Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Pig genome functional annotation enhances biological interpretations of complex traits and comparative epigenomics

Huaijun Zhou (hzhou@ucdavis.edu)
University of California, Davis https://orcid.org/0000-0001-6023-9521

Zhangyuan Pan
Department of Animal Science, University of California, Davis

Yuelin Yao
The University of Edinburgh

Hongwei Ying
Agricultural Genome Institute at Shenzhen, Chinese Academy of Agricultural Sciences

Zexi Cai
Center for Quantitative Genetics and Genomics, Faculty of Technical Sciences, Aarhus University
https://orcid.org/0000-0002-9579-3415

Ying Wang
UC Davis

Lijing Bai
Agricultural Genome Institute at Shenzhen, Chinese Academy of Agricultural Sciences

Colin Kern
University of California, Davis https://orcid.org/0000-0002-3249-6405

Michelle Halstead
University of California, Davis https://orcid.org/0000-0003-0168-2704

Kelly Chanthavixay
Department of Animal Science, University of California, Davis,

Nares Trakooljul
https://orcid.org/0000-0002-3214-2498

Klaus Wimmers
Leibniz-Institute for Farm Animal Biology (FBN) https://orcid.org/0000-0002-9523-6790

Goutam Sahana
Center for Quantitative Genetics and Genomics, Faculty of Technical Sciences

Guosheng Su
Center for Quantitative Genetics and Genomics, Faculty of Technical Sciences

Mogens Sandø Lund
Center for Quantitative Genetics and Genomics, Faculty of Technical Sciences

Merete Fredholm
University of Copenhagen
Peter Karlskov-Mortensen
University of Copenhagen https://orcid.org/0000-0002-2491-2763
Catherine Ernst
Agricultural Research Service https://orcid.org/0000-0003-2833-0995
Pablo Ross
University of California, Davis https://orcid.org/0000-0002-3972-3754
Christopher Tuggle
Iowa State University https://orcid.org/0000-0002-4229-5316
Lingzhao Fang
IGMM

Article

Keywords: livestock genome, epigenetics

Posted Date: March 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-253276/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Nature Communications on October 6th, 2021. See the published version at https://doi.org/10.1038/s41467-021-26153-7.

Pig genome functional annotation enhances biological interpretations of complex traits and comparative epigenomics

Zhangyuan Pan[1#], Yuelin Yao[2#], Hongwei Yin[3], Zexi Cai[4], Ying Wang[1], Lijing Bai[3], Colin Kern[1], Michelle Halstead[1], Kelly Chanthavixay[1], Nares Trakooljul[5], Klaus Wimmers[5], Goutam Sahana[4], Guosheng Su[4], Mogens Sandø Lund[4], Merete Fredholm[6], Peter Karlskov-Mortensen[6], Catherine W. Ernst[7], Pablo Ross[1], Christopher K. Tuggle[8], Lingzhao Fang[2]*, Huaijun Zhou[1]*

[1]Department of Animal Science, University of California, Davis, Davis, CA, USA

[2]MRC Human Genetics Unit at the Institute of Genetics and Molecular Medicine, The University of Edinburgh, Edinburgh EH4 2XU, UK

[3]Agricultural Genome Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China

[4]Center for Quantitative Genetics and Genomics, Faculty of Technical Sciences, Aarhus University, Tjele 8300, Denmark

[5]Leibniz-Institute for Farm Animal Biology, Dummerstorf, Germany

[6]Animal Genetics, Bioinformatics and Breeding, Department of Veterinary and Animal Sciences, University of Copenhagen, Frederikgsberg C 1870, Denmark

[7]Department of Animal Science, Michigan State University East Lansing, MI, USA

[8]Department of Animal Science, Iowa State University, Ames, IA, USA

[#]These authors contributed equally

*Corresponding authors

HZ: Department of Animal Science, University of California, Davis, CA, 95616, USA

E-mail: hzhou@ucdavis.edu

LF: MRC Human Genetics Unit at the Institute of Genetics and Molecular Medicine, The University of Edinburgh, Edinburgh EH4 2XU, UK

E-mail: Lingzhao.fang@igmm.ed.ac.uk

Abstract

The functional annotation of livestock genomes is crucial for understanding the molecular mechanisms that underpin complex traits of economic importance, adaptive evolution and comparative genomics. Here, we provide the most comprehensive catalogue to date of regulatory elements in the pig (*Sus scrofa*) by integrating 223 epigenomic and transcriptomic data sets, representing 14 biologically important tissues. We systematically describe the dynamic epigenetic landscape across tissues by functionally annotating 15 different chromatin states and defining their tissue-specific regulatory activities. We demonstrate that genomic variants associated with complex traits and adaptive evolution in pig are significantly enriched in active promoters and enhancers. Furthermore, we reveal distinct tissue-specific regulatory selection between Asian and European pig domestication processes. Compared with human and mouse epigenomes, we show that porcine regulatory elements are more conserved in DNA sequence, under both rapid and slow evolution, than those under neutral evolution across pig, mouse, and human. Finally, we provide novel biological insights on tissue-specific regulatory conservation and demonstrate that, depending on the traits, mouse or pig might be more appropriate biomedical models for different complex traits and diseases in humans through integrating comparative epigenomes with 47 human genome-wide association studies.

Main

Functional elements play essential roles in regulating gene expression in living cells and tissues[1]. There have been great efforts on the identification and annotation of functional elements in human and mouse genomes[1-11] as well as other model organisms including *Drosophila*[12] and *C. elegans*[13]. Significant enrichment of variants associated with human complex traits within regulatory elements has demonstrated the importance of Encyclopedia of DNA Elements (ENCODE) data[14]. Comparative analysis of epigenomes and transcriptomes across species could provide novel insights into the interrogation of underlying molecular mechanisms of human disease[8,15]. Genetic variants associated with common illnesses are enriched in human orthologues of mouse regulatory elements identified by ENCODE3[9], which validates the mouse could serve as a biomedical model for understanding some human diseases. However, compared with the mouse, pig (*Sus scrofa*) has more anatomical and physiological similarities to humans[16-18], and has been widely used as a human medical model[16,17,19-21]. The pig is also one of the most important farm animal species for meat production worldwide[22]. The genetic improvement of economically important complex traits such as growth, feed efficiency, and health contribute to efficient and sustainable production of animal protein ensuring a secure food supply for a growing world population. Functional annotation of regulatory elements in pig will lay a solid foundation for the identification of causative variants associated with phenotypic variations of complex traits, due to the enrichment of these variants in regulatory regions and the small genetic effect of each variant, making them difficult to discover[23].

Following ENCODE and Roadmap Epigenomics projects[8], the Functional Annotation of Animal Genomes (FAANG) initiative[24], although still in its infancy, has made great progress

86 towards annotating functional elements in many tissues across multiple domestic species including
87 pigs[25-30]. Here, we present 95 new genome-wide sequencing datasets from six gut-associated
88 porcine tissues and integrate them with 128 previously published FAANG datasets from eight
89 biologically distinct tissues. The collective interpretation of these datasets yields the most
90 comprehensive annotation of functional elements to date in any domesticated animal species. In
91 addition, we find that tissue-specific regulatory elements were enriched for the potential causative
92 variants of complex phenotypes by integrating a large scale of genome-wide association studies
93 (GWAS) and expression QTL (eQTL) datasets. Furthermore, by combining selection signature of
94 the pig genome, we show that tissue-specific regulatory elements play an important role during
95 the domestication. Finally, we compared porcine functional annotations with complementary
96 datasets from the human and mouse and integrated GWAS datasets concerning 47 human complex
97 traits. These comparisons demonstrate conservation of tissue-specific epigenetic signatures,
98 suggesting that, depends on diseases, pig or mouse may be a better animal model for them than
99 the other one.
100

101 # Results

102 **Data summary**

103 We integrated 223 genome-wide sequencing datasets from 14 major tissues in pig (Fig. 1a),
104 representing four histone modifications (H3K4me3, H3K4me1, H3K27ac and H3K27me3)
105 measured by Chromatin Immunoprecipitation sequencing (ChIP-seq), chromatin accessibility by
106 the Assay for Transposase-Accessible Chromatin (ATAC-seq), DNA methylation by Reduced
107 Representation Bisulfite sequencing (RRBS), and gene expression by RNA-seq (Supplementary
108 Fig. 1). We produced nearly 9 billion mapped reads with an average remaining rate of 68.81%

across samples (Supplementary Table 1). Among 14 tissues, we obtained an average of 32,387, 106,849, 72,252, 98,721, and 122,585 peaks for H3K4me3, H3K4me1, H3K27ac, H3K27me3, and ATAC, covering 1.56, 2.78, 2.37, 7.74, and 3.31% of the entire genome, respectively (Fig. 1b, c). Additionally, we utilized 16 CTCF ChIP-seq datasets from eight tissues[29] and four Hi-C datasets from liver[30] to identify topologically associating domains (TADs) and Hi-C loop for associating regulatory elements (enhancers) with potential target genes.

The hierarchical clustering of samples based on the signal intensity of epigenetic marks and gene expression profiles clearly recapitulated sequencing assays, followed by tissue types and biological replicates (Fig. 1d), which was consistent with results of principal component analysis (PCA) (Supplementary Fig. 2). The six assays formed three major clusters: (1) active regulatory regions (H3K4me3, H3K27ac, H3K4me1 and ATAC), (2) Polycomb repression (H3K27me3), and (3) gene expression (RNA-seq). The four active regulatory marks were positively correlated with each other, but were negatively correlated with H3K27me3, especially H3K27ac. The signal intensity of RNA-seq (within gene bodies) showed a weakly positive correlation with active regulatory marks, and a negative correlation with H3K27me3. Overall, all four active regulatory marks exhibited significant enrichments in the upstream of transcription start sites (TSS) of genes across tissues (Fig. 1e).

We present Myosin IA (*MYO1A*) gene, which is engaged in responses to *Escherichia coli* infection and microvillar membrane morphology in intestinal tissues[31,32], as an example to show the complex interplays of regulatory elements and gene expression in Figure 1f. *MYO1A* is specifically and highly expressed in intestine tissues, and showed specific enrichments of H3K27ac signals around its TSS in intestine tissues but not in other tissues (Fig. 1f). In addition,

the TSS of *MYO1A* is accessible and was enriched for other active regulatory marks (i.e., H3K27ac, H3K4me3 and H3K4me1) but not for Polycomb repression (H3K27me3) (Fig. 1f).

Prediction and characterization of chromatin states across 14 tissues

We defined 15 distinct chromatin states by combining all five epigenetic marks across 14 tissues. These states mainly represented promoters (TssA, TssAHet and TssBiv, covering 1.16% of the entire genome), TSS-proximal transcribed regions (TxFlnk, TxFlnkWk and TxFlnkHet, covering 0.92% of the genome), enhancers (EnhA, EnhAMe, EnhAWk, EnhAHet and EnhPois, covering 6.5% of the genome), repressed regions (Repr and ReprWk, covering 13.25% of the genome) and quiescent regions (73.39%) (Fig. 2a–e, Supplementary Table 2). Totally we identified 2,097,958 regulatory elements (exclude Qui) spanning 14 tissues including 39.351 active promoters (TssA), 188,827 active strong enhancers (EnhA), and 142,821 repressors (Repr) (Supplementary Fig. 4a-c). On average, 4.79% of the genome was accessible but did not coincide any other measured epigenetic marks (ATAC islands), indicating that additional epigenetic marks are required to further explore the biological function of such regions. TssA and TssBiv showed the highest enrichment of conserved DNA sequence elements, followed by TSS-proximal transcribed regions and accessible enhancers (EnhA and EnhAMe) (Fig. 2f). In general, TssA and TssBiv showed the highest enrichment at TSS, while other chromatin states showed enrichment at up- and down-stream of TSS (Fig. 2g). For instance, TssAHet and TSS-proximal transcribed states had the highest enrichment around 2kb up-stream of TSS, whereas enhancer states showed the highest enrichment around 20kb up-stream of TSS. Repressed states were enriched around 20kb up- and down-stream of TSS (Fig. 2g).

In general, different chromatin states showed distinct DNA methylation levels (Fig. 2h). Promoter and TSS-proximal transcribed states were hypomethylated compared to nearby sequence

(10kb up- and down-stream of TSS). Among promoter states, TssA had the lowest methylation level, confirming the well-known negative correlation between promoter methylation and gene expression[33]. The enhancer states showed intermediate methylation levels, among which EnhA and EnhAMe had lower methylation levels compared to other enhancers (Fig. 2g), where we also observed that EnhA and EnhAMe had more conserved sequence than other enhancers (Fig. 2f). This result suggests accessible enhancers may have more conserved sequences than non-accessible enhancers.

We took chromosome 7 (Chr7) as an example to explore the relationships among chromatin states, individual epigenetic marks, gene density, gene expression, DNA methylation and chromatin conformation (Fig. 2i). For instance, we observed that regions with higher density of genes were characterized by active chromatin states, higher gene expression, more chromatin accessibility, and lower methylation level. The chromatins were more physically interacted (measured by TADs from Hi-C data) within both gene desert and gene rich regions than the rest of genomic regions. To examine the associations of chromatin states with gene expression across tissues, we presented the *VIL1* locus (Villin-1), which participates in response to intestinal inflammation[34], as an example (Fig. 2j). *VIL1* exhibited tissue-specific active promoters and enhancers, as well as high expression in intestinal tissues compared to other tissues. Of particular note, despite the presence of TssA, *VIL1* was not expressed in stomach, possible due to the lack of enhancer activity in the upstream of its TSS, indicating that enhancers together with promoters may collectively regulate gene expression. Similar patterns were observed for *MYO1A* and *HNF4G* (Supplementary Fig. 3).

Dynamics of chromatin states across genome and tissues

We clustered the entire genome into 12 modules based on their relative frequency of chromatin states and observed that these modules exhibited distinct enrichments for protein-coding genes, non-coding genes and CpG islands (Fig. 3a). For instance, module 2 (M2) was characterized by active promoters and accessible enhancers, showed the highest enrichment for genes and CpG islands, and the lowest levels of DNA methylation, and the highest gene expression levels (Fig. 3b). Compared to modules 11 and 12, module 10 showed similar enrichment for Polycomb repression but higher enrichment for TssBiv, in which genes exhibited significantly lower expression levels, suggesting the crucial role of TssBiv for regulating gene repression (Fig. 3b). In addition, we noticed that module 1 had high enrichment for TssAHet, high levels of DNA methylation, and high representation of genes located on the X chromosome and these genes are relevant with histone modification Gene Ontology (GO) terms (Supplementary Table 3). This may indicate potential roles of TssAHet in heterochromatin on the Chromosome X[35].

By examining the distribution of chromatin states among all 14 tissues, we found that enhancer activity was the most variable between tissues, while promoters were least variable (Fig. 3c-d, Supplementary Fig. 4d,e). Among promoters, TssBiv was least constitutive and often switched to TSS-proximal transcribed or quiescent regions between tissues (Fig. 3d). Hierarchical clustering of samples using the signal intensity of H3K4me1 within EnhA clearly separated different tissue types (Fig. 3e), suggesting that the signal intensity of individual epi-mark in enhancers is highly indicative of tissue identity.

