Building an atlas of transposable elements reveals the extensive roles of young SINE in gene regulation, genetic diversity, and complex traits in pigs

Pengju Zhao[1,2], Lihong Gu[3], Yahui Gao[4], Zhangyuan Pan[5], Lei Liu[6], Xingzheng Li[6], Huaijun Zhou[5], Dongyou Yu[1,2], Xinyan Han[1,2], Lichun Qian[1,2], George E. Liu[4]*, Lingzhao Fang[7]*, Zhengguang Wang[2]*

[1]Hainan Institute, Zhejiang University, Yongyou Industry Park, Yazhou Bay Sci-Tech City, Sanya 572000, China.

[2]College of Animal Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China.

[3]Institute of Animal Science & Veterinary Medicine, Hainan Academy of Agricultural Sciences, No.14 Xingdan Road, Haikou 571100, China.

[4]Animal Genomics and Improvement Laboratory, Beltsville Agricultural Research Center, Agricultural Research Service, USDA, Beltsville, Maryland 20705, USA.

[5]Department of Animal Science, University of California, Davis, California, 95616, USA.

[6]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China.

[7]MRC Human Genetics Unit at the Institute of Genetics and Cancer, The University of Edinburgh, Edinburgh EH4 2XU, UK.

Email addresses:

Pengju Zhao: zhaopengju2014@gmail.com; Lihong Gu: nil2008@yeah.net;Yahui Gao: gyhalvin@gmail.com; Zhangyuan Pan: zhypan@ucdavis.edu; Lei Liu: liulei03@caas.cn; Xingzheng Li: lxz2019@alu.cau.edu.cn; Huaijun Zhou: hzhou@ucdavis.edu; Dongyou Yu: dyyu@zju.edu.cn; Xinyan Han: xyhan@zju.edu.cn; Lichun Qian: lcqian@zju.edu.cn; George E. Liu: george.liu@usda.gov; Lingzhao Fang: Lingzhao.fang@ed.ac.uk; Zhengguang Wang: wzhguang68@zju.edu.cn

* Corresponding author:

Zhengguang Wang, Ph.D.
Zhejiang University
College of Animal Sciences
866 Yuhangtang Rd, Hangzhou 310058, P.R. China
E-mail: wzhguang68@zju.edu.cn

Lingzhao Fang, Ph.D.
The University of Edinburgh
MRC Human Genetics Unit at the Institute of Genetics and Cancer
Edinburgh EH4 2XU, UK
E-mail: Lingzhao.fang@ed.ac.uk

George E. Liu, Ph.D.
Agricultural Research Service, USDA
Animal Genomics and Improvement Laboratory, Beltsville Agricultural Research Center
Beltsville, Maryland 20705, USA.
E-mail: george.liu@usda.gov

Abstract

Transposable elements (TEs) are an extensive source of genetic polymorphisms and play an indispensable role in chromatin architecture, transcriptional regulatory networks, and genomic evolution. The pig is an important source of animal protein and serves as a biomedical model for humans, yet the functional role of TEs in pigs and their contributions to complex traits are largely unknown. Here, we built a comprehensive catalog of TEs (n = 3,087,929) in pigs by a newly developed pipeline. Through integrating multi-omics data from 21 tissues, we found that SINEs with different ages were significantly associated with genomic regions with distinct functions across tissues. The majority of young SINEs were predominantly silenced by histone modifications, DNA methylation, and decreased accessibility. However, the expression of transcripts that were derived from the remaining active young SINEs exhibited strong tissue specificity through cross-examining 3,570 RNA-seq from 79 tissues and cell types. Furthermore, we detected 211,067 polymorphic SINEs (polySINEs) in 374 individuals genome-wide and found that they clearly recapitulated known patterns of population admixture in pigs. Out of them, 340 population-specific polySINEs were associated with local adaptation. Mapping these polySINEs to genome-wide associations of 97 complex traits in pigs, we found 54 candidate genes (e.g., *ANK2* and *VRTN*) that might be mediated by TEs. Our findings highlight the important roles of young SINEs in functional genomics and provide a supplement for genotype-to-phenotype associations and modern breeding in pigs.

