Research Article

For reprint orders, please contact: reprints@futuremedicine.com

Epigenomics

Profiling long noncoding RNA of multi-tissue transcriptome enhances porcine noncoding genome annotation

Pengju Zhao[‡,1], Xianrui Zheng[‡,1], Wen Feng[‡,1], Haifei Wang[1], Huimin Kang[1], Chao Ning[1], Heng Du[1], Ying Yu[1], Bugao Li[2], Yi Zhao[3] & Jian-Feng Liu[*,1]

[1]National Engineering Laboratory for Animal Breeding, Key Laboratory of Animal Genetics, Breeding & Reproduction, Ministry of Agriculture, College of Animal Science & Technology, China Agricultural University, Beijing 100193, China

[2]Department of Animal Sciences & Veterinary Medicine, Shanxi Agricultural University, Taigu 030801, China

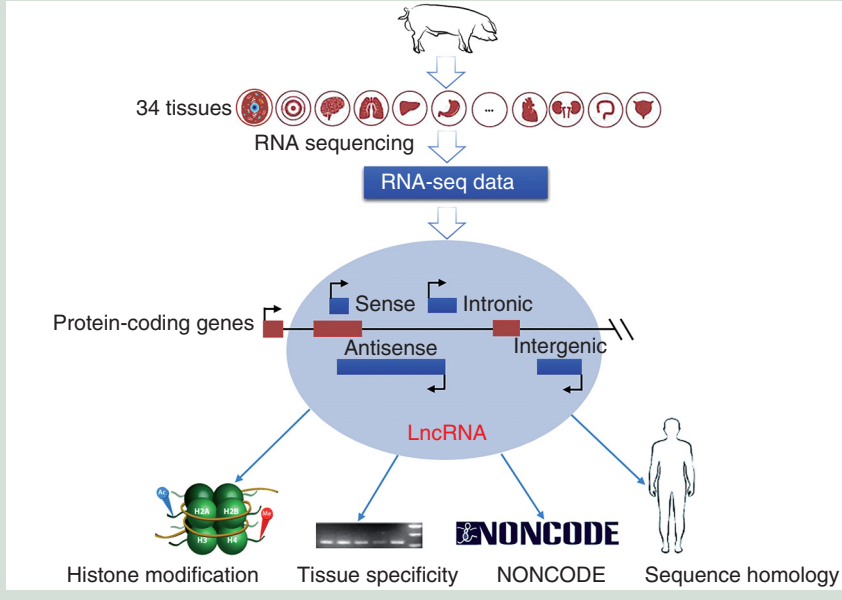
[3]School of Medicine, MOE Key Laboratory of Bioinformatics & Bioinformatics Division, Center for Synthetic & System Biology, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China

* Author for correspondence: Tel.: +86 10 62731921; liujf@cau.edu.cn

‡ Authors contributed equally

Aim: To construct a comprehensive pig noncoding transcriptome and further enhance porcine noncoding genome annotation. **Materials & methods:** We performed a tissue-based long noncoding RNA (lncRNA) profiling via exploiting 32,212 nonredundant lncRNA isoforms corresponding to 18,676 lncRNA loci across 34 normal pig tissues using high-throughput sequencing. Furthermore, the potential relationship between our identified lncRNAs and known protein-coding genes were globally assessed via a comprehensive computation-based strategy, developing a genome-wide lncRNA-targeted genome draft for further functional studies on noncoding genes. **Results & conclusion:** Among these lncRNAs, ubiquitously expressed lncRNA appeared at a higher level compared with tissue-specific one. Findings herein provide insight into comprehensive knowledge of porcine noncoding RNAs and further enhance pig noncoding annotation. For ease of accessing the information of the identified lncRNAs, we deposited those with high confidence in the publicly available NONCODE database, providing a valuable resource for facilitating pig noncoding genomic studies.

Graphical abstract:



Future Medicine

First draft submitted: 9 November 2017; Accepted for publication: 8 December 2017; Published online: 20 December 2017

Keywords: lncRNA • lncRNA function • NONCODE • pig

Long noncoding RNAs (lncRNAs) lack protein-coding potential and they nevertheless have important roles in a wide range of biological processes. For example, by acting as transcriptional co-regulators, lncRNAs can influence the process of transcription and regulate gene expression [1]. LncRNAs also participate in chromatin-modifying complexes, X-chromosome inactivation, epigenetic gene regulation and genomic imprinting [2–4]. Moreover, specific lncRNAs may underlie some human diseases such as cancer and Alzheimer's disease [5,6]. Some lncRNAs have been shown to have organ-specific sites [3]. To date, 141,353 lncRNA transcripts and 90,062 lncRNA genes have been identified in the human genome, while 117,405 lncRNA transcripts and 79,940 lncRNA genes have been found in the mouse genome. The identification of these transcripts and genes has advanced functional studies of lncRNAs in comparative genetic epidemiology and related fields.

The pig (*Sus scrofa*) is one of the most important domesticated animals and is also well-suited as a biomedical model for human disease because of similarities in anatomy and physiology between pigs and humans [7]. Previous studies have identified a number of roles for lncRNAs in the pig: a regulatory role in obesity development through co-expression with the gene *STEAP4*; embryo loss through differential expression of the *IL1R* gene in the MAPK pathway; development of pre-implantation embryos through cell cycle regulation; and regulation of transcription and of epigenetic marking of genes [8–10]. However, compared with comparable studies in humans, there is still a gap in our understanding of the roles of pig lncRNAs due to a lack of systematic profiling of transcriptome-wide lncRNAs in a wide range of tissues. To date, only 4515–4776 lncRNA loci in the pig genome have been identified [9,11]. The limited lncRNA resources in the pig genome weaken the use of these species as a model in functional comparative studies.

The present study was initiated to construct a global and comprehensive noncoding transcriptome in pigs and to further enhance porcine noncoding genome annotation. To achieve these aims, we performed genome-wide lncRNA identification in 34 normal tissues of the pig using over 340 Gb of sequence data generated from 116 RNA sequencing (RNA-seq) libraries. We identified 32,212 lncRNAs; of these, 25,733 were reported for the first time. The resulting large number of pig lncRNAs provides a resource to explore and characterize lncRNAs across the genome in pigs. Analysis of the lncRNAs revealed sequence orthologies and feature similarities between humans and pigs. Additionally, molecular combing and identification of recognizable characteristics of lncRNAs, including genomic location, expression pattern and sequence homology, enabled inferences to be drawn on potential target genes and lncRNA–RNA interaction patterns. The data reported here provide a valuable resource for enhancing the understanding and utilization of the pig noncoding transcriptome.

Materials & methods

Preparation of pig samples

For purpose of generating profiles of transcriptome of all major organs and tissues in the pig, we totally selected 33 different pool tissues and peripheral blood mononuclear cells (PBMC) from the nine unrelated Duroc pigs. An overview of all involved tissue and cell samples is provided in Supplementary Table 1. The whole sample collection and treatment were conducted in strict accordance with the protocol approved by the Institutional Animal Care and Use Committee of China Agricultural University.

All pig tissues were histologically confirmed to be carefully removed from the healthy pig, which was kept and fed according to the institutional guidelines and free from all major pig diseases (porcine reproductive and respiratory syndrome virus [PRRSV], porcine circovirus type 2, porcine parvovirus). Then, all these samples were snap frozen within the first 20 min after slaughter and stored in liquid nitrogen (-196°C) until usage. PBMCs were isolated using Ficoll–Hypaque PLUS (GE Healthcare, Beijing, China), following the manufacturer's instructions. In brief, the whole blood was first diluted by an equal volume of phosphate-buffered saline. Then, 20 ml of diluted blood was carefully added on top of 10 ml of Ficoll-Hypaque solution in a 50 ml conical tube and centrifuged at 460 × *g* for 20 min at room temperature. After centrifugation, the middle whitish interface containing mononuclear cells was transferred to a new tube, and washed by phosphate-buffered saline followed by centrifugation at 500 × *g* for 10 min twice.

Separation of RNA from tissues

Purification of total RNA from mixture of equally unrelated pig pool tissues via the Trizol method (Invitrogen, CA, USA) according to standard protocols was performed. RNA degradation and contamination was monitored on 1% agarose gels. The purity and contamination of total RNA was checked using NanoPhotometer® trophotometer (IMPLEN, CA, USA) and Qubit® RNA Assay Kit in Qubit® 2.0 Flurometer (Life Technologies, CA, USA). RNA integrity was measured using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA). All pig samples that met the criteria of having an RNA Integrity Number value of 7.0 or higher, optical density (OD) 260/280 value of 1.8–2.2 and at least 5 μg of total RNA, were included and batched for RNA-seq.

Library construction & RNA-seq

Total RNA of samples meeting quality control (QC) criteria were rRNA depleted and depleted QC using the RiboMinus™ Eukaryote System v2 (Thermo Fisher Scientific, MA, USA) and RNA 6000 Pico chip (Agilent Technologies) according to the manufacturer's protocol. RNA-seq libraries were constructed using the NEBNext® Ultra™ RNA Library Prep Kit (Illumina, CA, USA) with 3 μg rRNA depleted RNA according to the manufacturer's recommendation. RNA-seq library preparations were clustered on a cBot Cluster Generation System using HiSeq PE Cluster Kit v4 cBot (Illumina) and sequenced using the Illumina Hiseq 2500 platform according to the manufacturer's instructions, with data size of per sample to a minimum of 10G clean reads (corresponding to 126 bp paired-end reads). The sequenced RNA-seq raw data for 34 pig tissues is available from NCBI Sequence Read Archive with the BioProject number: PRJNA392949.

QC of RNA-seq

We conducted a QC step to raw data (raw reads) of fastq format for efficient and accurate RNA-seq alignment and analysis. In this step, raw reads were modified to clean reads for downstream analyses by following steps: BBmap automatically detect the adapter sequence of reads and remove reads containing illumina adapters; FASTQC calculated the Q20, Q30 and GC content of the clean data for QC and filtering; Fastx toolkit (version 0.0.14) carried out homopolymer trimming to 3′ end of fragments and removed the N bases of 3′ end.

Reads mapping & assembly

RNA-seq data were mapped and genome indexed with Hisat 0.1.6-beta 64-bit to the pig genome release version *Sus scrofa*10.2 [12]. Ensemble *Sus scrofa*10.2 version 82 [13] annotation was used as the transcript model reference for the alignment, splice junctions finding as well as for all protein-coding gene (PCG) and isoform expression-level quantifications. Besides, lncRNAs were classified to another group described in later. To obtain expression levels for all pig genes and transcripts across all 34 samples, fragments per kilobase of exon model per million mapped reads (FPKM [controlling for fragment length and sequencing depth]) values were calculated using Stringtie 1.0.4 (Linux_x86_64). A gene or transcript was defined as expressed if it was measured above 0.1 FPKM across all tissues and all final gene level models were produced in GTF format.

LncRNA annotation

Taking advantage of final gene level models from Stringtie, we filtered those transcripts with substantial expression levels and no overlap with PCGs with transcripts shorter than 200 nt and single exon. Then filtered transcripts were searched against Uniport pig protein database (version 20150717) and novel protein (confirmed by Mass spectrum in our research), any transcripts which showed sequence similarity with any of these proteins with cut-off E value of 10^{-5} were removed. After this filtration step, coding-non-coding index (CNCI; version 2) was used to predict coding potential of transcripts using frequency of adjoining nucleotide triplets (ANT) signature to classify protein coding and noncoding sequences. At last, candidate lncRNAs with no putative coding potential (CNCI score <0) were classified as final lncRNAs for each samples. The resulting lncRNAs from 34 tissues were merged to the final meta-lncRNAs (transcriptome GTF file and Fasta file) for pig using Cuffmerge tool.

