from 38 human and 40 macaque individuals covering the entire postnatal development interval. Using the human data set, we detected the expression of 5835 lincRNA annotated in GENCODE and further identified 1888 novel lincRNA. Most of these lincRNA show low DNA sequence conservation, as well as low expression levels. Remarkably, developmental expression patterns of these lincRNA were as conserved between humans and macaques as those of protein-coding genes. Transfection of development-associated lincRNA into human SH-SY5Y cells affected gene expression, indicating their regulatory potential. In brain, expression of these putative target genes correlated with the expression of the corresponding lincRNA during human and macaque PFC development. These results support the potential functionality of lincRNA in primate PFC development.

Keywords: lincRNA; prefrontal cortex; development; human; macaque

INTRODUCTION

Long intergenic noncoding RNA (lincRNA) is a cumulative RNA category describing transcripts of length >200 nt that have a low protein-coding potential. Common lincRNAs are transcribed by Pol II polymerase and, as with other Pol II transcripts, are capped, polyadenylated, and frequently spliced (Guttman et al. 2009). Numbers of lincRNA genes in humans and other species continue to increase steadily due to both computational (Jia et al. 2010) and experimental (Guttman et al. 2009, 2010; Khalil et al. 2009; Cabili et al. 2011; Ulitsky et al. 2011; Nam and Bartel 2012; Young et al. 2012) efforts. Thus, using chromatin-state maps, Guttman et al. discovered ~1600 lincRNAs in mouse cell lines (Guttman et al. 2009), while Khalil et al. identified 3269 lincRNAs in human cell lines, including human embryonic stem cells (Khalil et al. 2009). With the help of high-throughput RNA sequencing (RNA-seq), Guttman et al. (2010) and Cabili et al. (2011) identified thousands of further lincRNA gene candidates in the mouse and human genomes, respectively. Finally, detailed characterization of the human transcriptome and epigenome by the ENCODE project resulted in the annotation of a total of 13,248 lincRNA genes (Derrien et al. 2012).

Expression of long noncoding RNAs is not restricted to mammals. By integrating chromatin-state maps and RNA-seq data, Ulitsky et al. identified more than 500 lincRNA genes in zebrafish (Ulitsky et al. 2011). Further, Young et al. identified 1119 putative lincRNA genes in the fruit fly (*Drosophila melanogaster*) genome using modENCODE transcriptome data (Young et al. 2012), while Nam et al. found 170 lincRNA genes in the nematode worm (*Caenorhabditis elegans*) genome (Nam and Bartel 2012).

Across species, lincRNAs tend to share the same characteristic features: low DNA sequence conservation, relatively low expression levels, and high spatial and temporal expression specificity (Cabili et al. 2011; Ulitsky et al. 2011). Given these properties, the functional relevance of lincRNA expression has been questioned (Ponjavic et al. 2007). Nonetheless, the functional importance has been demonstrated for a number of lincRNAs, including *XIST*, *HOTTIP*, and *HOTAIR*

[4]These authors contributed equally to this work.
[5]**Corresponding author**
E-mail khaitovich@eva.mpg.de

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.043075.113. Freely available online through the *RNA* Open Access option.

(Brown et al. 1992; Rinn et al. 2007; Wang et al. 2011). These lincRNAs were shown to play roles in transcriptional regulation through changes in chromatin state. Furthermore, by examining the effects of lincRNA knock-downs in mouse embryonic stem cells, Guttman et al. demonstrated that 17.7% of the tested 147 lincRNA may serve as specific *trans*-regulators of protein-coding genes during cellular differentiation (Guttman et al. 2011). This, together with examples of regulatory roles for lincRNA *CRNDE* (Ellis et al. 2012) and *linc-MD1* (Cesana et al. 2011) in brain and muscle development, indicates that at least some lincRNAs may function as regulators during cell and tissue development and differentiation.

Previous studies have shown that multiple lincRNAs are expressed in the mouse (Mercer et al. 2008, 2010) and human brains (Ng et al. 2012). Here, we investigated patterns, conservation, and potential functional roles of lincRNA expression during human and macaque brain development and maturation in one specific brain region: prefrontal cortex (PFC).

RESULTS

LincRNA identification and characterization

We analyzed lincRNA expression in the human brain using RNA-seq time series data from the PFC region of 38 healthy human individuals with ages from 2 d to 61 yr, as well as 40 healthy macaque individuals with ages from 70 d before birth to 21 yr of age (Supplemental Table S1). The data set contains a total of 378 million human single-ended reads and 446 million macaque single-ended reads of 100 nt in length (Supplemental Table S1).

To identify novel lincRNA expressed in human PFC development, we first mapped the reads to the human genome (hg19) using TopHat (v.2.0.6) (Trapnell et al. 2009), allowing up to two mismatches. This resulted in 337 million (89.1%) uniquely mapped reads (Supplemental Table S1). Using the lincRNA identification procedure described in Cabili et al. (2011), we found 1888 novel putative human lincRNAs expressed in the PFC (Fig. 1A) that did not overlap with the GENCODE (v.16) annotation. Of them, 111 were also annotated by Cabili et al. (2011), while the others have not been previously reported.

