Genome-Wide Identification of Long Intergenic Noncoding RNA Genes and Their Potential Association with Domestication in Pigs

Zhong-Yin Zhou[1,2], Ai-Min Li[3], Adeniyi C. Adeola[2,4], Yan-Hu Liu[5], David M. Irwin[2,6,7], Hai-Bing Xie[2,*], and Ya-Ping Zhang[1,2,5,*]

[1]Department of Molecular and Cell Biology, School of Life Sciences, University of Science and Technology of China, Hefei, China

[2]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

[3]School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, China

[4]Kunming College of Life Science, University of Chinese Academy of Sciences, Kunming, China

[5]Laboratory for Conservation and Utilization of Bio-resources, Yunnan University, Kunming, China

[6]Department of Laboratory Medicine and Pathobiology, University of Toronto, Ontario, Canada

[7]Banting and Best Diabetes Centre, University of Toronto, Ontario, Canada

*Corresponding author: E-mail: zhangyp@mail.kiz.ac.cn; xiehb@mail.kiz.ac.cn.

Accepted: May 23, 2014

Abstract

Thousands of long intergenic noncoding RNAs (lincRNAs) have been identified in the human and mouse genomes, some of which play important roles in fundamental biological processes. The pig is an important domesticated animal, however, pig lincRNAs remain poorly characterized and it is unknown if they were involved in the domestication of the pig. Here, we used available RNA-seq resources derived from 93 samples and expressed sequence tag data sets, and identified 6,621 lincRNA transcripts from 4,515 gene loci. Among the identified lincRNAs, some lincRNA genes exhibit synteny and sequence conservation, including *linc-sscg2561*, whose gene neighbor *Dnmt3a* is associated with emotional behaviors. Both *linc-sscg2561* and *Dnmt3a* show differential expression in the frontal cortex between domesticated pigs and wild boars, suggesting a possible role in pig domestication. This study provides the first comprehensive genome-wide analysis of pig lincRNAs.

Key words: pig, lincRNA, domestication.

Introduction

Genome-wide analyses have uncovered more than 9,000 long intergenic noncoding RNA (lincRNA) genes in the human genome (Khalil et al. 2009; Jia et al. 2010; Cabili et al. 2011; Derrien et al. 2012), and more than 10,000 lincRNA transcripts in the mouse genome (Ravasi et al. 2006; Mitchell Guttman et al. 2009; Guttman et al. 2010). Several studies have indicated that some lincRNAs play important roles in fundamental biological processes, such as dosage compensation (Borsani et al. 1991; Brockdorff et al. 1992; Brown et al. 1992; Payer and Lee 2008), maintenance of pluripotency (Guttman et al. 2011), transcriptional regulation (Huarte et al. 2010; Orom et al. 2010; Hung et al. 2011), and epigenetic regulation (Martianov et al. 2007; Rinn et al. 2007).

The pig is an important domesticated animal and is a significant large-animal model for medical research. Thousands of years of selection have created considerable diversity in the phenotypes of pigs. Many protein-coding genes with major effects on diversity in pigs have been identified, including *IGF2* (Van Laere et al. 2003), *NR6A1* (Mikawa et al. 2007), *MC1R* (Fang et al. 2009), and *RYR1* (Fujii et al. 1991). The contribution of changes in lincRNAs to the domestication of pigs is currently unknown. To address this question, a comprehensive genome-wide identification of lincRNAs is required.

Here, we identified a total of 6,621 lincRNAs, encoded by 4,515 gene loci, and profiled the expression of these lincRNAs in various tissues. Several lincRNA sequences were found to share homology with sequences in the human and mouse

lincRNA data sets. Finally, we profiled changes in the expression of lincRNAs using RNA sequence (RNA-seq) data sets from the brain of domesticated pigs and wild boars, and found one lincRNA (*linc-sscg2561*), and its neighboring gene *Dnmt3a*, which might be associated with differences in emotional behavior between the domesticated pig and the wild boar.