Classification of lncRNA

The identified lncRNAs were further classified as sense, intergenic, intronic and antisense lncRNAs that are four basic categories based on spatial relationships of their gene loci with PCGs using Cuffcompare tool. Sense lncRNAs overlap with one or more exons of a transcript on the same strand, intergenic lncRNAs occur in the interval between two PCGs on the same strand, intronic lncRNAs derive from an intron within another protein-coding transcript

and antisense lncRNAs overlap with one or more exons of a transcript on the opposite strand. Therefore, the multiple distribution characteristics of lncRNAs indicate that lncRNAs have diverse roles in functions of lncRNAs in organism.

Comparisons of previously published pig lncRNA database

To construct a newly identified lncRNA database of pigs in current study, we integrated our final lncRNA results and compared it with the existing published repositories, including pig long intergenic noncoding RNAs (lincRNAs) datasets identified by Zhou *et al.* [11] and animal long noncoding database (ALDB) [14]. Comparison with genomic localization between different lncRNA datasets were proceeded using in-house perl scripts and Cuffcompare tool.

Permuting feature enrichment test

The permuting feature enrichment test is mainly used to explore the relationship between transcript sets (lncRNAs and PCGs) and feature regions (transposable elements and chromatin regions) in pig genomes. In the permutation testing, we randomly changed the location of regions of a specific set in the genome with 1000-times. And then we counted the number of overlaps with feature regions, and visualized their distribution with overlap counts between the feature regions and the specific sets. Finally, the student's *t*-test was used to check whether two sets of data differ significantly.

ChIP-seq data analysis

In this study, eight publicly available ChIP-Seq data were downloaded from GeneProf database with the accession number from SRX122653 to SRX122660, which were generated by the chromatin immunoprecipitation and followed by short sequencing in pig induced pluripotent stem cells, including H3K4me1/2/3, H3K9me3, H3K27ac, H3K27me3, H3K36me3 and H2AZ histone modifications. At first, the sequences with 75 bp were filtered by Fastx toolkit (version 0.0.14): the reads with the quality score of 70% bases lower than 20 were discarded and 3′ end of the reads with the quality score lower than 20 were trimmed. Then all the filtered sequences were aligned to the pig genome (*Sus scrofa*10.2) using Hisat (0.1.6-beta 64-bit) and the alignment files were transferred and sorted using samtools. Next, for each histone modifications, we used the Model-based Analysis of ChIP-seq (MACS) algorithm with the MACS (version 2) to capture the influence of genome complexity and to evaluate the significance of enriched ChIP regions. In this study, the parameters for MACS were: without control sample, effective genome size = 2.7×109 bps and cut-off of p-value = 0.01.

Enrichment analysis of transcription start site

The GenomicRanges, rtracklayer and IRanges packages from the Bioconductor were used to test the enrichment between peaks of ChIP-seq signal and the transcription start site (TSS) region of transcripts. We compared promoters of TSS ($\pm$200 bps around the TSS) with all chromatin signal regions based on enrichment test. Besides, as to each histone modifications, the heatmap and the plots of average values were used to exploit the distribution of chromatin around the TSS with $\pm$1000 bps.

Mapping significant SNPs of GWAS to lncRNAs

A total of 5,176 SNPs were obtained from National Animal Genome Research Program's (NAGRP) pig genome-wide association study (GWAS) catalog. According to the NAGRP's Pig Genome Coordination Program reports, we considered total five category traits associated with these significant SNPs in pig GWAS, including traits of reproduction, production, meat and carcass quality, health and exterior. The overlapping loci between GWAS SNPs and lncRNA datasets were determined based on the comparison of their chromosomal locations using in-house perl scripts.

Identification of lncRNAs & pre-mRNAs interactions

In order to identify the splicing regulation between lncRNAs and mRNAs within pig, we combined all detected lncRNAs in our study and pig protein-coding ensemble annotation (*Sus scrofa* 10.2 version 82) to find the potential lncRNA–RNA interactions based on similarity-search method. First, the lastal from LAST package was run to detect lncRNA–RNA base pairings. The detailed parameters are as follows: as to substitution matrix, G:C, A:T and G:T matches were scored as 4, 2 and 1, respectively. This was generally used in the field of RNA–RNA interactions (e.g., lncRNA–RNA interactions in human). Besides, the mismatch, gap opening and gap extension were scored as

-6, -20 and -8, respectively. We also set the threshold value to 108 for alignment scores to filter the final base-pairing regions. Finally, we applied in-house perl scripts to filter those regions that either overlapped base-pairing regions or spanned exon-intron borders.

Alternative splicing events in pig transcriptome

The alternative splicing events (ASE) in pig transcriptome were detected by jekroll-splicing express software [15], including exon skipping, alternative 3′ splice site, alternative 5′ splice site and intron retention.

Synteny analysis between human & pig

In order to explore putative homologous chromosomal regions between human and pig, we selected the protein sequences and lncRNA sequences as anchors to align these regions. In our research, MCScanX [16] software was used to effectively detect and visualize the genome synteny and collinearity with the default parameters.

Validation of lncRNA transcripts by quantitative real time-PCR

RNAs from 11 pig tissues including testis, uterus, stomach, kidney, liver, brain, ovary, breast, longissimus dorsi, heart and lung were transcribed into cDNA using PrimeScript™ RT reagent kit with gDNA Eraser (TaKaRa Bio, Beijing, China) for PCR reaction. Two reverse transcribed reaction systems were conducted. DNA-free contained 1.0 μg RNA, 2.0 μl 5× gDNAEraser Buffer, 1.0 μl gDNA Eraser, 7 μl RNase-free dH2O. After standing for 5 min, Reverse Transcribed Reaction System II contained 4.0 μl 5× PrimeScript® Buffer 2, 1.0 μl PrimeScript® RT Enzyme Mix I, 1.0 μl RT Primer Mix, 4.0 μl RNase-free dH2O with 10.0 μl of the liquid of Reverse transcribed reaction system I, whose reacting temperature is 37°C for 15 min, followed by 85°C for 5 s. The primers for PCR amplification were designed by Primer-blast and confirmed by Oligo 7.0. The details about the primers are listed in Supplementary Table 2. PCR was performed on 25 μl volumes containing 22 ml premix (Golden Star T6 Super PCR Mix, Beijing TsingKe Biotech Co., Ltd, Beijjing, China), 1 μl forward and 1 μl reverse primers (10 μmol/l) 0.5 μl and 1 μl cDNA. PCR conditions were 95°C for 8 min, followed by 35 cycles of 95°C for 30 s, 55–60°C for 30 s, 72°C for 30 s and 8 min at 72°C.

Heatmap drawing

Heat maps were produced using the heatmap package from the Bioconductor version 3.0 in R. RColorBrewer package provided in the palettes for drawing heatmap.

Results

Global identification of pig lncRNAs from 34 tissues

In order to obtain a comprehensive profile of lncRNAs, we developed and applied a custom-built pipeline (Figure 1A) containing four processing steps to detect lncRNAs in 34 pig tissues. An average of 48.48 million reads per tissue were sequenced from strand-specific and paired-end 126 bp cDNA libraries; from these sequences, an average of 43.97 million reads (90.7%) per sample passed the strict QC (Supplementary Table 3). The 1495 million high-quality reads (376.7 Gb, 135-fold genome coverage) were aligned to the pig genome (*Sus scrofa* 10.2) and 1230 million mapped fragments (310.1 Gb, 110.8-fold genome coverage) with an average alignment rate of 82.32% were recovered (Supplementary Table 1). The mapped reads were assembled and quantified as candidate transcripts using Stringtie software (Supplementary Table 4) [17]. This processing step produced 146,354 nonredundant transcript isoforms from 130,471 loci on average; of these, 10,981 loci (23,770 transcripts) were perfectly annotated in the pig Ensembl database. We determined known PCGs with FPKM <0.1 among these nonredundant transcript isoforms. The remaining 122,584 transcripts (119,490 loci), without overlaps with PCGs, were further filtered as candidate lncRNAs on two criteria: first, CNCI [18] was used to predict the coding potential of 5824 candidate lncRNA loci (7335 transcripts) on average per sample (CNCI score <0); second, 4136 lncRNA loci (5286 transcripts) per sample with a single exon and a length <200 nt were removed. As a result, 1688 lncRNA loci (2049 transcripts) on average per sample were filtered by searching against protein sequence data (32,944 known proteins in UniportDB and 7693 novel proteins, unpublished results).

Through the use of our global filtration pipeline and merging of multi-assemblies from the 34 tissue samples, we identified a final set of 32,212 nonredundant lncRNA isoforms from 18,676 lncRNA loci with an average of three exons and the average length of 1089 bp. The number of detected lncRNAs was approximately 1.5-fold greater than the 21,607 annotated PCGs and 805-fold greater than the 40 noncoding RNAs in Ensembl. When

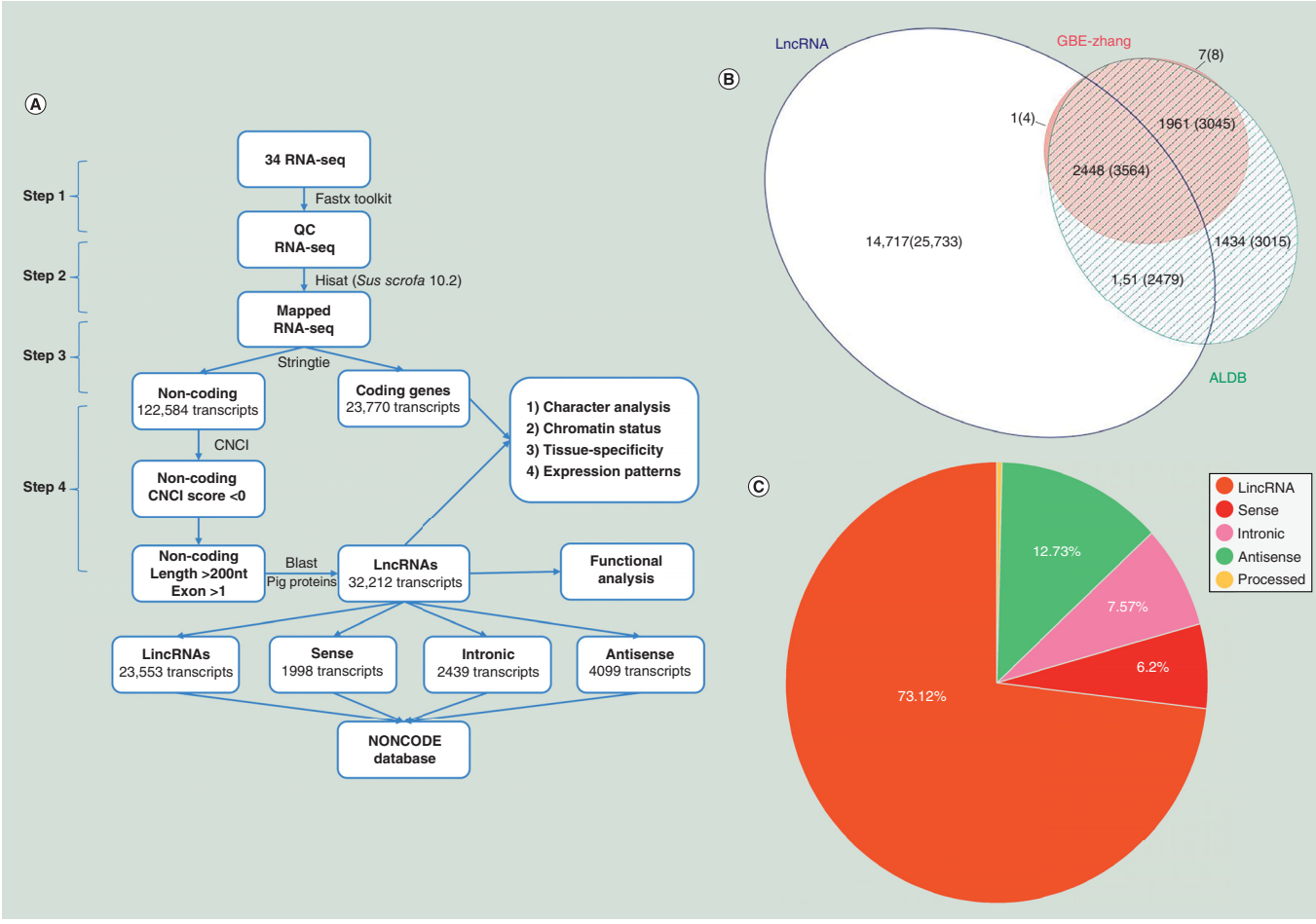


Figure 1. Global identification of pig long noncoding RNA. (A) A custom-built pipeline was used to identify transcripts from 34 pig RNA samples. This pipeline includes five steps: quality control of RNA sequencing; alignment of sequence reads; assembly and classification of new transcripts; identification and filtration of long noncoding RNAs (lncRNAs); and classification of lncRNAs. **(B)** Venn diagram for three different lncRNA databases. LncRNA: lncRNAs detected in this study; GBE-Zhang; ALDB [14]. **(C)** Pie charts for five categories of pig lncRNAs.
LncRNA: Long noncoding RNA.

compared with pig lncRNAs identified in previous studies [11,19], our findings overlapped with 53.89 and 49.93% of GBE-Zhang *et al.* [11] and the domestic-animal long noncoding database, respectively, at the transcript level (Figure 1B). Moreover, 3564 high-confidence lncRNAs (11.21% of our lncRNA set) were present in all three lncRNA sets while 25,733 novel lncRNAs were only detected in our study. We finally merged the 21,919 lncRNA loci from the three lncRNA sets using cuffcompare scripts; 14,717 (67.14%) of these lncRNA loci were detected here and 2448 (11.17%) lncRNA loci were present in all three lncRNA sets (Figure 1B).

