Construction of a transposase accessible chromatin landscape reveals chromatin state of repeat elements and potential causal variant for complex traits in pigs

Tao Jiang
Jiangxi Agricultural University
Ziqi Ling
Jiangxi Agricultural University
Zhimin Zhou
Jiangxi Agricultural University
Xiaoyun Chen
Jiangxi Agricultural University
Liqing Chen
Jiangxi Agricultural University
Sha Liu
Jiangxi Agricultural University
Yingchun Sun
Jiangxi Agricultural University
Jiawen Yang
Jiangxi Agricultural University
Bin Yang
Jiangxi Agricultural University
Jianzhen Huang
Jiangxi Agricultural University
Lusheng Huang (lushenghuang@hotmail.com)
Jiangxi Agricultural University

Research Article

Keywords: Chromatin accessibility, Pig, Tissue specific, Transcriptome, Transposable elements

Posted Date: April 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1558839/v1

Abstract

Background

A comprehensive landscape of chromatin states for multiple mammalian tissues is essential for elucidating the molecular mechanism underlying regulatory variants on complex traits. However, the genome-wide chromatin accessibility has been only reported in limited tissue types in pigs.

Results

Here we report a genome-wide landscape of chromatin accessibility of 20 tissues in two female pigs at ages of 6 months using ATAC-seq, and identified 557,273 merged peaks, which greatly expanded the pig regulatory element repository. We revealed tissue-specific regulatory elements, which were associated with tissue-relevant biological function. We identified both positive and negative significant correlations between the regulatory elements and gene transcripts, which showed distinct distributions in terms of their strength and distances from corresponding genes. We investigated the presence of transposable elements in open chromatin regions across all tissues, these included identifications of porcine endogenous retroviruses exhibiting high accessibility in liver and homology of porcine specific virus sequences to universally accessible TEs. Furthermore, we prioritized a potential causal variant for polyunsaturated fatty acid in the muscle.

Conclusions

Our data provides a novel multi-tissue accessible chromatin landscape that serve as an important resource for interpreting regulatory sequences in tissue specific and conserved biological function, as well as regulatory variants underlying variation in complex traits associated loci in pigs.

Background

Eukaryotic genomes are usually packed into nucleosomes, which comprise of 147bp DNA wrapping around a histone octamer, forming the structural units of chromatin. Chromatin is categorized as active euchromatin or inactive heterochromatin based on its accessibility [1, 2], which is linked to fundamental cellular processes including gene transcription, DNA replication and repair [3]. The accessible chromatin often reflects binding of transcriptional machinery to cis-regulatory elements such as promoters and enhancers [4–8]. Profiling the accessible chromatin in different tissues could help to interpret regulatory basis underlying spatiotemporal transcription of genes [9–11], as well as genomic variants causing diseases susceptibility and complex trait variations. [12–15].

Over the past decades, several chromatin accessibility profiling methods including DNase I hypersensitive site sequencing (DNase-seq), micrococcal nuclease sequencing (MNase-seq) and assay for Transposase-

Accessible Chromatin using sequencing (ATAC-seq) have been developed. Among these, ATAC-seq has become increasingly popular because of its simplicity, high sensitivity and low cell number requirement [16], and has been applied to construct regulatory elements landscapes in human and mouse [17–20]. The international consortiums like, Encyclopedia of DNA Elements (ENCODE) [21], Roadmap Epigenomics projects [22] and the Functional Annotation of Animal Genomes (FAANG) [23] have played important roles in generating the epigenetic data in multiple species. There have been efforts to profile accessible chromatin in pigs [24–26]. However, only few tissues including fat, cerebellum, cortex, heart, Liver, muscle, spleen, lung and hypothalamus were covered in these studies. Moreover, studies have investigated chromatin states of TEs including LINE, SINE, LTR and DNA transposons in human and mouse [27–29], allowing better understanding the regulatory function of transposable elements (TEs). However, the chromatin landscape of pig genomic TEs has not been explored yet.

Here, we performed ATAC-seq on 40 samples from 20 tissues in 2 female Duroc-Landrace-Yorkshire (DLY) commercial pigs, including 10 tissues (medulla spinalis, bronchia, parotid gland, pharynx, stomach, small intestine (SI), kidney, ovary, cervix, thymus) that have not been investigated in previous studies. We inferred the open chromatin regions displaying tissue specific or tissue shared accessibility. By integrating transcriptomic data (RNA-seq), we revealed open chromatin regions that significantly correlated with gene transcription. In addition, we examined TEs including the PERVs and different TEs in the open chromatin regions, providing insights into the regulatory role of the TEs and their evolutionary history. The data and analyses in this study provide vital resources to explore the function of cis-regulatory elements in determining tissue specific biological function and complex traits in pigs.

Methods And Materials

Sample collection

The pigs used in this study were obtained from a slaughter house in Nanchang, Jiangxi province. 20 tissues (brain, cerebellum, hypothalamus, medulla spinalis, lung, bronchia, parotid gland, pharynx, stomach, small intestine, liver, kidney, cervix, ovary, heart, thymus, longissimus dorsi, semimembranosus, backfat, leaf fat) was collected from 2 healthy commercial sows at age of 6 months. After slaughter, tissue samples were collected in sterile tube and stored at -80℃.

ATAC-seq library preparation and sequencing

ATAC-seq libraries were prepared as previously described with minor modifications [16]. Briefly, samples were ground to powder with liquid nitrogen using mortar. Then, we homogenized the tissue powder in 2ml ice-cold homogenization buffer (10% NP40, 500 mM EDTA, 1M Sucrose, 100mM PMSF, 14.3M β-mercaptoethanol, 1M CaCl2, 1M Mg(Ac)2, 1M Tris[pH 7.8], nuclease-free H2O) using a 7mL Dounce ho mogenizer for 20 strokes. We filtered the solution through a 70μm cell strainer and gradient centrifugation with different concentrations of iodoxanol. After that, approximately 60,000 nuclei from each sample were collected, centrifuged at 500g for 5 min at 4℃, washed the pellet with 1 mL ATAC-RSB(5M NaCl, 1M MgCl2, 1M Tris-HCl[pH 7.4], nuclease-free H2O) containing 0.1%Tween-20, centrifuged at 500g for 10 min

at 4℃; and resuspended the pellet in 50 μL of tagmentation reaction mix (2×Tagmentation buffer, 10×PBS, 0.5% Digitonin, 10% Tween20, Tn5 transposase and nuclease-free H2O), incubated for 30 min at 37℃. Libraries were prepared using active motif kit (ATAC-Seq Kit #53150) as previously described [30]. The ATAC-seq DNA was sequenced in the 150 bp paired-end sequencing mode with a NovaSeq 6000 platform. Two biological replicates were performed per tissue.

ATAC-seq processing

According to the Additional file 1: Fig. S1a, the raw data with sequence adapters and low-quality reads were trimmed with Trim Galore v0.6.6 (available at https://github.com/FelixKrueger/TrimGalore) using default parameters. The high-quality reads were then aligned to the pig genome (Sus scrofa 11.1) using Bowtie2 v2.3.5.1 [31]. Duplicate alignments were removed using Picard toolkit v1.119 (http://broadinstitute.github.io/picard/) and alignments with low mapping quality (mapq < 30) and mitochondrial DNA were removed with samtools v1.9 [32]. We used MACS2 v2.1.1 [33] to generate the accessible chromatin regions (peaks) for each sample with the parameter –qvalue 0.01 –nomodel –shift − 100 –extsize 200 -B -SPMR –keep-dup all, and then we merged peaks across all samples within the same tissue in two individuals overlapping peaks using BEDTools merge [34]. We recalculated the number of reads in each merged peak using samtools bedcov and Peaks Per Kilobase Million values were determined to measure the peak intensity using R program v3.5.1 (https://www.r-project.org/). The clustering of the samples based on peak intensity were performed using hcluster function with ward.D method in R program, and visualized using R package circlize and dendextend. Genomic annotation for peaks was performed with GAT (https://github.com/AndreasHeger/gat/blob/master/doc/contents.rst). And tissue-specific peaks are defined as peak intensity in specific tissue at least twice higher than in the others.

RNA-seq library preparation, sequencing and processing

Total RNA was extracted from 40 frozen samples using Trizol and sequenced using MGI-2000 system with paired end and strand specific. RNA-seq reads of each sample were mapped to pig genome (Sus scrofa 11.1) using STAR v2.7.1a [35]. Expression level of each gene was determined with stringtie [36] and featureCounts [37] and normalized with R. Only genes expressed (TPM > 0) across the 40 samples in more than 2 samples were included in the subsequent analyses. The clustering of the samples based on gene expression levels were performed using hcluster function with ward.D method in R program, and visualized using R package circlize and dendextend. And two samples (stomach_sample1 and backfat_sample2) with poor clustering were discarded. And tissue-specific genes are defined as expression level in specific tissue at least twice higher than in the others.

Gene ontology enrichment analysis

The gene ontology enrichment analysis was performed using ClueGO [38]. The P values for the enrichment of GO terms were corrected using Benjamini-Hochberg approach.

Transcription factor motifs

We tested for enrichment of vertebrates known transcription factor motifs in peaks using HOMER [39] with default parameters.

Correlation of Peaks with gene expression

To clarify the regulatory relationship between peaks and genes, three other data were used in this study. The first study was from Zhou et al [40]. A total of 16 samples were collected from 8 tissues (fat, cerebellum, brain, hypothalamus, liver, lung, muscle and spleen) of yorkshire with 2 biological replicates per tissue. The second was designed by Zhao et al [41], which include 5 tissues (muscle, liver, fat, spleen, heart) of MeiShan, LargeWhite, Enshi and Duroc with 1–2 biological replicates per tissue. And the last one was executed by Tang et al [42], which we downloaded and only included one tissue (skeletal muscle) of Luchuan and Duroc pigs with 3 biological replicates per breed (Supplementary Table 3). In total, combined with our own data, there were 85 samples of ATAC-seq and corresponding RNA-seq. After the above uniform ATAC-seq process analysis, we identified 796,506 peaks with p-value less than 10^{-5}, and 14 samples with poor tissue clustering were discarded. And then 527,718 peaks exist in at least 2 samples were retained. For RNA-seq, after uniform RNA-seq process analysis above, 24,974 genes were identified, then, we selected 17,634 genes with transcripts per million (TPM) > 1 in at least 10% samples for further analysis. Using 527,718 peaks and 17,634 genes across 71 samples, we predicted target genes for peaks by implementing correlation between peak intensity and gene expression within 500 kb. We identified 827,942 significant correlations at q-value less than 0.01.