To quantify the expression of novel and annotated lincRNAs in human and macaque samples in an unbiased manner, we mapped human and macaque RNA-seq reads to a consensus genome, constructed based on a pairwise whole-genome alignment of the human (hg19) and the rhesus macaque (rheMac3) genomes. Using the STAR mapping software and default parameters, 268 million (70.67%) human reads and 311 million (69.77%) macaque reads could be mapped uniquely to the consensus genome. Expression levels of annotated and novel lincRNAs were quantified as the number of reads per kilobase per million mapped reads (RPKM). Low expressed lincRNAs (with a maximum expression below 1 RPKM or detected in less than half of the samples within one species) were removed from further analyses. In humans, 1061 lincRNAs, including 167 novel ones, were expressed above this detection threshold.

In agreement with previous studies (Guttman et al. 2010), expression levels of lincRNAs were on average lower than expression levels of protein-coding genes. Furthermore, the expression levels of novel lincRNAs were lower than expression levels of annotated lincRNA (Fig. 1B). Novel lincRNAs were also less conserved at the DNA sequence level than annotated lincRNAs (one-sided Wilcoxon test, $P < 0.0001$) (Fig. 1C). Notably, both 894 annotated and 167 novel lincRNAs expressed in PFC were significantly more conserved than the annotated or novel lincRNAs with low or no detectable expression in the PFC (one-sided Wilcoxon test, $P < 0.0001$) (Fig. 1C). Similar results were obtained by mapping human RNA-seq data to the reference human genome (hg19) instead of the human-macaque consensus genome (Supplemental Fig. 1). This finding is noteworthy as it parallels previous observations of higher sequence conservation of protein-coding genes expressed in brain (Jordan et al. 2005; Tuller et al. 2008), which are further confirmed using this RNA-seq data set (Supplemental Fig. 2). Overall sequence conservation of lincRNAs measured as the average phastCon scores of

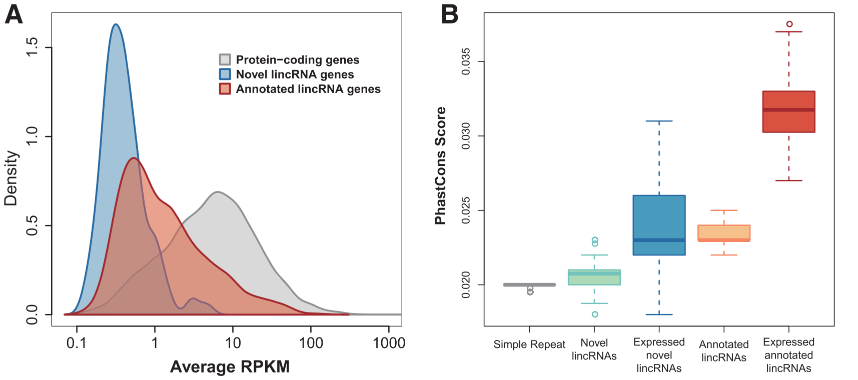

FIGURE 1. Expression and conservation of novel and annotated human lincRNAs. (*A*) Distributions of the average expression levels of novel (blue) and annotated (red) lincRNAs, as well as annotated protein-coding genes (gray), in human PFC samples measured in RPKM (reads per kilobase of exon model per million mapped reads). (*B*) Sequence conservation of lincRNA genes. Shown are median phastCon scores of novel lincRNAs expressed (blue) and not expressed (green) in the human PFC, annotated lincRNAs expressed (red) and not expressed (orange) in the human PFC, as well as median phastCon scores of simple repeats (gray) used as a nonconserved sequence reference. PhastCon scores were calculated based on 11 primate genomes. "Expressed lincRNAs" are defined as lincRNA detected in at least half of the samples with maximal expression > 1 RPKM. The box plots show variation of the median conservation estimates calculated by bootstrapping over phastCon score values 1000 times.

transcribed regions was still substantially lower than the sequence conservation of protein-coding gene exons (median = 0.139, one-sided KS-test, $P < 0.0001$).

In agreement with lower DNA sequence conservation of the lincRNA genes, a smaller proportion of lincRNA was detected to be expressed in both species compared to the protein-coding genes. Using the consensus genome, 13,722 (93%) of 14,745 protein-coding genes expressed in human PFC at RPKM > 1 passed this detection threshold in macaque PFC. In contrast, only 514 (48%) of 1061 lincRNAs expressed in human PFC at RPKM > 1 passed this detection threshold in macaque PFC. Mapping human and macaque RNA-seq data to the respective genomes, resulted in similar observations: 10,052 (93%) of 10,820 protein-coding genes and 391 (51%) of 770 lincRNAs expressed in human PFC at RPKM > 1 passed this detection threshold in macaque PFC. These results support previous observations of lower sequence conservation and greater turnover rate for lincRNA genes (Kutter et al. 2012).

LincRNA expression changes with age

Among 1061 lincRNAs expressed in the human PFC time series, 409 (38.5%) showed significant expression level changes with age (age-test) (Somel et al. 2009)—$P < 0.05$ after Benjamini-Hochberg correction, permutation-based FDR < 5%. Similarly, 45.9% of protein-coding genes expressed in PFC showed significant expression level changes with age. The lower proportion of age-related lincRNAs was partly due to their lower average expression: after subsampling the protein-coding genes based on the distribution of lincRNA expression levels, the proportion of age-related protein-coding genes (42.6 ± 1.4%) was closer to that of lincRNA (Supplemental Fig. 3).