Keywords: Transposable elements; Regulatory networks; Polymorphic SINEs; Pig

Introduction

Transposable elements (TEs) or common repeats are ubiquitous sequences that can copy and insert themselves throughout the eukaryotic and prokaryotic genome[1-7]. The movement of TEs is often accompanied by an increase in their abundance, comprising a large fraction of genomic sequence[8-9]. According to the mechanism of transposition, TEs can be generally classified into 1) RNA-mediated class I elements (retrotransposons), including long terminal repeats (LTRs), long interspersed nuclear elements (LINEs), and short interspersed nuclear elements (SINEs); and 2) RNA-independent class II elements (DNA transposons)[10]. TE classes could be further divided into distinct families or subfamilies in terms of their age (active period) and DNA sequence characteristics.

At the predominant view of the 1960s -1990s, TEs were described as “selfish” or “junk” DNA[11]. Thanks to the availability of whole-genome sequences of various species and the ongoing development of bioinformatics tools[12-15], our knowledge of TEs has progressed at a fast pace. TEs are known to play an essential role in shaping genomic sequences and contributing to the diversity in genome size and chromosome structure[1,16-17]. Most of TEs, in fact, are fixed, inactive, and not randomly distributed in the genome[18-19]. However, several TE families are still actively transposing and serving as a major source of genetic polymorphisms between individuals, such as the Alu, L1, and SVA TE families in the human genome[20]. It is evident in many species (e.g., human, rice and bird) the impacts of active TEs on genome evolution are wide-ranged, including admixture, adaption, footprints of selection, and population structure[21-24]. For example, the polymorphic TEs (polyTEs) detected in the 1000 Genomes Project, consisting of 16,192 loci in 2,504 individuals across 26 human populations, successfully recapitulated the human evolution and captured the sign for positive selection on recent human TE insertions[20,25-26].

In addition to their direct influences on DNA sequence, there is also emerging evidence that TEs have important functional contributions to gene regulatory networks and epigenome variation. For instance, TEs can directly affect gene transcriptional structure by provoking various forms of alternative splicing, including exonization, exon skipping, and intron retention (3’S and 5’S), to generate novel protein-coding sequences or premature ends[27-30]. TEs can disrupt the existing *cis*-regulatory elements, e.g., promoter, enhancer, and insulator, or provide novel ones[31-34]. They can also serve as a rich source of non-coding RNAs, including LncRNAs, circRNA, small RNA, and microRNA targets[35-38]. Moreover, the silence of TEs has a close connection with epigenetic regulatory mechanisms, such as DNA methylation, piRNA, histone modifications, and RNA interference (RNAi)[18,39-42]. These epigenomic landscapes, together with the TE landscapes, vary from breed to breed in plants, e.g., angiosperms[43]. Importantly, it has been reported that the complex interactions between TEs and epigenetic elements could allow a rapid phenotypic adaptation to environmental

changes[40,42,44].

Pig (*Sus scrofa*), one of the earliest domestic animals, is estimated to be domesticated approximately 10,000 years ago in Asia and Europe independently[45]. It serves as an indispensable source of animal protein and an important biomedical model for humans[46-47]. Currently, a total of 22 pig assemblies are publicly available in NCBI[48-52], accompanied by the availability of massive high-throughput whole-genome sequences (WGS), providing researchers with ideal materials to boost the current development of genomic research in pigs. However, the study of TEs in the pig genome is still in its infancy. A few previous studies mainly focused on its diversity and distribution[48-52], yet the functional and evolutionary significance of TEs is largely overlooked in pigs. In our recent study[53], we identified novel introgressions in Eurasian boars from Asian and European pig populations using the SINE (PRE-1 subfamily) polymorphisms, suggesting that a part of TEs are still active in the current pig genome. Nevertheless, these studies are far from sufficient to comprehensively understand the important roles of TEs in pigs.