To explore the relationship between proximal regulatory elements (within 2kb of TSS of genes) and tissue-specific gene expression, we identified genes with tissue-specific expression (TSE), which were significantly engaged in known biological functions of specific tissues (Supplementary Fig. 5, Supplementary Table 4). We also observed that TSE were enriched for active states

(promoters, transcribed regions and enhancers) and depleted for repressed states in the 2kb regions around their TSS in the corresponding tissue compared to other tissues (Fig. 3f). Furthermore, we found that predicted target enhancers of TSE in a tissue were more constitutive among biologically similar tissues compared with other tissues (Fig. 3g), which was consistent with promoters of TSE (Supplementary Fig. 6).

Functional characterization of tissue-specific chromatin states

As enhancers were most variable among tissues compared to other chromatin states, we identified an average of 6,895 tissue-specific EnhAs among 14 tissues, ranging from 1,393 in jejunum to 14,811 in skeletal muscle (Fig. 4a). To further investigate the biological functions of such enhancers, we defined three other types of EnhA, including all-common EnhA (shared among all tissues), gut-common EnhA (shared among gut tissues) and brain-common EnhA (shared among brain tissues). Gene Ontology (GO) analysis of putative target genes of these different types of EnhAs revealed distinct biological functions (Fig. 4b, Supplementary Table 5). For instance, all-common EnhAs were involved in fundamental biological processes (e.g., regulation of mRNA catabolic processes and responses to wounding), whereas gut-common EnhAs were significantly involved in intestinal development, digestion and absorption, and immune response. EnhAs that were specifically active in individual gut tissues showed distinct functions, clearly matching the known biological functions of the tissue in question. For example, jejunum-specific EnhAs were involved in biological processes relevant to T cell and lymphocyte function[36], while colon-specific EnhAs were mainly engaged in stress-activated MAPK cascades[37] (Fig. 4b). We observed that intestine- and spleen-specific EnhAs shared many immune functions, and brain-specific EnhAs were significantly involved in memory and learning (Fig. 4b). Furthermore, we observed that genes whose topologically associated with tissue-specific EnhAs (Methods) were specifically

highly expressed in the corresponding tissues (Fig. 4c), and that methylation levels of tissue-specific EnhAs were lower in the corresponding tissues (Supplementary Fig. 7a), indicating that these tissue-specific enhancers and their methylation level were accurately predicted to regulate the expression of associated target genes.

To explore potential tissue-specific transcription factors (TF), first we identified motifs that were significantly enriched in tissue-specific EnhAs (Fig. 4d, Supplementary Fig. 8a), such as MEF2A, HNF1B, and HNF4A1 in muscle, liver and intestinal tissues, respectively, which was in line with previous findings in humans[8]. In addition, we found the binding motif of HNF4G, which participates in the renewal of intestinal stem cells in mice[38] and is specifically active in intestine (Supplementary Fig. 3b), and is enriched in most intestine-specific EnhAs, whereas CDX2, a major regulator of intestine-specific genes involved in cell growth and differentiation and is highly expressed in jejunum over duodenum and ileum[39,40], and its motif is specifically enriched in jejunum-specific EnhAs. The expression levels of the inferred TFs were higher in the corresponding tissue than in other tissues (Supplementary Fig. 8b,c), indicating that these tissue-specific enhancers are hotspots for TF activity and play important roles in the tissue-specific regulation of gene expression. We further observed that genes linked to tissue-specific EnhAs were significantly associated with biologically relevant complex diseases in humans and mice (Fig. 4e, Supplementary Fig. 7b, Supplementary Table 6). For example, colon-specific EnhAs were associated with recurrent bacterial infection diseases, and cecum-specific EnhAs were significantly associated with bruising susceptibility diseases.

We also explored potential function for tissue-specific promoters (TssA), and found that promoters also showed tissue-specific regulatory (TSR) function, but to a lesser degree than enhancers (Supplementary Fig. 9, Supplementary Table 7).

Chromatin states predictions enhanced the biological interpretations of adaptive evolution and complex traits in pigs

To determine whether genomic regions associated with adaptive evolution are significantly enriched in regulatory elements (REs), we first identified selection signatures (the top 5% of regions measured by Fst) by comparing wild with domesticated pigs in Asian and European populations separately (Supplementary Table 8). We found that genomic regions under selective pressure were most enriched for TssA and TSS-proximal transcribed regions, followed by enhancers, with similar patterns in both Asian and European populations (Fig. 5a, Supplementary Fig. 10a). In examining tissue-specific regulation, our analysis revealed that the all-common TssA were significantly enriched within regions under selective pressure in both populations (Fig. 5b). Interestingly, spleen-specific REs were most enriched in Asian pig domestication, wheras cortex-specific REs were most enriched in European pig domestication (Fig. 5b). Consequently, tissue-specific gene regulation may have played an essential role in the adaptive selection processes that resulted in Asian and European pig domestication. This result was also in agreement with the observation that Asian domesticated pigs being more disease resistant[41], whereas European domesticated pigs are more active and aggressive[42,43].

To ask whether SNPs associated with complex traits in pigs are enriched in regulatory regions, we integrated GWAS signal enrichment analysis for 44 complex traits (Supplementary Table 9) with all 15 chromatin states, and demonstrated that GWAS signals were most enriched in TssA (Fig. 5c), which was consistent with previous findings in humans[44]. We also found that enrichment for variants associated with complex traits was significantly positively correlated with signatures of selection (Supplementary Fig. 10b,c). We then asked if tissue-specific REs were involved in genetic control of specific complex traits. To answer this question, we conducted GWAS signal

enrichment analysis for average daily gain (ADG) in three separate breeds (i.e., Duroc, Landrace and Yorkshire), with emphasis on tissue-specific TssA and EnhA. As we expected, muscle, adipose, liver, and gut-common regulatory elements were the most relevant for ADG (Fig. 5d). In further examining the top ADG QTLs in Landrace (Fig. 5e), we found that the top hit SNPs that are within a muscle-specific EnhA (Fig. 5f) that appears to target two genes (*ZNF532* and *ALPK2*) based on TAD and Hi-C loop derived from CTCF and Hi-C data. Among all seven genes within this QTL, *ALPK2* plays important roles in cardiogenesis and was upregulated in the *longissimus dorsi* muscle in Wannanhua compared with Yorkshire pig[45,46], and was the only gene specifically expressed in muscle (Fig. 5f-h). Additional evidence from the eQTLs in muscle showed the highest enrichment in accessible enhancers (EnhA and EnhAMe) compared to other chromatin states (Supplementary Fig. 10d), suggesting that genetic regulatory variants are more likely to influence gene expression through perturbing enhancers. In summary, these results together demonstrated the important role of functional genome annotation for interpreting the molecular mechanisms underpinning complex traits, adaptive evolution and gene regulation.

Comparative analysis of pig, mouse and human epigenomes

The distribution of individual histone epigenetic marks and chromatin accessibility with respect to genomic features (e.g., 5'UTRs and exons) was consistent between pig, mouse, and human (Supplementary Fig. 11). To determine if the chromatin states are similarly conserved between these species, we predicted 15 chromatin states in mouse and human based on the same epigenetic marks in pig. The resulting chromatin state predications demonstrated general similarity among the three species in terms of genome coverage, genomic distribution and sequence conservation (Fig. 6a, Supplementary Fig. 12).

To explore the relationship between the epigenome and DNA sequence conservation among three species, we divided each genome into regions corresponding to 50 different levels of sequence conservation (0^{th}-49^{th}) (Methods). Our results revealed that the majority of chromatin states showed higher conservation levels in sequences under both rapid and slow evolution than those under neutral evolution, following a U-shaped distribution[47] (Fig. 6b, Supplementary Fig. 13a). We also found that the densities of chromatin states and gene elements followed the similar U-shaped distribution (Supplementary Fig. 13b,c), supporting the hypothesis that conserved epi-modifications may buffer negative selective pressures by providing the genome more elastic room to adapt[47]. Furthermore, we categorized orthologous genes into 50 groups based on the degree of conservation of gene expression between species and observed that genes with more conserved expression levels also demonstrated more conserved TssA and TssBiv signatures (Fig. 6c). In further examining sequence extremely conserved (49^{th}) or extremely variable regions (0^{th}), genes linked to TssA shared by human and pig are involved in basic biological processes, such as ncRNA metabolic process and mRNA catabolic process (Supplementary Fig.14). For the sequence extremely conserved region(49^{th}), we found that genes proximal (± 2 kb) by human-specific (comparing with pig) TssA in brain (e.g., *FOXG1*[48]) are engaged in neuron fate commitment, cerebral cortex development, learning and memory (Fig. 6d, Supplementary Fig. 15, Supplementary Table 10).

Next, we evaluated the evolutionary basis of complex traits in humans. Heritability enrichment analysis of 47 complex traits across 15 chromatin states that were mapped from pigs to orthologous regions in humans found that promoters and TSS-proximal transcribed regions were most enriched for variants (Fig. 6e). We further revealed that the more conserved (species-shared) chromatin states showed significantly higher enrichment of complex traits heritability than the more

313 divergent (species-specific) chromatin states (Fig. 6f). Then we further examined the role of tissue-
314 specific gene regulation on human complex traits. Our heritability enrichment analysis of complex
315 traits, based on human orthologous regions of tissue-specific EnhAs identified in pigs,
316 demonstrated that tissue-specific enhancers were significantly enriched for the corresponding
317 human complex traits relevant to biological functions of specific tissues (Fig. 6g). For instance,
318 the lung-specific EnhAs were significantly enriched for the heritability of lung forced expiratory
319 volume 1 (FEV1), liver for fasting glucose and cholesterol, colon for Crohn's disease, and cortex
320 for intelligence (Fig. 6g).

321 Finally, we sought to determine if this annotation of regulatory elements substantiated the use
322 of pig as an appropriate animal model for different human diseases by comparing human, mouse
323 and pig epigenomes in specific tissues. In brain cortex, the mouse-human shared EnhAs exhibited
324 significantly higher heritability enrichment than the pig-human shared EnhAs for most brain-
325 relevant traits, such as attention deficit hyperactivity disorder (ADHD), intelligence, depression
326 and reaction time, with the exception of Alzheimer's disease, for which heritability was
327 significantly enriched in pig-human shared EnhAs rather than the mouse-human shared EnhAs
328 (Fig. 6h). This was in line with previous findings that pigs have been used as a biomedical model
329 for studying Alzheimer's disease[21,49]. Similar observations were found in intestine (Crohn's disease
330 and inflammatory bowel disease (IBD), but not colorectal cancer, which demonstrated more
331 heritability in the pig-human shared EnhAs) (Fig. 6i) and in adipose (body mass index (BMI),
332 body fat percentage, waist-hip ratio and weight with significantly higher heritability enrichments
333 in the pig-human shared EnhAs) (Fig. 6j). Similar results showed in promoter (TssA)
334 (Supplementary Fig.16). Our findings suggest the pig could be a better biomedical model for
335 certain human traits and diseases than mouse, and vice versa.

Discussion

In this study, we provided the most comprehensive catalog of porcine regulatory elements to date, spanning 14 tissues including six gut-associated tissues, and characterized the dynamic chromatin state landscape across these tissues and uncovered extensive tissue-specific regulation of gene expression.

The annotation of functional elements in human and mouse has proven highly effective for the identification of causative variants of complex traits[23,27]. Our results also demonstrated that variants of complex traits and eQTLs of growth-related traits were significantly enriched in the active promoters and enhancers annotated by this study. Specifically, we speculate that a potential causative SNP, which was associated with average daily gain and which was found within a muscle-specific enhancer, may regulate the expression of *ALPK2*[45,46], a gene demonstrating muscle-specific expression (0.5Mb away). In addition, our annotation of functional elements in pigs allows us to evaluate the potential role of regulatory elements on pig domestication. Our analysis illustrated that signatures of domestication were significantly enriched in porcine regulatory elements. Specifically, genetic variants in the spleen-specific promoters were enriched during Asian pig domestication, whereas variants within cortex-specific promoters were enriched during European pig domestication. This novel insight may reflect the observed distinct phenotypic difference between Asian (more disease resistance[41]) and European domesticated pigs (more active and aggressive[42,43]). Further investigation is warranted to deepen our understanding of genetic selection and domestication in the pig. This regulatory element atlas will serve as a valuable source for the livestock community to inform GWAS and eQTL findings, genomic

selection program, and genome editing strategies, as well as to enhance our understanding of genome evolution and adaptation. With continued efforts by the FAANG Consortium[50], more epigenomic data will be available from diverse samples such as reproduction related tissues, additional developmental stages, and different physiological states. The systemic integration of "omics" data, for instance, the on-going pig GTEx effort will contribute additional insight into the biological mechanisms that underpin agronomic traits, and thereby enhancing genetic improvement of economically important phenotypes[50].

Finally, this atlas of functional elements provided a unique opportunity for comparative epigenomic analysis between human, mouse and pig, the results of which can inform which species constitute the most appropriate biomedical model(s) for specific human diseases. We observed that regions under positive or negative selective pressure demonstrated higher conservation of epigenetic signatures (such as TssA, TssBiv and TxFlnk) than those under neutral selective pressure, further confirming the hypothesis that elasticity of regulatory conservation may play an important role in the evolution of the less conserved regions (impact of negative selection pressure)[47]. Recently evolved liver enhancers (i.e., species-specific) are often associated with genes that show evidence for being under positive selection[51]. Such enhancers have been further demonstrated to actively affect gene expression, although they have a lower effect than those enhancers shared across species when the comparison is controlled for number of enhancer elements acting on that gene[52]. However, the human-specific promoters in brain tissues were enriched in intelligence related genes, which suggests a critical role for epigenomic regulation of novel biological function in humans in the sequence most evolutionarily conserved regions. It is widely accepted that neither mouse nor pig is universally appropriate to serve as an animal model for every human disease[18,53]. Gene regulatory networks play significant roles in controlling

phenotypic variance of complex traits, including most human diseases. In examining heritability enrichment of 47 complex traits in humans, our epi-conservation analysis among three species by comparing pig-human vs. mouse-human shared enhancers in different tissues revealed novel insights and potential underlying molecular mechanisms as to why pig might be a more appropriate animal model for certain human diseases than mouse and vice versa. This new line of evidence aligns well with many studies of human diseases using either mouse or pig as an animal model[18]. Our study provides a new basis for understanding genetic regulation of complex traits, such as human diseases by focusing on regulatory network conservation across different mammalian species. While the findings from our study are exciting, they are not yet conclusive. More epigenomic data from additional tissues and cell types, as well as additional species such as non-human primates, along with more experimental studies will be needed to extend and functionally validate these mechanisms that underpin complex traits and diseases[9,47].

Methods

Animals and tissues

Procedures for tissue collection followed the Animal Care and Use protocol (#18464) from the Institutional Animal Care and Use Committee (IACUC), University of California, Davis. We collected five gut-associated tissues (stomach, jejunum, duodenum, ileum, and colon) of two Yorkshire littermate male pigs at six months of age from Michigan State University[29]. Cecum from two female hybrid pigs (Yorkshires X Hampshires, five months) were obtained at University of California, Davis meat lab. Tissues were first flash frozen in liquid nitrogen, and then stored at –80 °C until further processing.