Results and Discussion

Identification of lincRNAs Based on Expressed Sequence Tag and RNA Sequencing Data Sets

Only 47 pig lincRNA transcripts are annotated in the Ensembl database (version 73), a quantity far lower than that known for the human or mouse. As the human genome has identified about 9,000 lincRNA genes (Derrien et al. 2012), and the pig genome is of comparable size and contains a similar number of protein-coding genes (Groenen et al. 2012), one might expect that the pig will also should have similar number of lincRNA genes. Hence, a large number of pig lincRNAs are likely undetermined. To comprehensively identify pig lincRNAs, we used expressed sequence tag (EST) (UniGene) and RNA-seq data sets and performed searches using the following criteria that provide a strict definition for lincRNAs: 1) Transcript must include $\geq$2 exons, 2) length should be $\geq$200 nt, 3) must be located at least 500 bp away from any protein-coding genes or house-keeping ncRNAs genes annotated in the Ensembl *Sus scrofa* 10.2 gene set (GTF), and 4) Coding Potential Calculator (CPC) score of less than −1, as calculated using the CPC tool (Kong et al. 2007) to assess the protein-coding potential for every transcript (fig. 1*A*). A total of 1,125 lincRNA transcripts from 1,090 intergenic regions were identified in the pig genome from the EST data set.

High-throughput RNA sequencing has been used to identify lincRNAs in diverse species (Cabili et al. 2011; Ulitsky et al. 2011; Liu et al. 2012). To identify novel pig lincRNA, we used ten RNA sequencing data sets derived from various tissues of the pig. RNA sequencing reads were aligned to the Sus scrofa 10.2 genome (Groenen et al. 2012) using TopHat (Trapnell et al. 2009). Mapped reads were assembled into transcripts using Cufflinks and Cuffcompare (Trapnell et al. 2010, 2012). The number of transcripts identified in the intergenic regions from these ten studies ranged from 2,999 to 48,272 (supplementary table S1, Supplementary Material online). Using our criteria, the number of lincRNAs for each of the ten studies ranged from 222 to 3,010 (supplementary table S1, Supplementary Material online), and could be merged into a single data set of 5,594 lincRNAs encoded by 3,753 gene loci. Of these lincRNAs, 328 genes were detected from both the RNA-seq and EST data sets, with a final total of 6,621 unique lincRNA transcripts being identified.

To determine the basic features of pig lincRNAs, we compared our identified lincRNAs with mRNAs identified by Ensembl. LincRNAs are shorter in length than protein-coding transcripts (supplementary fig. S1*A*, Supplementary Material online), and their genes tend to contain fewer exons (supplementary fig. S1*B*, Supplementary Material online). The length and number of exons for lincRNAs might have been overestimated in our study as transcripts with only a single exon were excluded as lincRNAs. Despite having shorter transcript lengths, exons for pig lincRNAs were on average larger (average 451 nt) than those for protein-coding genes (average 221 nt). The distance between lincRNA genes and their closest protein-coding genes was greater than the median distance between adjacent protein-coding genes (median 80,818 nt for mRNA–lincRNA intervals, compared with 36,072 nt for mRNA–mRNA intervals; Mann–Whitney $P < 2.2 \times 10^{-16}$; fig. 1*B*); 1,354 of the lincRNA genes are located within 10 kb of a protein-coding gene. Gene ontology (GO) enrichment analyses were conducted for the set of protein-coding genes proximal ($\leq$10 kb) to these lincRNAs. These closest neighbors of pig lincRNAs are enriched for GO terms associated with transcriptional regulation processes (supplementary table S2, Supplementary Material online), which is consistent with a previous report in other species (Ulitsky et al. 2011). The distances between lincRNA genes and their closest protein-coding genes were larger than the lengths of the introns in the protein-coding genes (Mann–Whitney $P < 2.2 \times 10^{-16}$; fig. 1*B*), indicating that these lincRNAs are independent transcripts, rather than being unannotated exons of these protein-coding genes.

Tissue Expression Profile of Pig lincRNAs

We used RNA-seq data sets from ten tissues (ERA178851) (Farajzadeh et al. 2013) of wild boars to characterize the expression pattern of the lincRNA genes. The expression level of lincRNA genes is lower than that of protein-coding genes (Kolmogorov–Smirnov test, $P < 2.2 \times 10^{-16}$; fig. 1*C*), which has also been observed in other mammalian species (Ravasi et al. 2006; Cabili et al. 2011). This feature implies that lincRNAs and mRNAs have a number of differences in their biogenesis, processing, stability, and spatial–temporal expression patterns.