We compared the relative genomic locations of the identified lncRNAs and known PCGs, and further classified putative lncRNAs into five categories, including sense lncRNAs, intronic lncRNAs, antisense lncRNAs, lincRNAs and processed transcripts (Materials & methods, Supplementary Figure 1). As shown in Figure 1C, almost 73.12% of nonredundant lncRNA isoforms were categorized as lincRNAs (n = 23,553), 12.73% as antisense lncRNAs (n = 4099), 7.57% as intronic lncRNAs (n = 2439), 6.2% as sense lncRNAs (n = 1998) and 0.38% as processed lncRNAs (n = 123). In general, these different types of pig lncRNA most likely interact differently with their target genes, as sense lncRNAs and antisense lncRNAs are correlated differently with coding genes [20].

Characteristics of lncRNAs & PCGs in the pig transcriptome

A comparison of the lengths of lncRNAs and PCGs in the pig transcriptome indicated that the former spanned 1078 nt on average, which was far less than the average length of 4198 nt for PCGs; we also found that different

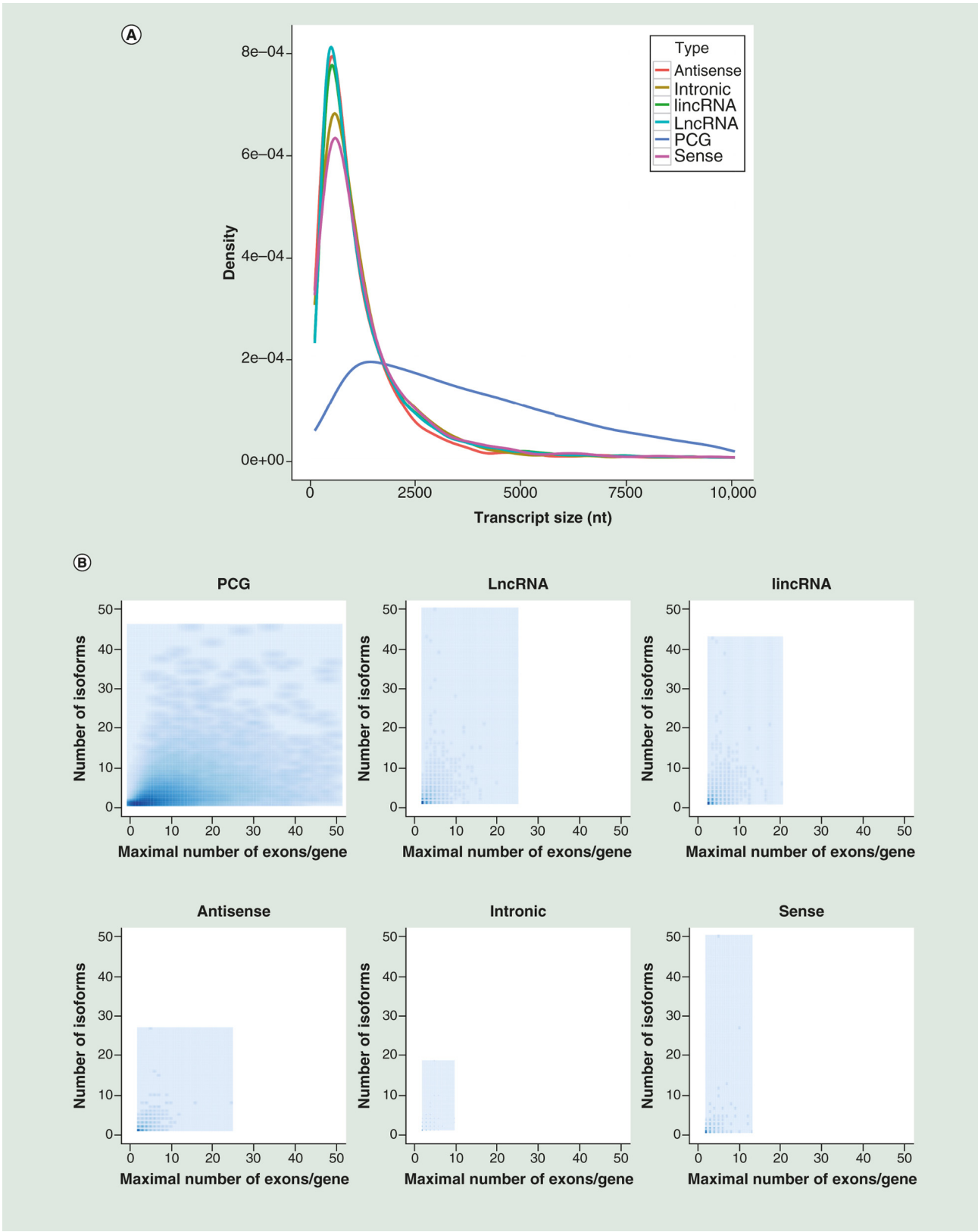


Figure 2. Structural characteristics of long noncoding RNAs. (A) Density distribution of transcript sizes among different types of lncRNAs and PCGs. **(B)** Scatterplot smoothing shows the relationship between the number of isoforms and the maximal number of exons among different transcripts. **(C)** Percentage of different expression levels for different types of lncRNAs and PCGs. **(D)** Comparison of a number of overlaps to repeat elements of the genome between permutated sets and the real lncRNA set.
LncRNA: Long noncoding RNA; PCG: Protein-coding gene.

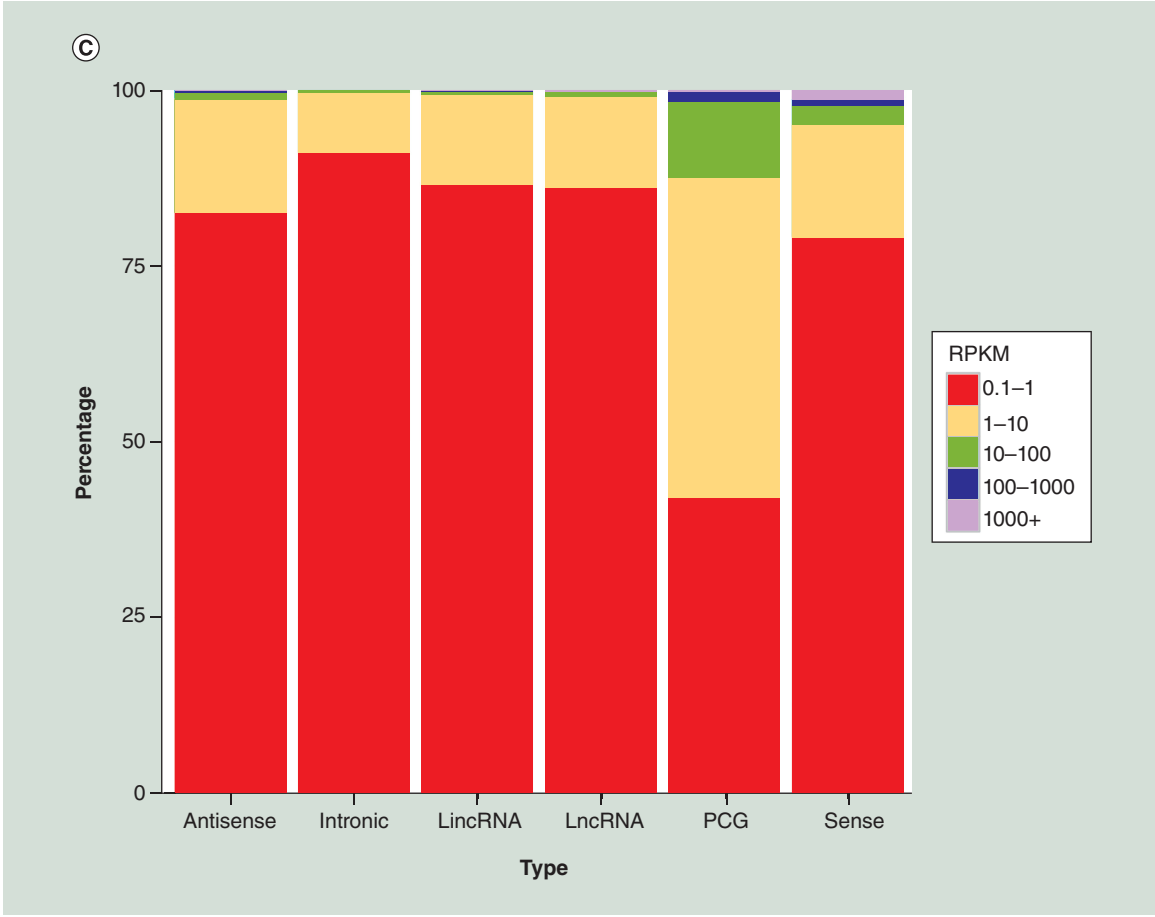


Figure 2. Structural characteristics of long noncoding RNAs (cont.). (A) Density distribution of transcript sizes among different types of lncRNAs and PCGs. **(B)** Scatterplot smoothing shows the relationship between the number of isoforms and the maximal number of exons among different transcripts. **(C)** Percentage of different expression levels for different types of lncRNAs and PCGs. **(D)** Comparison of a number of overlaps to repeat elements of the genome between permutated sets and the real lncRNA set.
LncRNA: Long noncoding RNA; PCG: Protein-coding gene.

types of lncRNA had similar distributions with regard to transcript size (325?Figure 2A). PCGs had 11 exons on average and 70.7% of PCGs had more than five exons; by contrast, lncRNAs had three exons on average and approximately 81.8% of lncRNAs had only two or three exons (Supplementary Table 5). Scatterplot smoothing further highlighted the differences between PCGs and different types of lncRNAs (Figure 2B). The lncRNAs and PCGs had similar spans within the number of isoforms but had great differences in the maximum number of exons ($n = 25$ in lncRNAs vs $n = 119$ in PCGs). This pattern of fewer exons and shorter transcripts in lncRNAs than PCGs has also been reported for the human transcriptome [21].