Library construction and sequencing

We performed ChIP-seq (H3K4me3, H3K4me1, H3K27ac and H3K27me3) experiments on flash-frozen tissue samples using the iDeal ChIP-seq kit for histones (Diagenode Cat.#C01010059, Denville, NJ), as previously described[29]. Libraries were sequenced on the Illumina's HiSeq 4000 with 50 bp single-end reads. ATAC-seq libraries were generated from frozen tissue samples by a modified protocol (https://figshare.com/articles/dataset/Final_ATAC_protocol_docx/13891268) according to the protocol of Omni-ATAC[54] and cryopreserved nuclei[55]. The sequencing was performed on Illumina's NextSeq with paired-end 40 bp reads (PE40). For the RRBS-seq experiments, DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) was used for extraction of DNA from frozen tissues. The samples were sent to Novogene (Sacramento, CA, USA) for library construction and sequencing by Illumina HiSeq 4000 PE150. Total RNA isolated from flash-frozen tissue by Zymo Quick-RNA™ Miniprep Kit (Irvine, CA, USA). RNA-seq libraries were constructed by NEBNext® Poly(A) mRNA Magnetic Isolation Module kit (NEB #E7490) and NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina (NEB #E7720, New England Biolabs (NEB), Ipswich, MA) and sequenced on Illumina HiSeq 4000 with PE100.

Data processing and data summary

In total, 95 new datasets including ChIP-seq (H3K4me3, H3K4me1, H3K27ac, H3K27me3, input control), ATAC-seq, RRBS, RNA-seq in two biological replicates of six gut-associated tissues were generated. We also integrated additional 144 existing pig epigenomic datasets including ChIP-seq (H3K4me3, H3K4me1, H3K27ac, H3K27me3, CTCF, input control), ATAC-seq, RRBS, RNA-seq in the same two biological replicates of eight core tissues (Adipose, Cerebellum, Cortex, Hypothalamus, Liver, Lung, Muscle, Spleen) from our FAANG pilot project (PRJEB14330)[29], and four Hi-C pig liver datasets from publicly available dataset (PRJEB27364)[30].

UC Davis FAANG Functional Annotation Pipeline (https://github.com/kernco/functional-annotation) was applied to process the ChIP-seq, ATAC-seq, and RNA-seq as previously described[29]. Briefly, the susScr11 genome assembly and Ensembl genome annotation (v100) were used as references for pig. Sequencing reads were trimmed with Trim Galore![56](v.0.6.5), and aligned with STAR[57](v.2.5.4a) or BWA[58] (v0.7.17) to the respective genome assemblies. Alignments with MAPQ scores less than 30 were filtered using Samtools[59](v.1.9). For RNA-seq, gene counts were determined using htseq-count[60](v.0.13.5), and then trimmed mean of M-values (TMM) and transcript per million (TPM) normalization were performed using EdgeR (v3.32.0) and StringTie2 (v.1.3.3), respectively[61]. For ChIP-seq, after the filtering, duplicates were marked and removed using Picard (v.2.18.7). Regions of signal enrichment ("peaks") were called by MACS2[62] (v.2.1.1). Various quality metrics (e.g., JSD, Supplementary Table 1) were calculated following the method described in our previous study[29]. RRBS data were processed using Bismark[63] (v.0.22.1) with parameter set in RRBS pipeline (https://github.com/zhypan/Functional-Annotation-of-Pig). We called the Hi-C contacts using the Juicer pipeline[64] with default parameters.

The global correlations among assays, tissues, and biological replicates were performed by deepTools[65](v.3.5.0). Briefly, the Z-score normalized read signals of all samples with step of 1k bp window were calculated by multiBigwigSummary and were presented by plotCorrelation. The signal of marks along with protein coding genes were generated by deepTools[65] (computeMatrix scale-regions function) with parameter -a 2500 -b 2500. The Z-score was used to normalize bigWig of five marks as input files.

Annotation of chromatin states

ChromHMM[66] (v.1.20) was used to train the chromatin state prediction model by integrating ChIP-seq (H3K4me3, H3K4me1, H3K27ac, H3K27me3, and input control) and ATAC-seq data

from two biological replicates of 14 tissues. The same tissue of two biological replicates were collectively considered as one tissue epigenome. The 15-state model was chosen, as it presented maximum number of states with distinct epigenetic mark combinations. We labelled these 15 chromatin states based on their combinations of histone modifications and enrichment around TSS[8,27]. Then the fold enrichment of each chromatin state for each external gene element (e.g., exon, CpG islands) was calculated by (C/A)/(B/D), where A, B, C, D are the number of bases in a chromatin state, a gene element, overlapped between a chromatin state and a gene element, in the genome, respectively. In addition, we also computed chromatin state fold enrichment in mammalian conserved elements which identified from Multiple Sequence Alignments (MSA) using the Genomic Evolutionary Rate Profiling (GERP) software based on 103 mammals (ftp://ftp.ensembl.org/pub/release-100/bed/ensembl-compara/103_mammals.gerp_constrained_element/). The methylation level of each state and its up- and down- stream 10kb was calculated by the computeMatrix scale-regions function of deepTools with parameters --binSize 500, --regionBodyLength 2000 and --skipZeros.

Clustering of large-scale chromatin structure

To examine genome-wide chromatin structure, we first divided the genome (excluding chrUn) into 1,224 fragments of 2Mb in length. Then we calculated the state frequencies (state bin/total bin) in each 2Mb fragment for each tissue and then the average frequency across tissues. To identify modules, column clustering was performed by k-means=12, and rows were clustered using k=3. In addition, we calculated number of protein-coding, lncRNA, and CpG islands for each 2Mb fragment by BEDTools[67] (v.2.29.2). We also calculated the average TPM of protein coding gene and average methylation level across14 tissues in each 2Mb fragment. Then the average gene

expression and methylation level in each of the 12 modules were calculated and a Student's T-test was performed with parameter setting ref.group = "M3".

Chromatin state variability

For each state, we first obtained regulatory regions across 14 tissues (RRATs) (Supplementary Fig. 4a-c) using BEDtools merge function (any regulatory region between two tissues overlapped by 1 bp was merged), then we calculated the total genomic length for each tissue (GL) and the total combined genomic length (TGL) for RRATs. The relative state coverage per tissue was derived by GL/ TGL (Supplementary Fig. 4d). Finally, followed by the order from high to low based on the GL/TGL value in each tissue, we calculated the total genomic length of accumulated tissues (aGL) by adding one tissue each time until all 14 tissues were added, and the cumulative state coverage was calculated as aGL/ TGL. States whose cumulative coverage changed faster than others were considered to be less constitutive (more variable) states.

Chromatin state switching between tissues

Chromatin state switching between tissues was calculated by pairing two tissues. Given a pairing of A and B tissues, we first counted total bins of chromatin state "e" in A (TbAe), then obtained the overlap bins of chromatin state "e" (Obe) in A and B, then computed the state switching probabilities using Obe/TbAe for the tissue A to B transition and Obe/TbBe for the tissue B to A transition. By averaging these calculations for a pair of tissues, we obtained the pair switching probabilities. We calculated the state switching probabilities in between intestinal tissues, between brain tissues （Supplementary Fig. 6a,b) and between 8 distinguishable tissues (jejunum, cortex, adipose, liver, lung, muscle, spleen).

Hierarchical epigenome clustering

We first calculated the an epigenetic mark's signal confidence scores ($-\log_{10}$(Poisson *P* value)) within 200 bp of the genomic regions for each mark of each sample as described in http://jvanheld.github.io/stats_avec_RStudio_EBA/practicals/02_peak-calling/peak-calling_report.html#data_sets. Then, we extracted a specific mark's signal confidence score of each sample for specific state of RRATs regions. For example, we extract H3K4me1 signal confidence scores for EnhA. After combining all samples' mark confidence scores for each tissue and each state, we constructed a distance matrix using the ward.D2-linkage hierarchical clustering following by Euclidean distance method in R.

Promoter enrichment analysis of tissue-specific expressed genes among 14 tissues

To evaluate how chromatin state changes at promoter regions of TSE genes across tissues, we first performed a Student's *t*-test among 14 tissues to identify tissue-specific expressed genes based on TPM. We further grouped some tissues into different sub-groups such as small intestine (Jejunum, Ileum, Duodenum), large intestine (Cecum, Colon), and brain (Cortex, Cerebellum, Hypothalamus), and identified tissue-specific expressed genes by excluding the tissues in the same sub-group. Then we selected genes with the top 5% *t*-value as TSE genes[68]. The biological process of GO enrichment for these TSE genes were identified by WebGestalt2019[69] using the default significance level (FDR< 0.05). Then we calculated the chromatin state fold enrichment of TSE (up and down stream 2000bp around TSS) in each tissue and the change in enrichment by TSE enrichment in specific tissue minus other tissues.

Chromatin state switching of target enhancer (EnhA) of TSE gene

To evaluate how enhancers of TSE genes switch among tissues, we first identified the target enhancers of TSE genes following the method described in our recent study[29]. Briefly, we generated the predicted TADs from CTCF ChIP-seq data by FIMO[70] following the method described in Oti, et al.[71]. Then we predicted the enhancer-gene pairs according to the Spearman's rank correlation of every possible combination of regulatory element H3K27ac signal and gene expression value within each TAD. Benjamini-Hochberg adjustment (FDR < 0.05) was used to define putative interacting pairs. The enhancers in the enhancer-gene pairs that corresponded to TSE genes were considered as TSE genes' target enhancers. Finally, we computed enhancer state switching probabilities of TSE genes among tissues using the method described above.

TSR of enhancer, promoter and their putative functional regulation

For strong enhancer (EnhA) identified in each tissue, we counted the bins of overlapping RRATs by comparing to other tissues. If the number of bins >= 1, the tissue of this RRATs region would be assigned 1, otherwise it was assigned 0. We generated a total of 17 modules of tissue-specific regulatory elements (TSR) enhancers. The 17 modules included all-common (presented in all tissues), gut-common (presented in all 5 intestinal tissues), brain-common (presented in all 3 brain tissues) and 14 tissue-specific modules. The same method was used to obtain TSR for promoters (1_TssA). In addition, we performed enrichment analyses (GO, Human Phenotype Ontology (HPO), Mouse Phenotype) based on genes proximal to TSR using the GREAT[72] tool with default parameters except for TSR promoters (proximal 2kb upstream, 1kb downstream, plus distal up to 3kb). We used a cut-off of FDR<0.05 for both the binomial and the hypergeometric distribution-based tests.

The motifs of tissue-specific EnhAs were identified by HOMER[73] (v.4.11) with cutoff FDR<0.05. We selected the top three enriched or tissue function relevant motifs for each tissue as the candidate tissue-specific EnhAs motifs and generated a total of 51 motifs enriched in tissue-specific EnhAs. In addition, we used these 51 motifs as known TF motifs to conduct the enrichment for all tissues by HOMER. The mRNA expression of corresponding TFs in pigs were used to calculate the correlation with motif enrichment.

Selection signature enrichment analysis of chromatin state

A total of 406 whole genome sequence datasets (Supplementary Table 8) in pigs (Asian wild (58) and domestic pigs (129), European wild (35) and domestic pigs (184)) were trimmed by Trimmomatic[74] (v.0.39), mapped by BWA (0.7.17), and marked duplicates by GATK[75] (v4.1.4.1) MarkDuplicates with default parameters. The SNPs of Gvcf for each sample were called by GATK HaplotypeCaller. All Gvcf were then combined and the variants for each sample were called by GenotypeGVCFs. After SNP calling, the variants were filtered using VariantFiltration (QD < 2.0, MQ < 40.0, FS > 60.0, SOR > 3.0, MQRankSum < -12.5, ReadPosRankSum < -8.0) to remove low-quality SNPs. We then performed Fst analysis between Asian wild and domestic pigs, and between European wild and domestic pigs, and calculated the fold enrichment of selection signature for chromatin states using the same method for gene elements enrichment described above.

GWAS and eQTL signal enrichment of chromatin state

The pig GWAS data of 44 traits was described previously[76,77](Supplementary Table 9). First, we filtered out all SNPs with minor allele frequency below 0.5%, with a large deviation from Hardy–

Weinberg proportions ($P< 1.0^{-6}$), or with a R2 value of the imputation accuracy estimated by Minimac4 of less than 0.4. We performed GWAS signal enrichment of 44 pig complex traits (3 daily gain related, 20 lipid related, and 21 feed efficiency related) for each chromatin state across 14 tissues using a 10,000 times genotype cyclical permutation tests[68]. The eQTLs data in pig muscle[78] with FDR<0.05 were used to calculate the fold enrichment for the chromatin states using the same method above.

Interspecies conservation of chromatin state

We collected data from ENCODE[4,9], Roadmap Epigenomics[8] and published articles[79] (9 tissues in human and 7 tissues in mouse, Supplementary Table 11,12), including ChIP-seq (H3K4me3, H3K4ac, H3K4me1, H3K27me3, Input), ATAC-seq, DNase-seq, and RNA-seq. In total, we obtained six matched tissues (small intestine, liver, spleen, lung, adipose, cortex) among pig, human, and mouse. All the data were processed following the same pipeline used in pig. The GRCh38 (human) and GRCm38 (mouse) assemblies with Ensembl annotations (v100) were used for data analysis. Chromatin states of human and mouse were also trained by ChromHMM and 15 chromatin states were identified. To explore the relationship between sequence conservation and epi-conservation among the three mammals, we first divided the genome into 50 equal sized sets (0^{th}-49^{th}) with increasing average PhyloP scores using the method detailed by Xiao *et a*[47]. Briefly, the human genome was divided into 15 million 200 bp segments. Then average PhyloP score (100 vertebrate genomes[80]) was computed for each 200 bp segment. These genomic segments were divided into 50 equal sized sets from the fastest changing sequence (smallest PhyloP scores) to the most conserved (greatest PhyloP scores). (Supplementary Fig. 13d). To quantify epigenomic conservation, we downloaded the whole genome alignments UCSC chain files among human

(hg38), pig (SusScr11), mouse (mm10) and processed as described in the UCSC Genome Wiki website (http://genomewiki.ucsc.edu/index.php/HowTo:_Syntenic_Net_or_Reciprocal_Best) to derive reciprocal best chains. Then we converted genomic coordinates between assemblies using the UCSC Liftover tool (https://genome.sph.umich.edu/wiki/LiftOver) based on 0.65 sequence identity. All the chromatin states in pig and mouse were lifted over to human. The conservation rate (0~1) of each region of each state from pig to human was calculated based on state region coverage of pig over human. If there was no overlap it was assigned 0, if completely occupied it was assigned 1. The same analysis was conducted for pig to mouse and mouse to human. Furthermore, we performed genomic and epigenomic conservations for every pair of mammalian species in each tissue. Finally, we conducted the same analysis on mammalian conserved score based Genomic Evolutionary Rate Profiling (GERP) using 103 mammalian genomes (ftp://ftp.ensembl.org/pub/release-100/compara/conservation_scores/103_mammals.gerp_conservation_score/)

To examine the biological relevance of sequence extremely variable (0^{th}-2^{th} sets) and conserved regions (47^{th}-49^{th} sets), we extracted the human-pig shared and human-specific chromatin state TssA from these regions. Then the GREAT tool with parameter of proximal 2kb upstream, 1kb downstream, plus distal up to 3kb was used to conduct GO function enrichment analysis.