Protein-coding genes proximal to lincRNAs are enriched in specific gene functions. Previous studies have indicated that in some mammalian species, lincRNAs may act in *cis* to regulate the expression of their neighboring protein-coding genes (Mercer et al. 2009; Orom et al. 2010; Wang et al. 2011). To determine whether lincRNAs in the pig had a similar effect on expression, we focused on pig lincRNAs that are located within 10 kb of a protein-coding gene and tested to see whether there was a correlation in the expression patterns between the lincRNAs and their neighboring protein-coding genes. Across ten tissues, expression of the closely linked lincRNAs tended to correlate with that of their protein-coding neighbors (average Spearman correlation $r^2 = 0.31$).

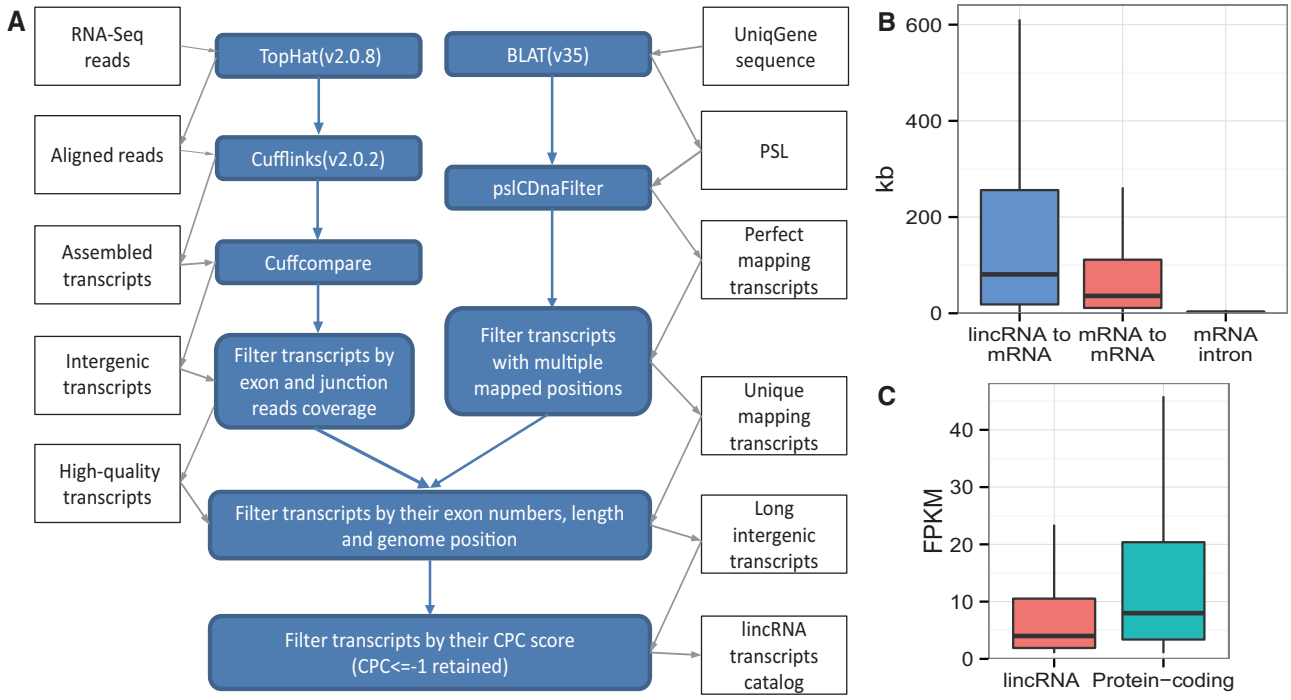

Fig. 1.—Identification and characterization of lincRNA genes in the pig. (*A*) Pipeline for the identification of lincRNAs. (*B*) Comparison of the mRNA–lincRNA intervals, mRNA–mRNA intervals, and sizes of mRNA introns. (*C*) Expression levels of lincRNA and protein-coding genes detected using RNA-seq data from ten tissues (ERA178851).