Spatial correlation between peaks and TEs

The transposable elements (TEs) data of pig (Sus scrofa 11.1) were analyzed by RepeatMasker v2.9.0 [43] with parameter -s. And correlated interval sets between peaks and TEs were calculated by GenometriCorr (Genometric Correlation) [44] which was used to calculate the spatial correlation of genome-wide interval datasets. The permutation tests with 100 times were performed to verify whether TEs enriched in the regulatory elements or not.

Identification of accessible TEs

The transposable elements (TEs) data of pig (Sus scrofa 11.1) were obtained from the

study (unpublished, DOI: 10.24272/j.issn.2095-8137.2021.379). The accessible TEs were identified by the overlapping between TEs and peaks, and the evaluation criterion is that at least 50% of the overlap should be TE while at least 20% of the overlap should be peak region [45]. Accessible TEs that were only identified in one tissue was defined as tissue-specific accessible TEs while accessible TEs that were identified in all tissues was defined as shared accessible TEs. GO enrichment analysis of tissue-specific accessible TEs and shared accessible TEs were performed using Cluego [38].

Multiple sequence alignment

Virus sequences were obtained from a study (unpublished) by self-assembly and comparison with the database (https://img.jgi.doe.gov/cgi-bin/vr/main.cgi). The conserved accessible TEs sequences and viral sequences were put together. Then, we used MAFFT v7.407 [46] with the automatically (–auto) detected parameters to identified sequence homology. FastTree v2.1.11 [47]was used to infer approximately-maximum-likelihood phylogenetic trees from alignments of nucleotide sequences. Finally, experiment visualization features in FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

The enrichment of GWAS signals

The pig GWAS data of fatty acid was collected from a published study [48]. A total of 191 significant GWAS loci were obtained. we performed these GWAS signal enrichment for peaks across 20 tissues.

Results

Mapping the open chromatin states of 20 pig tissues

To map genome-wide regulatory elements in pig, we performed ATAC-seq and RNA-seq on 40 samples representing 20 tissues, namely, brain, cerebellum, hypothalamus, medulla spinalis, lung, bronchia, parotid gland, pharynx, stomach, liver, SI, kidney, ovary, cervix, heart, thymus, *longissimus* muscle (LD), semimembranosus muscle (SM), backfat and leaf fat) from two female DLY pigs at 6 months (Fig. 1a), representing two biological replicates per tissue. After quality control, we obtained a total of 2.43G uniquely mapped reads from the 40 samples (Supplementary Table 1). The reads showed enrichment around transcription start sites (TSS) and a periodical fragment size distribution corresponding to the nucleosome-free regions (NFR) (< 100 bp) and mono-, di-, and tri-nucleosomes (~ 200, 400 and 600 bp, respectively) (Additional file 1: Fig. S1b, S1c). Furthermore, we computed quality related statistics FRIP (fraction of reads in peaks), TSS score, library complexity (NRF, PBC1 and PBC2) and cross correlation (NSC and RSC)) (Supplementary Table 1). These statistics suggested that the data generated hereby are of good quality. We identified an average of 77,096 peaks with average size of 448 bp per tissue (Additional file 1: Fig. S1e). Peaks from all samples were merged into 557,273 non-redundant peaks, among which 363,638 (65.25%) overlapped with peaks identified in Zhou et al [40] (Fig. 1d), while the rest of peaks were newly identified owing to more types of tissues were profiled in this study.

To obtain a set of peaks with higher confidence, we only kept peaks that were evidenced in both biological replicates for each sample, and finally retained 267,836 merged peaks for further analysis. Most of the 267,836 peaks located in intronic (54.23%) and intergenic (36.75%) regions, meanwhile, 4.23%, 2.59% and 2.20% peaks were evidenced in CDS, 5'UTR and 3'UTR regions (Fig. 1f). Similar to the results obtained in the previous study [25], we found that accessible chromatin region varies greatly across tissues (range: 9,729 to 94,946), covering 0.56% to 3.21% sequences of the reference genome (Fig. 1c), with a good overlap with those reported in independent studies by tissues (Supplementary Table 4).

a Schematic diagram showing tissues assayed in this study. The ring colors represent different organ systems. Starting with the location of the brain tissue, clockwise around the circle, corresponding to the

nervous system (brain, cerebellum, hypothalamus, medulla spinalis), respiratory system (lung, bronchia), digestive system (parotid gland, pharynx, stomach, liver, small intestine (SI)), urinary system (kidney), reproductive system (ovary, cervix), circulatory system (heart), lymphatic system (thymus), locomotor system (longissimus muscle (LD), semimembranosus (SM)), endocrine system (backfat and leaf fat). **b** Hierarchical clustering of samples using ATAC-seq signal intensity. **c** Distribution of the peak number and genome coverage in each tissue. **d** Venn plot showing the overlap of peaks identified in this study with those reported in Zhou et al. 2021. **e** Pearson correlation heatmaps of 20 tissues by density of ATAC-seq distal peaks (left) and ATAC-seq proximal peaks (right). **f** Pie chart represents the proportion of peaks located in introns, intergenic, 5'UTR, 3'UTR and CDS regions.

Hierarchical clustering based on the peak intensity clearly separated different tissue types (Fig. 1b), indicating that the peak intensities largely reflect tissue identity. We grouped the 267,836 peaks into 20,488 proximal peaks (within 1Kb of TSS) and 247,248 distal peaks (away from TSS 1Kb). As expected, the distal peaks displayed stronger tissue specificity comparing to proximal peaks (Fig. 1e and Additional file 1: Fig. S1d), agreeing with the reported characteristics of promoters and enhancers [49].

To investigate the effect of regulatory elements on gene expression, RNA-seq was successfully performed on 38 out of the 40 ATAC-Seq samples. The summary statistics of each RNA library is listed in Supplementary Table 2. A total of 17,624 genes were identified. The reproducibility of the 38 RNA-seq data was verified by hierarchical clustering of gene expression (Additional file 1: Fig. S2a).

Tissue specificity of the open chromatin peaks

To explore the key regulatory elements that determine specialized function of tissues, we defined peaks whose average intensity was at least two folds in the target tissue relative to any other tissues as tissue-specific peaks, and identified 16,704, 7,887, 6,742, 5,960, 5,286, 4,278, 3,089, 884, 649, 407, 326, 277, 261, 177, 118, 84, 21, 16, 6 and 3 tissue-specific peaks in cerebellum, kidney, liver, heart, parotid gland, cervix, thymus, leaf fat, stomach, brain, SI, LD, pharynx, SM, ovary, hypothalamus, medulla spinalis, lung, backfat and bronchia, respectively (Fig. 2a, 2b). We exemplified a liver specific peak located close to *ARG1* (Fig. 2c and Additional file 1: S2e), a protein coding gene expressed primarily in the liver and involved in the urea cycle [50]. The genes locate closest to the tissue-specific peaks were enriched in tissue-specific pathways, such as the heart and circulatory system development for heart specific peaks (Additional file 1: Fig. S2f), such as heart development and circulatory system development in heart. Given that the numbers of tissue-specific peaks were largely affected by the inclusion of physiologically similar tissues, we further determined peaks that specific to a certain tissue system using a similar rule applied on the individual tissues, and identified 11,114, 434, 310, 28, 5,053 and 50 specific peaks for locomotor system, reproductive system, endocrine system, respiratory system, nervous system and digestive system, respectively, located close to genes with function matching the biology of corresponding tissues (Fig. 2a). For example, genes expressed specifically in nervous system were significantly involved in synaptic membrane (q-value = 1.39E-32), while those expressed specifically in locomotor system were significantly associated with muscle structure development (q-value = 5.68E-28) (Additional file 1: Fig. S2d).

Next, we investigated enrichment of TF binding motifs in the tissue-specific peaks using HOMER [39]. We defined the TF motifs with enrichment strength (-log (p-value)) in one tissue were at least two folds relative to the other tissues as tissue-specific, and identified 205 tissue-specific TF motifs (Fig. 2d). These included motif of ATOH1 which controls primary cilia formation that facilitates SHH-Triggered granule neuron progenitor proliferation [51] was evidenced to be cerebellum specific, and motif of MEF2C which controls cardiac morphogenesis [52] was enriched in heart.

Ubiquitously accessible chromatin peaks

Meanwhile, we also identified 8,734 peaks that consistently active in at least 90% of investigated tissues (Fig. 2e). The genes locate nearest to the conserved peaks were enriched for pathways like cytoskeleton and mitotic cell cycle that have housekeeping functions (Additional file 1: Fig. S2d). The top 5 enriched TF motifs (SP1, NFY, SP5, ELK4 and KLF3) (Supplementary Table 5) participate in chromatin remodeling and transcriptional activation (Fig. 2f), which are indispensable transcription factors in growth and development [53].

Tissue-specific gene expressions

We applied similar approach to investigate tissue specificity of gene expressions, and identified 135, 528, 439, 711, 228, 181, 117, 467, 40, 75, 469, 486, 94, 581, 124, 117, 411, 34, 336 and 1,590 tissue-specific genes were found to be specific in backfat, brain, bronchia, cerebellum, cervix, heart, hypothalamus, kidney, LD, leaf fat, liver, lung, medulla spinalis, ovary, parotid gland, pharynx, SI, SM, stomach and thymus, respectively (Additional file 1: Fig. S2b). For example, *MYH7*, a gene described to be responsible for hypertrophic cardiomyopathy, was highly expressed in heart than in other tissues (Additional file 1: Fig. S2c). The GO analysis showed these tissue-specific genes were associated with tissue-specific biological processes (Additional file 1: Fig. S2b). For example, genes significantly expressed in brain were related to the function of synaptic signaling, and genes significantly expressed in thymus were related to the function of T cell activation (Additional file 1: Fig. S2b).