Expression conservation versus epi-conservation

The TPM of 14302 orthologous genes from pig, human, and mouse were used to identify differentially expressed genes in each tissue using the Student's *t*-test. We sorted the genes by p-value within each species and divided them into 50 equally sized sets. Then we calculated the average epi-conservation score of states in the 20kb region around TSS of gene in each set.

600 **Heritability enrichment of human complex traits in chromatin state**

601 To explore how conserved or species-specific chromatin states affects complex traits in humans,
602 we extracted six types of species-share or species-specific regulatory elements (all_shared,
603 human_mouse_shared, human_pig_shared, human_specific, mouse_specific, pig_specific). We
604 applied stratified linkage disequilibrium score regression (LDSC) to partition heritability of 47
605 human complex traits into distinct functional categories[44], which revealed which functional regions
606 explained more genetic variation of complex traits from an evolutionary point of view. These
607 functional categories included six types of species-shared/specific regulatory elements, chromatin
608 state of each tissue, and TSR of EnhA/TssA. We calculated the stratified LD scores using 1000G
609 Phase 3 European human samples, where only HapMap3 SNPs with INFO≥0.9 and MAF>0.05 in
610 1000G European samples were used (the 1000G samples and default SNP weights were obtained
611 from https://github.com/bulik/ldsc).

612 The GWAS summary statistics for 47 human complex traits were obtained from public
613 databases (Supplementary Table 13), with an average sample size of 321,978 (all European
614 ancestry) and a high quality overlap with HapMap3 panel. In addition, these GWAS results have,
615 a mean χ^2 statistics of > 1.02 and a heritability Z-score of > 4[81]. We also performed default
616 quality control for each GWAS summary statistics by LDSC to remove GWAS SNPs that are with
617 MAF ≤ 0.01, genotype call rate ≤ 0.75, INFO ≤ 0.9, out-of-bounds *P*-value, duplicated Rsid, strand
618 ambiguous variants and extreme large χ^2 statistics[81]. The results of LDSC regression for base
619 model, which has not been partitioned heritability, are available in Supplementary Table 14.

620

Data availability

All high-throughput sequencing data in this study were deposited in European Nucleotide Archive (ENA) with accession number PRJEB37735 and PRJEB14330. All raw data are also available through the FAANG portal (https://data.faang.org/dataset). All processed data are publicly available at https://doi.org/10.6084/m9.figshare.13480425. Chromatin states of pig, mouse, and human are available in Genome Browser: http://genome.ucsc.edu/s/zhypan/susScr11_15_state_14_tissues_new . http://genome.ucsc.edu/s/zhypan/mm10_7tissues_chr_state; http://genome.ucsc.edu/s/zhypan/hg38_9tissue_chr_state.

Code availability

The pipeline for RNA-seq, ATAC-seq, DNase-seq and ChIP-seq processing is available at https://github.com/kernco/functional-annotation. RRBS pipeline and other processing codes are publicly available at https://github.com/zhypan/Functional-Annotation-of-Pig.

Acknowledgements

This study was supported by Agriculture and Food Research Initiative Competitive Grant no. 2018-67015-27501 (CKT, HZ CE and PR) and no. 2015-67015-22940 (HZ and PR) from the

USDA National Institute of Food and Agriculture, Multistate Research Project NRSP8 and NC1170 (HZ), and the California Agricultural Experimental Station (HZ).

Author contributions

H.Z., L.F. Z.P. P.R., C.W. E. and C.K.T. conceived and designed the study. C.E., Y.W., K.C. and Z.P. were responsible for sample collection. Z.P., Y.W. and M.H. performed ChIP-seq, ATAC-seq, RNA-seq. Z.P., N.T. and K.W. contribute for RRBS data collection. Z.P., Y.Y. L.F. and C. K. conducted bioinformatic analysis. Z.X., G.S., GS. S., MS. L., M.F., and P. KM. were responsible for pig GWAS data analysis. H.Y. and L.B. responsible for pig selection signature collection. Z.P., L.F., Y.Y. and H.Z. wrote the initial draft of the manuscript. M.H. C.W. E., P. R., and C.K. T. revised manuscript. All co-authors contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.Z., and L.F.

Reprints and permissions information is available at www.nature.com/reprints.

References

1. Consortium, E.P. *et al*. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710 (2020).
2. Consortium, E.P. The ENCODE (ENCyclopedia of DNA elements) project. *Science* **306**, 636-640 (2004).
3. Consortium, E.P. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799 (2007).
4. Dunham, I. *et al*. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).
5. Filion, G.J. *et al*. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143**, 212-224 (2010).
6. Ernst, J. *et al*. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49 (2011).

7. Pang, B. & Snyder, M.P. Systematic identification of silencers in human cells. *Nat. Genet.* **52**, 254-263 (2020).
8. Roadmap Epigenomics, C. *et al*. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).
9. Gorkin, D.U. *et al*. An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744-751 (2020).
10. Zabidi, M.A. *et al*. Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556-559 (2015).
11. He, Y. *et al*. Spatiotemporal DNA methylome dynamics of the developing mouse fetus. *Nature* **583**, 752-759 (2020).
12. Roy, S. *et al*. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**, 1787-1797 (2010).
13. Gerstein, M.B. *et al*. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**, 1775-1787 (2010).
14. Maurano, M.T. *et al*. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195 (2012).
15. Boix, C.A., James, B.T., Park, Y.P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature*, 1-8 (2021).
16. Zhang, Q., Widmer, G. & Tzipori, S. A pig model of the human gastrointestinal tract. *Gut Microbes* **4**, 193-200 (2013).
17. Bassols, A. *et al*. The pig as an animal model for human pathologies: A proteomics perspective. *PROTEOMICS–Clinical Applications* **8**, 715-731 (2014).

18. Meurens, F., Summerfield, A., Nauwynck, H., Saif, L. & Gerdts, V. The pig: a model for human infectious diseases. *Trends Microbiol*. **20**, 50-57 (2012).

19. Sullivan, T.P., Eaglstein, W.H., Davis, S.C. & Mertz, P. The pig as a model for human wound healing. *Wound Repair Regen*. **9**, 66-76 (2001).

20. Gieling, E.T., Schuurman, T., Nordquist, R.E. & van der Staay, F.J. The pig as a model animal for studying cognition and neurobehavioral disorders. in *Molecular and Functional Models in Neuropsychiatry* 359-383 (Springer Press, 2011).

21. Kragh, P.M. *et al*. Hemizygous minipigs produced by random gene insertion and handmade cloning express the Alzheimer's disease-causing dominant mutation APPsw. *Transgenic Res*. **18**, 545-558 (2009).

22. de Almeida, A.M. & Bendixen, E. Pig proteomics: a review of a species in the crossroad between biomedical and food sciences. *Journal of Proteomics* **75**, 4296-4314 (2012).

23. Xiang, R. *et al*. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proc. Natl. Acad. Sci. USA* **116**, 19398-19408 (2019).

24. Andersson, L. *et al*. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol*. **16**, 1-6 (2015).

25. Burns, E.N. *et al*. Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. *Anim. Genet*. **49**, 564-570 (2018).

26. Kingsley, N. *et al*. Functionally annotating regulatory elements in the equine genome using histone mark chip-seq. *Genes* **11**, 3 (2020).

27. Fang, L. *et al*. Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations. *BMC Biol*. **17**, 68 (2019).
28. Halstead, M.M. *et al*. A comparative analysis of chromatin accessibility in cattle, pig, and mouse tissues. *BMC Genomics* **21**, 1-16 (2020).
29. Colin Kern, Y.W., Xiaoqin Xu, Zhangyuan Pan, Michelle Halstead, Kelly Chanthavixay, Perot Saelao, Susan Waters, Ruidong Xiang, Amanda Chamberlain, Ian Korf, Mary E. Delany, Hans H. Cheng, Juan F. Medrano, Alison L. Van Eenennaam, Chris K. Tuggle, Catherine Ernst, Paul Flicek, Gerald Quon, Pablo Ross, Huaijun Zhou. Functional genome annotations of three domestic animal species provide a vital resource for comparative and agricultural research. *Nat. Commun*. (2021).
30. Foissac, S. *et al*. Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biol*. **17**, 1-25 (2019).
31. Tyska, M.J. *et al*. Myosin-1a is critical for normal brush border structure and composition. *Mol. Biol. Cell* **16**, 2443-2457 (2005).
32. Shifrin Jr, D.A. *et al*. Enterocyte microvillus-derived vesicles detoxify bacterial products and regulate epithelial-microbial interactions. *Curr. Biol*. **22**, 627-631 (2012).
33. Wagner, J.R. *et al*. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol*. **15**, R37 (2014).
34. Roy, S. *et al*. Villin-1 and gelsolin regulate changes in actin dynamics that affect cell survival signaling pathways and intestinal inflammation. *Gastroenterology* **154**, 1405-1420. e1402 (2018).

35. Wutz, A. Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nat. Rev. Genet.* **12**, 542-553 (2011).
36. Savilahti, E., Reunala, T. & Mäki, M. Increase of lymphocytes bearing the gamma/delta T cell receptor in the jejunum of patients with dermatitis herpetiformis. *Gut* **33**, 206-211 (1992).
37. Hommes, D. *et al*. Inhibition of stress-activated MAP kinases induces clinical improvement in moderate to severe Crohn's disease. *Gastroenterology* **122**, 7-14 (2002).
38. Chen, L. *et al*. HNF4 regulates fatty acid oxidation and is required for renewal of intestinal stem cells in mice. *Gastroenterology* **158**, 985-999. e989 (2020).
39. Silberg, D.G., Swain, G.P., Suh, E.R. & Traber, P.G. Cdx1 and cdx2 expression during intestinal development. *Gastroenterology* **119**, 961-971 (2000).
40. Mach, N. *et al*. Extensive expression differences along porcine small intestine evidenced by transcriptome sequencing. *PLoS ONE* **9**, e88515 (2014).
41. Chen, H. *et al*. Introgression of Eastern Chinese and Southern Chinese haplotypes contributes to the improvement of fertility and immunity in European modern pigs. *GigaScience* **9**, giaa014 (2020).
42. Chu, Q., Liang, T., Fu, L., Li, H. & Zhou, B. Behavioural genetic differences between Chinese and European pigs. *J. Genet.* **96**, 707-715 (2017).
43. Li, L.-A. *et al*. Erhualian and Pietrain pigs exhibit distinct behavioral, endocrine and biochemical responses during transport. *Livestock Science* **113**, 169-177 (2008).
44. Finucane, H.K. *et al*. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228 (2015).

766 45. Hofsteen, P. *et al*. ALPK2 promotes cardiogenesis in zebrafish and human pluripotent
767 stem cells. *iScience* **2**, 88-100 (2018).
768 46. Li, X.J., Zhou, J., Liu, L.Q., Qian, K. & Wang, C.L. Identification of genes in
769 longissimus dorsi muscle differentially expressed between Wannanhua and Yorkshire
770 pigs using RNA-sequencing. *Anim. Genet.* **47**, 324-333 (2016).
771 47. Xiao, S. *et al*. Comparative epigenomic annotation of regulatory DNA. *Cell* **149**, 1381-
772 1392 (2012).
773 48. Kumamoto, T. & Hanashima, C. Evolutionary conservation and conversion of Foxg1
774 function in brain development. *Development, growth & differentiation* **59**, 258-269
775 (2017).
776 49. Lee, S.-E. *et al*. Production of transgenic pig as an Alzheimer's disease model using a
777 multi-cistronic vector system. *PLoS ONE* **12**, e0177933 (2017).
778 50. Clark, E.L. *et al*. From FAANG to fork: application of highly annotated genomes to
779 improve farmed animal production. *Genome Biol.* **21**, 1-9 (2020).
780 51. Villar, D. *et al*. Enhancer evolution across 20 mammalian species. *Cell* **160**, 554-566
781 (2015).
782 52. Berthelot, C., Villar, D., Horvath, J.E., Odom, D.T. & Flicek, P. Complexity and
783 conservation of regulatory landscapes underlie evolutionary resilience of mammalian
784 gene expression. *Nature ecology & evolution* **2**, 152-163 (2018).
785 53. Peters, L.L. *et al*. The mouse as a model for human biology: a resource guide for complex
786 trait analysis. *Nat. Rev. Genet.* **8**, 58-69 (2007).
787 54. Corces, M.R. *et al*. An improved ATAC-seq protocol reduces background and enables
788 interrogation of frozen tissues. *Nat. Methods* **14**, 959-962 (2017).

55. Halstead, M. *et al*. Systematic alteration of AtAc-seq for profiling open chromatin in cryopreserved nuclei preparations from livestock tissues. *Sci. Rep*. **10**, 1-12 (2020).
56. Krueger, F. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. (2015).
57. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2012).
58. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **1303.3997**(2013).
59. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
60. Anders, S., Pyl, P.T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2014).
61. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol*. **20**, 1-13 (2019).
62. Zhang, Y. *et al*. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. **9**, 1-9 (2008).
63. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572 (2011).
64. Durand, N.C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems* **3**, 95-98 (2016).
65. Ramírez, F. *et al*. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*. **44**, W160-W165 (2016).
66. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215-216 (2012).

67. Quinlan, A.R. BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12. 11-11.12. 34 (2014).

68. Fang, L. *et al*. Comprehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle. *Genome Res.* **30**, 790-801 (2020).

69. Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199-W205 (2019).

70. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018 (2011).

71. Oti, M., Falck, J., Huynen, M.A. & Zhou, H. CTCF-mediated chromatin loops enclose inducible gene regulatory domains. *BMC Genomics* **17**, 252 (2016).

72. McLean, C.Y. *et al*. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495-501 (2010).

73. Heinz, S. *et al*. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576-589 (2010).

74. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

75. Van der Auwera, G.A. *et al*. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10. 11-11.10. 33 (2013).

76. Zexi Cai, O.F.C., Mogens Sandø Lund, Tage Ostersen, Goutam Sahana. Large-scale association study on daily weight gain in pigs reveals overlap of genetic factors for growth in humans. *Genetics Selection Evolution* (2021).

77. Zhang, Q. *et al*. Inclusion of endophenotypes in a standard GWAS facilitate a detailed mechanistic understanding of genetic elements that control blood lipid levels. *Sci. Rep.* **10**, 1-14 (2020).

78. Velez-Irizarry, D. *et al*. Genetic control of longissimus dorsi muscle gene expression variation and joint analysis with phenotypic quantitative trait loci in pigs. *BMC Genomics* **20**, 3 (2019).

79. Liu, C. *et al*. An ATAC-seq atlas of chromatin accessibility in mouse tissues. *Scientific data* **6**, 1-10 (2019).

80. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110-121 (2010).