A similar magnitude of correlation was observed for adjacent protein-coding genes ($r^2 = 0.28$). The correlated expression of the lincRNAs and their adjacent protein-coding genes suggests that they may share *cis*-regulatory modules or chromatin domains.

Based on the hierarchically clustering of the gene expression profiles, many lincRNA genes exhibit a tissue preferential expression pattern, which is similar to that of the protein-coding genes (supplementary fig. S2, Supplementary Material online). The differential expression patterns of the lincRNAs were further analyzed using Deseq2 with a cutoff of 2-fold change and padj < 0.1 (Anders and Huber 2010). This analysis identified 581 tissue preferential lincRNAs based on the RNA-seq data set (supplementary fig. S3, Supplementary Material online). Interestingly, 261 lincRNAs are preferentially expressed in the frontal cortex and occipital cortex (supplementary fig. S3, Supplementary Material online).

Identification of Sequence Homology with Human and Mouse lincRNAs

To identify homologs of the pig lincRNAs in humans and mice, we aligned the pig lincRNAs with human and mouse lincRNAs using BLASTn and identified 2,630 (40%) of the pig lincRNAs that had detectable homology with human lincRNAs, and 2,598 (39%) with mouse lincRNAs, of which 1,660 were shared between human and mouse. In comparison, 3,672 (31%) human lincRNAs had detectable homology with mouse lincRNAs when compared using the same approach. Among the pig lincRNA transcripts that align to the human lincRNAs, 187 have one-side or two-side synteny that extends to at least one neighboring protein-coding gene. Similarly, 244 of the pig lincRNAs have one-side or two-side synteny to a protein-coding neighbor in the mouse. These results imply that the pig may be an excellent model for research on lincRNA function.

Differential Expression of lincRNAs in Domesticated Pigs and Wild Boars

Domesticated pigs differ from wild boars in several behavioral traits, such as lower levels of aggressive behavior and reduced fear of humans (Price 1999). Therefore, we considered whether changes in the level of lincRNAs expression in the brain occurred during pig domestication. We analyzed the expression profile of lincRNAs in the published RNA-seq data set derived from the brains of five domesticated pigs and five wild boars (ERA209456) (Albert et al. 2012) and found 30 lincRNAs that show significant differential expression between pigs and wild boars (padj<0.1) (fig. 2*A*). Of these 30 differentially expressed lincRNA genes, 18 have higher expression in the domesticated pig, and 12 in the wild boar.