Joint profiling of chromatin accessibility and gene expression

To investigate the regulatory elements that link to gene transcription, we examined correlations of gene expressions with intensity of ATAC peaks within 500Kb from the TSS of corresponding genes, as majority of promoter-anchored chromatin interactions were within 500 kb [54, 55]. The analysis was performed on the integrated dataset of 527,718 peaks and 17,634 genes from 71 samples, including 33 samples from public datasets (Supplementary Table 3, Additional file 1: fig. S3). We identified a total of 827,942 significant peak-gene associations (q-value < 0.01) that involves 255,166 peaks and 17,493 genes (Additional file 1: Fig. S4). Majority of the significant correlations (406,896/827,942, 49.15%) are positive only, while we also identified 10.54% negative correlations only. For example, a peak (chr5:10,663,155-10,664,871) has high intensity in liver showed positive correlation with TMPRSS6 (Spearman r2 = 0.77, q-value = 6.43e-12) (Fig. 3f), which encodes a type II transmembrane serine protease exclusively produced by liver [56]. We also exemplified negative peak – gene correlation where a peak (chr1:128,083,325-

128,085,034) show strong negative correlation with expression levels of PDIA3 (Spearman r^2 = - 0.60, q-value = 1.42e-6) (Fig. 3g). In addition, there are 40.31% peak – gene correlations showed both positive and negative correlations, indicating that the identity of regulatory elements was mutable.

On average, there were 2.87 genes per peak and 35.86 peaks per gene in positive correlations while 2.13 genes per peak and 14.49 peaks per gene in negative correlations (Fig. 3a, 3c). Interestingly, we observed that the strength of positive and negative correlations showed distinct distribution against peak-gene distances (fig. 3b). The positive correlations were enriched around the TSS of genes, while the negative correlations appeared to be underrepresented in TSS regions (fig. 3d). We identified 1,786 tissue-specific genes were associated with 38,574 tissue-specific peaks within 500 kb from the TSS regions (Fig. 3e, 3h and Fig. S4), suggesting important roles of tissue-specific chromatin accessibility in driving the tissue-specific genes expression thereby encoded tissue-specific biological function.

Open chromatin states of transposable elements in different pig tissues

Over 40% of the sequences in the non-coding region of the mammalian genome are TEs whose function were largely unexplored. To understand the function of TEs in pig tissues, we examined the chromatin accessibility of genome wide TEs. We found 26.91% of the 267,836 open chromatin regions were overlapped with at least one TE.

The spatial correlation analysis between TEs and open chromatin calculated by GenometriCorr [44] suggested underrepresentation of ratio of TEs located in the open chromatin regions than random expectation, indicating that TEs are tend to be epigenetically silenced. Despite this, each tissue contains an average of 17.96% peaks that were overlapped with at least one TE, range from 9.49% to 22.33% (Fig. 4a). The SINE, LINE and LTR accounted for approximately 40%, 30% and 23% of the accessible TEs, respectively, across the 20 tissues (Fig. 4b). We found over 50% of these TEs existed in a single tissue were located close to genes involved in tissue specific functions (Fig. 4c, 4d). For example, we observed enrichment of T cell receptor signaling pathway enriched in thymus while muscle system process enriched in LD (Fig. 4d). These results suggested that accessible TEs existed in one tissue exhibit strong tissue-specific regulatory roles, which was consistent with the previous reports [57, 58]. Meanwhile we observed approximate 0.19% of accessible TEs were commonly accessible in all 20 tissues (Fig. 4c). The conserved accessible TEs were more associated with housekeeping functions (e.g., ribosome assembly) (Fig. 4d).

Mounting evidence suggests that specific TEs in pigs, especially PERV, could potentially lead to immunodeficiency and tumorigenesis, making it difficult for xenotransplantation [59, 60]. Upon further inspection, we found a total of 10 unique PERVs overlapped with 13 peaks (Supplementary Table 6). For example, the PERV (sPERV-JX2like-12-23.0M) was overlapped with the peak (chr12:22,990,219-22,991,950) which showed high intensity in liver (Fig. 4e), significantly correlated multiple genes (NEUROD2, ENSSSCG00000017507, CACNB1, PLXDC1, PIP4K2B). Among them, PIP4K2B, which participates in 1-phosphatidylinositol-4-phosphate 5-kinase activity and IP6 [61], the product of 1-

phosphatidylinositol-4-phosphate 5-kinase, can facilitates assembly of the immature HIV-1 Gag lattice [62], shows highly correlation (Fig. 4f), suggesting potential role of the PERVs in facilitating virus infection.

Besides this, we also focused on the function of other genes regulated by PERV. For example, GTPBP8 [63], BCKDK [64], IDH3B [65] and IDH1 [66] were related to the function of mitochondrion; CD2BP2 [67]and FUS [68] involved in mRNA splicing; BOC [69] and FZD5 [70] participated in signal transduction pathway. Taken together, these results suggested that PERV which act as regulatory element involved in a variety of biological functions.

Endogenous retroviruses (ERVs) are LTR retrotransposons (class I of TEs), which originated from exogenous retroviruses that infected the germ line throughout evolution [71, 72], which could manifest their function across multiple organs in an animal, thereby could have fundamental functional impact. In this study, we identified a total of 162 TEs that are accessible in all investigated tissues. To trace the origin or evolutionary history of these ubiquitously accessible TEs, we calculated sequence similarity between 124 viral sequences (Supplementary Table 8) and 162 conserved accessible TEs sequences, and constructed phylogenetic trees to reconstruct evolutionary histories. Interestingly, we observed some conserved accessible TEs sequences were grouped with some viruses whose host was pig (Fig. 5), such as PBoV, TTSuVK2a, TTSuVK2b, PADV-A, PoBuV, PPV7 and Po-Circo-Like Virus 21 (Supplementary Table 9). We performed GO and KEGG enrichment analysis of all genes associated with conserved accessible TEs, and found that the biologic process of mitochondrial depolarization was highly enriched (Supplementary Table 10), which indicated that accessible TEs derived from viral sequences have a potential regulatory function in clearance of excess or impaired mitochondrial. Altogether, these evidences suggested that some special viral sequences have been integrated into the host genome and retained as regulatory elements that could play constitutive regulatory functions.

An annotation resource for regulatory mechanisms in complex traits

To demonstrate the utility of such resource in annotating candidate casual variants for complex traits, we examined overlap of 191 significant variants for fatty acid composition identified in a genome sequence based imputation GWAS [48] with the ATAC-Seq peak regions revealed in this study. A total of 22 variants were found to be coincident with the ATAC-Seq peaks (Supplementary Table 7). Among these, a lead SNP (chr2:9,736,686, P = 2.10 × 10-10) for C20:3n-6/C18:2n-6 was located in an ATAC-Seq peak (chr2:9,736,357-9,737,907) that showed high intensity in brain and moderate intensity in muscle (Fig. 6a). This peak showed significant correlation with three genes (FADS1, FADS2 and RAB3IL1) (Supplementary Table 7). Among them, FADS1 (q = 1.56 × 10-04) which exhibited strongest correlation (Fig. 6b), encodes delta-5 desaturase, one of the rate-limiting enzymes in the endogenous synthesis of polyunsaturated fatty acids (PUFAs) [73]. FADS2 which converts linoleate and alpha-linolenate into PUFAs is one of the key limiting enzymes in the lipid metabolic pathway [74] (Fig. 6c). In addition, several other genes related to fatty acid have been identified. Osteoglycin (OGN) is involved in matrix assembly, cellular growth, and migration. Previous study showed that OGN was associated with fat acids composition traits to some

extent [75] and regulated lipid differentiation through the Wnt/β-catenin signaling pathway [76]. TEAD3 which participates in the fatty acid, triacylglycerol, and ketone body metabolism pathway from the Reactome Pathways dataset[77] is a member of the transcriptional enhancer factor (TEF) family of transcription factors. Further experiment is needed to validate effect of the variants on the accessibility of the peak, and in turn the expression of relevant genes and fatty acid composition phenotypes.

Discussion

Regulatory elements are crucial for gene expression regulation, which involved in the initiation and regulation of transcription in different tissues and organs and provided precise control of genes and restrict gene expression to certain cells and tissues throughout life [78]. Activation of regulatory elements needs 'pioneer' factor opening the compacted chromatin to recruit transcription coactivators and chromatin remodeling proteins [79, 80]. Regulatory elements usually contain multiple transcription factors binding sites (TFBS) which recognized by TFs. Genetic variations in regulatory elements affect gene expression usually through disrupting TFBS [81]. Therefore, the generation of multi-tissues atlas of open chromatin enabled exploration of gene expression regulation and non-coding genetic variations.

In this study, we performed ATAC-seq and RNA-seq across 20 tissues in pig at the same development stage to characterized chromatin state landscape, uncovering extensive tissue-specific regulation of gene expression. However, not a single study until now, as far as we are aware, has comprehensively and systematically investigated chromatin accessibility in multi-tissues by ATAC-seq and conjunction with transcriptomic analysis. In our study, we observed that the patterns of TSS proximal regulatory elements are more similar across tissues while distal regulatory elements were more tissue-specific, reflecting the architecture and function of distal regulatory elements have strong tissue specificity [82].

Consistent with the previous reports that regulatory elements mediate specific differentiation of tissues or cells [83, 84], we identified a large number of tissue-specific open chromatin regions which regulate the genes with function matching the biology of corresponding tissues. In addition, we identified tissue-specific TF motifs based on the enrichment of their cognate binding motifs, such as ATOH1 enriched in cerebellum and FOXA1 enriched in liver. And we acknowledge that TF motifs enrichment analysis needs to verify using other approaches such as ChIP-seq for specific TFs.

Connecting peaks to their target genes remains challenge given that Hi-C data is not available for the samples under study, particularly for distal peaks which are far from target genes, and regarding genes nearest to peaks as targets is error prone [85–87]. To solve this problem, we combine other published data to reveal the association between peak intensity and gene expression, which has been proved effective [83]. We observed significant negative peak-gene pairs showed strikingly different peak-gene distances compared to that in positive correlation pairs, which suggested that positive and negative correlation pairs may have different regulatory mechanism. To further understand tissue-specific regulatory relationships between peak intensity and gene expression, we filtered for tissue-specific correlation pairs (Additional file 1: Fig. S4). Our results suggest tissue-specific gene expression may occur

through the activation of tissue-specific regulatory elements. Whereas, the above predictions require further support using 3D chromosome conformation data or validated through CRISPRi experiments.