81. Bulik-Sullivan, B.K. *et al*. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291-295 (2015).

Figures and legends:

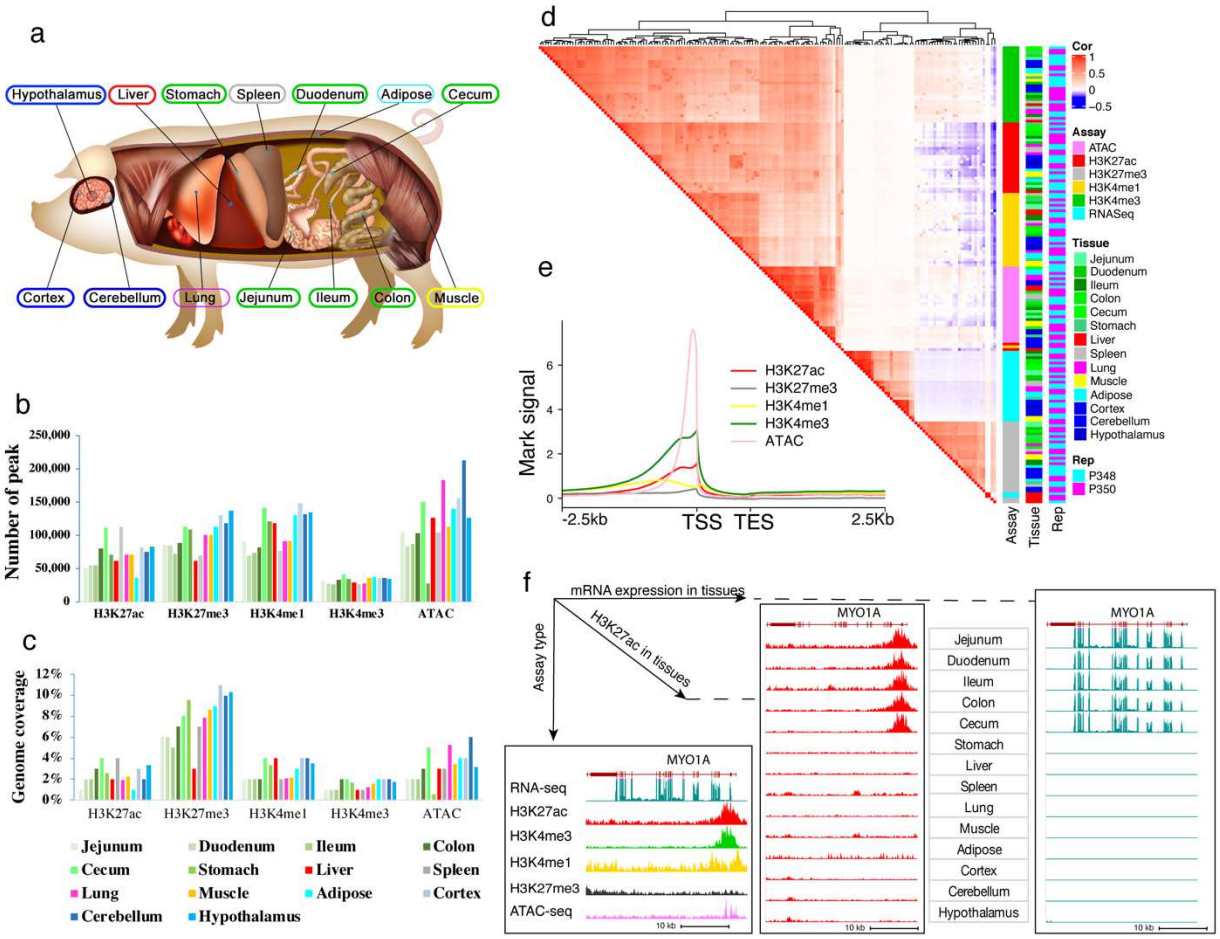
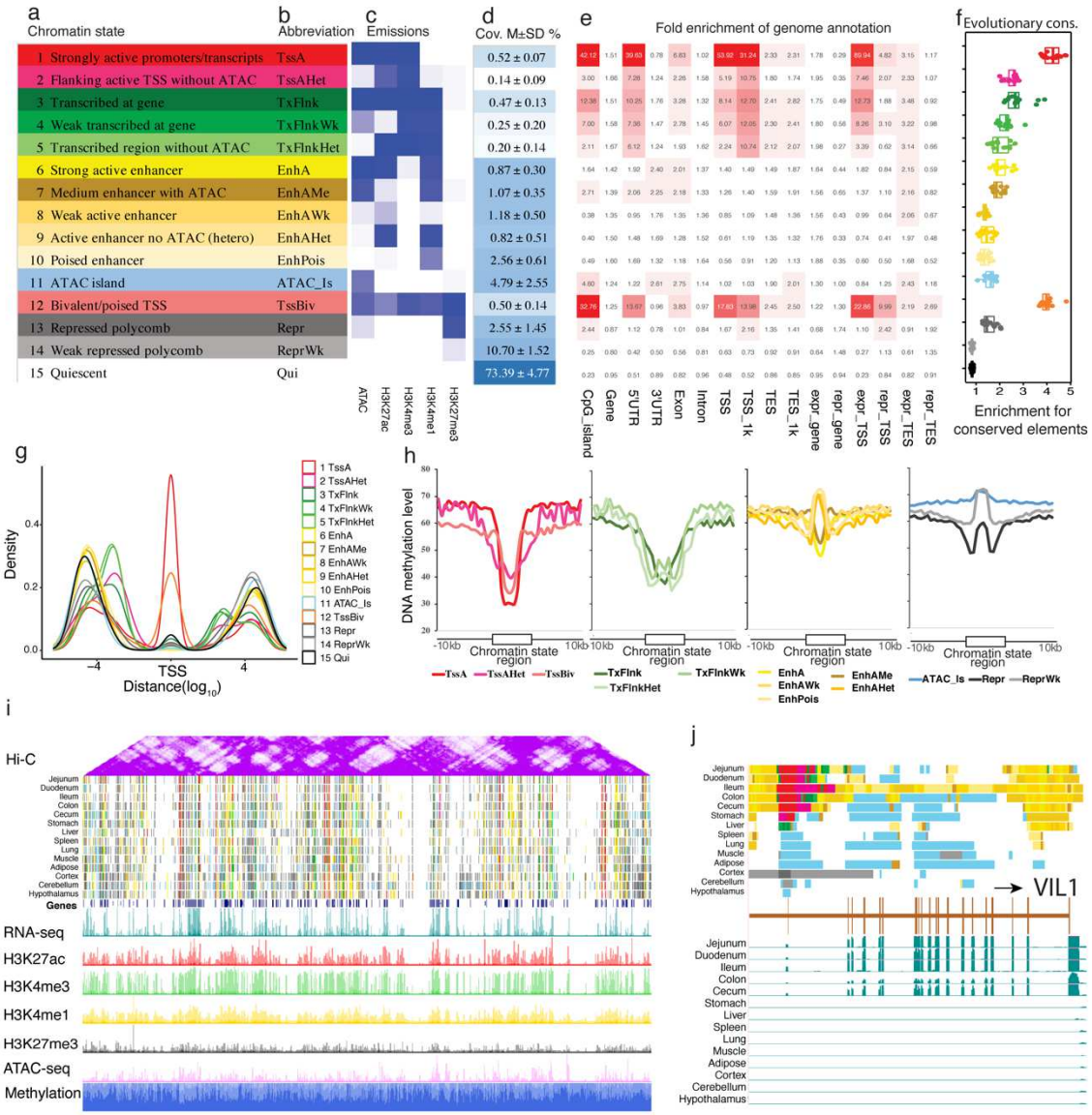


Fig.1 Data summary of epigenomic information across tissues and marks.

a, Tissues assayed by this study**. b,c,** Average peak number and genome coverage for each epigenetic mark in each tissue. **d,** The Pearson correlations among assays, tissues, and biological replicates (P348 and P350) based on the normalized signal in 1kb windows stepped across the whole genome. **e,** Average epigenetic mark signals proximal to protein coding genes. TSS, transcription start site. TES, transcription end site. **f,** Epigenetic signal at the *MYO1A* locus according to different assays and in different tissues. Vertical scale UCSC reads signal 0-200 for RNA-seq, 0-100 for H3K27ac and H3K4me3, 0-50 for other marks and ATAC-seq.

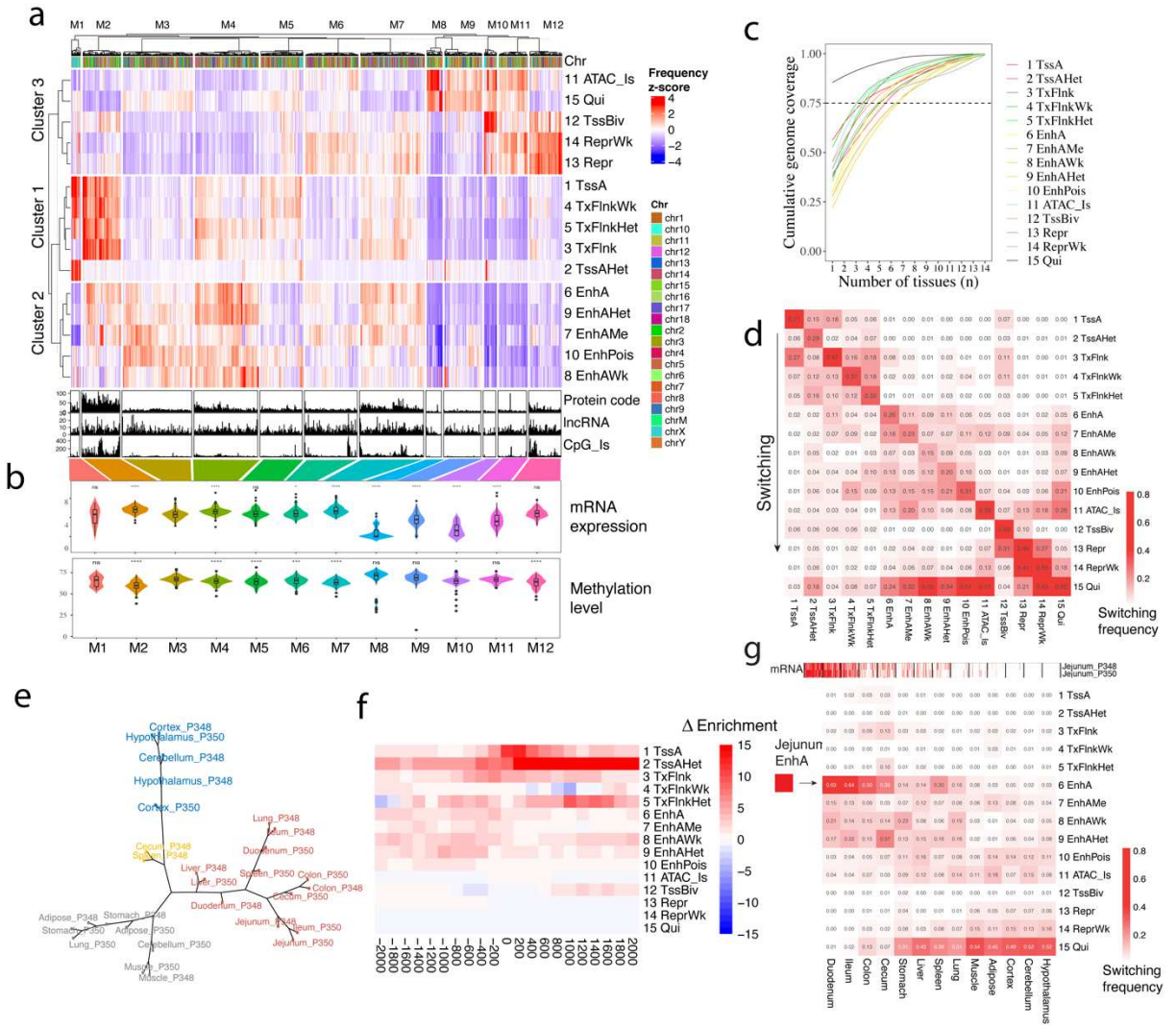


870

871 **Fig.2 Chromatin landscape across 14 tissues.**

872 **a, b,** Definitions and abbreviations of 15 chromatin states. **c,** Emission probabilities of individual
873 epigenetic marks for each chromatin state. The color from white to deep blue indicate emission
874 probability (0-1). **d,** Genomic coverages of each chromatin state. M±SD, means ± standard
875 deviation **e,** Average enrichments of chromatin states for genomic annotations, including CpG
876 islands, genes, TSS/TES_1K (± 1 kb around transcription start/end sites), expressed genes
877 (TPM >=0.1), repressed genes (TPM < 0.1) in each tissue. **f,** Fold enrichments of chromatin states

878 for non-coding mammalian conserved elements from Genomic Evolutionary Rate Profiling
879 (GERP). **g,** Density of each chromatin state in positions relative to gene transcription start sites
880 (TSS). **h,** Average methylation level of chromatin states in jejunum. **i,** Hi-C (250kb resolution),
881 predicted chromatin states, and epigenetic signal, and normalized methylation level landscape in
882 jejunum across the Chr7. **j,** Chromatin state landscapes and mRNA expression at *VIL1* locus
883 (chr15:120,459,825-120,493,312, susScr11) across 14 tissues in pig. Vertical scale UCSC reads
884 signal 0-200 for RNA-seq.
885
886
887
888

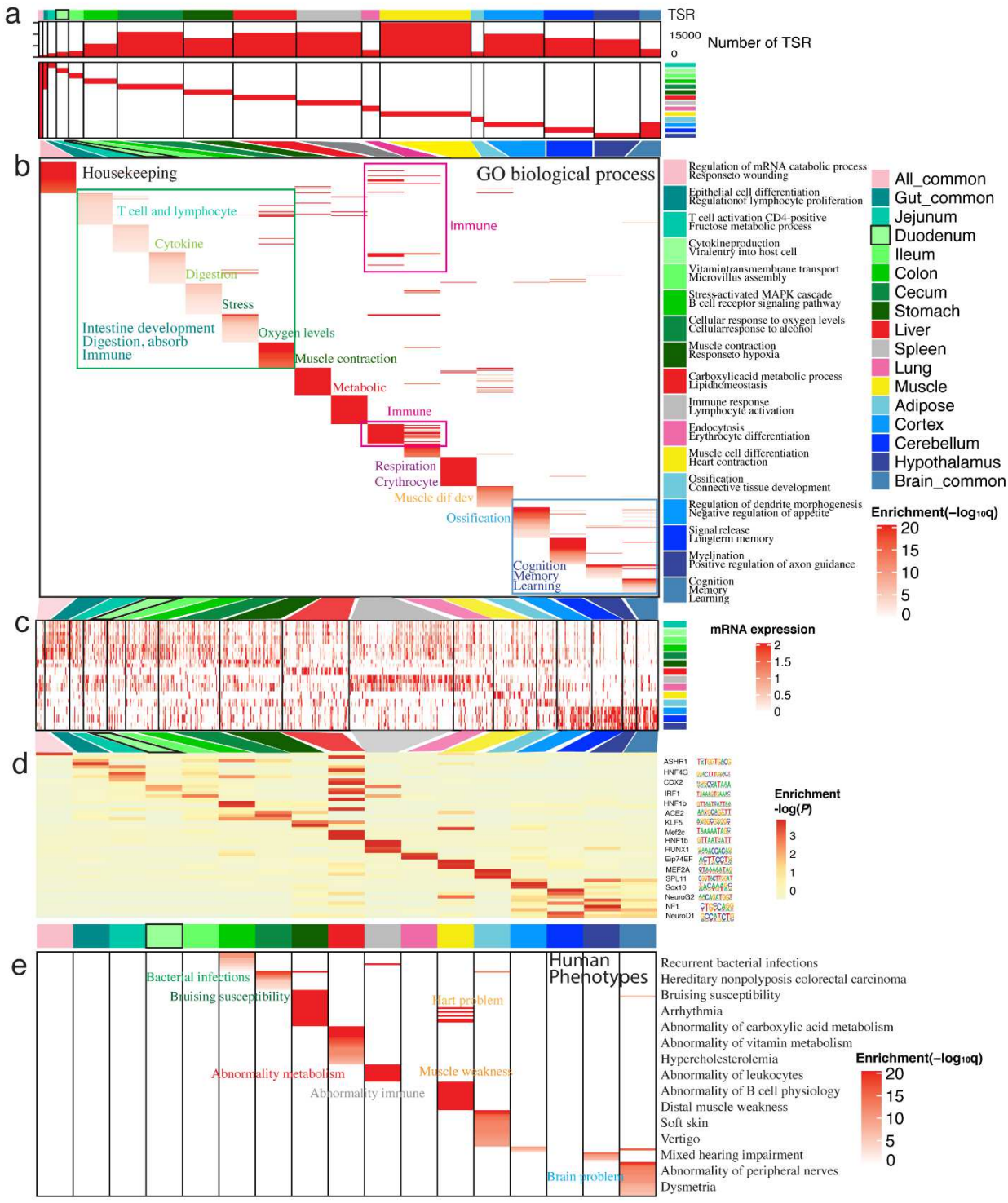


889

890 **Fig.3 Genome-wide chromatin state dynamics across tissues.**

891 **a,** Clustering of 2-Mb intervals (1,224 columns) into modules (M1-M12) based on average
892 chromatin state frequency across tissues in each interval. **b,** Average mRNA expression (TPM) of
893 genes and average methylation level of 2-Mb intervals belonging to each module. The statistical
894 significances for comparisons were calculated using t test, where "*" means $P < 0.05$, "**" means
895 $P < 0.01$, "***" means $P < 0.001$. **c,** Chromatin state variability based on cumulative genome
896 coverage fraction. Dash line = 0.75. **d,** Chromatin state switching between all tissues. **e,**
897 Hierarchical epigenome clustering using H3K4me1 signal in EnhA states. **f,** Chromatin state