A screen of lncRNA expression patterns across the 34 tissues showed that 22.27% had tissue-specific expression and only 21 lncRNA loci were expressed in all tissues. By comparison, the number of ubiquitously expressed PCGs was 2472. Thus, the analysis demonstrated that lncRNAs showed more specific patterns of expression in different tissues than the PCGs. In addition, the expression of lncRNAs (16.14 FPKM on average) was about 1.5-fold lower than PCGs (26.78 FPKM per transcript on average). As shown in Figure 2C, the number of transcripts with a high expression level (above 10 FPKM) were 7721 and 17,244 for lncRNAs and PCGs, respectively. Approximately 86% of the lncRNAs had a low expression level (<1 FPKM) across all tissues. This differentiation between lncRNAs and PCGs in multiple pig tissues was also seen for human lncRNAs and PCGs [22].

It has been shown that evolutionary pressures on repeat regions do not have as great an effect on functional regions of the genome [23]. To explore the relationship between lncRNAs and the repeat regions of pig genome, we compared the transcripts of all PCGs and lncRNAs with three types of transposable elements in the pig genome,

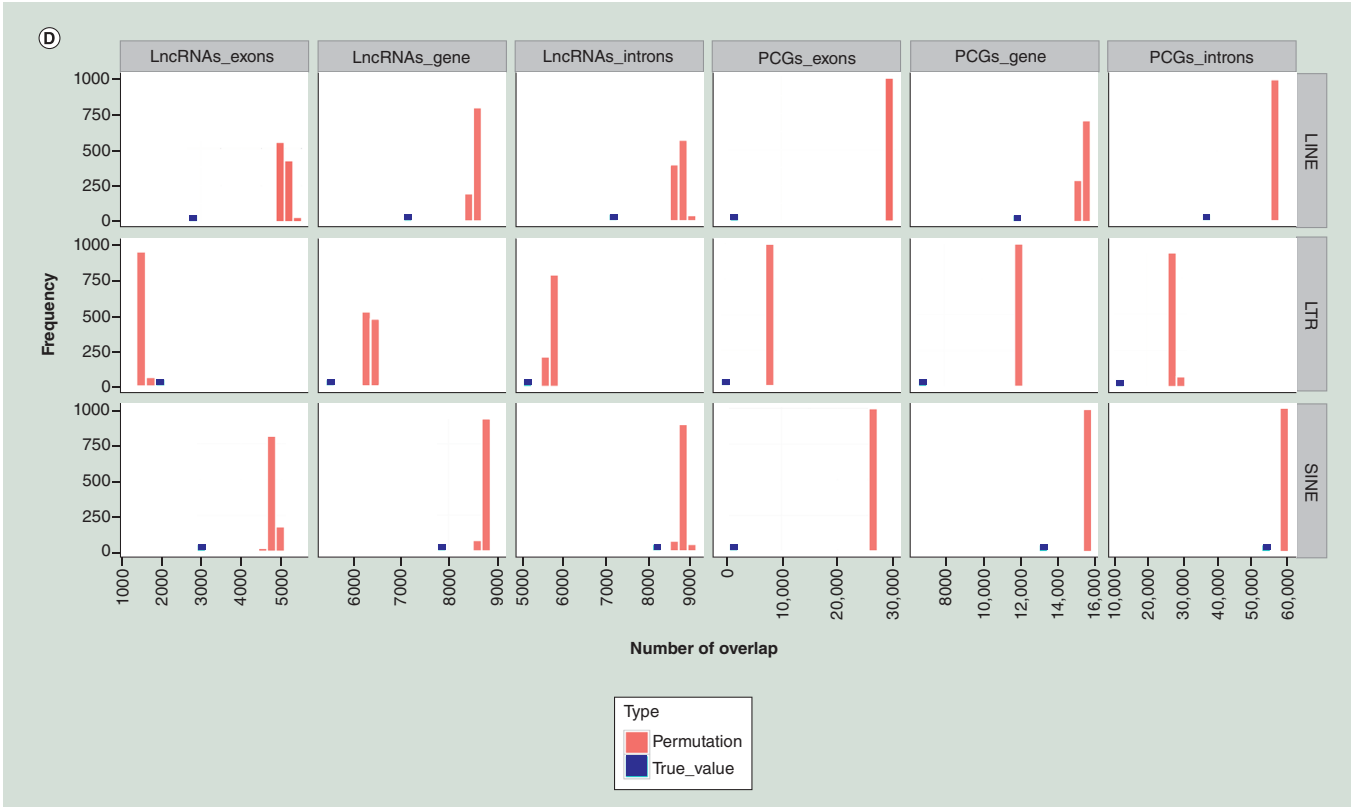


Figure 2. Structural characteristics of long noncoding RNAs (cont.). (A) Density distribution of transcript sizes among different types of lncRNAs and PCGs. **(B)** Scatterplot smoothing shows the relationship between the number of isoforms and the maximal number of exons among different transcripts. **(C)** Percentage of different expression levels for different types of lncRNAs and PCGs. **(D)** Comparison of a number of overlaps to repeat elements of the genome between permutated sets and the real lncRNA set.
LncRNA: Long noncoding RNA; PCG: Protein-coding gene.

namely short interspersed elements, simple repeats, long terminal repeats and long interspersed nuclear elements. We separately permuted lncRNA and mRNA datasets (repeats of permutation = 1000) as a control for testing potential feature enrichment. As shown in Figure 2D, almost all transcript regions had a negative enrichment (p-value < 2.2×10^{-16}) with transposable elements; this effect was consistent in both PCG and lncRNA loci (Figure 2D), in other words, a high proportion of repetitive elements occurred in introns while they were infrequent in exons (9% exons vs 38% introns with repeats in lncRNA loci; 2% exons vs 54% introns with repeats in PCG regions). This might be explained by the occurrence of relatively higher conserved features and functional importance of exons. Additionally, more functional regions overlapped with short interspersed elements (n = 88,377), long interspersed nuclear elements (n = 66,600) and long terminal repeats (n = 32,044), which might be due to the larger numbers of the smaller transposable elements in the pig genome (Figure 2D).

Chromatin status of lncRNAs

It is known that lncRNAs usually interact with chromatin signature and play an essential role in chromatin modification. To date, eight publicly available epigenetic modification databases [24] have been generated using Chromatin Immunoprecipitation followed by short sequencing (ChIP-Seq) technology in pig induced pluripotent stem cells. Here, these databases were used to investigate the relationship between the chromatin status and the pig lncRNAs identified in our study.

First, we sought to explore the association of diverse chromatin signals at TSS loci with both PCGs and lncRNAs. The chromatin signatures for various histone modifications (H3K4me1/2/3, H3K9me3, H3K27me3, H3K27ac, H3K36me3 and H2AZ) were identified and used in a permutation model to test the enrichment of PCG and lncRNA sets (Supplementary Tables 6–13). For each permutation test, we generated 1000 permuted transcript sets. Additionally, for each permutation, equal numbers and lengths of transcripts of the PCG and

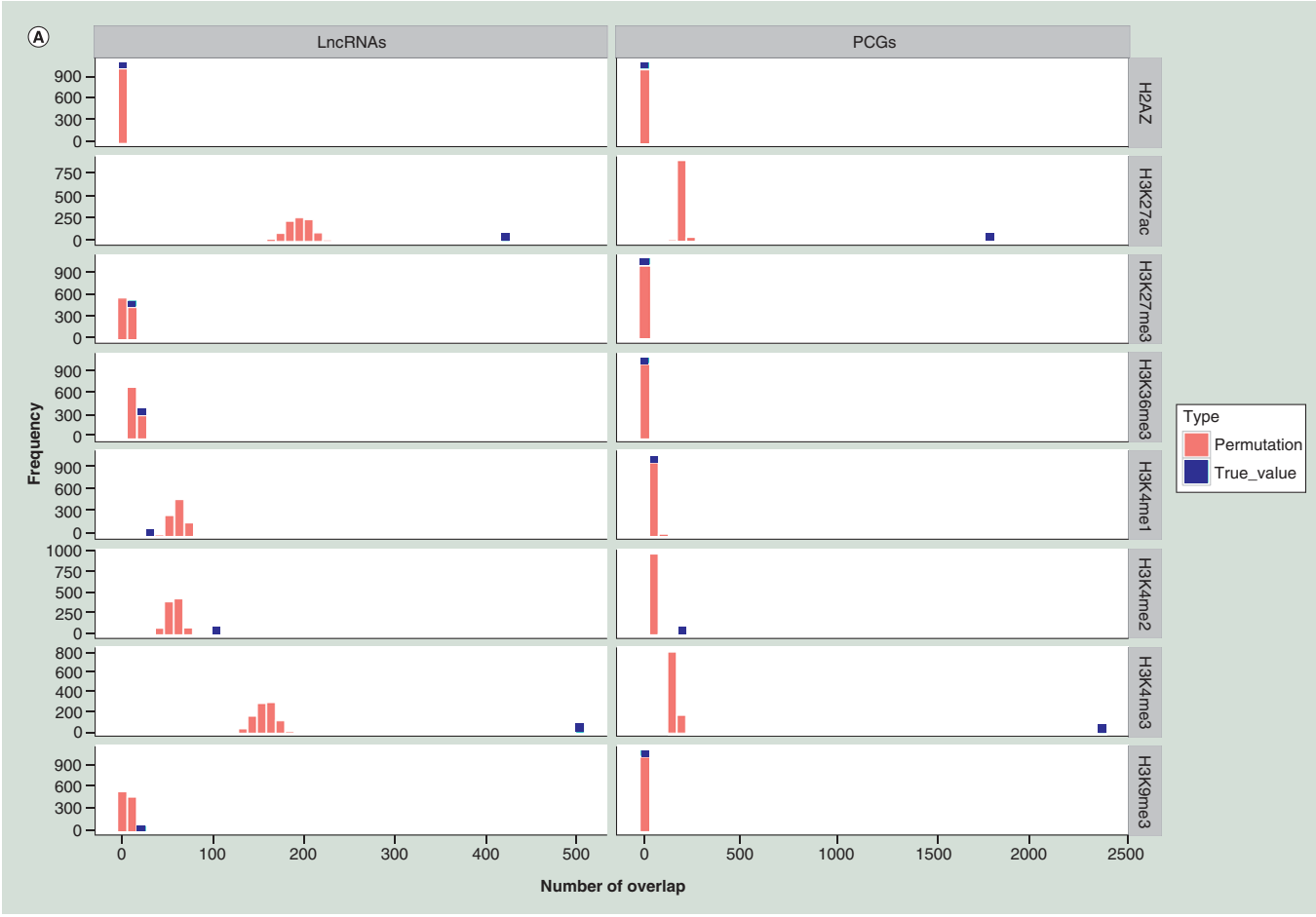


Figure 3. Chromatin status of long noncoding RNAs. (A) Comparison of enriched numbers to chromatin signals of the genome between permutated sets and real lncRNA set. **(B)** Distribution of chromatin around the corresponding TSS. The adjoining region is set to ±1000 bps.
LncRNA: Long noncoding RNA; TSS: Transcription start site.

lncRNA sets were randomly distributed in the pig genome. The simulation tests showed strong statistical evidence (p-value $< 2.2 \times 10^{-16}$) according to the empirical distribution that both PCGs and lncRNAs were significantly associated with H3K4me2/3 and H3K27ac regions of the pig genome (Figure 3A).

Next, we further swept all chromatin signatures with the promoter sequences (the regions of ±200 bp around the TSS) of PCGs and lncRNAs. As shown in Table 1, almost all chromatin-enriched regions had a significant overlap with promoter regions; most of these were confirmed to be strongly enriched by binomial tests, with the exception of H3K9me3 in PCGs and H2AZ in lncRNAs.

Finally, to investigate the distribution of chromatin around the corresponding TSS, we extend the adjoining region to ±1000 bps around the TSS and compared the fluctuation of surrounding signal peaks for each chromatin signature (Figure 3B). As expected, both PCGs and lncRNAs were associated with a series of chromatin signals, including H3K4me1/3, H3K9me3, H3K27me3, H3K27ac and H2AZ. However, the H3K36me3 and H3K4me2 signals were stronger across intervals for lncRNAs around TSSs but weaker for PCGs. Interestingly, it was also obvious that PCGs displayed a negative enrichment of H3K4me2 around TSSs.