To test the reproducibility of observed age-related changes in lincRNA expression, we used a published RNA-seq time series data set containing measurements from the PFC of 14 healthy human individuals (Mazin et al. 2012). Of 409 lincRNAs showing age-related expression profiles in the original time series data, 292 were expressed in both data sets. Developmental expression trajectories of these lincRNA were consistent between the two data sets, as shown by the Pearson correlation coefficient distribution (Supplemental Fig. 4). Specifically, 209 of the 292 lincRNAs (71.6%) showed significant positive correlation between the data sets (one-sided Pearson correlation test, $P < 0.05$). Importantly, reproducibility of lincRNA expression profiles was comparable to that of the age-related protein-coding genes sampled from the same expression level distribution: 75.8 ± 2.3% of them showed significant positive correlation between the data sets (Supplemental Fig. 4).

To obtain an overview of expression patterns formed by the 409 age-related lincRNAs and the 6771 age-related protein-coding genes, we grouped them into clusters according to eight main expression patterns (Fig. 2A). When compared to protein-coding genes, age-related expressed lincRNAs were enriched in two patterns, Group 3 and Group 8, both showing decreases in gene expression levels in late development (χ^2 test, $P < 0.05$ after multiple testing correction) (Supplemental Fig. 5). Protein-coding genes in these two groups were significantly enriched in Gene Ontology (GO) terms that included "plasma membrane," "neurological system process," "synaptic transmission," "calcium ion binding," "transmission of nerve impulse," and "cell-cell signaling" (Supplemental Table S2).

We next tested whether the human age-related lincRNA expression patterns are conserved in macaques. Based on macaque RNA-seq reads mapped to the consensus genome, expression of 478 annotated and 36 novel human lincRNAs could be detected above the expression level cutoff of one RPKM, in both human and macaque samples. Of these, 196 and 11 showed significant age-related expression changes in the human time series, respectively. Despite the relatively small overlap of lincRNA genes expressed in human and macaque PFC development that could be observed above the detection threshold, the developmental expression profiles of these 196 annotated and 11 novel lincRNAs correlated positively and significantly between the human and macaque time series (median $r = 0.57$ and $r = 0.48$ for annotated and novel lincRNAs, respectively, permutation $q < 0.001$). Importantly, this level of developmental profile conservation was comparable to the one observed for protein-coding genes sampled from the same expression level distribution (median $r = 0.60 \pm 0.05$) (Fig. 2B). This result was not caused by the use of the consensus genome, as mapping RNA-seq data to the human and rhesus macaque genomes separately yielded consistent observations (Supplemental Fig. 6). Thus, despite the lower DNA sequence conservation and higher turnover rate of lincRNAs, developmental expression trajectories of lincRNAs and protein-coding genes expressed in both species showed comparable conservation levels.

For protein-coding genes, expression level conservation is known to correlate with the DNA sequence conservation of the transcribed region, as well as the core promoter elements (Khaitovich et al. 2005; Donaldson and Gottgens 2006). To test whether this holds true for lincRNAs, we compared the conservation of lincRNA developmental expression profiles with the DNA sequence conservation of lincRNA transcribed regions, as well as the core promoter regions situated 200 bp upstream of and 50 bp downstream from the transcription start site. Using average phastCon scores as a proxy for DNA sequence conservation, we, indeed, found a significant correlation between the conservation of lincRNAs' developmental expression profiles and the sequence of the transcribed regions (Spearman correlation, $\rho = 0.177$, $P = 0.01$) and core promoter regions (Spearman correlation, $\rho = 0.150$, $P = 0.03$). The significance of these effects was additionally validated by 1000 permutations of gene labels ($P < 0.02$) (Supplemental Fig. 7). Notably, the strength of the correlation between the conservation of developmental