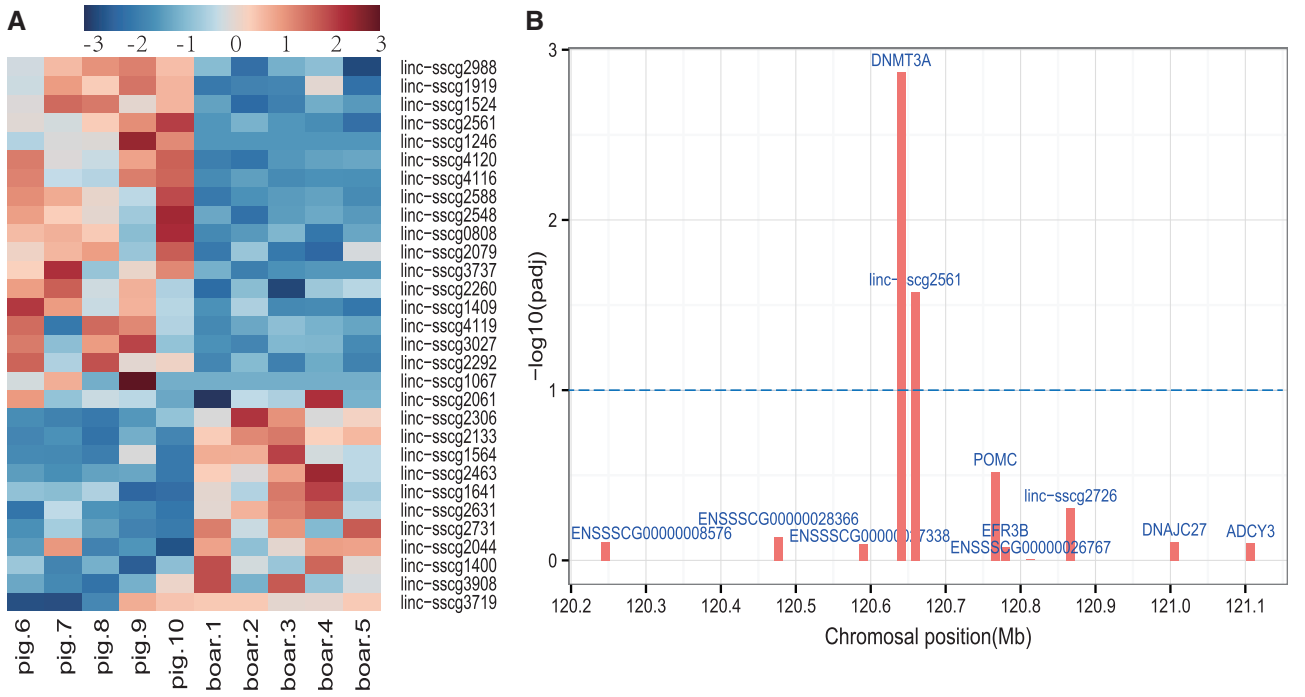

Fig. 2.—Expression differences between domesticated pigs and wild boars. (*A*) Heatmap showing expression abundance of lincRNA genes showing significant differences in expression. Expression levels (FPKM) were measured by RNA-seq. Genes were clustered by hierarchical clustering. (*B*) Expression differences of *linc-sscg2561* and genes in the surrounding 500 kb of genomic DNA. The *x* axis shows the genomic positions of these genes. A threshold of Padj = 0.1 is indicated by the dashed line.

Interestingly, two lincRNA genes (*linc-sscg1409* and *linc-sscg2561*) that have two-sided synteny extending to protein-coding neighbors in humans were found to show differential expression between the domesticated pigs and boars. *Linc-sscg1409* is a 1,200-nt transcript encoded by four exons and neighbors the *VWA2* and *FAM160B1* protein-coding genes. *Linc-sscg2561* is a 3,067-nt transcript encoded by two exons and shows tissue-specific expression in the pig brain (frontal cortex and occipital cortex). Our BLASTn search identified a conserved 367-nt match between *linc-sscg2561* and a human lincRNA (ENSG00000272048.1). The PhastCons plot from the UCSC 99 vertebrate whole-genome alignment to human showed a conserved region within the terminal exon, which includes the approximately 300-nt region that is conserved between pigs and humans (fig. 3). *Linc-sscg2561* displays 1.4-fold higher expression in domesticated pigs compared with boars. As lincRNAs are known to interact with chromatin proteins to positively and negatively regulate expression of neighboring genes (Wang et al. 2011), we conducted an analysis of the protein and lincRNA genes in the 500-kb window surrounding this lincRNA gene. *Dnmt3a* is the only gene adjacent to this lincRNA gene that displays differential expression, with 1.4-fold higher expression in the domesticated pig (fig. 2*B*). This observation implies that the *linc-sscg2561* may be a regulation element for *Dnmt3a* gene. *Dnmt3a* is an important protein with functions in DNA methylation, and a previous study had shown that *Dnmt3a* regulates behavioral plasticity to emotional stimuli (LaPlant et al. 2010), indicating that *linc-sscg2561* and *Dnmt3a* may influence the methylation of genes in the pig nervous system, and thus contribute to changes in emotional behavior during the domestication of the pig. In addition, experimental studies are needed to unravel the functions of lincRNA genes to understand the domestication of the pig.

Materials and Methods

We used two types of data sets from the pig for the identification of pig lincRNAs. The first data set included 50,136 UniGene transcripts, which was downloaded from the National Center for Biotechnology Information (NCBI) UniGene database build 42. Blat was used to align the UniGene transcripts against the Sus scrofa 10.2 genome sequence (Groenen et al. 2012) and pslCDnaFilter was used to filter the blat results. A total of 43,942 UniGene transcripts that had unique matches to the genome were retained. The second set of data included ten RNA-seq data sets was downloaded from the NCBI SRA database. RNA-seq reads were mapped to the Sus scrofa 10.2 genome using TopHat version 2.0.8 (Trapnell et al. 2009). Aligned reads for each sample were assembled using Cufflinks version 2.0.2. We then used Cuffcompare to generate intergenic transcripts for each