Our results here suggested that most histone modifications had significant enrichment of PCGs and lncRNAs; in addition, lncRNAs appeared to show a stronger relevance than PCGs, which was consistent with the biological function of lncRNAs. The findings here are consistent with those in a previous study [25].

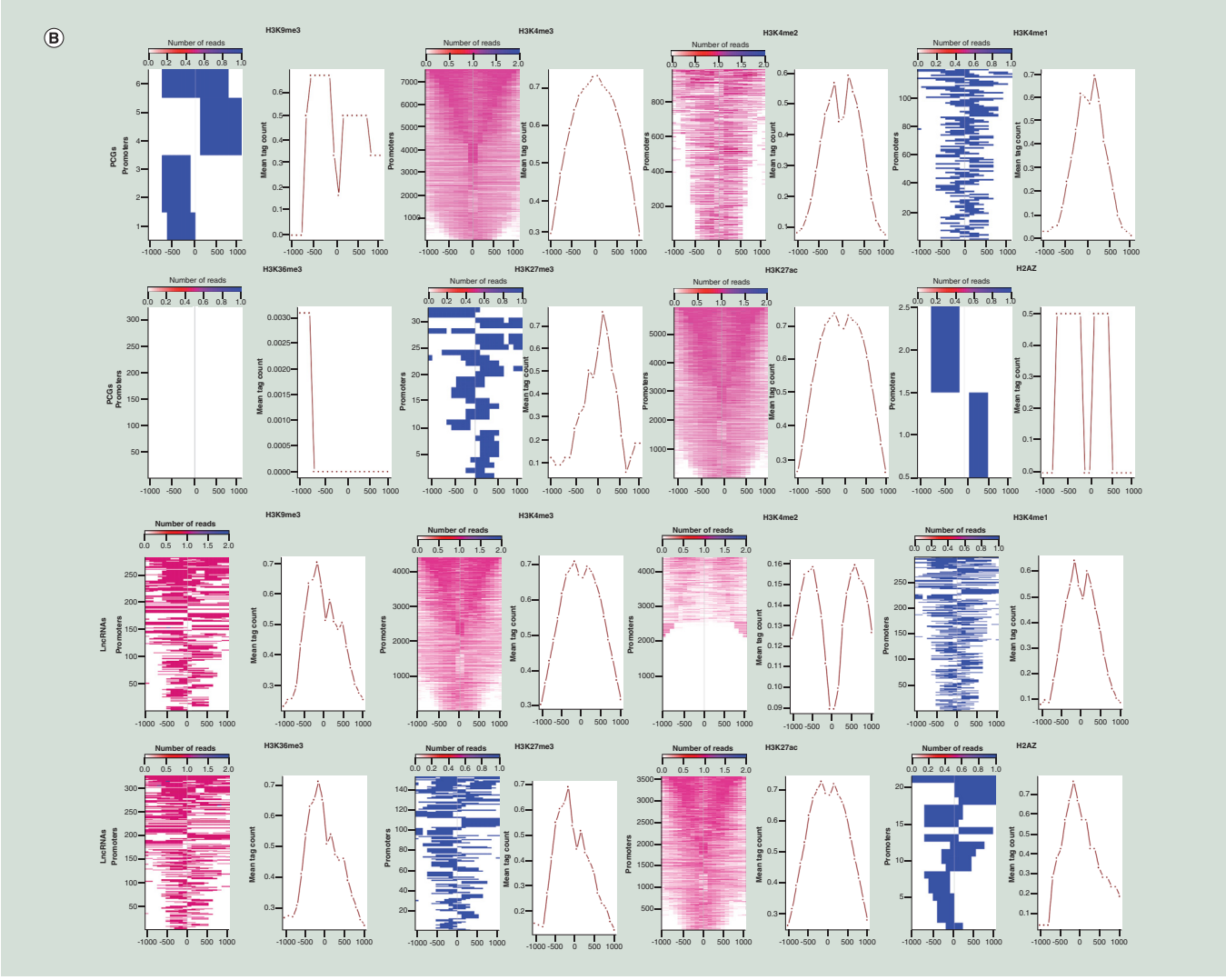


Figure 3. Chromatin status of long noncoding RNAs (cont.). (A) Comparison of enriched numbers to chromatin signals of the genome between permutated sets and real lncRNA set. **(B)** Distribution of chromatin around the corresponding TSS. The adjoining region is set to ±1000 bps.
LncRNA: Long noncoding RNA; TSS: Transcription start site.

Table 1. The enrichment analysis of chromatin signatures with the promoters.

Histone modification	PCGs			LncRNAs		
	Promoters overlapped by an enriched region	Enriched regions overlap a promoter	p-value	Promoters overlapped by an enriched region	Enriched regions overlap a promoter	p-value
H3K4me1	119/26598	119/13,625	$<2.2 \times 10^{-16}$	298/26,778	213/13,625	$<2.2 \times 10^{-16}$
H3K4me2	980/26598	907/11,528	$<2.2 \times 10^{-16}$	976/26,778	673/11,528	$<2.2 \times 10^{-16}$
H3K4me3	7554/26598	6864/19,339	$<2.2 \times 10^{-16}$	4415/26,778	2753/19,339	$<2.2 \times 10^{-16}$
H3K9me3	6/26598	6/914	0.1489	281/26,778	281/914	$<2.2 \times 10^{-16}$
H3K27me3	32/26598	32/753	$<2.2 \times 10^{-16}$	152/26,778	81/753	$<2.2 \times 10^{-16}$
H3K27ac	5940/26598	5385/27,928	$<2.2 \times 10^{-16}$	3555/26,778	2144/27,928	$<2.2 \times 10^{-16}$
H3K36me3	96/26598	96/2546	$<2.2 \times 10^{-16}$	324/26,778	161/2546	$<2.2 \times 10^{-16}$
H2AZ	2/26598	2/51	0.01432	21/26,778	9/51	1.06×10^{-13}

LncRNA: Long noncoding RNA; PCG: Protein-coding gene.

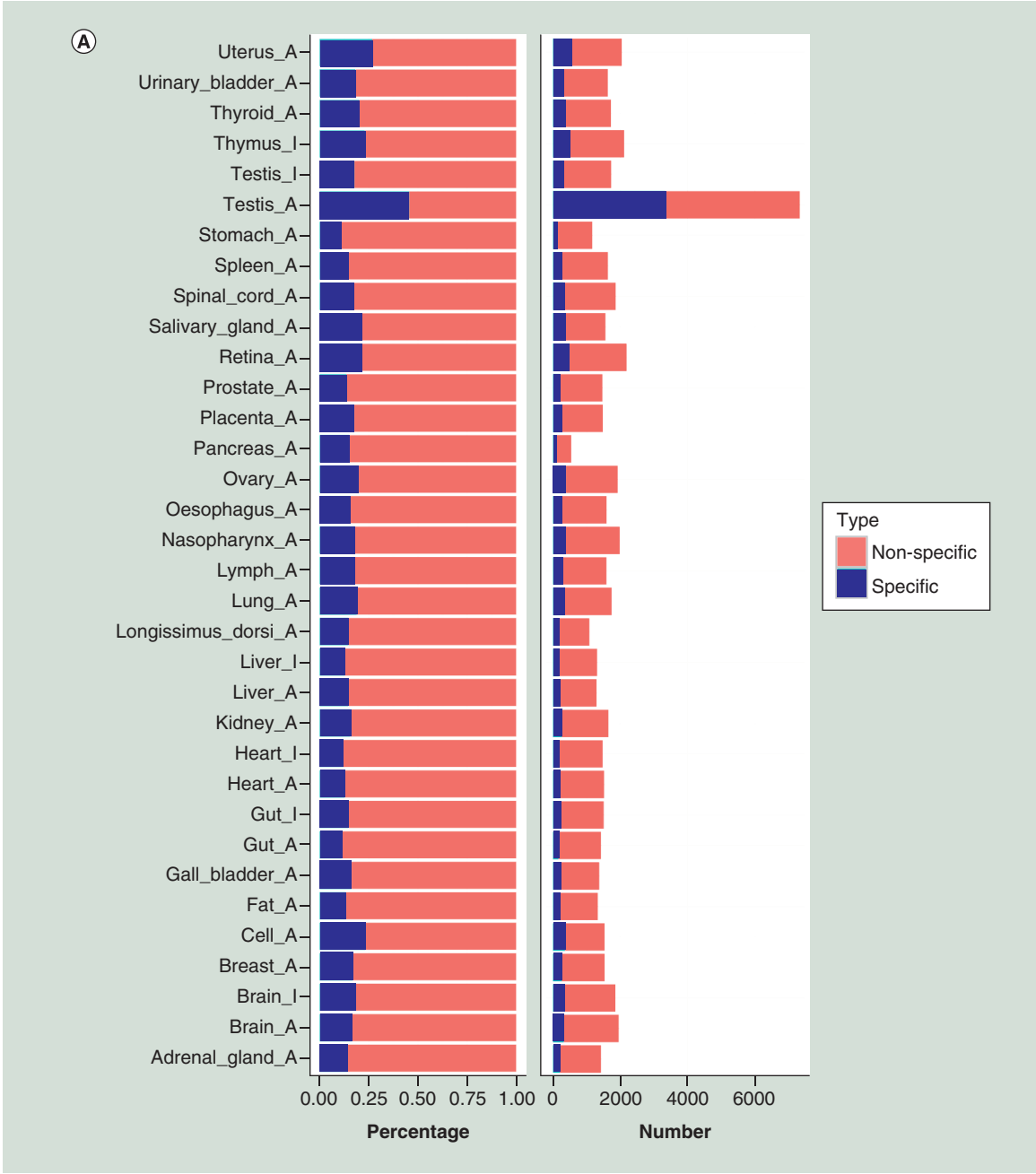


Figure 4. Expression patterns of long noncoding RNAs in different tissues. (A) The number and proportion of tissue-specific lncRNAs within all lncRNAs per sample. **(B)** The heatmap reveals the relationship between the tissues from a noncoding viewpoint. **(C)** Box plots show expression level of whole lncRNA sets in different tissues. **(D)** The heatmap for expression patterns in different tissues. Three categories of expression level are displayed: green bar shows high expression (FPKM <10); pink bar shows medium expression (FPKM: 1–10); and blue bar shows low expression (FPKM: 0.1–1).
FPKM: Fragments per kilobase of exon model per million mapped read; LncRNA: Long noncoding RNA.