To deepen our understanding of regulatory elements, here we performed profiles in TEs as they can provide binding sites for TFs and behave as alternative promoters and enhancers [88]. In this study, we observed a small proportion of accessible TEs, which were in good accordance with the classical polarization theory that TEs promote genetic innovation but also threaten genome stability [89]. Similar to previous study [45], we observed most of accessible TEs only present in one tissue. And tissue-specific accessible TEs were enriched in tissue-specific biological pathways, which support the findings of previous studies [45, 90–92]. In order to trace the potential source of accessible TE, we performed multiple sequence alignments between the viral sequences and the conserved accessible TEs sequences, and found some candidate viral sequences (PBoV, TTSuVK2a, TTSuVK2b, PADV-A, PoBuV, PPV7 and Po-Circo-Like Virus 21) might be the source of conserved accessible TEs.

The annotation of regulatory elements has proven highly effective for the identification of candidate causative variants and screening candidate genes for complex traits [93–95]. Similarly, our results demonstrated that variants of fatty acid were enriched in active regulatory elements annotated by this study. Specifically, we speculate that a potential causative SNP (chr2:9,736,686, $P = 2.10 \times 10–10$) that was associated with C20:3n-6/C18:2n-6 and found within a tissue-shared promoter, may regulate the expression of FADS1 which encodes the Δ5 desaturase enzyme - one of the rate-limiting enzymes in the endogenous synthesis of polyunsaturated fatty acids (PUFAs) [96]. Taken together, our data unveiled the chromatin landscapes for studying the regulation of gene expression, provided potential sources of chromatin accessibility and deconstructed underlying mechanisms of complex traits in pig. Further investigations with additional complementary data - such as single cell and different development stage data - are warranted to fully dissect biological mechanisms of gene expression.

Conclusions

We performed an integrated analysis of transcriptome sequencing and transposase-accessible chromatin with high throughput sequencing and provided a comprehensive understanding of tissue-specific accessible chromatin regions and tissue-specific transcription factor motifs. Then, we constructed the relationship between regulatory elements and their target genes. This large regulatory network will serve as a foundation for understanding the precise regulatory mechanisms in different tissues of pig. We also analyzed accessible TEs which act as candidate regulatory elements. In addition, we analyzed the sequence characteristics of regulatory elements, and found some virus sequences similar to regulatory element sequences, which provided a certain basis for the essential characteristics of regulatory elements. Finally, the regulatory elements identified in this paper were used to annotate the GWAS loci that affect fatty acid traits, providing a reference for genetic analysis of complex shapes.

Abbreviations

TF: transcription factor

ATAC-seq: transposase-accessible chromatin with high throughput sequencing

DNase-seq: DNase I hypersensitive site sequencing

ERVs: Endogenous retroviruses

FRIP: Fraction of reads in peaks

LD: longissimus muscle

MNase-seq: micrococcal nuclease sequencing

NFR: nucleosome-free regions

NSC: Normalized strand cross-correlation coefficent

OGN: Osteoglycin

PADV-A: Porcine mastadenovirus A

PBC1: PCR Bottlenecking Coefficient 1

PBC2: PCR Bottlenecking Coefficient 2

PBoV: Parvoviridae Bocaparvovirus Porcine bocavirus

PoBuV: Protoparvovirus Zsana/2013/HUN

PoCv21: Po-Circo-Like Virus 21

PPV7: Porcine parvovirus 7

PUFAs: polyunsaturated fatty acids

RSC: Relative strand cross-correlation coefficient

SI: small intestine

SM: semimembranosus

TE: transposable elements

TEF: transcriptional enhancer factor

TFBS: transcription factors binding sites

TPM: transcripts per million

TSS: transcription start sites

TTSuVK2a: Torque teno sus virus k2a

TTSuVK2b: Torque teno sus virus k2b

Declarations

Availability of data and materials

The data that support the findings of this study are available on request from the corresponding author.

Acknowledgements

We are grateful to colleagues in State Key Laboratory of Pig Genetic Improvement and Production Technology, Jiangxi Agricultural University for sample collection.

Funding

This work was supported by the National Natural Science Foundation of China (32160781).

Author information

Tao Jiang and Ziqi Ling contributed equally to this work.

Bin Yang, Jianzhen Huang and Lusheng Huang contributed equally to this work.

Affiliations

State Key Laboratory of Pig Genetic Improvement and Production Technology, Jiangxi Agricultural University, Nanchang 330045, China.

Tao Jiang, Ziqi Ling, Zhimin Zhou, Xiaoyun Chen, Liqing Chen, Sha Liu, Yingchun Sun, Jiawen Yang, Bin Yang**,** Jianzhen Huang, Lusheng Huang

Author Contributions

LSH designed this study, wrote and edited the manuscript; JZH and BY analysed the data and wrote the manuscript; TJ performed the experiments, analysed the data and wrote the manuscript; ZQL performed the experiment and wrote the manuscript; ZMZ, XYC, LQC, SL, YCS and JWY assisted with experiments and data analysis.

Corresponding authors

Correspondence to Bin Yang, Jianzhen Huang or Lusheng Huang.

Ethics declarations

All of the procedures involving animals are in compliance with the care and use guidelines of experimental animals established by the Ministry of Science and Technology of China. The Ethics Committee of Jiangxi Agricultural University approved this study.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Simpson RT. Nucleosome positioning can affect the function of a cis-acting DMAelementin vivo. Nature. 1990;343(6256):387–9.
2. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K. Dynamic regulation of nucleosome positioning in the human genome. Cell. 2008;132(5):887–98.
3. Kouzarides T. Chromatin modifications and their function. Cell. 2007; 128.
4. Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD. Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat. Genet. 36, 900–905. Nature Genetics. 2004; 36(8): 900–905.
5. Ozsolak F, Song JS, Liu XS, Fisher DE. High-throughput mapping of the chromatin structure of human promoters. Nat Biotechnol. 2007;25(2):244.
6. Nathan S, Terrence F. Identifying and Characterizing Regulatory Sequences in the Human Genome with Chromatin Accessibility Assays. Genes. 2012; 3(4).
7. A comparative encyclopedia. of DNA elements in the mouse genome. Nature. 2014;515(7527):355–64.
8. Thurman RE, Rynes E, Humbert R, Vierstra J, Stamatoyannopoulos JA. The accessible chromatin landscape of the human genome. Nature. 2012;489(7414):75–82.
9. Rivera CM, Bing R. Mapping Human Epigenomes. Cell. 2013;155(1):39–55.
10. Lettice LA, Williamson I, Devenney PS, Kilanowski F, Dorin J, Hill RE. Development of five digits is controlled by a bipartite long-range cis-regulator. Development. 2014;141(8):1715–25.
11. Reddi PP, Urekar CJ, Abhyankar MM, Ranpura SA. Role of an insulator in testis-specific gene transcription. 1120: Annals of the New York Academy of Sciences; 2010.

12. Smith RP, Eckalbar WL, Morrissey KM, Luizon MR, Hoffmann TJ, Sun X, Jones SL, Aldred SF, Ramamoorthy A, Desta Z. Genome-Wide Discovery of Drug-Dependent Human Liver Regulatory Elements. PLoS Genetics. 2014; 10(10).
13. Wang D, Chen H, Momary KM, Cavallari LH, Sadée W. Regulatory polymorphism in vitamin K epoxide reductase complex subunit 1 (VKORC1) affects gene expression and warfarin dose requirement. Blood. 2008;112(4):1013–21.
14. Smemo S, Tena JJ, Kim KH, Gamazon ER, Nóbrega M. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature. 2014;507(7492):371–5.
15. Kapoor A, Sekar RB, Hansen NF, Fox-Talbot K, Morley M, Pihur V, Chatterjee S, Brandimarto J, Moravec CS, Pulit SL. An Enhancer Polymorphism at the Cardiomyocyte Intercalated Disc Protein NOS1AP Locus Is a Major Regulator of the QT Interval. Am J Hum Genet. 2014;94(6):854–69.
16. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, Satpathy AT, Rubin AJ, Montine KS, Wu B. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods. 2017;14(10):959.
17. Markenscoff-Papadimitriou E, Whalen S, Przytycki P, Thomas R, Rubenstein JL. A Chromatin Accessibility Atlas of the Developing Human Telencephalon. Cell. 2020; 182(3).
18. Wu J, Huang B, Chen H, Yin Q, Liu Y, Xiang Y, Zhang B, Liu B, Wang Q, Xia W, Li W, Li Y, Ma J, Peng X, Zheng H, Ming J, Zhang W, Zhang J, Tian G, Xu F, Chang Z, Na J, Yang X, Xie W. The landscape of accessible chromatin in mammalian preimplantation embryos. Nature. 2016;534(7609):652–7.
19. Gorkin DU, Barozzi I, Zhao Y, Zhang Y, Huang H, Lee AY, Li B, Chiou J, Wildberg A, Ding B, Zhang B, Wang M, Strattan JS, Davidson JM, Qiu Y, Afzal V, Akiyama JA, Plajzer-Frick I, Novak CS, Kato M, Garvin TH, Pham QT, Harrington AN, Mannion BJ, Lee EA, Fukuda-Yuzawa Y, He Y, Preissl S, Chee S, Han JY, Williams BA, Trout D, Amrhein H, Yang H, Cherry JM, Wang W, Gaulton K, Ecker JR, Shen Y, Dickel DE, Visel A, Pennacchio LA, Ren B. An atlas of dynamic chromatin landscapes in mouse fetal development. Nature. 2020;583(7818):744–51.
20. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, Silva TC, Groeneveld C, Wong CK, Cho SW, Satpathy AT, Mumbach MR, Hoadley KA, Robertson AG, Sheffield NC, Felau I, Castro MAA, Berman BP, Staudt LM, Zenklusen JC, Laird PW, Curtis C, Cancer Genome Atlas Analysis N, Greenleaf WJ, Chang HY. The chromatin accessibility landscape of primary human cancers. Science. 2018; 362(6413).
21. Consortium T. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004;306(5696):636–40.
22. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Al E. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.
23. Coordinated international action. to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biol. 2015;16(1):57.
24. Pan Z, Yao Y, Yin H, Cai Z, Wang Y, Bai L, Kern C, Halstead M, Chanthavixay G, Trakooljul N, Wimmers K, Sahana G, Su G, Lund MS, Fredholm M, Karlskov-Mortensen P, Ernst CW, Ross P, Tuggle CK, Fang L,

Zhou H. Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. Nat Commun. 2021;12(1):5848.