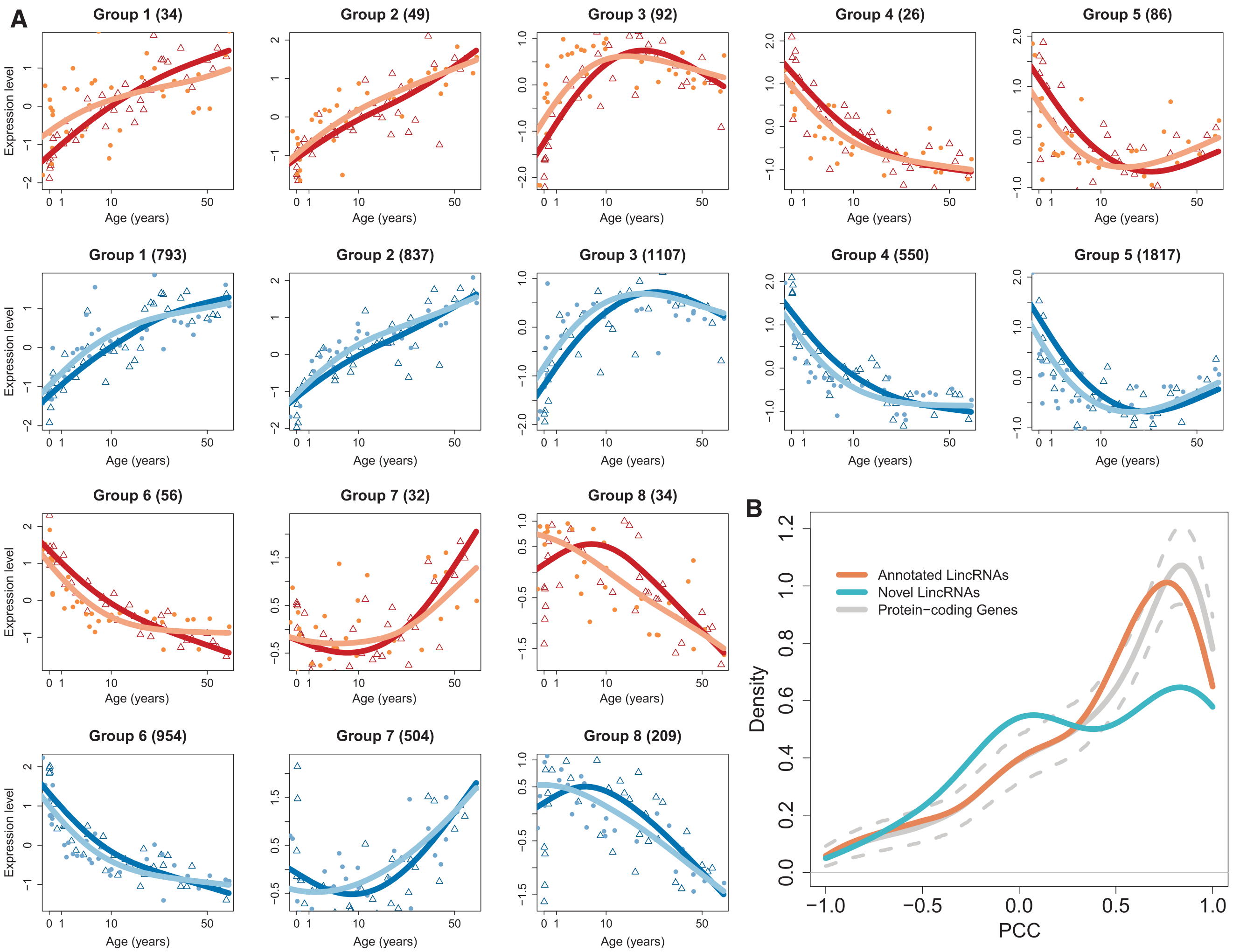

FIGURE 2. Developmental expression patterns of lincRNAs and protein-coding genes in human and macaque PFC. (*A*) Expression patterns of lincRNA (red: human, light red: macaque) and protein-coding genes (blue: human, light blue: macaque) defined as age-related in human PFC development. Expression values were z-transformed before plotting. Numbers in brackets *above* each panel indicate numbers of lincRNAs or protein-coding genes in the cluster. On all plots, macaque ages are linearly transformed to the human age scale. (*B*) Distribution of Pearson correlation coefficients (PCC) based on the comparison of developmental expression profiles between humans and macaques for annotated lincRNAs (orange), novel lincRNAs (green), and protein-coding genes (gray). Confidence intervals of the PCC distribution for protein-coding genes are shown by the dashed gray lines and were calculated by subsampling protein-coding genes according to the expression distribution of lincRNA genes 1000 times.

expression profiles and the DNA sequence observed for lincRNAs was comparable to that observed in our data for protein-coding genes: ρ = 0.163 for transcribed regions and ρ = 0.162 for core promoter regions ($P < 0.001$ calculated in 1000 permutations of gene labels) (Supplemental Fig. 7).

LincRNA transfection

To obtain deeper insights into the functions of lincRNAs in PFC development, we overexpressed three lincRNA transcripts showing age-related expression in the human PFC in a human neuroblastoma cell line (SH-SY5Y) (Supplemental Table S3). We tested the effect of lincRNA overexpression by examining the transcriptomes of cells transfected with lincRNA constructs or with an empty vector (mock transfection). All transfection experiments were carried out in triplicate. The cell line transcriptomes were examined 24 h after transfection using RNA-seq. The expression levels of the three transfected lincRNAs were within the expression range of endogenous transcripts (Fig. 3A).

The overexpression of each of the three lincRNA transcripts showed significant effects on gene expression in the cell line, affecting the expression levels of 42, 56, and 103 genes—19, 30, and 87 of them protein-coding ([edgeR] [Robinson et al. 2010], FDR < 10% after Benjamini-Hochberg correction) (Fig. 3B). To assess whether overexpression itself results in the observed changes in cell transcript expression, we compared genes affected by overexpression of