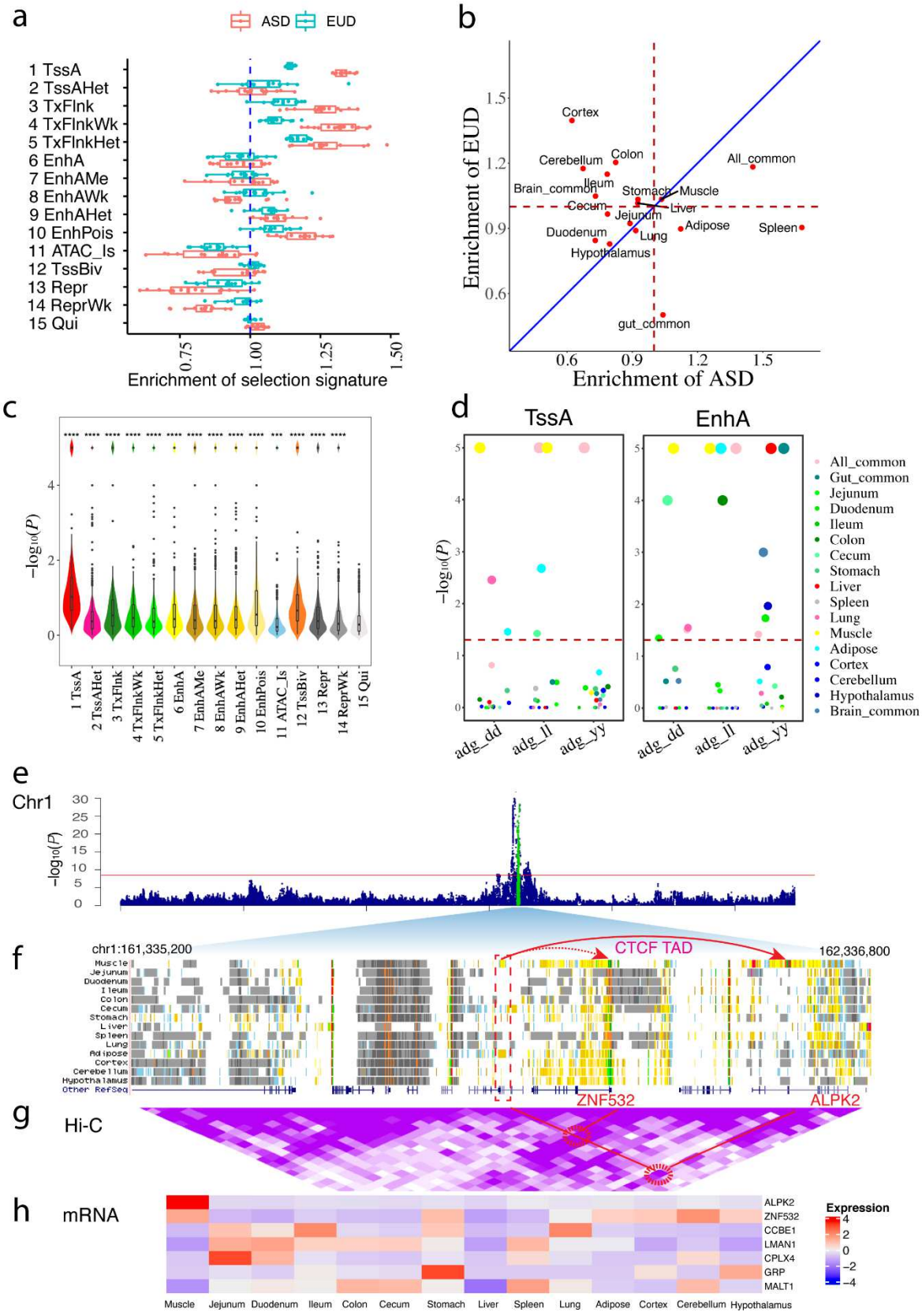
enrichment relative to promoters of genes with jejunum-specific expression genes, relative to muscle. **g,** Chromatin state switching of target enhancers (EnhA) of jejunum-specifically expressed genes in other tissues.



902

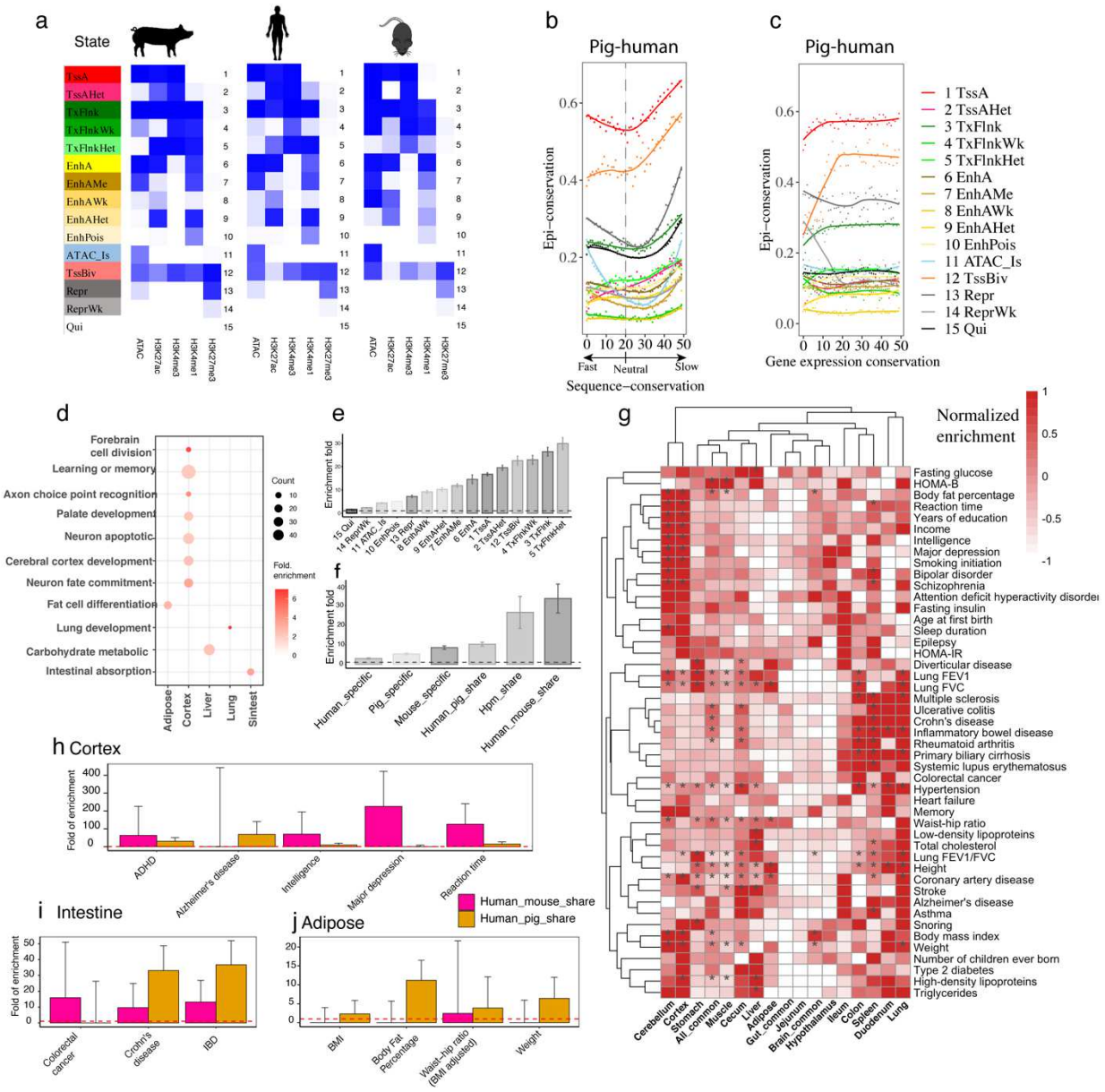
903 **Fig.4 Tissue-specific strong enhancers (EnhA) and their potential functions in 14 tissues.**

a, The number and enrichment distribution of 17 modules of TSR (strong enhancers (EnhA)) in tissues. TSR: tissue-specific regulatory elements. The top colors represent 17 modules of strong enhancers (column) referred to the legend on the right side. The side colors represent 14 tissues (row) referred to the legend on the right side. **b,** Functional enrichment of proximal genes for each module based on gene ontology (GO) biological processes. The columns represent 17 modules of strong enhancers. The rows represent GO terms in each module. All GO terms are presented in Supplementary table 5 (The notes in heatmap are summary function of nearby GO terms enrichment (up-noted from jejunum to spleen, down-noted for lung, muscle and adipose)). **c,** The mRNA expression (TPM) of EnhAs' putative target genes in each module. The columns represent the genes in each module, the rows represent each tissue. **d**, The enrichment of transcription factor motifs in each module. The columns represent 17 modules of EnhAs. The rows represent motif in each module. All enriched motifs are presented in Supplementary Fig. 8a. **e,** Enrichment for human phenotypes in each module, based on proximal genes. The columns represent 17 modules of EnhAs. The rows represent phenotypes in each module. All phenotypes' enrichment are presented in Supplementary table7. The notes in heatmap are summary function of nearby phenotypes enrichment (each color stands for each tissue).


a
ASD
EUD
1 TssA
2 TssAHet
3 TxFlnk
4 TxFlnkWk
5 TxFlnkHet
6 EnhA
7 EnhAMe
8 EnhAWk
9 EnhAHet
10 EnhPois
11 ATAC_Is
12 TssBiv
13 Repr
14 ReprWk
15 Qui
0.75
1.00
1.25
1.50
Enrichment of selection signature
b
Enrichment of EUD
Enrichment of ASD
Cortex
Cerebellum
Colon
Ileum
Brain_common
Stomach
Muscle
Cecum
Jejunum
Liver
All_common
Adipose
Spleen
Duodenum
Lung
Hypothalamus
gut_common
c
−log10(P)
d
TssA
EnhA
adg_dd
adg_ll
adg_yy
All_common
Gut_common
Jejunum
Duodenum
Ileum
Colon
Cecum
Stomach
Liver
Spleen
Lung
Muscle
Adipose
Cortex
Cerebellum
Hypothalamus
Brain_common
e
Chr1
f
chr1:161,335,200
CTCF TAD
162,336,800
ZNF532
ALPK2
g
Hi-C
h
mRNA
Expression
Muscle
Jejunum
Duodenum
Ileum
Colon
Cecum
Stomach
Liver
Spleen
Lung
Adipose
Cortex
Cerebellum
Hypothalamus

920

Fig.5 Chromatin state plays an important role in pig domestication and complex traits. **a,** Domestication selection signature enrichment of chromatin states in Asian and European pigs. ASD: Asian pig domestication. EUD: European pig domestication. Dash-line = 1, above dash line means significant enrichment. **b,** Domestication selection signature enrichment in tissue-specific promoters (TssA) between Asian and European pigs. Dash-line = 1, above 1 dash line means significant enrichment. **c,** Genome-wide association studies (GWAS) signal enrichment within chromatin states across 14 tissues and 44 complex traits in pigs. The statistical significances for comparisons were calculated using a t test, where "***" means $P < 0.001$. **d,** GWAS signal enrichment of promoter (TssA) and strong enhancer (EnhA) tissue-specific regulatory elements (TSR) in average daily gain (adg) of three pig populations (dd: Duroc, ll: Landrace, yy: Yorkshire). Dash line=$-\log_{10}(P=0.05)$, over dash line means significantly high enrichment. **e,** Manhattan plot of average daily gain in the Landrace population. **f,** Chromatin states in the genomic region where GWAS hits for each tissue (dashed rectangle box includes a muscle-specific enhance where SNPs of GWAS hits locate; two arrows in red were predicted from CTCF TAD and H3K27ac signal that suggest the muscle-specific enhancer may target *ZNF532* and *ALPK2*). **g.** Hi-C loop (25kb resolution) depiction between a muscle-specific enhancer and putative target genes. Purple shading for the Hi-C data means loop intensity (auto-scale). Two highlighted Hi-C loop with the red circles are potential contacts between a muscle-specific enhancer and *ZNF532* and *ALPK2*. **h,** Expression (normalized and centered TPM) of genes proximal to the muscle-specific enhancer.



941

942 **Fig.6 Interspecies conservation of chromatin states.**

943 **a,** 15 chromatin states predicated in three species. The colors from white to deep blue indicate

944 emission probability from 0 to1. **b,** Relation between sequence conservation and epigenomic

945 conservation across six tissues. 50 genomic regions were ordered from the fastest changing (0^{th}),

946 to neutral (20^{th}), and to slowest changing (49^{th}) in terms of sequence conservation

947 (Supplementary Fig. 13d). Epigenome conservation (see method) of chromatin states (b) within

these regions was calculated for pig to human. Value in each region for each chromatin state was then plotted. **c,** Relation between expression conservation and epigenomic conservation across six tissues. Expression conservation was based on expression of 14,302 orthologous genes among 3 species. Regions were ordered from the biggest difference in expression (0th), to the smallest difference (49^{th}). **d,** GO enrichment was based on genes proximal to (± 2 kb) human-specific TssA in sequences of extreme conservation sets (49^{th}). Count refers to the number of genes. **e,** Human GWAS signal enrichment in different chromatin states. dash line=1, over dash line means significant high enrichment. Error bars represent standard error around the estimates of enrichment. Same meaning for dash line and error bars in following sub-figures. **f,** Human GWAS enrichment in 6 groups of species-specific or shared EnhA. hpm_share stands for human-pig-mouse shared. **g,** GWAS enrichment of pig tissue-specific enhancer (EnhA) in humans. *means significant enrichment (FDR<0.05). **h,i,j,** Different GWAS enrichments between human-pig and human-mouse shared strong enhancers (EnhA) in brain cortex, small intestine, and adipose, respectively.