Tissue-specificity & expression patterns of lncRNAs in different tissues

To elucidate the functional roles of lncRNAs during tissue differentiation, we identified and integrated all lncRNA expression signatures for each pig tissue (Supplementary Table 14). In total, 12,342 of 32,212 lncRNAs were identified to be tissue-specific in the 34 pig tissues (Figure 4A). The testis of the adult pig had the highest number of tissue-specific lncRNAs (n = 3345; 45.6%), followed by the uterus (n = 546; 26.7%), thymus (n = 487; 23.1%)

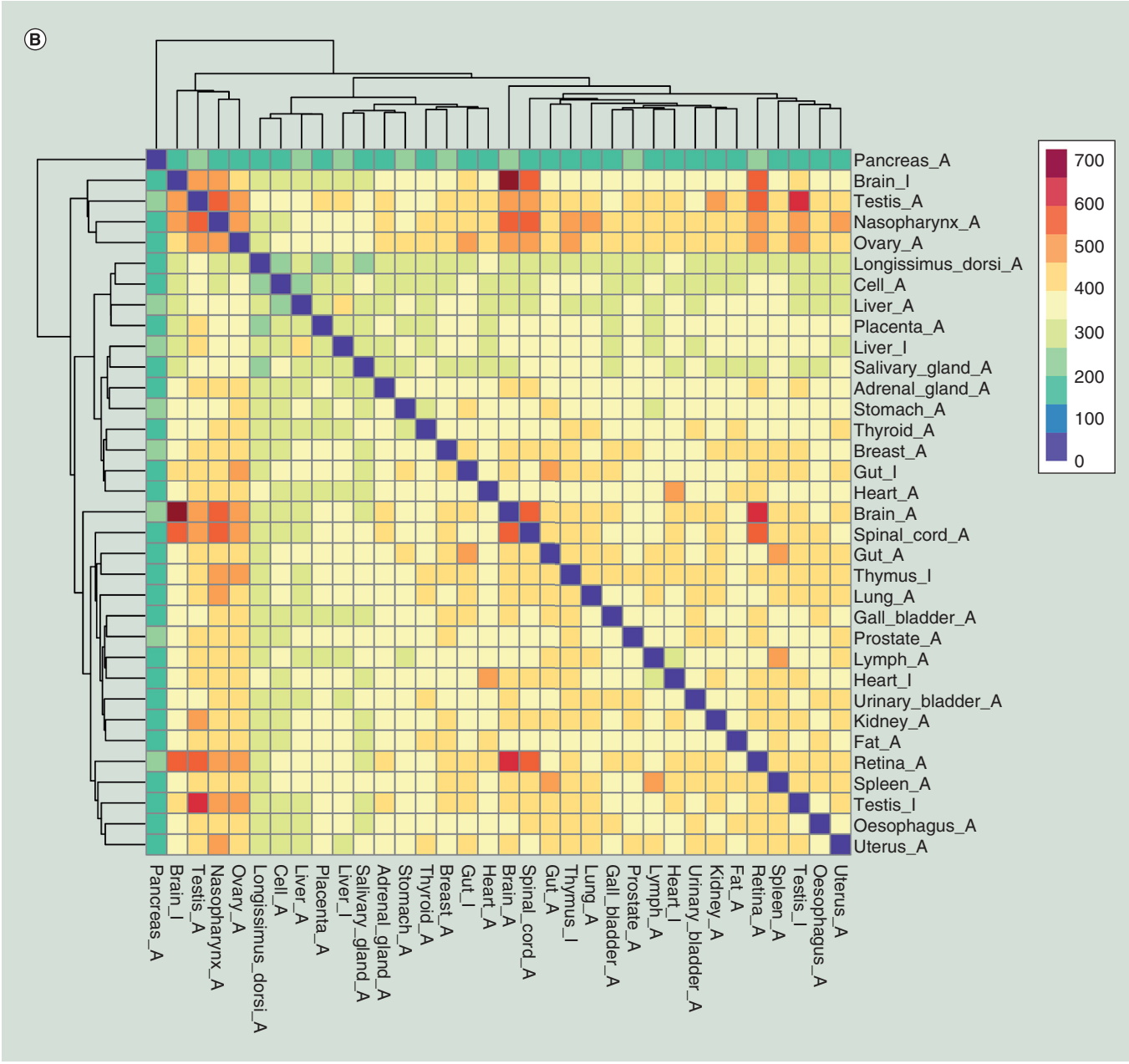


Figure 4. Expression patterns of long noncoding RNAs in different tissues (cont.). (A) The number and proportion of tissue-specific lncRNAs within all lncRNAs per sample. **(B)** The heatmap reveals the relationship between the tissues from a noncoding viewpoint. **(C)** Box plots show expression level of whole lncRNA sets in different tissues. **(D)** The heatmap for expression patterns in different tissues. Three categories of expression level are displayed: green bar shows high expression (FPKM <10); pink bar shows medium expression (FPKM: 1–10); and blue bar shows low expression (FPKM: 0.1–1).
FPKM: Fragments per kilobase of exon model per million mapped read; LncRNA: Long noncoding RNA.

and retina (n = 469; 21.5%). By contrast, the pancreas (n = 81; 15%) had the lowest number of tissue-specific lncRNAs, followed by the stomach (n = 128; 11%), longissimus dorsi (n = 157; 14.6%) and adult gut (n = 160; 11.2%).

In theory, the tissue specificity of lncRNAs reflects the functions of the tissue. We calculated the degree of overlap using shared lncRNAs from different tissues to further reveal the relationship between the various tissues at the noncoding viewpoints (Supplementary Table 15). As shown in Figure 4B, there was no clear cluster for tissues with similar biological functions; nevertheless, the same tissue with different physiological status (adult or infant) still

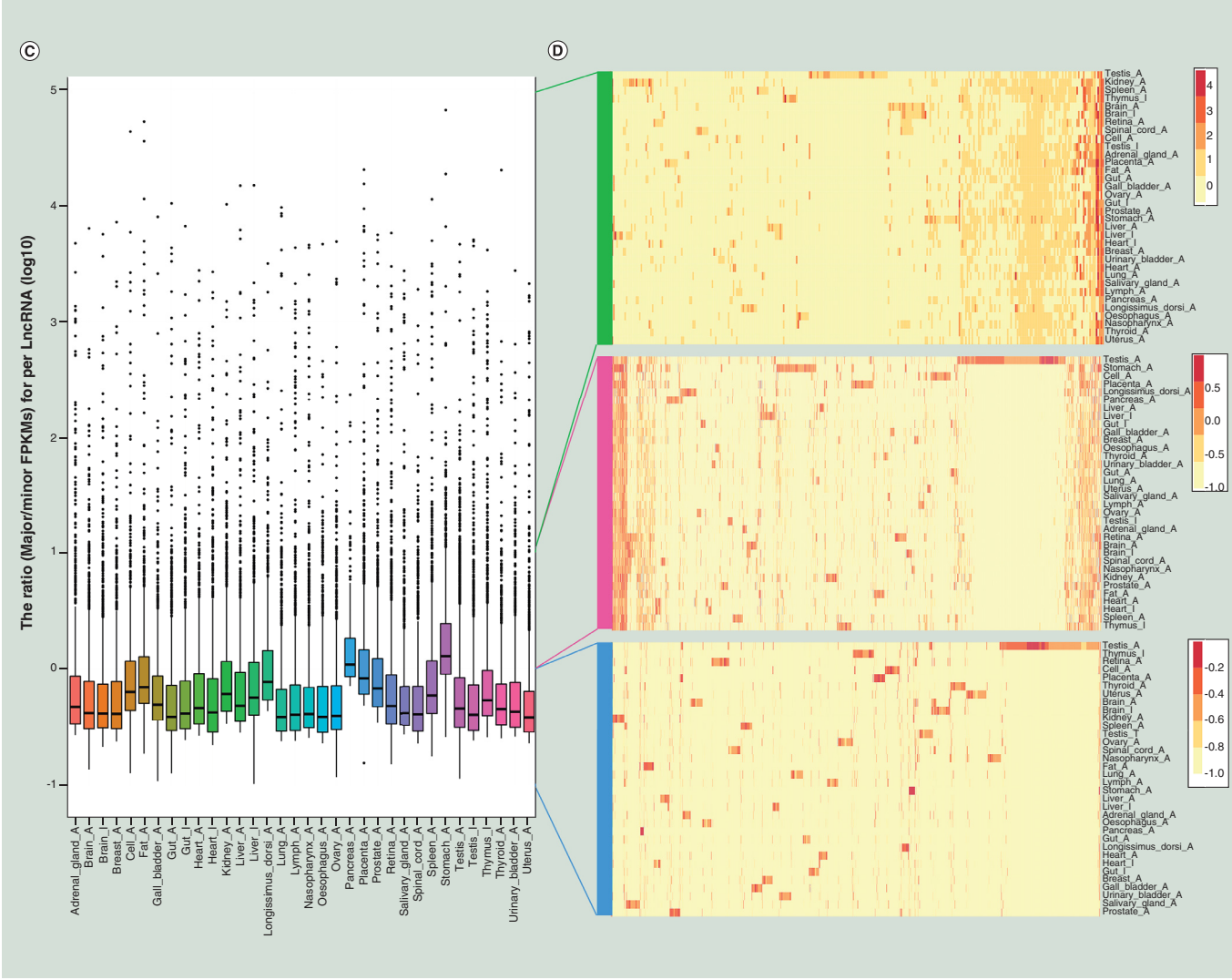


Figure 4. Expression patterns of long noncoding RNAs in different tissues (cont.). (A) The number and proportion of tissue-specific lncRNAs within all lncRNAs per sample. **(B)** The heatmap reveals the relationship between the tissues from a noncoding viewpoint. **(C)** Box plots show expression level of whole lncRNA sets in different tissues. **(D)** The heatmap for expression patterns in different tissues. Three categories of expression level are displayed: green bar shows high expression (FPKM <10); pink bar shows medium expression (FPKM: 1–10); and blue bar shows low expression (FPKM: 0.1–1).
FPKM: Fragments per kilobase of exon model per million mapped read; LncRNA: Long noncoding RNA.

had more shared lncRNAs, as in the heart, liver and brain. Therefore, we speculate that these shared lncRNAs from the same tissues at different stages were more likely to be associated with tissue specificity. Interestingly, even in the same tissue, the number of tissue-specific lncRNAs in the testis of immature pigs (n = 305) was approximately 11-fold smaller than the testis of adult pigs. This implies that these additional lncRNAs in the testis of adult pigs were associated with the development of the testis and the sexual maturity of the males.

In addition to tissue specificity of lncRNAs, different expression patterns were also present in the different tissues with a wide range of expression levels of lncRNAs in each tissue (Figure 4C). Interestingly, although the stomach had the lowest proportion of tissue-specific lncRNAs (10.96%), it showed the highest levels of expression levels (85.6 FPKM on average). In addition, we split expressed lncRNAs into three major categories, namely high expression (FPKM >10), medium expression (FPKM: 1–10) and low expression (FPKM: 0.1–1). As shown in Figure 4D, clearly different expression patterns among these categories were found. Tissue-specific lncRNAs from different tissues usually had relatively low expression levels, while most lncRNAs expressed in all tissue types were highly expressed. This finding clearly demonstrated a negative correlation between lncRNA expression specificity