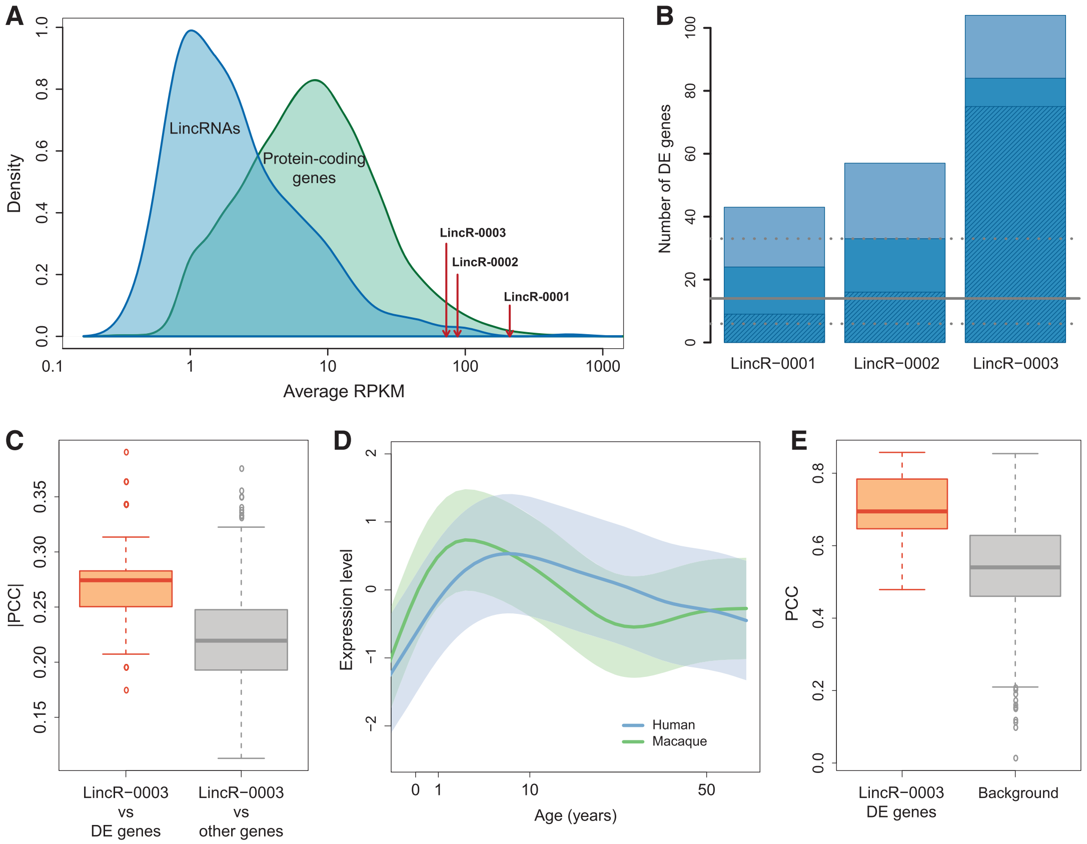

FIGURE 3. The effect of lincRNA overexpression on putative target genes in SY5Y cells. (*A*) Expression level distributions of protein-coding genes (green) and lincRNAs (blue) in the SH-SY5Y human neuroblastoma cell line. The red arrows mark the average expression levels of the three transfected lincRNAs after overexpression. (*B*) Number of genes showing differential expression (DE) between mock and lincRNA transfection in the SH-SY5Y cell line plotted for each of the three development-associated lincRNAs. Genes affected by more than one lincRNA transfection are shown in light blue, with independently affected genes in dark blue. The shaded bars show the number of independent protein-coding genes. The solid gray line shows the number of differentially expressed genes expected by chance. The two gray dotted lines show the 90% confidence interval for the chance distribution. (*C*) Regulatory effects of *lincR-0003* in brain. The red box plots show the medians of the absolute values of the Pearson correlation coefficients based on the developmental expression profile of a lincRNA and its independent protein-coding target genes, identified in SY5Y cells. The variance of the median estimates was calculated by bootstrapping target genes 1000 times. The gray boxes show the distribution of the median values of the absolute Pearson correlation coefficients based on the developmental expression profiles of lincRNAs and the same number of nontarget protein-coding genes sampled 1000 times. (*D*) Expression profile of *lincR-0003* in human (blue) and macaque (green) PFC development. Macaque ages are linearly transformed to the human scale. Expression values were z-transformed. Shaded areas indicate standard deviation of expression level estimates. (*E*) Conservation of *lincR-0003* target profiles between human and macaque PFC development. The red box plots show the medians of the Pearson correlation coefficients based on developmental expression profiles of independent *lincR-0003* protein-coding target genes in human and macaque PFC. The variance estimate and the background distribution (gray box plot) are as in panel *C*.

the three lincRNAs. We, indeed, find a greater than expected overlap between affected genes: 18 of 42 transcripts affected by *lincR-0001*, 23 of 56 transcripts affected by *lincR-0002*, and 19 of 103 transcripts affected by *lincR-0003* overlapped across at least two transfection experiments. These genes were excluded from further analyses.

To test the functional relevance of putative lincRNA targets detected in the cell line, we examined lincRNA/target coexpression in the human PFC time series. Putative targets of *lincR-0001* and *lincR-0002* did not show significant coexpression with their presumed regulatory lincRNA in human PFC (Supplemental Fig. 8). In contrast, there was stronger than expected coexpression of *lincR-0003* and its 67 independent protein-coding putative target genes expressed in the human PFC time series (one-sided Wilcoxon test, $P = 0.014$) (Fig. 3C). This effect was not caused by the higher expression of putative *lincR-0003* targets, as selecting background genes with the same expression level distribution did not abolish the result (Supplemental Fig. 9).

We next asked whether the regulatory relationship between *lincR-0003* and its protein-coding putative target genes was preserved in macaque PFC development. Indeed, the expression profile of *lincR-0003* was more conserved between the human and macaque PFC time series than the expression profiles of other genes expressed in PFC ($r = 0.68$) (Fig. 3D). Notably, *lincR-0003* target genes, predicted based on the cell line experiment, also showed more conserved developmental expression trajectories in human and macaque PFC development than other genes expressed in both species (median $r = 0.69$, one-sided Wilcoxon test, $P = 0.0002$) (Fig. 3E). This observation was not caused by differences in expression levels between lincRNA targets and other genes expressed in PFC, as selecting background genes with the same expression level distribution did not abolish the result (Supplemental Fig. 9). Thus, the conservation of *lincR-0003* expression between human and macaque PFC may have contributed to the conservation of the expression trajectories of its putative target genes.