25. Yue J, Hou X, Liu X, Wang L, Zhang L. The landscape of chromatin accessibility in skeletal muscle during embryonic development in pigs. J Anim Sci Biotechnol. 2021;12(1):56.
26. Kern C, Wang Y, Xu X, Pan Z, Halstead M, Chanthavixay G, Saelao P, Waters S, Xiang R, Chamberlain A, Korf I, Delany ME, Cheng HH, Medrano JF, Van Eenennaam AL, Tuggle CK, Ernst C, Flicek P, Quon G, Ross P, Zhou H. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. Nat Commun. 2021;12(1):1821.
27. Kunarso G, Chia NY, Jeyakani J, Hwang C, Bourque G. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010;42(7):631–4.
28. Jumpei I, Ryota S, Hirofumi N, Shiro Y, Tetsuaki K, Takahide H, Ituro I, Cédric F. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. PLoS Genet. 2017;13(7):e1006883.
29. Jang HS, Shah NM, Du AY, Dailey ZZ, Pehrsson EC, Godoy PM, Zhang D, Li D, Xing X, Kim S. Transposable elements drive widespread expression of oncogenes in human cancers. Nature Genetics. 2019.
30. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Current protocols in molecular biology / edited by Frederick M Ausubel [et al]. 2015; 109(21 29): 21.29.21-29.
31. Langmead B, Salzberg SL. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359. Nature Methods. 2012; 9(4): 357–359.
32. Li H. The Sequence Alignment/Map format and SAMtools. BIOINFORMATICS -OXFORD-. 2009.
33. Zhang Y. Model-based analysis of ChIP-Seq (MACS). 2008.
34. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Current Protocols in Bioinformatics. 2014; 47(1).
35. Valencia A. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2014.
36. 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. 2015;33(3):290–5.
37. Yang L, Smyth GK, Wei S. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. (7): 923–930.
38. Gabriela B, Bernhard M, Hubert H, Pornpimol C, Marie T, Amos K, Wolf-Herman F, Franck P, Zlatko T, Jérôme G. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25(8):1091–3.
39. Sven. Heinz, and, Christopher, Benner, and, Nathanael, Spann, and, Eric. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities - ScienceDirect. Mol Cell. 2010;38(4):576–89.

40. Kern C, Wang Y, Xu X, Pan Z, Zhou H. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. Nature Communications.
41. Zhao Y, Hou Y, Xu Y, Luan Y, Zhou H, Qi X, Hu M, Wang D, Wang Z, Fu Y, Li J, Zhang S, Chen J, Han J, Li X, Zhao S. A compendium and comparative epigenomics analysis of cis-regulatory elements in the pig genome. Nat Commun. 2021;12(1):2217.
42. Yalan Y, Xinhao F, Junyu Y, Muya C, Min Z, Yijie T, Siyuan L, Zhonglin T. A comprehensive epigenome atlas reveals DNA methylation regulating skeletal muscle development. Nucleic Acids Research.
43. Tarailo-Graovac M, Chen N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Current Protocols in Bioinformatics. 2009; 25(1).
44. Favorov A, Mularoni L, Cope LM, Medvedeva Y, Mironov AA, Makeev VJ, Wheelan SJ. Exploring Massive, Genome Scale Datasets with the GenometriCorr Package. PLoS Comput Biol. 2012;8(5):1245–53.
45. Miao B, Fu S, Cheng L, Gontarz P, Zhang B. Tissue-specific usage of transposable element-derived promoters in mouse development. Genome Biol. 2020;21(1):255.
46. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform (describes the FFT-NS-1, FFT-NS-2 and FFT-NS-i strategies). 2002.
47. Price*† MN. FastTree. Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Molecular Biology and Evolution. 2009; 26(7).
48. Junjie Z, Yifeng, Huanfa, Gong, Leilei, Cui nwuJ, Congying. Chen. Landscape of Loci and Candidate Genes for Muscle Fatty Acid Composition in Pigs Revealed by Multiple Population Association Analysis. Front Genet. 2019;10:1067–7.
49. Ko JY, Oh S, Yoo KH. Functional Enhancers As Master Regulators of Tissue-Specific Gene Regulation and Cancer Development. Molecules & Cells. 2017;40(3):169–77.
50. Iyer R, Jenkinson CP, Vockley JG, Kern RM, Grody WW, Cederbaum S. The human arginases and arginase deficiency. J Inherit Metab Dis. 1998;21(1):86–100.
51. Chang CH, Zanini M, Shirvani H, Cheng JS, Yu H, Feng CH, Mercier AL, Hung SY, Forget A, Wang CH. Atoh1 Controls Primary Cilia Formation to Allow for SHH-Triggered Granule Neuron Progenitor Proliferation. Dev Cell. 2019;48(2):184–99.
52. Stone NR, Gifford CA, Thomas R, Pratt K, Srivastava D. Context-Specific Transcription Factor Functions Regulate Epigenomic and Transcriptional Dynamics during Cardiac Reprogramming. Cell Stem Cell. 2019;25(1):87–102.e109.
53. Halstead MM, Ma X, Zhou C, Schultz RM, Ross PJ. Chromatin remodeling in bovine embryos indicates species-specific regulation of genome activation. Nat Commun. 2020;11(1):4654.
54. Wu Y, Qi T, Wang H, Zhang F, Zheng Z, Phillips-Cremins JE, Deary IJ, McRae AF, Wray NR, Zeng J, Yang J. Promoter-anchored chromatin interactions predicted from genetic analysis of epigenomic data. Nat Commun. 2020; 11(1): 2061.

55. Javierre B, Burren O, Wilder S, Kreuzhuber R, Hill S, Sewitz S, Cairns J, Wingett S, Várnai C, Thiecke M. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016;167(5):1369–84.e1319.
56. Finberg KE, Heeney MM, Campagna DR, Ayd?Nok Y, Pearson HA, Hartman KR, Mayo MM, Samuel SM, Strouse JJ, Markianos K. Mutations in TMPRSS6 cause iron-refractory iron deficiency anemia (IRIDA). Nat Genet. 2008;40(5):569–71.
57. Evolutionary rate of. human tissue-specific genes are related with transposable element insertions. Genetica. 2012;140(10–12):513.
58. Marco T, Aurélie K, Brown CD. Transposable elements generate regulatory novelty in a tissue-specific fashion. BMC Genomics. 2018;19(1):468.
59. Bendinelli M, Matteucci D, Friedman H. Retrovirus-Induced Acquired Immunodeficiencies. Adv Cancer Res. 1985;45(6):125.
60. Lauren J, Wegman-Points, Melissa LT, Teoh-Fitzgerald. Gaowei, Mao, Yueming, Zhu. Retroviral-infection increases tumorigenic potential of MDA-MB-231 breast carcinoma cells by expanding an aldehyde dehydrogenase (ALDH1) positive stem-cell like population - ScienceDirect. Redox Biol. 2014;2(1):847–54.
61. John D, York. Regulation of nuclear processes by inositol polyphosphates. Biochimica Et Biophysica Acta Molecular & Cell Biology of Lipids; 2006.
62. Dick RA, Zadrozny KK, Xu C, Schur FKM, Lyddon TD, Ricana CL, Wagner JM, Perilla JR, Ganser-Pornillos BK, Johnson MC. Inositol phosphates are assembly co-factors for HIV-1. Nature. 2018;560(7719):509–12.
63. Verma Y, Mehra U, Pandey DK, Kar J, Datta K. MRX8, the conserved mitochondrial YihA GTPase family member is required for de novo Cox1 synthesis at suboptimal temperatures in Saccharomyces cerevisiae. Molecular Biology of the Cell. 2021: mbc.E20-07-0457.
64. Sato Y, Tate H, Yoshizawa F, Sato Y. Data on the proliferation and differentiation of C2C12 myoblast treated with branched-chain ketoacid dehydrogenase kinase inhibitor. Data in Brief. 2020: 105766.
65. Duncan DM, Kiefel P, Duncan I. Mutants for Drosophila Isocitrate Dehydrogenase 3b Are Defective in Mitochondrial Function and Larval Cell Death. G3-Genes Genomes Genetics. 2017; 7(3): g3.116.037366.
66. Gonzalez-Menendez P, Romano M, Yan H, Oburoglu L, Tardito S, Daumur M, Dume AS, Phadke I, Qu X, Bories PN. An IDH1-Vitamin C Crosstalk DrivesHuman Erythroid Development ByInhibiting Pro-Oxidant Mitochondrial Metabolism. Social Science Electronic Publishing.
67. Kofler M, Heuer K, Zech T, Freund C. Recognition Sequences for the GYF Domain Reveal a Possible Spliceosomal Function of CD2BP2. J Biol Chem. 2004;279(27):28292–7.
68. Orozco D, Edbauer D. FUS-mediated alternative splicing in the nervous system: consequences for ALS and FTLD. J Mol Med. 2013;91(12):1343–54.
69. Echevarría-Andino M, Allen BL. The Hedgehog Co-Receptor BOC Differentially Regulates SHH Signaling During Craniofacial Development. 2020.