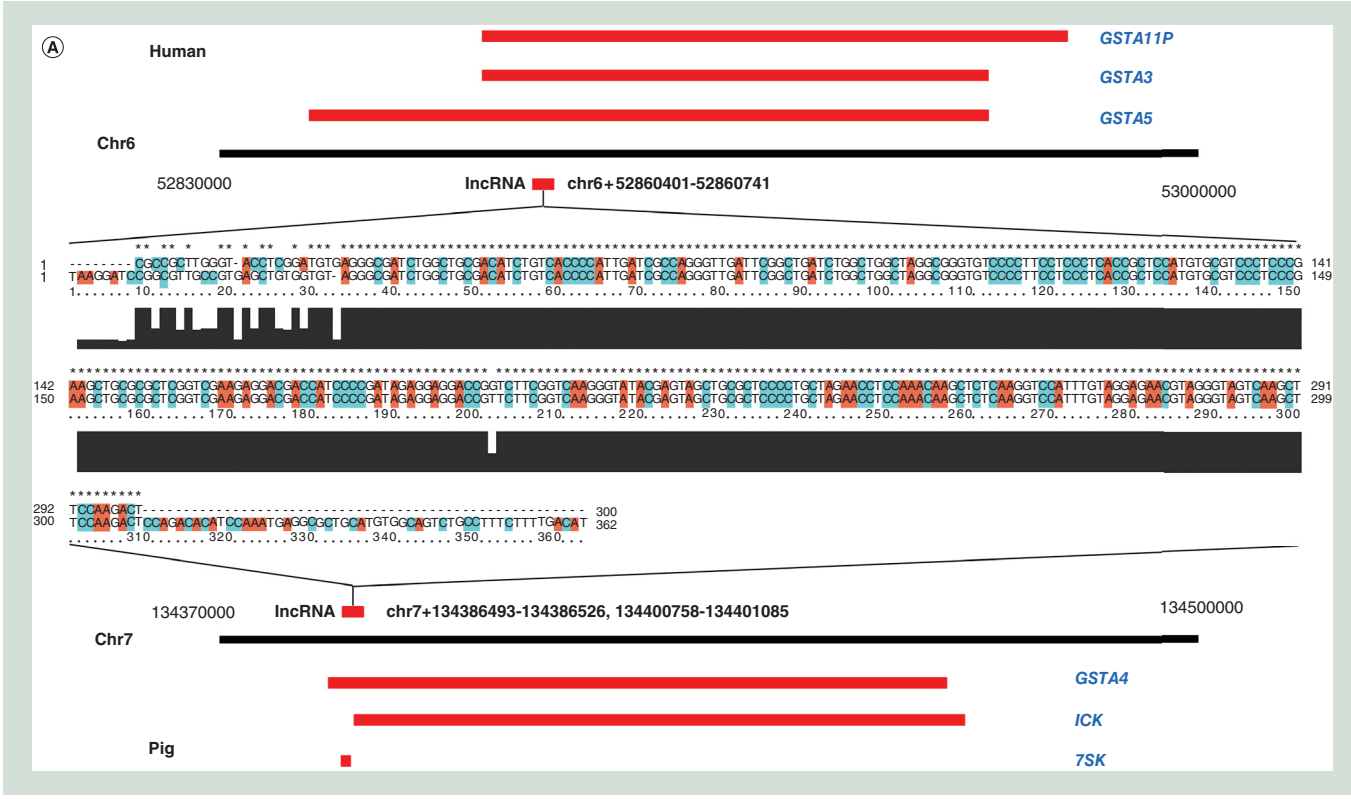


Figure 5. Functional prediction of long noncoding RNAs. (A) Example of a homologous lncRNA between human and pig. **(B)** Genome colinearity between human and pig at the lncRNA level. Various color bars represent different chromosomes in the human genome; the same color indicates colinear regions in pig chromosomes. **(C)** The validation of lncRNA transcripts using RT-PCR.
C: Control and 1–11 represent testis, uterus, stomach, kidney, liver, brain, ovary, breast, longissimus dorsi, heart and lung, respectively; LncRNA: Long noncoding RNA; M: Marker; RT: Real time.

and relative expression value and indicates that ubiquitously expressed lncRNAs might be a significant portion of the pig transcriptome [26].

We also determined whether tissue specificity differed among the four types of lncRNA. We found no significant differences among the four lncRNA types; moreover, 78–80% of lncRNAs belonged to a specific lncRNA. The levels of expression of the four lncRNA types were then examined. Sense lncRNAs were found to be expressed at the highest level (172.87 FPKM per lncRNA) on the basis of overlapping with PCGs; lincRNAs (6.61 FPKM per lncRNA) and antisense lncRNAs (5.62 FPKM per lncRNA) showed lower levels of expression in all tissues. Interestingly, intronic lncRNAs showed the lowest levels of expression (0.90 FPKM per lncRNA). Thus, intronic lncRNAs were expressed in a highly tissue-specific pattern and potentially acted as novel functional regulatory ncRNAs [27].

Functional prediction of lncRNAs

The function of the detected lncRNAs was investigated by first comparing them with the pig quantitative trait loci (QTL) database. This database includes five categories of trait-associated QTLs (reproduction, production, meat & carcass quality, health and exterior) and contains 5176 SNPs; it was downloaded from the National Animal Genome Research Program. By extracting overlapping regions, we found 65 lncRNAs that were located within 140 trait-associated regions from the published GWAS catalog (Supplementary Tables 16 &17). The annotation of these lncRNAs might provide an important supplement to previous studies of pig traits. For example, the lincRNA (NONSUSG015976) overlapped a reproduction-associated SNP in chromosome 9 (rs81413949, p-value = 8.21×10^{-7}) that is associated with the total number of piglets produced; the identification of overlapping lncRNAs may further help to pinpoint potential mechanisms related to reproductive traits.

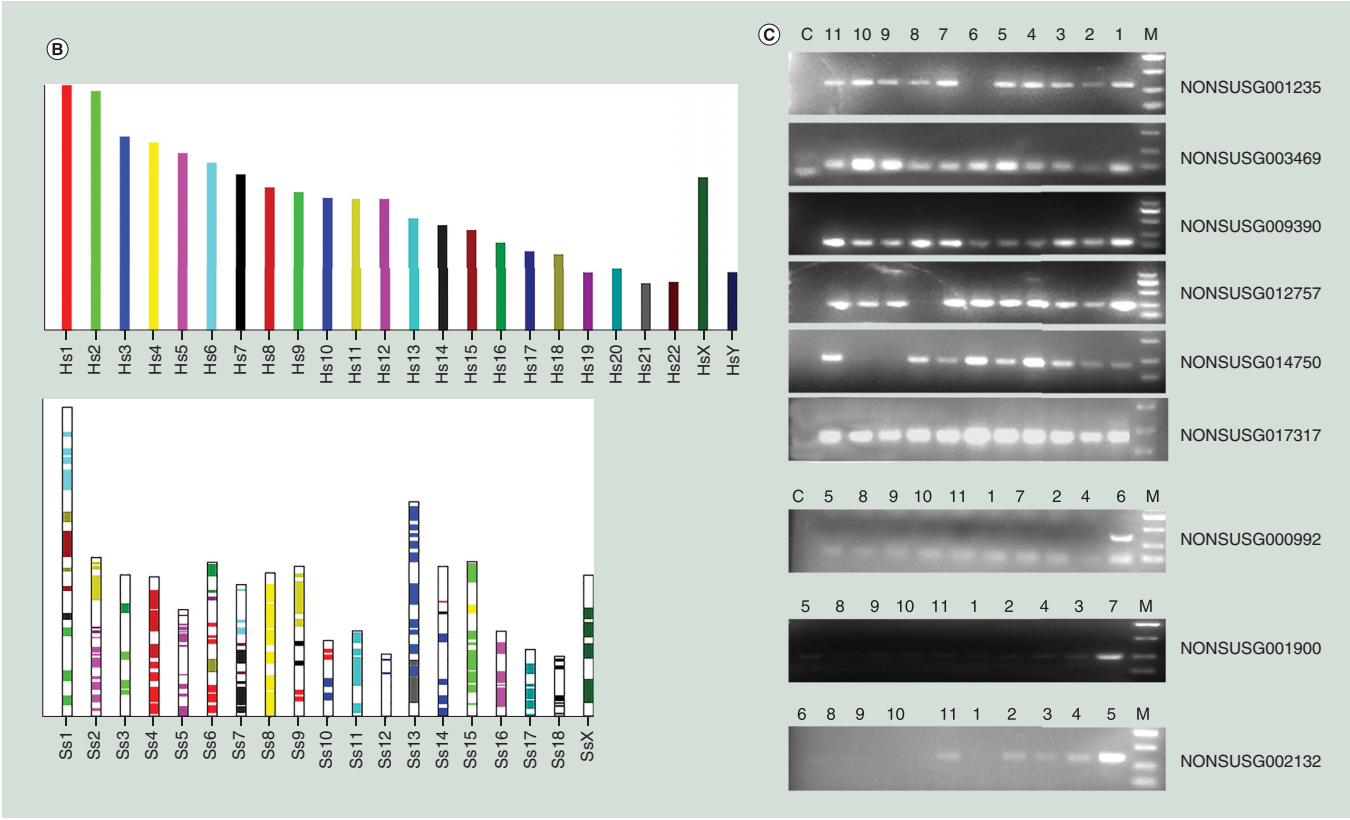


Figure 5. Functional prediction of long noncoding RNAs (cont.). (A) Example of a homologous lncRNA between human and pig. **(B)** Genome colinearity between human and pig at the lncRNA level. Various color bars represent different chromosomes in the human genome; the same color indicates colinear regions in pig chromosomes. **(C)** The validation of lncRNA transcripts using RT-PCR. C: Control and 1–11 represent testis, uterus, stomach, kidney, liver, brain, ovary, breast, longissimus dorsi, heart and lung, respectively; LncRNA: Long noncoding RNA; M: Marker; RT: Real time.

Functional prediction of lncRNAs based on known trait-associated regions was confined to current pig research. Two computation-based strategies were used to identify the regulatory function of lncRNAs that correspond to the 'cis' and 'trans' functions, respectively. First, we focused on lncRNAs that are usually regarded as cis-acting elements regulating gene expression in their chromosomal neighborhood. We defined PCGs that are located within 10 kb of an lncRNA as neighboring PCGs and performed a genome-wide scan of all identified lncRNAs. In total, 16,145 lncRNAs were identified as gene neighbors and 9930 PCGs as neighboring genes. Of these lncRNAs, 8280 had one-to-one type lncRNA-gene pairs, including 1914 sense, 2384 intronic and 3982 antisense. In addition, we found that 7748 lincRNAs (32.9%) were coding neighbors; the remainder of the lincRNAs (n = 15,805) was speculated to be potentially trans-acting factors that act in the control of gene expression.

Next, we evaluated correlations in patterns of expression between lincRNAs and PCGs in order to identify potential trans-acting factors. To this end, we identified Pearson correlation coefficients with a p-value $< 10^{-5}$ between each PCG and nontissue-specific lincRNA (n = 4975) based on their FPKM values in the 34 tissues. Overall, 97,651 lncRNA-transcript pairs (114 one-to-one type lncRNA-transcript pairs), involving 3042 lincRNAs and 5764 PCGs, were identified to be strongly correlated ($R^2 > 0.9$). In addition, we found that the correlation coefficients of neighboring lincRNA-PCG pairs ($cor = 0.83$) were larger than for long-distance lincRNA-PCG pairs ($cor = 0.78$, p-value $= 2.20 \times 10^{-5}$).

Next, we investigated the possible mechanism of lncRNA–RNA interactions in the pig by determining sequence similarities (covering the breakpoints between exons and introns) between each lncRNA and each PCG. For this purpose, we used a pipeline of human lncRNA–RNA interactions [28]. Overall, 2,789,645 potential RNA–RNA interactions were detected between 9006 lncRNAs and 13,934 PCGs that might influence pre-mRNA splicing. By combining 5606 detected ASE, we identified 6716 lncRNAs (74.57%) that were tissue-specific; 3717 of these

lncRNAs (55.35%) were predicted to associate with tissue-specific alternative splicing (Supplementary Table 18). This suggests that tissue-specific lncRNAs predominantly interact with PCGs via regulating alternative splicing of pre-mRNAs [29].

Conserved lncRNAs between pig & human

The pig is generally considered a promising medical model for human disease and pig orthologs of many human disease-associated genes have been identified. In order to investigate the orthologous relationship between humans and pigs at the lncRNA level, we first determined the level of conservation between pig lncRNAs and the whole human genome (Hg19). Overall, 8507 (26.41%) pig lncRNA isoforms of 5402 lncRNA loci had detectable orthology with the human genome; these were distributed evenly across the human genome, totaled 7.27M homologous sequences, and included the largest lncRNA of 39.22 kb.

Next, we aligned the detected pig lncRNAs with the human lncRNAs using blast software and the criteria of a minimum length ≥50 bp and identity ≥80%. We identified 5743 (17.83%) pig lncRNA isoforms with orthology to 31,259 (17.78%) human lncRNA isoforms. Orthology for lncRNAs was far lower than for PCGs; this result is consistent with those reported elsewhere for other organisms [30]. Moreover, 1437 lncRNA isoform pairs appeared highly orthologous with coverage of more than 80%, identity of more than 90% and minimum length ≥50 bp. More intriguingly, we found that highly orthologous lncRNAs usually showed similar biological functions and had the same adjoining genes (Figure 5A). These characteristics will allow us to explore lncRNA function in humans using the pig model.