DISCUSSION

In this study, we quantified the expression of 5835 lincRNAs annotated in GENCODE in a human PFC developmental time series and further identified 1888 novel lincRNA transcripts that are not listed in the GENCODE annotation. Most of the lincRNAs, both annotated and novel, show low DNA sequence conservation, as well as low expression levels across the developmental time series. Still, lincRNAs showing substantial expression in PFC were significantly more conserved than their nonexpressed counterparts. This is interesting, as protein-coding genes expressed in brain and, specifically in PFC, were shown to be more conserved at the DNA sequence level and have greater conservation of their promoter region sequences (Jordan et al. 2005; Tuller et al. 2008). Our results indicate that the sequences of lincRNAs expressed in PFC are also more constrained than the sequences of other human lincRNAs.

More remarkably, while overall sequence conservation and ortholog numbers of lincRNAs expressed in human and macaque PFC development remain low compared to that of protein-coding genes, for lincRNAs expressed in both species, the conservation of developmental expression profiles between humans and macaques was comparable to that of protein-coding genes. This finding indicates that despite the low conservation at the DNA sequence level and high turnover rate of expressed lincRNA genes (Kutter et al. 2012), some lincRNAs might play conserved functional roles in primate PFC development. Alternatively, this result may show that PFC developmental expression patterns are guided by the same conserved regulatory mechanisms in humans and macaques, resulting in shared lincRNA expression as an evolutionary neutral by-product. The higher conservation of lincRNAs expressed in PFC may, in this case, reflect purifying selection removing lincRNA variants that are sufficiently deleterious.

Our cell line results appear, however, to argue against full evolutionary neutrality of lincRNA expression in the human and macaque PFC. The strongest argument toward the relevance of expression effects of *lincR-0003* identified in the cell line comes from the conserved and correlated expression of *lincR-0003* and its putative target genes in human and macaque PFC development. It has to be noted, however, that, in the absence of further functional studies, this result provides no evidence of physiological and functional significance of these regulatory effects in either cell line or in the primate brain. Still, these findings suggest that at least some of the lincRNAs showing conserved expression profiles between species may serve as *trans*-regulators of specific gene groups during PFC development. This notion coincides with previous results showing extensive lincRNA-driven *trans*-regulation in mouse embryonic stem cells (Guttman et al. 2011). More generally, our study illustrates the general usefulness of an evolutionary approach to the investigation of lincRNA functionality. Further studies of lincRNA expression across species and tissues combined with painstaking functional studies of individual lincRNAs will be needed to assess the functionality of these transcripts.

MATERIALS AND METHODS

Ethics statement

Informed consent for the use of human tissues for research was obtained in writing from all donors or their next of kin. All nonhuman primates used in this study suffered sudden deaths for reasons other than their participation in this study and without any relation to the tissue used. The Biomedical Research Ethics Committee of the Shanghai Institutes for Biological Sciences completed the review of the use and care of the animals in the research project (approval ID: ER-SIBS-260802P).

RNA-seq data sets

Sample material was dissected post-mortem from the prefrontal cortex of 38 human and 40 rhesus macaque individuals, with ages from 2 d to 61 yr (humans) and 70 d before birth to 21 yr old (macaques) (Supplemental Table S1). The poly(A)+ RNA fraction was sequenced on the Illumina HiSeq 2000 platform, using a standard cDNA library preparation protocol. Approximately 378 million human and 446 million macaque single-end reads, each 100 nt in length, were obtained in total. All data are deposited in the GEO database under accession number GSE51264. This data set was used in transcriptome reconstruction and gene expression level estimation analyses. To validate developmental expression trajectories in humans, we further used published RNA-seq time series data from the PFC region of 14 healthy humans with ages from 2 d to 98 yr (Mazin et al. 2012).

Human-macaque consensus reference genome construction

Chained and netted alignment files for human (hg19) and macaque (rheMac3), aligned using BLASTZ (Schwartz et al. 2003), were downloaded from the UCSC Genome Browser. On the basis of these alignments, we constructed a human-macaque consensus genome. Specifically, we replaced the human genomic sequences with "**N**"s for all discordant genomic sites, as well as replacing insertions and deletions with "N"s, in the consensus genome. Furthermore, the 6-bp regions flanking each insertion or deletion site were masked by "N"s in the resulting human-macaque consensus reference genome.

Read mapping

To identify novel human lincRNA genes, sequence reads from the prefrontal cortex time series, containing 38 human PFC samples, were mapped to the *hg19* reference genome using TopHat v.2.0.6 (Trapnell et al. 2009), allowing up to two mismatches. No annotations for genes or splice junctions were used during TopHat mapping. The same procedure was applied to the published strand-specific RNA-seq time series containing 14 human PFC samples (Mazin et al. 2012). Here, we used STAR v.2.3.0 (Dobin et al. 2013) with default parameters, allowing up to nine mismatches, to map sequence reads from the 38 human and 40 macaque PFC samples to the human-macaque consensus reference genome. To test whether mapping to the consensus reference genome produces any biases or artifacts, we further mapped 40 macaque PFC samples to the rheMac3 reference genome using TopHat with the above listed parameters. For every mapping procedure, only uniquely mapped reads were used in further analyses.

De novo discovery of novel lincRNA candidates

The pipeline for the de novo identification of lincRNA candidates was based on the procedure developed by Cabili et al. (2011). All reads uniquely mapped to the human genome were combined using SAMtools and assembled to transcripts using Cufflinks v.2.1.1. (Trapnell et al. 2010). In order to maximize the sensitivity of transcript detection, the number of required reads for transcript assembly was lowered to one. In doing so, every mapped read was considered in the assembly procedure. The assembled transcripts were further filtered using the GENCODE (v.16) annotation: all annotated transcripts and transcripts with exonic overlap to the GENCODE annotation were excluded. We further excluded transcripts that were shorter than 200 bp, as well as transcripts with only one exon.