Supplementary information

Supplementary Information

Supplementary Figs. 1-16 and Tables 1–13

Figures

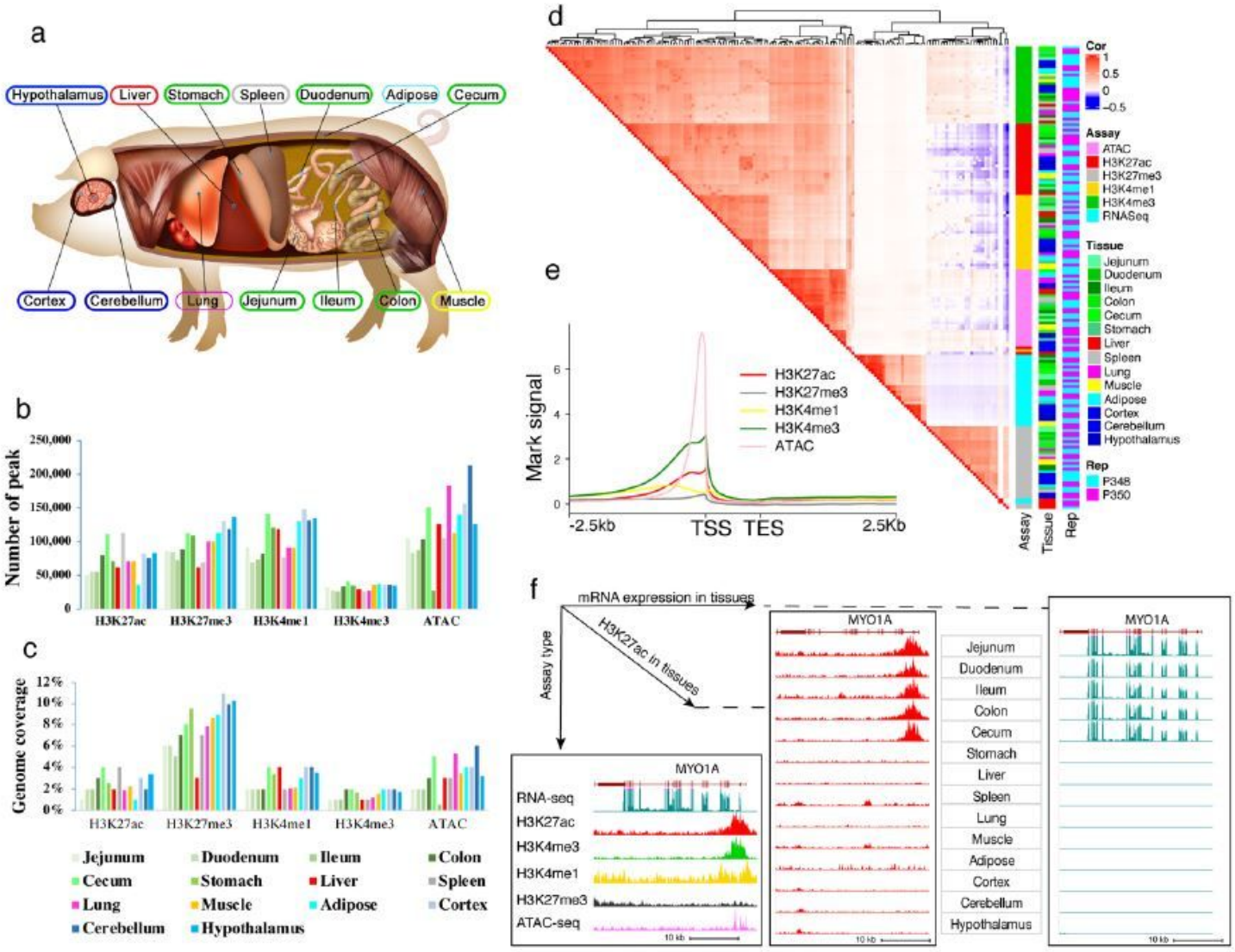


Figure 1

Fig.1 Data summary of epigenomic information across tissues and marks. a, Tissues assayed by this study. b,c, Average peak number and genome coverage for each epigenetic mark in each tissue. d, The Pearson correlations among assays, tissues, and biological replicates (P348 and P350) based on the normalized signal in 1kb windows stepped across the whole genome. e, Average epigenetic mark signals proximal to protein coding genes. TSS, transcription start site. TES, transcription end site. f, Epigenetic signal at the MYO1A locus according to different assays and in different tissues. Vertical scale UCSC reads signal 0-200 for RNA-seq, 0-100 for H3K27ac and H3K4me3, 0-50 for other marks and ATAC-seq.

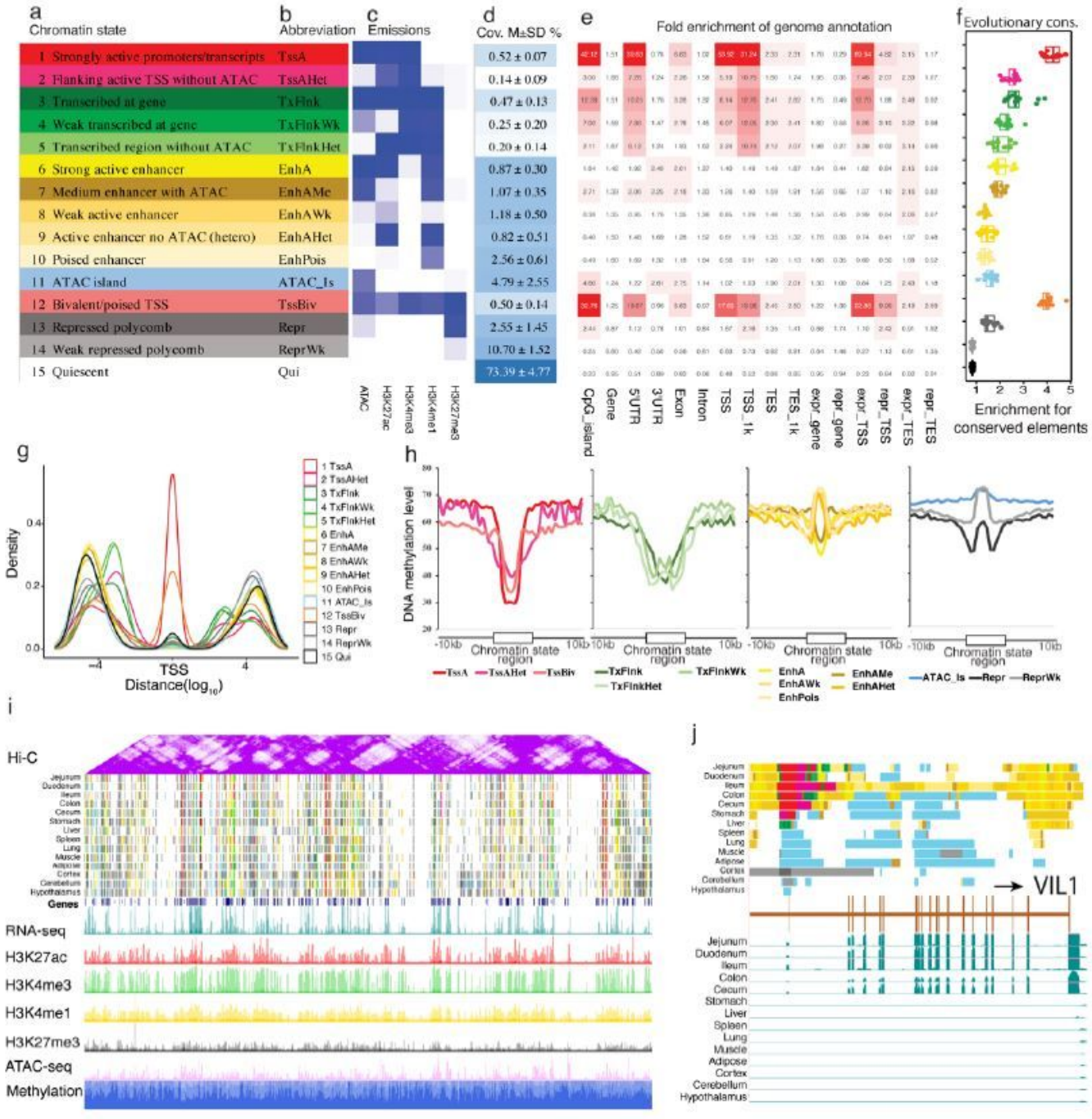


Figure 2

Chromatin landscape across 14 tissues. a, b, Definitions and abbreviations of 15 chromatin states. c, Emission probabilities of individual epigenetic marks for each chromatin state. The color from white to deep blue indicate emission probability (0-1). d, Genomic coverages of each chromatin state. M±SD, means ± standard deviation e, Average enrichments of chromatin states for genomic annotations, including CpG islands, genes, TSS/TES_1K (± 1 kb around transcription start/end sites), expressed genes (TPM >=0.1), repressed genes (TPM < 0.1) in each tissue. f, Fold enrichments of chromatin states 42 for non-coding mammalian conserved elements from Genomic Evolutionary Rate Profiling (GERP). g, Density of each chromatin state in positions relative to gene transcription start sites (TSS). h, Average

methylation level of chromatin states in jejunum. i, Hi-C (250kb resolution), predicted chromatin states, and epigenetic signal, and normalized methylation level landscape in jejunum across the Chr7. j, Chromatin state landscapes and mRNA expression at VIL1 locus (chr15:120,459,825-120,493,312, susScr11) across 14 tissues in pig. Vertical scale UCSC reads signal 0-200 for RNA-seq.

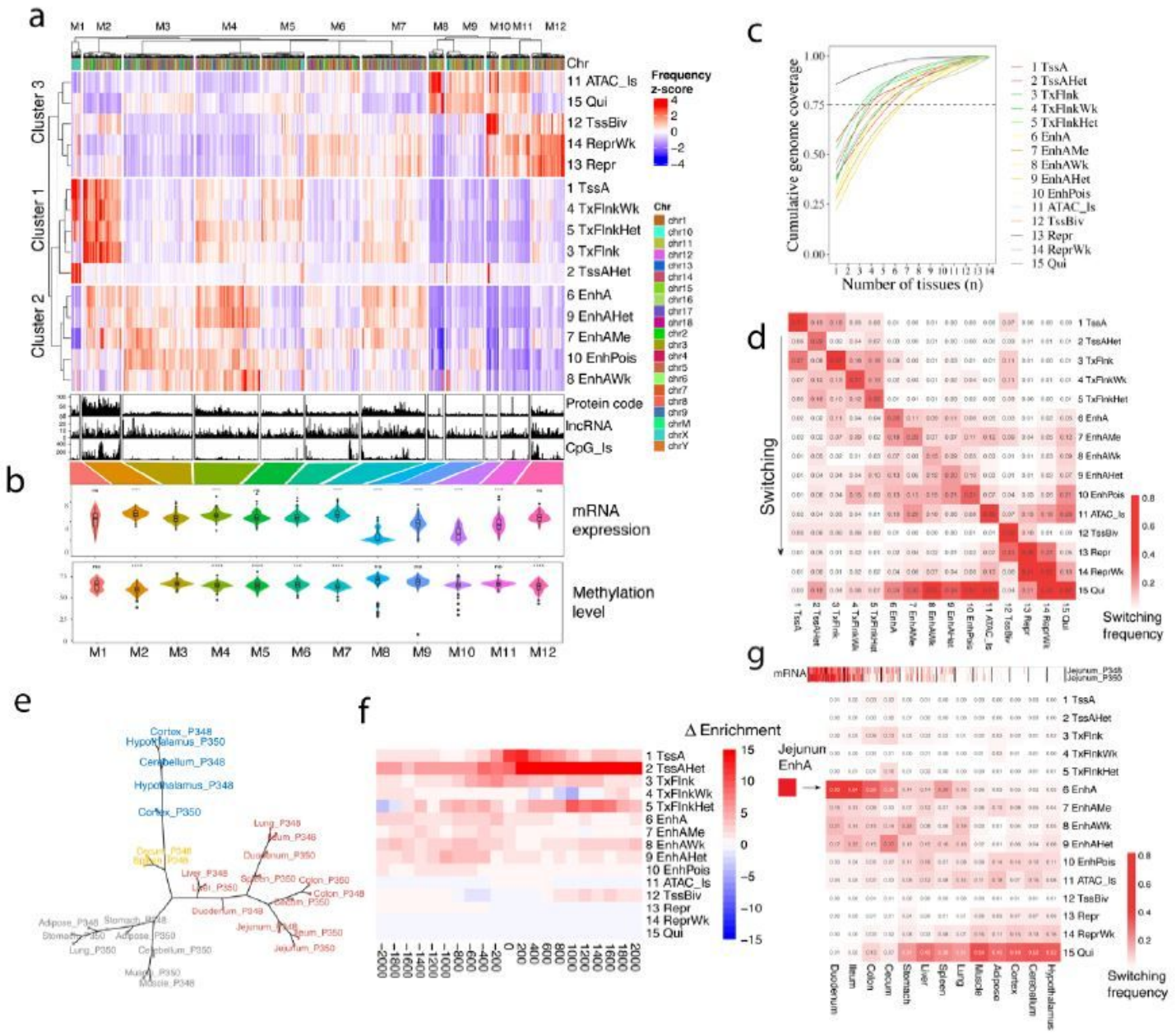


Figure 3

Genome-wide chromatin state dynamics across tissues. a, Clustering of 2-Mb intervals (1,224 columns) into modules (M1-M12) based on average chromatin state frequency across tissues in each interval. b, Average mRNA expression (TPM) of genes and average methylation level of 2-Mb intervals belonging to each module. The statistical significances for comparisons were calculated using t test, where “*” means P < 0.05, “**” means P < 0.01, “***” means P < 0.001. c, Chromatin state variability based on cumulative genome coverage fraction. Dash line = 0.75. d, Chromatin state switching between all tissues. e,

Hierarchical epigenome clustering using H3K4me1 signal in EnhA states. f, Chromatin state 44 enrichment relative to promoters of genes with jejunum-specific expression genes, relative to muscle. g, Chromatin state switching of target enhancers (EnhA) of jejunum-specifically expressed genes in other tissues.