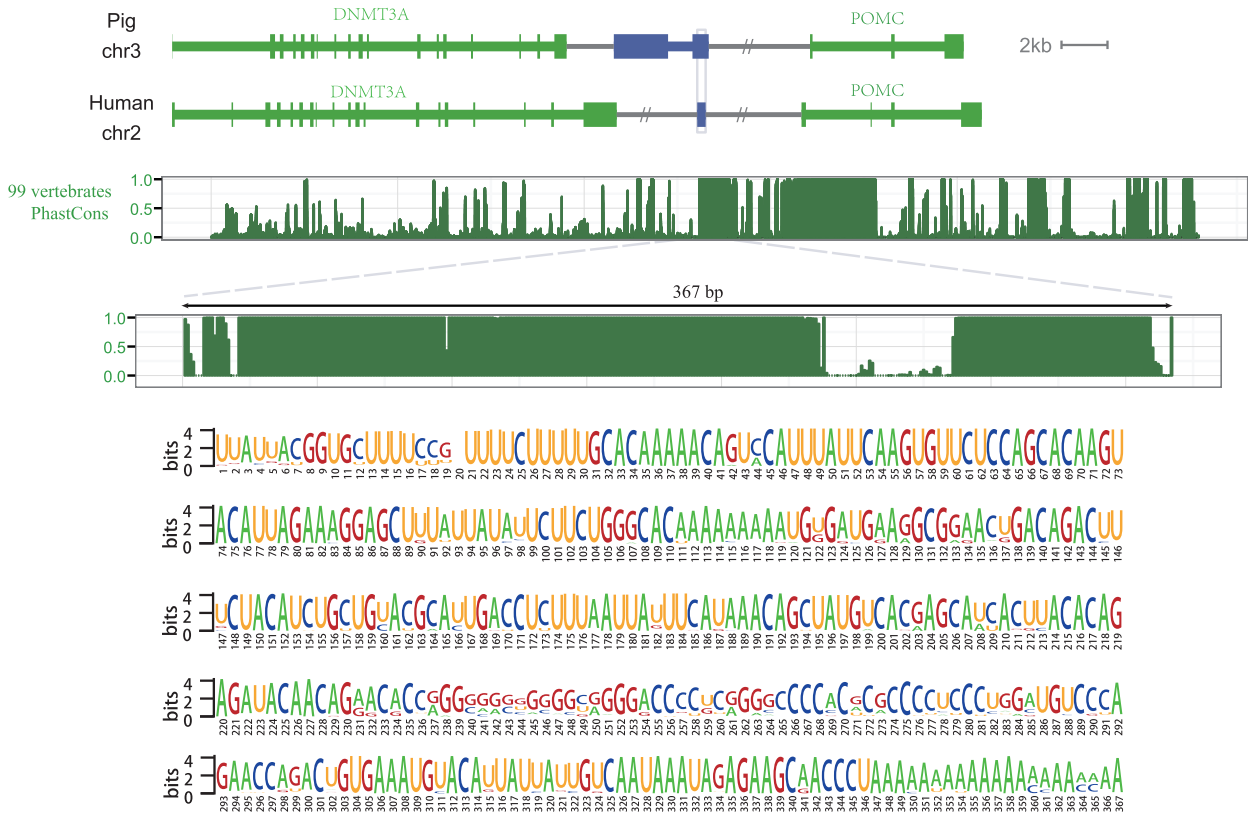

Fig. 3.—*Linc-sscg2561* shows synteny and sequence conservation. The gray box shows the region with sequence conservation. The PhastCon plot is relative to loci in the human genome and is derived from 99 vertebrate whole-genome alignments. The consensus logo highlights the 367-nt conserved sequence, which was identified from the 99 vertebrate genome alignments. A score of 4 bits indicates that these bases are perfectly conserved in the 99 vertebrate genomes.

sample assembly. To obtain high confidence transcripts, we used two criteria to filter the transcripts: RNA-seq reads must cover at least 80% of the predicted exon nucleotides for a transcript, and there must be at least three RNA-seq reads mapping to the predicted splice structure in at least one sample. We used strict criteria to identify lincRNAs as figure 1*A*. Tophat and Cufflinks were used to obtain FPKM (fragments per kilobase of exon per million fragments mapped) value. For each pairwise comparison of the samples, differentially expressed genes were identified based on the integer count data using Deseq2 version 1.2.8. (Anders and Huber 2010). SummarizeOverlaps was used to calculate counts of reads for each gene with the default mode of "Union." We downloaded human lincRNAs from the Gencode database (v19) (Harrow et al. 2012) and mouse lincRNAs from the NONCODE database (v4) (Xie et al. 2014). NCBI BLASTn was used to identify lincRNA sequence homology.