Protein-coding potential of the obtained novel transcripts was estimated by PhyloCSF (Lin et al. 2011). Transcripts with an open reading frame (ORF) scoring higher than 100 in one of all six possible reading frames were considered as potentially protein-coding and excluded from the list of putative novel human lincRNA. Further, potential lincRNA genes were translated in all possible reading frames to protein sequences with transseq on the Galaxy web server (Giardine et al. 2005; Blankenberg et al. 2010; Goecks et al. 2010). The translated sequences were used to search for motifs associated with protein-coding genes, using the Pfam database (Punta et al. 2012). All transcripts with a significant Pfam-A domain in any of the reading frames were excluded from the list of putative novel human lincRNAs.

LincRNA annotation and gene expression level estimation

GENCODE v.16 (Dobin et al. 2013) annotation was used to define protein-coding and annotated lincRNA genes. Novel lincRNA gene candidates without any overlap to the GENCODE annotation were annotated as described above.

In order to estimate the expression level of each gene in each sample, we counted the number of reads that had at least 1 nt overlap with at least one exon of this gene. The expression level of gene *i* in sample *j* can be represented as RPKM:

$$\mathrm{RPKM}_{i,j} = \frac{n_{i,j} \times 10^3 \times 10^6}{l_i \times N_j}$$

Here, $n_{i,j}$ is the number of reads overlapping with the exons of gene *I*, l_i is the length of gene *I*, and N_j is the total number of uniquely mapped reads in sample *j*. Only genes with maximum RPKM ≥ 1 and reads detected in more than half of the samples were classified as expressed and used in the following analysis.

Mapping of human transcript annotation to the macaque genome using liftOver

For each gene in the human annotation (GENCODE v.16 including novel lincRNAs identified in this study), we first merged overlapping exons within genes. We then mapped resulting human exons to the macaque reference genome (rheMac3) using reciprocal liftOver (Hinrichs et al. 2006). The ortholog gene regions within the macaque genome were required to contain more than half of the merged exons located on the same chromosome and in the correct order.

LincRNAs with age-related expression and their expression categories

To test the effect of age on the expression level of each gene, the polynomial regression model developed by Somel et al. (2009) was applied. Briefly, for each gene the best polynomial regression model, with age as a predictor and RPKM as response, was selected in a hierarchical manner from a linear model up to the third-degree polynomial. In order to produce a more uniform sample age distribution, the original sample ages were transformed using the square root. The significance of the regression model was estimated by *F*-test. Genes with age-test *P*-values smaller than 0.05 after Benjamini-Hochberg correction were considered as age-related. The FDR of the age-test was estimated by 1000-time random permutation of ages. The median of the permutation distribution was used as a null expectation.

Age-related lincRNAs and protein-coding genes were grouped into eight groups, according to all basic expression patterns with expression trends separated into early and late developmental intervals. Specifically, the patterns were designed to discriminate between increasing, decreasing, and stable expression during early and late developmental stages. Stable expression over the whole lifespan was not considered, as it is not age-related. The expression trajectory of each gene was correlated to each of the eight patterns using Spearman's rank correlation. Each gene was assigned to the best-correlated pattern.

Expression profile comparison between species and data sets

In order to compare expression profiles between humans and macaques for a given gene with age-related expression in the human PFC, we applied cubic spline interpolation, as implemented in the *smooth.spline* function in R, to both the human and the macaque time series data to obtain the regressive trajectory. A generalized cross-validation procedure was also implemented in the *smooth. spline* R function to choose suitable smoothing parameters to avoid overfitting (Green and Silverman 1994). The age points of macaque PFC data were transformed to the human scale by multiplying the age by three, as the approximate human lifespan is about three times longer than the lifespan of a macaque. The regression curves were compared between species by calculating the Pearson correlation coefficient based on values interpolated from the fitted spline curves at 40 uniformly distributed time points. The same procedure was used to estimate the reproducibility between two human data sets.

Taking the expression differences between lincRNAs and protein-coding genes into account, we subsampled protein-coding genes according to the $\log_{10}$-transformed RPKM distribution of age-related human lincRNA genes. Therefore, the interval between minimum and maximum $\log_{10}$-transformed RPKM values of lincRNA was split into 20 equal bins, and the number of lincRNA in each bin was counted. Protein-coding genes were picked according to their RPKM values, to construct a gene subset resembling the expression distribution of the lincRNA data set. This subsampling was applied 1000 times in order to estimate variance and confidence interval.

LincRNA transfection experiments

The lincRNAs were overexpressed as previously described (Tsai et al. 2010; Yang et al. 2011). Briefly, three lincRNA genes showing age-related expression patterns in human PFC were randomly chosen among all 409 age-related human lincRNAs. The three full-length lincRNA sequences were separately cloned into the pRNAT-CMV3.2 Puro vector under control of a CMV promoter. SH-SY5Y neuroblastoma cells were cultured in a 1:1 mixture of Ham's F12 medium and Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (Hyclone) at 37°C in a 5% CO_2 atmosphere. Cells were transiently transfected with 1 µg plasmids using Lipofectamine 2000 reagent (Invitrogen) according to the manufacturer's protocol at ~80% confluency in six-well plates. After 4 h incubation, the transfection mixture was replaced with 2 mL complete medium.