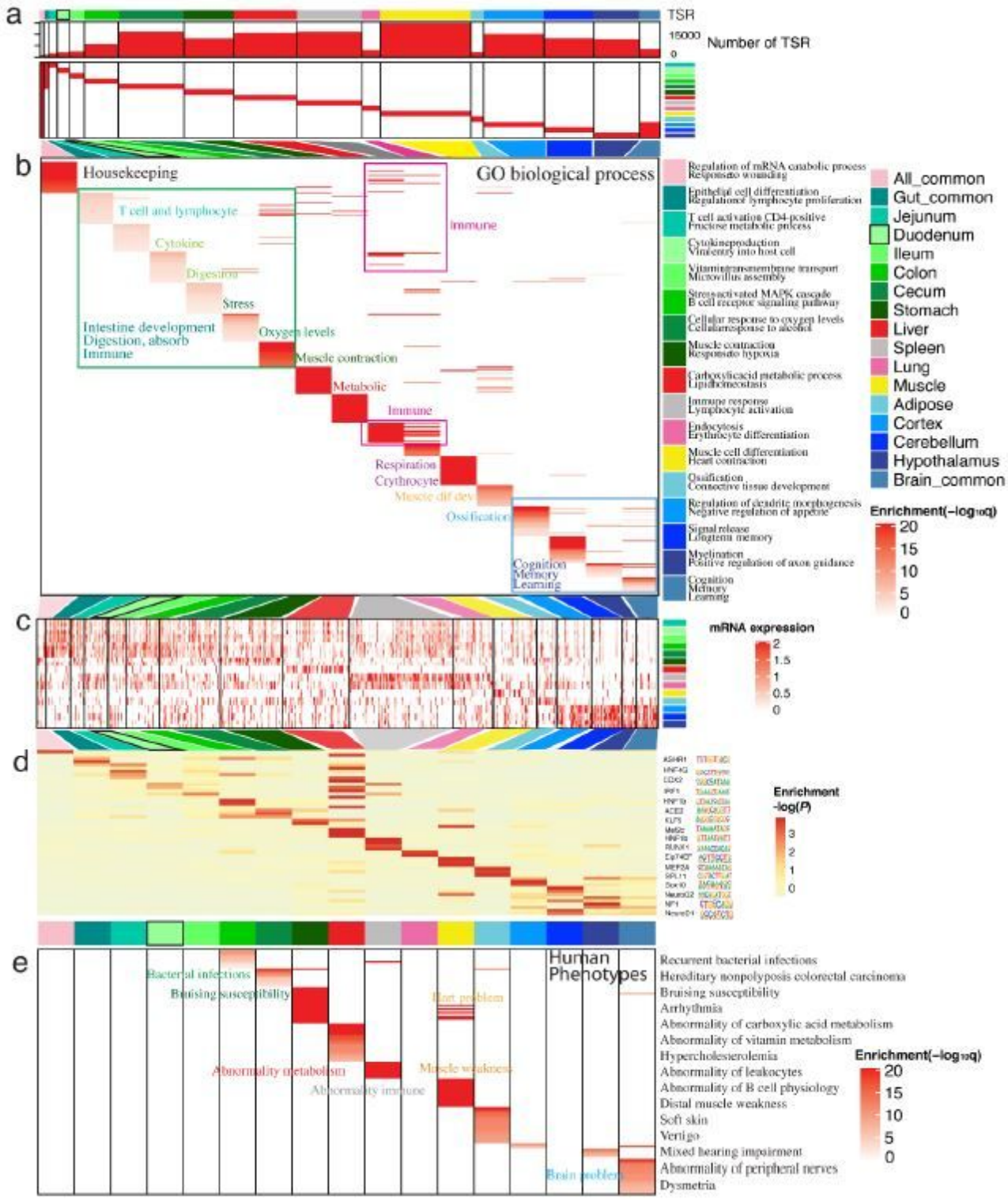


Figure 4

Tissue-specific strong enhancers (EnhA) and their potential functions in 14 tissues. 46 a, The number and enrichment distribution of 17 modules of TSR (strong enhancers (EnhA)) in tissues. TSR: tissue-specific regulatory elements. The top colors represent 17 modules of strong enhancers (column) referred to the legend on the right side. The side colors represent 14 tissues (row) referred to the legend on the right side. b, Functional enrichment of proximal genes for each module based on gene ontology (GO) biological processes. The columns represent 17 modules of strong enhancers. The rows represent GO terms in each module. All GO terms are presented in Supplementary table 5 (The notes in heatmap are summary function of nearby GO terms enrichment (up-noted from jejunum to spleen, down-noted for lung, muscle and adipose)). c, The mRNA expression (TPM) of EnhAs' putative target genes in each module. The columns represent the genes in each module, the rows represent each tissue. d, The enrichment of transcription factor motifs in each module. The columns represent 17 modules of EnhAs. The rows represent motif in each module. All enriched motifs are presented in Supplementary Fig. 8a. e, Enrichment for human phenotypes in each module, based on proximal genes. The columns represent 17 modules of EnhAs. The rows represent phenotypes in each module. All phenotypes' enrichment are presented in Supplementary table7. The notes in heatmap are summary function of nearby phenotypes enrichment (each color stands for each tissue).

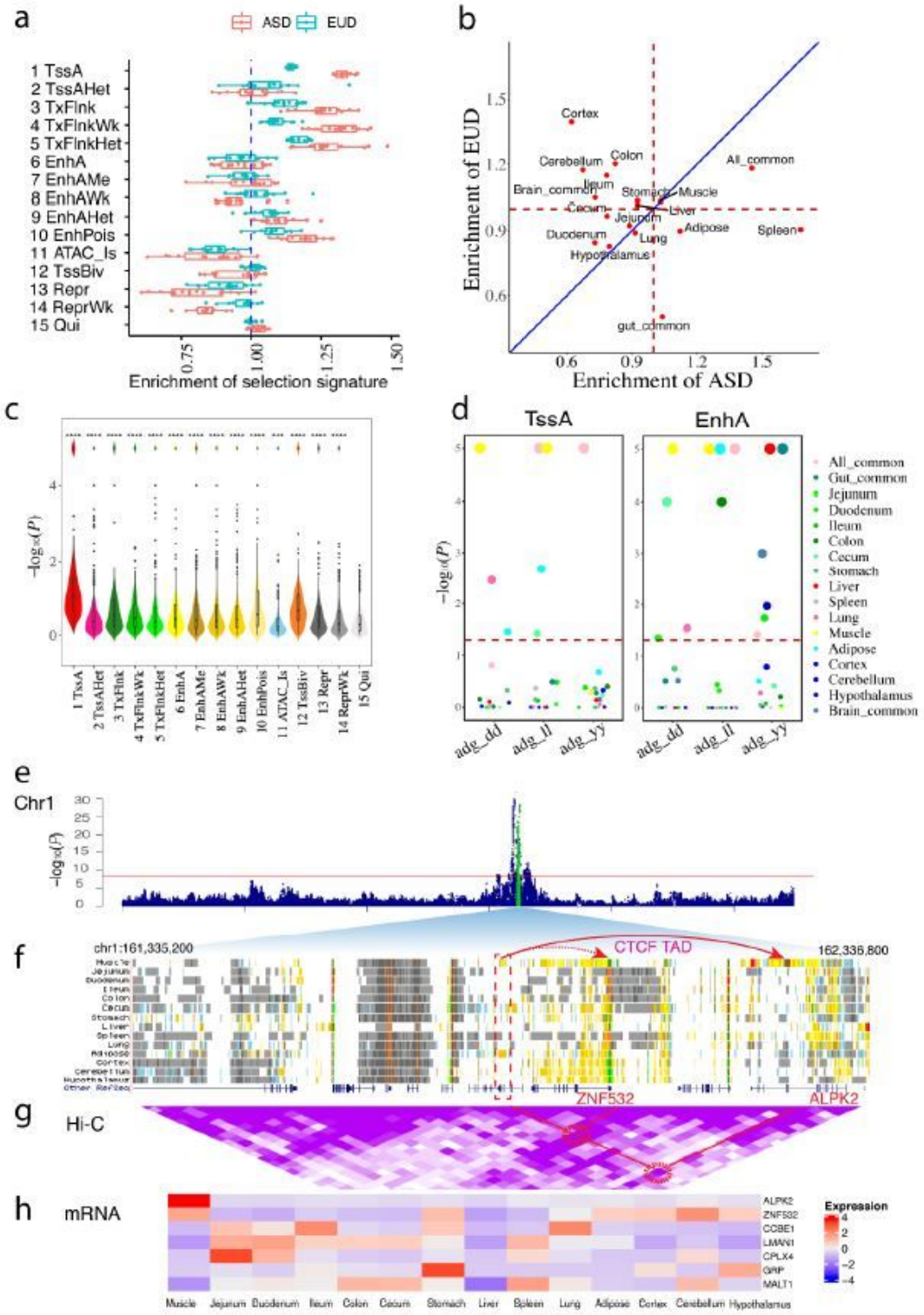


Figure 5

Chromatin state plays an important role in pig domestication and complex traits. a, Domestication selection signature enrichment of chromatin states in Asian and European pigs. ASD: Asian pig domestication. EUD: European pig domestication. Dash-line = 1, above dash line means significant enrichment. b, Domestication selection signature enrichment in tissue-specific promoters (TssA) between Asian and European pigs. Dash-line = 1, above 1 dash line means significant enrichment. c, Genome-wide

association studies (GWAS) signal enrichment within chromatin states across 14 tissues and 44 complex traits in pigs. The statistical significances for comparisons were calculated using a t test, where “***” means P < 0.001. d, GWAS signal enrichment of promoter (TssA) and strong enhancer (EnhA) tissue-specific regulatory elements (TSR) in average daily gain (adg) of three pig populations (dd: Duroc, ll: Landrace, yy: Yorkshire). Dash line=-log10(P=0.05), over dash line means significantly high enrichment. e, Manhattan plot of average daily gain in the Landrace population. f, Chromatin states in the genomic region where GWAS hits for each tissue (dashed rectangle box includes a muscle-specific enhance where SNPs of GWAS hits locate; two arrows in red were predicted from CTCF TAD and H3K27ac signal that suggest the muscle-specific enhancer may target ZNF532 and ALPK2). g. Hi-C loop (25kb resolution) depiction between a muscle-specific enhancer and putative target genes. Purple shading for the Hi-C data means loop intensity (auto-scale). Two highlighted Hi-C loop with the red circles are potential contacts between a muscle-specific enhancer and ZNF532 and ALPK2. h, Expression (normalized and centered TPM) of genes proximal to the muscle-specific enhancer.

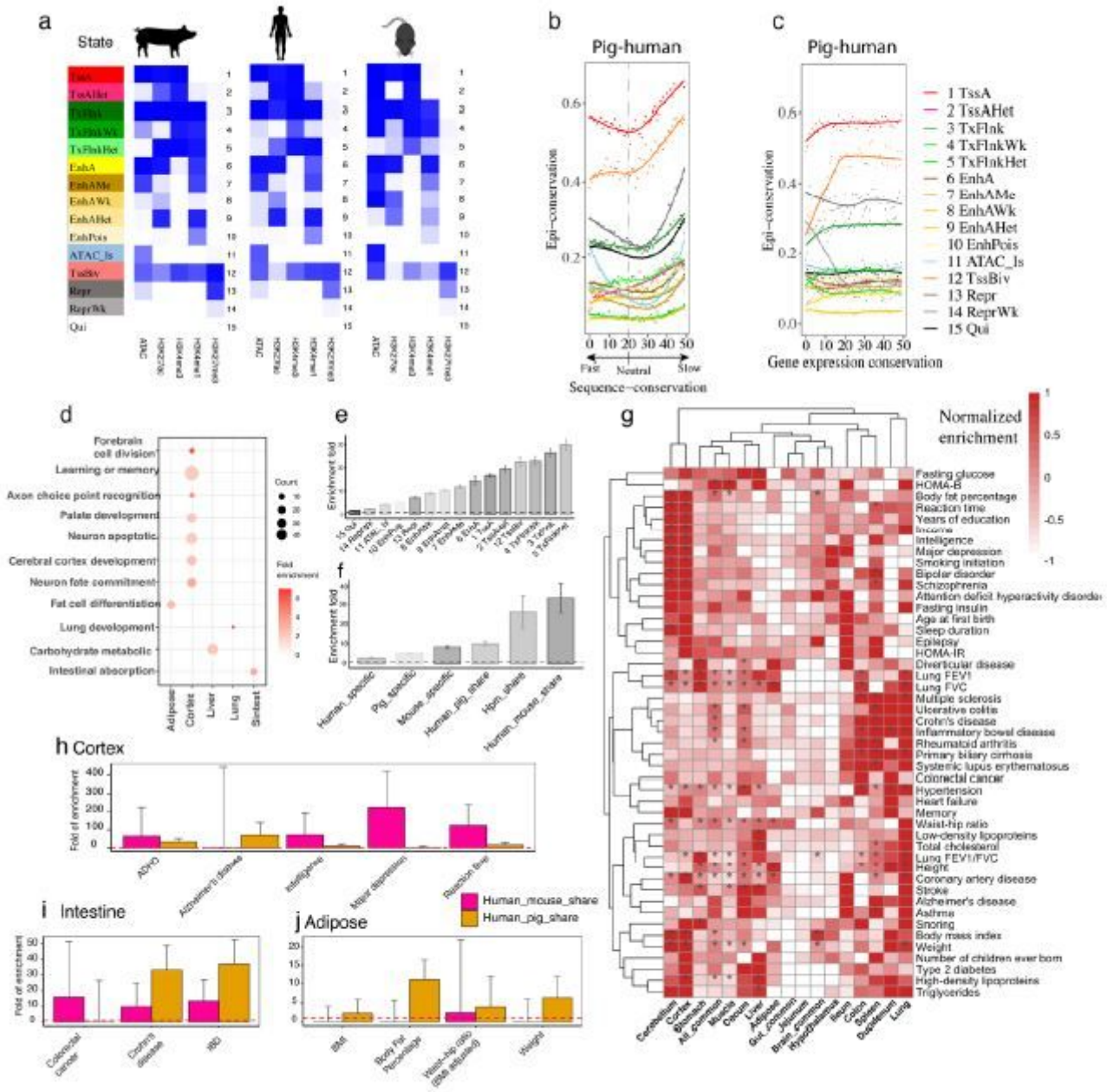


Figure 6

Interspecies conservation of chromatin states. a, 15 chromatin states predicated in three species. The colors from white to deep blue indicate emission probability from 0 to1. b, Relation between sequence conservation and epigenomic conservation across six tissues. 50 genomic regions were ordered from the fastest changing (0th), to neutral (20th), and to slowest changing (49th) in terms of sequence conservation (Supplementary Fig. 13d). Epigenome conservation (see method) of chromatin states (b) within 50 these regions was calculated for pig to human. Value in each region for each chromatin state was then plotted. c, Relation between expression conservation and epigenomic conservation across six tissues. Expression conservation was based on expression of 14,302 orthologous genes among 3 species. Regions were ordered from the biggest difference in expression (0th), to the smallest difference (49th). d, GO enrichment was based on genes proximal to (± 2 kb) human specific TssA in sequences of

extreme conservation sets (49th). Count refers to the number of genes. e, Human GWAS signal enrichment in different chromatin states. dash line=1, over dash line means significant high enrichment. Error bars represent standard error around the estimates of enrichment. Same meaning for dash line and error bars in following sub-figures. f, Human GWAS enrichment in 6 groups of species-specific or shared EnhA. hpm_share stands for human958 pig-mouse shared. g, GWAS enrichment of pig tissue-specific enhancer (EnhA) in humans. *means significant enrichment (FDR<0.05). h,i,j, Different GWAS enrichments between human pig and human-mouse shared strong enhancers (EnhA) in brain cortex, small intestine, and adipose, respectively.

Supplementary Files

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

- 3SupplementaryFig.pdf
- 4SupplementaryTable.xlsx