Evolutionary conservation and genome colinearity between human and pig lncRNAs were further investigated by selecting these orthologous lncRNAs as anchors to identify putative homologous chromosomal regions. In parallel with PCGs (Figure 5B, Supplementary Figures 2–6), lncRNAs showed synteny for specific chromosomes between the human and pig, such as human chromosome 4 and pig chromosome 8. These lncRNAs that showed genomic synteny and colinearity provide an indication of genome evolution and allow us to further explore the potential functional similarity between human and pig genomes.

Validation of lncRNA transcripts by RT-PCR

The reliability of lncRNA identification in this study was verified by RT-PCR of nine randomly selected lncRNA transcripts in 11 tissues (Figure 5C). For lncRNAs identified in all tissues, the RT-PCR analysis showed three (NONSUSG003469, NONSUSG009390 and NONSUSG017317) which were expressed in all tissues, and three others (NONSUSG001235, NONSUSG012757 and NONSUSG014750) were detected in nine or ten tissues. NONSUSG000992, NONSUSG002132 and NONSUSG001900 were successfully validated as tissue-specific lncRNAs in brain, liver and ovary, respectively. These results confirm the validity of our identification and characterization of the pig lncRNAs.

Discussion

To date, a considerable number of high-confidence lncRNAs have been deposited in the publicly available NONCODE database. Although 527,336 lncRNAs from 16 different species have been collected and annotated in NONCODE 2016, none are available from the pig despite its importance economically and biomedically. This dearth of information can be ascribed to poor lncRNA annotation in the pig transcriptome; only 40 transcripts for pig lncRNA loci are annotated in Ensembl, restricting in-depth research into pig functional genomics. Nevertheless, despite the lack of sufficient resources, the pig has been used for lncRNA-based species evolutionary analyses [3,30]. The databases GBE-Zhang [11] and ALDB [19] include 4515 and 7194 lncRNA loci, respectively; however, the lncRNAs have been characterized in a limited number of tissues and the majority of the sequences belong to the lincRNA category. One of the aims of this study was to enrich noncoding genomic annotation for the pig. We identified 32,212 nonredundant lncRNA isoforms from 18,676 lncRNA loci in 34 normal tissues, and all identified lncRNAs were successfully deposited in the NONCODE database to enable other researchers to make use of the information for functional genomic studies in pigs.

Profiling the expression of these lncRNAs in multiple tissues provided a means to better comprehend their tissue specificity and their potential functions. Almost 38.3% of pig lncRNAs were identified in a single sample, indicating a high level of specific expression in different tissues or developmental stages. Interestingly, tissue specificity of lncRNAs showed a negative relationship with expression level; thus, lncRNAs expressed in all tissues always appeared to have high expression levels (FPKM >10). We speculate that lncRNAs expressed in all tissues

are involved in basic functions necessary for the sustenance or maintenance of the cell, in a similar manner as housekeeping genes. Tissue-specific lncRNAs were temporally restricted in specific functional activity and biological pathways. Identification of ubiquitously expressed versus tissue-specific lncRNAs is crucial for the interpretation of their functionality. The testis is a specialized tissue and has been identified as having the maximum number of expressed lncRNAs and protein-coding RNAs in both human and mouse [31,32]. Our results showed a similar feature in the pig adult testis, in other words, a high proportion of lncRNAs in the adult testis appeared to be tissue-specific, reflecting the fact that these lncRNAs may be associated with testis development and function.

The findings in this study provide a valuable resource for further enhancing noncoding genome annotation in pigs. In addition, a comprehensive computation-based strategy was employed to identify potential relationships between the identified lncRNAs and known PCGs. To our knowledge, this is the first genome-wide functional prediction of lncRNAs in the pig that can be used as a reference for further investigation of target lncRNAs. We also analyzed multiple pig tissue samples to profile the lncRNAs and this information provides an opportunity to explore the potential functions of tissue-specific lncRNAs, such as the association of tissue-specific lncRNAs with known ASE. The potential relationships between lncRNAs and PCGs still need to be confirmed by experimental methods in future studies, such as through the use of RNA pull-down, RNA-RIP and ChIP-seq strategies.

Conclusion

Our study provides the first comprehensive and high-quality pig lncRNA resource, which has been deposited in the publicly available NONCODE database. The comprehensive functional predictions for pig lncRNAs provide new insights for the development of a genome-wide lncRNA-targeted genome draft for further functional studies on pig noncoding genes. A large number of newly annotated sequences for pig noncoding genes will be of value for many current and future pig research projects.

Summary points

- We identified a final set of 32,212 nonredundant long noncoding RNA (lncRNA) isoforms from 18,676 lncRNA loci with average three exons and average length of 1089 bp.
- The number of detected lncRNAs showed approximately 1.5-times and 805-times of 21,607 annotated protein-coding genes (PCGs) and 40 noncoding RNAs in Ensembl, respectively.
- The comprehensive computation-based strategy was employed to pinpoint the potential relationship between the identified lncRNAs and known PCGs.
- Both PCGs and lncRNAs had the significant enrichment of most histone modification, as well as this lncRNAs displayed a stronger enrichment than PCGs.
- We found 6716 lncRNAs (74.57%) were tissue-specific and 3717 of tissue-specific lncRNAs (55.35%) were predicted to associate with tissue-specific alternative splicing.
- LncRNAs also shared synteny between human and pig genome, such as chromosome 4 of human versus chromosome 8 of pig.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: https://www.futuremedicine.com/doi/suppl/10.2217/epi-2017-0149

Availability of data & materials

The sequenced RNA-seq raw data for 34 pig tissues are available from NCBI Sequence Read Archive with the BioProject number: PRJNA392949.

Author contributions

JF Liu conceived and designed the experiments. P Zhao performed lncRNAs detection and transcriptome analyses. Y Zhao provided NONCODE platform for depositing pig lncRNAs database. W Feng and X Zheng assisted the lncRNA experimental validations. P Zhao, X Zheng, H Wang, H Kang, C Ning and H Du collected samples and prepared for sequencing. JF Liu, P Zhao, X Zheng, W Feng, Y Yu and B Li wrote and revised the paper. All authors declare that they have read and approved the final manuscript.

Acknowledgements

We thank Z Fan, Y Dong and K Yang for sample collection.

Financial & competing interests disclosure

This work was supported by the National High Technology Research and Development Program of China (863 Program, 2013AA102503), the National Natural Science Foundations of China (31661143013) and the Program for Changjiang Scholar and Innovation Research Team in University (IRT1191). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethics approval and consent to participate

The whole sample collection and treatment were conducted in strict accordance with the protocol approved by the Institutional Animal Care and Use Committee (IACUC) of China Agricultural University (permit number: DK1023).

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Quan M, Chen J, Zhang D. Exploring the secrets of long noncoding RNAs. *Int. J. Mol. Sci.* 16(3), 5467–5496 (2015).
2. Wang L, Tang H, Xiong Y, Tang L. Differential expression profile of long noncoding RNAs in human chorionic villi of early recurrent miscarriage. *Clin. Chim. Acta* 464, 17–23 (2017).
3. Necsulea A, Soumillon M, Warnefors M *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505(7485), 635–640 (2014).
4. Wei S, Wang K. Long noncoding RNAs: pivotal regulators in acute myeloid leukemia. *Exp. Hematol. Oncol.* 5(1), 30 (2016).
5. Wapinski O, Chang HY. Long noncoding RNAs and human disease. *Trends Cell Biol.* 21(6), 354–361 (2011).
6. Faghihi MA, Modarresi F, Khalil AM *et al.* Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase expression. *Nat. Med.* 14(7), 723 (2008).
7. Bassols A, Costa C, Eckersall PD, Osada J, Sabria J, Tibau J. The pig as an animal model for human pathologies: a proteomics perspective. *Proteomics Clin. Appl.* 8(9–10), 715–731 (2014).
8. Yu L, Tai L, Zhang L, Chu Y, Li Y, Zhou L. Comparative analyses of long non-coding RNA in lean and obese pig. *Oncotarget* 8(25), 41440–41450 (2017).
9. Li J, Gao Z, Wang X, Liu H, Zhang Y, Liu Z. Identification and functional analysis of long intergenic noncoding RNA genes in porcine pre-implantation embryonic development. *Sci. Rep.* 6, 38333 (2016).
10. Wang Y, Hu T, Wu L, Liu X, Xue S, Lei M. Identification of noncoding and coding RNAs in porcine endometrium. *Genomics* 109(1), 43–50 (2017).
11. Zhou Z-Y, Li A-M, Adeola AC *et al.* Genome-wide identification of long intergenic noncoding RNA genes and their potential association with domestication in pigs. *Genome Biol. Evol.* 6(6), 1387–1392 (2014).

•• **The first paper about genome-wide long noncoding RNA (lncRNA) in pigs.**

12. Index of ftp://ftp.ensembl.org/pub/release-67/fasta/sus_scrofa/dna/. ftp://ftp.ensembl.org/pub/release-67/fasta/sus_scrofa/dna/
13. Index of /pub/release-82/gff3/sus_scrofa/. http://ftp.ensembl.org/pub/release-82/gff3/sus_scrofa/
14. ALDB: a domestic-animal long noncoding RNA database. http://res.xaut.edu.cn/aldb/index.jsp
15. Splicing express. https://bitbucket.org/jekroll/splicingexpress/
16. MCScanX: Multiple Collinearity Scan toolkit. http://chibba.pgml.uga.edu/mcscan2/
17. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33(3), 290–295 (2015).
18. Sun L, Luo H, Bu D *et al.* Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* 41(17), e166 (2013).
19. Li A, Zhang J, Zhou Z, Wang L, Liu Y, Liu Y. ALDB: a domestic-animal long noncoding RNA database. *PLoS ONE* 10(4), e0124003 (2015).

• **Important lncRNA analysis in domestic animals.**

20. Ma L, Bajic VB, Zhang Z. On the classification of long non-coding RNAs. *RNA Biol.* 10(6), 925–933 (2013).
21. Derrien T, Johnson R, Bussotti G *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22(9), 1775–1789 (2012).

22. Kornienko AE, Dotter CP, Guenzl PM *et al.* Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol.* 17, 14 (2016).

•• **Provides the expression variation of lncRNAs in human.**

23. Kellis M, Wold B, Snyder MP *et al.* Defining functional DNA elements in the human genome. *Proc. Natl Acad. Sci. USA* 111(17), 6131–6138 (2014).

24. Xiao S, Xie D, Cao X *et al.* Comparative epigenomic annotation of regulatory DNA. *Cell* 149(6), 1381–1392 (2012).

25. Peschansky VJ, Wahlestedt C. Non-coding RNAs as direct and indirect modulators of epigenetic regulation. *Epigenetics* 9(1), 3–12 (2014).

26. Jiang C, Li Y, Zhao Z *et al.* Identifying and functionally characterizing tissue-specific and ubiquitously expressed human lncRNAs. *Oncotarget* 7(6), 7120–7133 (2016).

•• **Shows the tissue-specific expression of lncRNAs in human.**

27. Louro R, Smirnova AS, Verjovski-Almeida S. Long intronic noncoding RNA transcription: expression noise or expression choice? *Genomics* 93(4), 291–298 (2009).

28. Szczesniak MW, Makalowska I. lncRNA-RNA interactions across the human transcriptome. *PLoS ONE* 11(3), e0150353 (2016).

•• **Shows the function of lncRNA through interactions with RNA to regulate and diversify the human transcriptome.**

29. Bardou F, Ariel F, Simpson CG *et al.* Long noncoding RNA modulates alternative splicing regulators in Arabidopsis. *Dev. Cell* 30(2), 166–176 (2014).

30. Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* 11(7), 1110–1122 (2015).

31. Lindskog C. The potential clinical impact of the tissue-based map of the human proteome. *Expert. Rev. Proteomics* 12(3), 213–215 (2015).

32. Zhao Y, Liu W, Zeng J *et al.* Identification and analysis of mouse non-coding RNA using transcriptome data. *Sci. China Life Sci.* 59(6), 589–603 (2016).

•• **Provides lncRNA database platform across different species.**