Transfection was carried out in triplicate for each lincRNA, as well as for a mock transfection (empty vector). Cells were harvested, and total RNA was extracted with Trizol reagent (Invitrogen) 24 h after transfection. In addition, cells were harvested, and RNA was isolated at the 0 h transfection time point (immediately before transfection). Isolated total RNA from each of the 15 samples (three lincRNA transfection triplicates and one mock transfection triplicate collected 24 h after transfection, as well as one triplicate of cells collected at 0 h) was poly(A)-enriched and converted into an Illumina sequencing library using the TruSeq RNA Sample Prep Kit. Libraries were quantified with the Qubit dsDNA BR Assay Kit (Invitrogen) and qualified using an Agilent Technologies 2100 Bioanalyzer (Agilent). Pooled barcoded libraries were sequenced on the Illumina HiSeq 2000 platform using the 100-bp single-read module.

After mapping and gene expression estimation (in read counts) as described above, edgeR (Robinson et al. 2010), a Bioconductor package for differential expression analysis, was applied to identify genes differentially expressed between the mock transfection and each of the lincRNA transfections (FDR < 10%, Benjamini-Hochberg correction).

To estimate the FDR of potential lincRNA targets in the transfection experiments, we randomly selected two groups each with three samples from the 12 transfected samples and including the mock transfection. Assuming no significant difference between these two random groups, edgeR was applied with the same *P*-value cutoff to obtain the number of differentially expressed genes. This process was repeated 1000 times to estimate the FDR for each transfection.

DATA DEPOSITION

All RNA-seq data are deposited in the Gene Expression Omnibus (GEO) database under accession number GSE51264.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank all members of the Comparative Biology Group in Shanghai for helpful discussions and suggestions. We thank Jerome Boyd-Kirkup for editing the manuscript. This work was supported by the National Natural Science Foundation of China (grant number 31171232, 91331204), the 1000 Plan for Foreign Experts (WQ20123100078), and the International Cooperation Program of the Bureau of International Cooperation of the Chinese Academy of Sciences (grant number GJHZ201313).

Received October 17, 2013; accepted April 9, 2014.

REFERENCES

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. 2010. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* Chapter 19: Unit 19.10. 1–21.

Brown CJ, Hendrich BD, Rupert JL, Lafreniere RG, Xing Y, Lawrence J, Willard HF. 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.

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.

Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M, Tramontano A, Bozzoni I. 2011. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* **147:** 358–369.

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

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21.

Donaldson IJ, Gottgens B. 2006. Evolution of candidate transcriptional regulatory motifs since the human-chimpanzee divergence. *Genome Biol* **7:** R52.

Ellis BC, Molloy PL, Graham LD. 2012. CRNDE: a long non-coding RNA involved in CanceR, Neurobiology, and DEvelopment. *Front Genet* **3:** 270.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, et al. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15:** 1451–1455.

Goecks J, Nekrutenko A, Taylor J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11:** R86.

Green PJ, Silverman BW. 1994. *Nonparametric regression and generalized linear models*. Chapman and Hall/CRC, London.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458:** 223–227.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, et al. 2010. *Ab initio* reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28:** 503–510.

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, et al. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477:** 295–300.

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**(Database issue): D590–D598.

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.

Jordan IK, Marino-Ramirez L, Koonin EV. 2005. Evolutionary significance of gene expression divergence. *Gene* **345:** 119–126.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309:** 1850–1854.

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, et al. 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci* **106:** 11667–11672.

Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC. 2012. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet* **8:** e1002841.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27:** i275–i282.

Mazin P, Xiong J, Liu X, Yan Z, Zhang X, Li M, He L, Somel M, Yuan Y, Chen YP, et al. 2012. Widespread splicing changes in human brain development and aging. *Mol Syst Biol* **9:** 633.

Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS. 2008. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci* **105:** 716–721.

Mercer TR, Qureshi IA, Gokhan S, Dinger ME, Li G, Mattick JS, Mehler MF. 2010. Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation. *BMC Neurosci* **11:** 14.

Nam JW, Bartel DP. 2012. Long noncoding RNAs in *C. elegans*. *Genome Res* **22:** 2529–2540.

Ng SY, Johnson R, Stanton LW. 2012. Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *EMBO J* **31:** 522–533.

Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17:** 556–565.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al. 2012. The Pfam protein families database. *Nucleic Acids Res* **40**(Database issue): D290–D301.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, et al. 2007. Functional demarcation of active and silent chromatin domains in human *HOX* loci by noncoding RNAs. *Cell* **129:** 1311–1323.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. 2003. Human–mouse alignments with BLASTZ. *Genome Res* **13:** 103–107.

Somel M, Franz H, Yan Z, Lorenc A, Guo S, Giger T, Kelso J, Nickel B, Dannemann M, Bahn S, et al. 2009. Transcriptional neoteny in the human brain. *Proc Natl Acad Sci* **106:** 5743–5748.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25:** 1105–1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511–515.

Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329:** 689–693.

Tuller T, Kupiec M, Ruppin E. 2008. Evolutionary rate and gene expression across different brain regions. *Genome Biol* **9:** R142.

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.

Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, et al. 2011. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472:** 120–124.

Yang F, Zhang L, Huo XS, Yuan JH, Xu D, Yuan SX, Zhu N, Zhou WP, Yang GS, Wang YZ, et al. 2011. Long noncoding RNA high expression in hepatocellular carcinoma facilitates tumor growth through enhancer of zeste homolog 2 in humans. *Hepatology* **54:** 1679–1689.

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.