70. Katoh M, Katoh M. Molecular genetics and targeted therapy of WNT-related human diseases (Review). International Journal of Molecular Medicine. 2017.
71. Marta G-M, Tara Doucet-O'Hare. Lisa, Henderson, Avindra, Nath. Human endogenous retrovirus-K (HML-2): a comprehensive review. Critical reviews in microbiology. 2018.
72. Dixie L, Mager, Jonathan P, Stoye. Mammalian Endogenous Retroviruses. Microbiology spectrum. 2015.
73. Je L, Hyungjae L, Kang SB, Woo P. Fatty Acid Desaturases, Polyunsaturated Fatty Acid Regulation, and Biotechnological Advances. Nutrients. 2016;8(1):23.
74. Chen-Xi XU, Wang MQ, Zhu XR, Zhang YF, Xia HL, Liu XH, Wang XL, Zhang HM, Yang ZP, Mao YJ. Effects of SNPs in the 3' Untranslated Regions of FADS2 on the Composition of Fatty Acids in Milk of Chinese Holstein. Scientia Agricultura Sinica.
75. Mroz T, Pryszcz L, Skarzynska A, Kielkowska A, Havey M, Bartoszewski G. RNA-seq reveals differentially expressed genes in cucumber MSC lines possessing mitochondrial DNA rearrangements. 2013.
76. Giacomo D, Valentina P, Lichtenberger MB. Ken, Natsuga, Rodney, Sinclair. Epidermal Wnt/β-catenin signaling regulates adipocyte differentiation via secretion of adipogenic factors. Proceedings of the National Academy of Sciences of the United States of America. 2014.
77. Rouillard AD, Gundersen GW, Fernandez NF, Zichen W, Monteiro CD, Mcdermott MG, Avi MA. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database. 2016: baw100.
78. Hernandez-Garcia CM, Finer JJ. Identification and validation of promoters and cis-acting regulatory elements. Plant ence. 2014;217–218:109–19.
79. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009;461(7261):199–205.
80. Zhu F, Farnung L, Kaasinen E, Sahu B, Taipale J. The interaction landscape between transcription factors and the nucleosome. 2017.
81. Functional genomics. reveal gene regulatory mechanisms underlying schizophrenia risk. Nature communications. 2019.
82. Ong CT, Corces VG. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet. 2011;12(4):283–93.
83. Zhu Y, Zhou Z, Huang T, Zhang Z, Huang L. The swine spatiotemporal H3K27ac spectrum provides novel resources for exploring gene regulation related to complex traits and fundamental biological process. 2021.
84. Smith AD, Sumazin P, Zhang MQ. Tissue-specific regulatory elements in mammalian promoters. Molecular Systems Biology. 2007; 3(1).
85. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA. Mapping long-range promoter contacts in human cells with high-resolution capture

Hi-C. Nat Genet. 2015;47(6):598–606.
86. The pluripotent regulatory. circuitry connecting promoters to their long-range interacting elements. Genome Research. 2015; 25(4).
87. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489(7414):109–13.
88. Pehrsson EC, Chou D, Hary M, Sundaram V, Wang T. The epigenomic landscape of transposable elements across normal human development and anatomy. Nature Communications. 2019.
89. Senft AD, Macfarlan TS. Transposable elements shape the evolution of mammalian development. Nature Reviews Genetics. 2021: 1–21.
90. Pierre-tienne. Jacques, Justin, Jeyakani, Guillaume, Bourque. The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements. PLoS Genet. 2013;9(5):1–12.
91. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 2014;24(12):1963–76.
92. Hutchins AP, Pei D. Transposable elements at the center of the crossroads between embryogenesis, embryonic stem cells, reprogramming, and long non-coding RNAs. Science Bulletin. 2015.
93. Xiang R, Berg IV, Macleod IM, Hayes B, Goddard ME. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. 2019.
94. Fang L, Liu S, Liu M, Kang X, Li CJ. Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations. BMC Biology. 2019; 17(1).
95. Huo Y, Li S, Liu J, Li X, Luo XJ. Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. Nature Communications. 2019; 10(1).
96. Zhao R, Tian L, Zhao B, Sun Y, Liu M. FADS1 promotes the progression of laryngeal squamous cell carcinoma through activating AKT/mTOR signaling. Cell Death & Disease. 2020; 11(4).

Figures

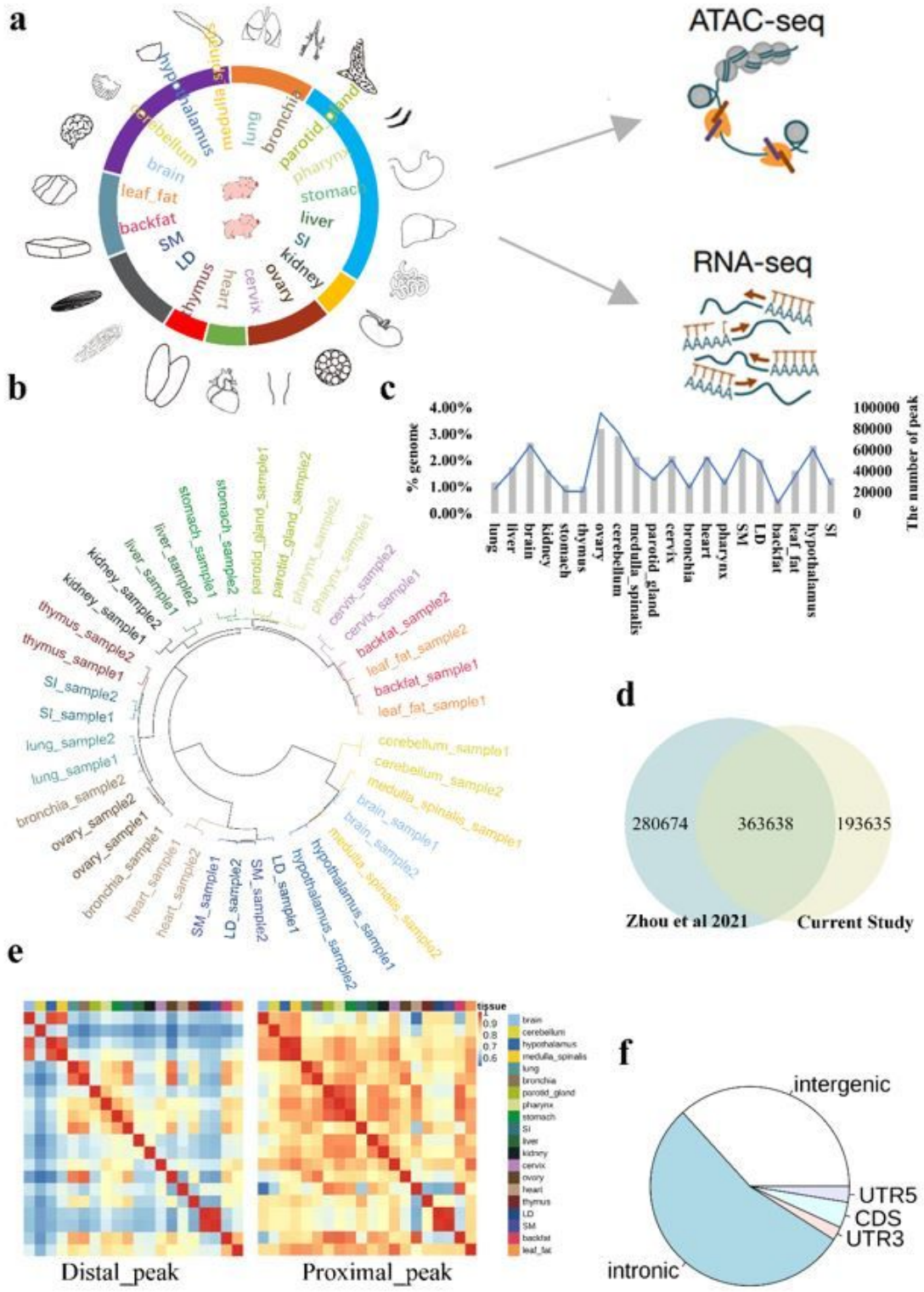

Figure 1

Identifying the landscape of chromatin accessibility via ATAC-seq in pigs.

a Schematic diagram showing tissues assayed in this study. The ring colors represent different organ systems. Starting with the location of the brain tissue, clockwise around the circle, corresponding to the nervous system (brain, cerebellum, hypothalamus, medulla spinalis), respiratory system (lung, bronchia),

digestive system (parotid gland, pharynx, stomach, liver, small intestine (SI)), urinary system (kidney), reproductive system (ovary, cervix), circulatory system (heart), lymphatic system (thymus), locomotor system (longissimus muscle (LD), semimembranosus (SM)), endocrine system (backfat and leaf fat). **b** Hierarchical clustering of samples using ATAC-seq signal intensity. **c** Distribution of the peak number and genome coverage in each tissue. **d** Venn plot showing the overlap of peaks identified in this study with those reported in Zhou et al. 2021. **e** Pearson correlation heatmaps of 20 tissues by density of ATAC-seq distal peaks (left) and ATAC-seq proximal peaks (right). **f** Pie chart represents the proportion of peaks located in introns, intergenic, 5'UTR, 3'UTR and CDS regions.