Supplementary Material

Supplementary tables S1 and S2 and figure S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank all of the contributors of the RNA-seq data sets. They also thank Jun Liu from Rockefeller University for his suggestions. This work was supported by grants from the National 863 Program of China (2011AA100304-5), the Ministry of Agriculture of China (2011ZX08009-003-006), the National 973 Program of China (2013CB835203), the National Natural Science Foundation of China (31061160189), and the Yunnan Provincial Science and Technology Department (2011AB008).

Literature Cited

Albert FW, et al. 2012. A comparison of brain gene expression levels in domesticated and wild animals. PLoS Genet. 8:e1002962.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biol. 11:R106.

Borsani G, et al. 1991. Characterization of a murine gene expressed from the inactive X chromosome. Nature 351:325–329.

Brockdorff N, et al. 1992. The product of the mouse *Xist* gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell 71:515–526.

Brown CJ, et al. 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, et al. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 25:1915–1927.
Derrien T, 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.
Fang M, Larson G, Ribeiro HS, Li N, Andersson L. 2009. Contrasting mode of evolution at a coat color locus in wild and domestic pigs. PLoS Genet. 5:e1000341.
Farajzadeh L, et al. 2013. Pairwise comparisons of ten porcine tissues identify differential transcriptional regulation at the gene, isoform, promoter and transcription start site level. Biochem Biophys Res Commun. 438:346–352.
Fujii J, et al. 1991. Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. Science 253:448–451.
Groenen MA, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491:393–398.
Guttman M, 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, et al. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477:295–300.
Harrow J, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22:1760–1774.
Huarte M, et al. 2010. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell 142: 409–419.
Hung T, et al. 2011. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nat Genet. 43:621–629.
Jia H, et al. 2010. Genome-wide computational identification and manual annotation of human long noncoding RNA genes. RNA 16: 1478–1487.
Khalil AM, et al. 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A. 106:11667–11672.
Kong L, et al. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 35:W345–W349.
LaPlant Q, et al. 2010. *Dnmt3a* regulates emotional behavior and spine plasticity in the nucleus accumbens. Nat Neurosci. 13:1137–1143.
Liu J, et al. 2012. Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis*. Plant Cell 24:4333–4345.
Martianov I, Ramadass A, Barros AS, Chow N, Akoulitchev A. 2007. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. Nature 445:666–670.
Mercer TR, Dinger ME, Mattick JS. 2009. Long non-coding RNAs: insights into functions. Nat Rev Genet. 10:155–159.
Mikawa S, et al. 2007. Fine mapping of a swine quantitative trait locus for number of vertebrae and analysis of an orphan nuclear receptor, germ cell nuclear factor (*NR6A1*). Genome Res. 17:586–593.
Mitchell Guttman IA, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458:223–227.
Orom UA, et al. 2010. Long noncoding RNAs with enhancer-like function in human cells. Cell 143:46–58.
Payer B, Lee JT. 2008. X chromosome dosage compensation: how mammals keep the balance. Annu Rev Genet. 42:733–772.
Price EO. 1999. Behavioral development in animals undergoing domestication. Appl Anim Behav Sci. 65:245–271.
Ravasi T, et al. 2006. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res. 16:11–19.
Rinn JL, et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129: 1311–1323.
Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111.
Trapnell C, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 28:511–515.
Trapnell C, et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 7: 562–578.
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.
Van Laere A-S, et al. 2003. A regulatory mutation in *IGF2* causes a major QTL effect on muscle growth in the pig. Nature 425:832–836.
Wang KC, et al. 2011. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472:120–124.
Xie C, et al. 2014. NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res. 42:D98–D103.

Associate editor: Takashi Gojobori