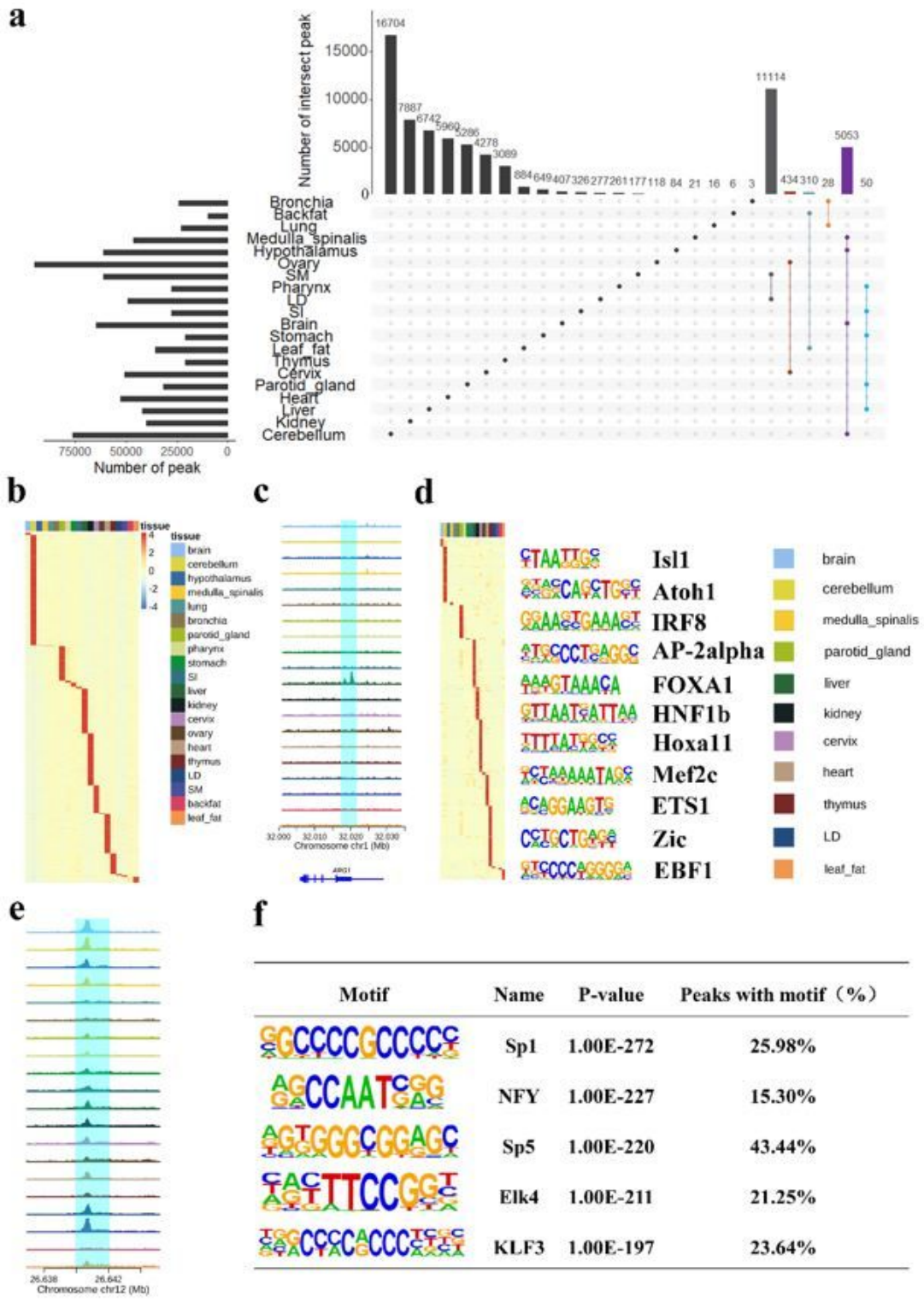

Figure 2

Functional annotation of system-specific peaks. a Intervene plot showing the number of system-specific peaks and tissue-specific peaks. **b** Heatmap of tissue-specific peaks in 20 tissues. Each row of the heatmap shows normalized density of read-depth at one peak**. c** Representative examples of liver specific peaks which located close to ARG1(A protein coding gene expressed primarily in the liver and involved in the urea cycle). **d** Enrichment of tissue specific transcription factor motifs in each tissue (left). The

columns represent 20 tissues. The rows represent motifs. The P values were generated by HOMER. Representative examples of tissue specific transcription factor motifs were displayed (right). **e** Representative examples of tissue conserved peak. **f** Top 5 transcription factors with binding motifs enriched in tissue conserved peaks.

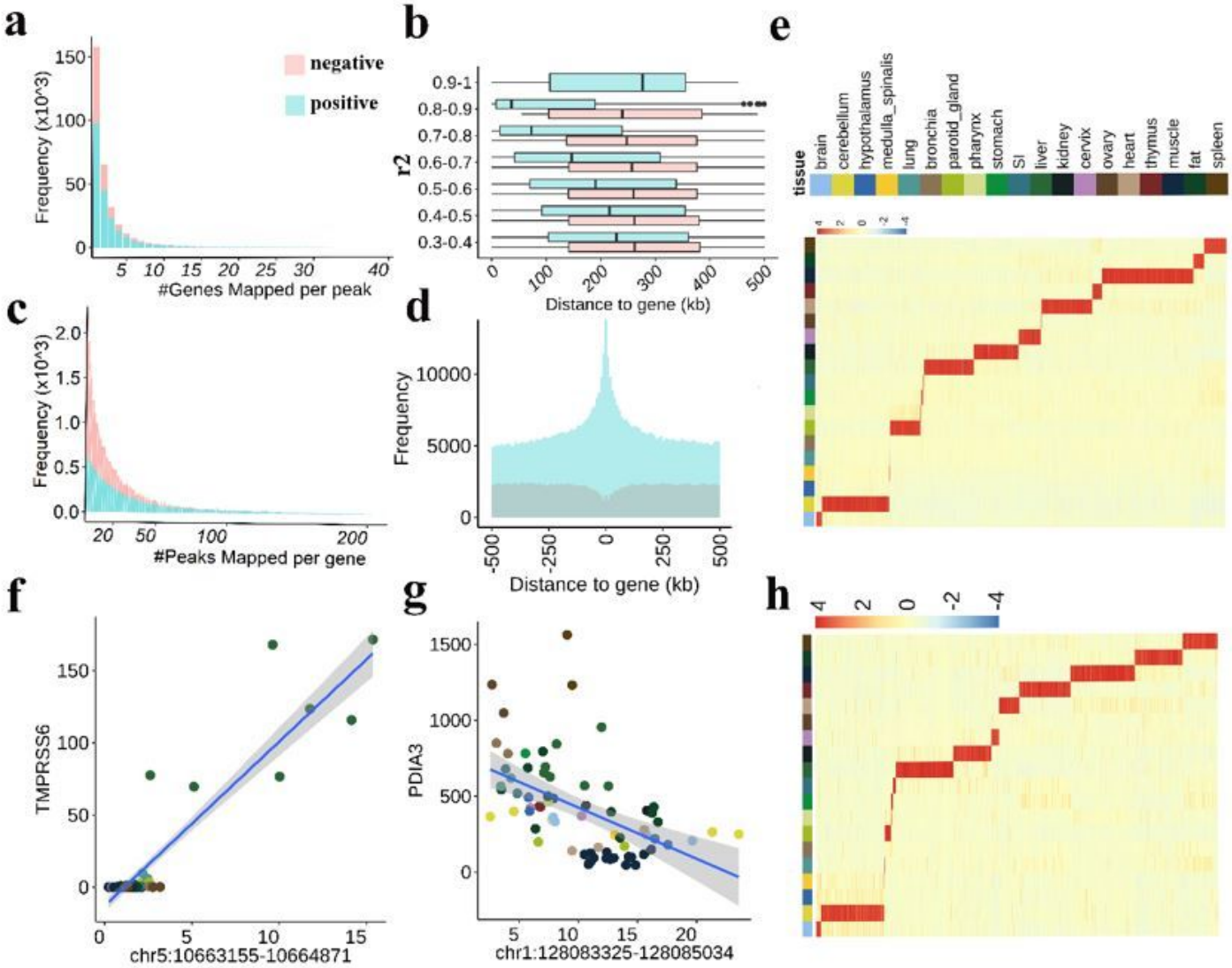

Figure 3

Characterizing the correlations between peak intensity and gene expression of nearby genes. a, c The number of genes regulated by one peak, and the number of peaks regulated one gene. The X-axis of the histogram represents different groups that were classified according to the associated gene(peak) numbers per peak(gene), and the Y-axis represents the gene(peak) count for each group. Positive peak-gene pairs and negative peak-gene pairs are distinguished by different colors. **b** Barplot of the relationship between the correlation and distance of peak-gene pairs. **d** Distribution of the distance for significant positive peak-gene pairs and negative peak-gene pairs. **e, h** Heatmap of tissue-specific peaks (genes) from all significant peak-gene pairs. **f** Representative examples of a significant tissue-specific

peak-gene pair which peak and gene are both tissue-specific. The dots represent peak intensity and gene expression from each sample. Different colors indicate different tissue. **g** Representative examples of a significant conserved peak-gene pair. The peak and gene are both expressed ubiquitously across tissues. The dots represent peak intensity and gene expression from each sample. Different colors indicate different tissue.

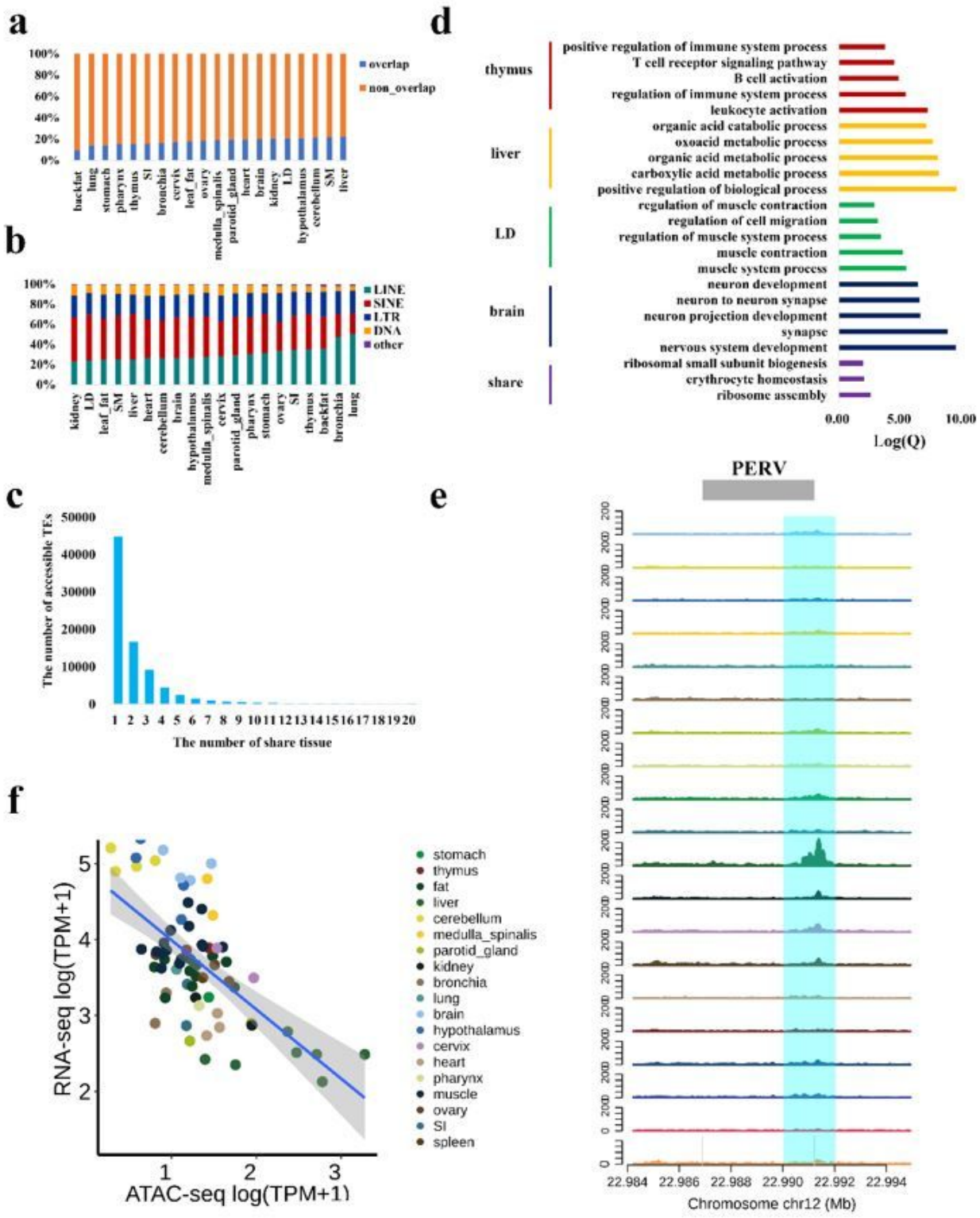

Figure 4

Accessible transposable elements (TEs) in 20 pig tissues. a Distribution of ATAC-seq peaks identified in 20 pig tissues associated with TEs. **b** TE class distribution of accessible TEs in 20 pig tissues. **c** Distribution of accessible TEs among 20 pig tissues. **d** Functional enrichment of genes for 4 tissue-specific and tissue-conserved accessible TEs based on gene ontology (GO) biological processes. **e** Comparison the tracks of peak activity across all tissues; Y axis corresponding to the normalized read depths of ATAC-seq data. **f** Scatter plot showing the correlation between the peak at chr12:22,990,219-22,991,950 with expression of PIP4K2B gene.

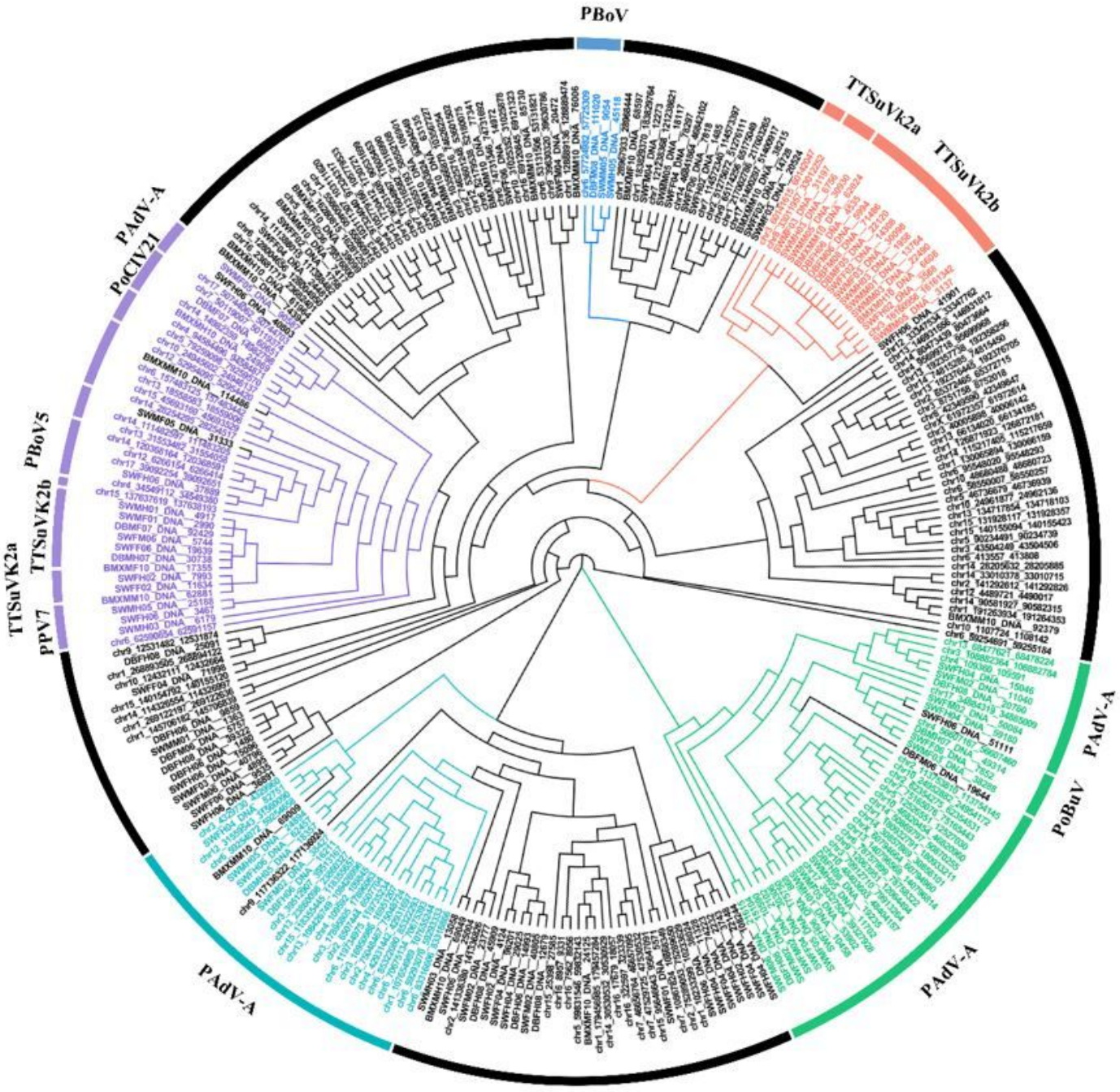

Figure 5

Phylogram of 124 viruses and 164 accessible TEs sequences. Color-highlighted clusters indicate pig virus sequences. PBoV: porcine bocavirus; TTSuVK2a(b): Torque teno sus virus k2a(b); PADV-A: porcine mastadenovirus; PoBuV: Protoparvovirus Zsana/2013/HUN; PPV7: Porcine parvovirus 7; PoCIV21: Po-Circo-like virus 21.

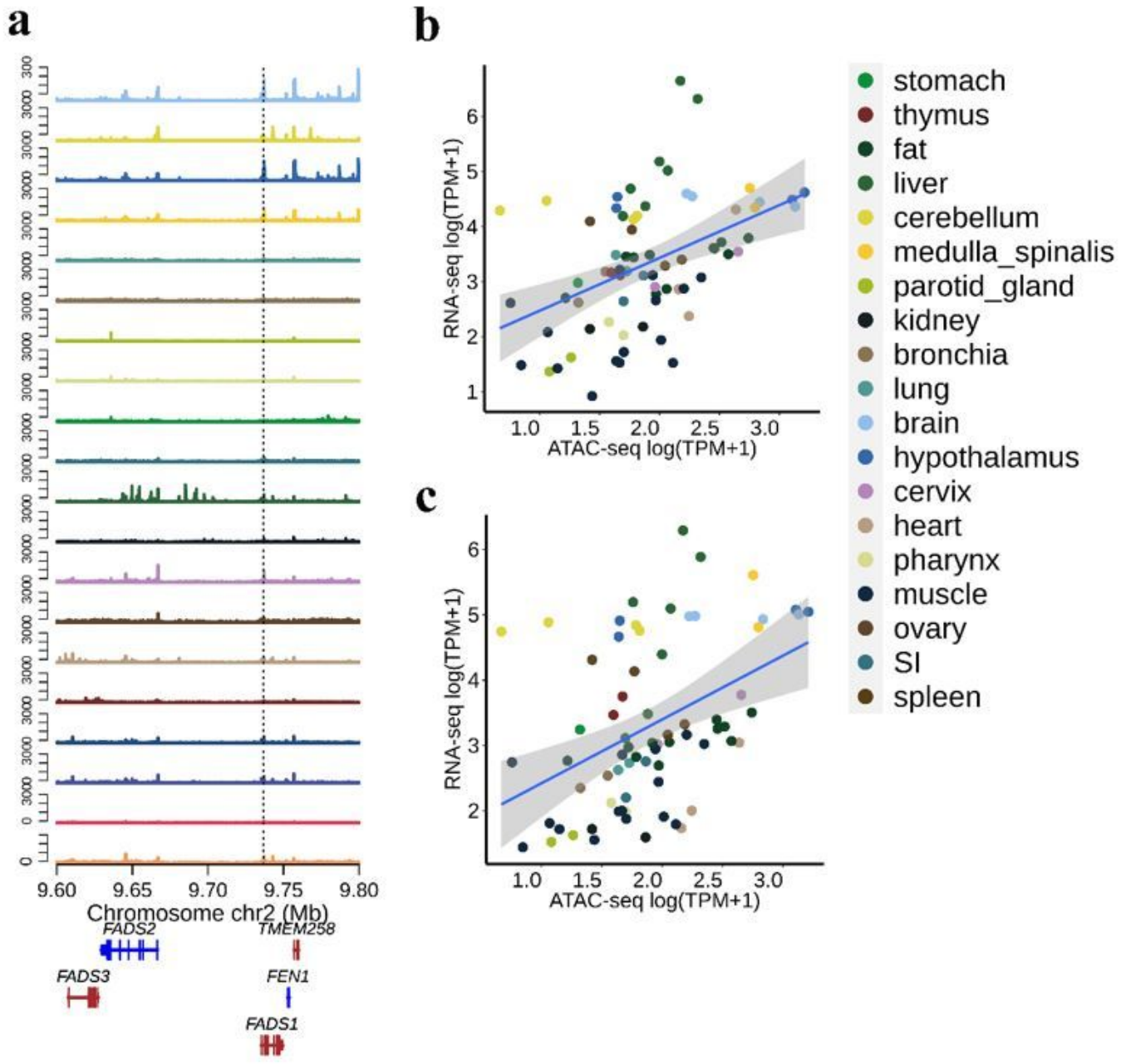

Figure 6

Overlap with GWAS SNPs. a The GWAS signal for fatty acid located within a peak in intergenic region. **b** Scatter plot showing the correlation between the peak at chr2:9,736,357-9,737,907 with expression of FADS1 gene. **c** Scatter plot showing the correlation between the peak at chr2:9,736,357-9,737,907 with expression of FADS2 gene.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile2.xlsx