In this study, we built the most comprehensive and high-quality atlas of TEs so far in pigs using the newly built pipelines and further defined the SINE families into four categories based on their ages. We then systematically explored the functional aspects of these SINE categories by combining large-scale multi-omics data from 21 tissues, including three-dimensional (3D) chromatin architecture, chromatin accessibility, histone modifications, transcription factor binding sites (TFBS), and DNA methylation. We estimated the contribution of active SINEs to tissue-specific gene expression by cross-examining 3,570 published RNA-seq samples from 52 tissues and 27 cell types. Furthermore, we built a comprehensive atlas of polymorphic SINEs using 374 publicly available WGS data to investigate the roles of young SINEs in pig population admixture and local adaptation. The TE-mediated adaptation has been found in functional regions, such as the almost fixed polySINEs observed in laboratory-inbred Bama Xiang pigs at the upstream region of the *LEP* gene. Finally, by mapping these polySINEs to 4,072 loci associated with 97 complex traits in pigs, we proposed 54 candidate genes that might regulate complex traits through TEs.

Results

Composition of young SINE families in the pig genome

To detect TEs as thoroughly as possible, we developed the **Pig TE D**etection and **C**lassification (PigTEDC) pipeline (**Fig. 1a**, Supplementary Information) and applied it to the high-quality pig genome (*Sus scrofa* 11.1). The PigTEDC pipeline employed a combination of similarity-, structure-, and *de novo*-based methods. Based on the existing TE repository (RepBase update and Dfam 2.0 databases), we further classified all potential pig TEs into classes/superfamilies and families and derived their consensus sequences.

Excluding the nested TEs, we detected a total of 3,087,929 TEs, occupying 37.9% (947 MB) of the total pig genome. Two-thirds of those TEs were assigned to a specific family. Similar to previous studies[48-52], the most abundant TEs in the pig genome were retrotransposons (~90% of TEs), including LTR (9.25%), LINE (27.57%), and SINE (54.95%), whereas DNA transposons only accounted for 8.12% of TEs (**Fig. 1b**). Although SINE had the largest count, its genome coverage was still less than LINE's (**Fig. 1c**). Out of 532 families of TE, 65 (more than 3,000 TE copies in each family) consisted of 84.6% classified TEs (**Fig. 1d**), particularly for PRE1f in SINE/tRNA (170,511 copies), MIR in SINE/MIR (45,927 copies), L1B-SSc in LINE/L1 (35,819 copies), and MLT1D in LTR/ERVL-MaLR (7,866 copies).

We performed the divergence analysis for all classified TEs using RepeatMasker after the CpG content correction. The stacking plots show the divergence distribution for either superfamilies or families (**Fig. 1e**). Similar to the divergence distribution of TEs in the human genome[54], we observed two bursts at 10% and 30% in pig TE amplification that were estimated to occur 20 and 60 million years ago (Mya), assuming a substitution rate of 5×10^{-9} substitutions/site per year[55]. Obviously, the amplification of most TE families occurred 70-50 Mya (divergence at 30 ± 5%), consistent with that the Paleocene Epoch (65-54 Mya) opened up vast ecological niches for surviving mammals, birds, reptiles, and marine animals[56]. The latest obvious burst of TEs was mainly concerned with SINE/tRNA, LINE/L1, and LTR/ERV1 families, of which SINE/tRNA was still most active in the modern pig genome[57-59]. Further exploring the ages of highly homologous subfamilies in SINE classes (**Fig. 1f** with an average divergence of 4%, labelled in purple), we found that 3 out of 26 SINE families (PRE1-SS, PRE0-SS, and PRE1a) were recently active and had been proved to be polymorphic within pig breeds in a previous study[53], and thus can be viewed as the young SINE families.

We next focused on these young SINE families (PRE1-SS, PRE0-SS, and PRE1a) to further reclassify them into subfamilies at a high resolution. After processing all of the full-length young SINEs from 14 publicly available pig genomes (Supplementary Information), we retained 978,506 non-redundant young SINEs and created their consensus sequence by multiple sequence alignment (**Fig. S1-S2**). Subsequently, a minimum spanning (MS) tree analysis recategorized them into 90 new SINE subfamilies, including 17 large (size > 10000), 12 medium (size ≥ 3000 and ≤ 10,000), and 61 small (size < 3000) subfamilies, with *P*-values for subfamily partition ranging from 1×10^{-53186} to 1×10^{-4} (**Fig. S3-S4**, Table S1).

When we compared locations of the polymorphic young SINEs with those of medium-length structure variants (SVs, range from 200 bp to 300 bp) detected from the 14 assemblies as compared to the pig reference assembly (Supplementary Information), we

found that polymorphic SINEs mainly belonged to PRE1-SS, PRE1a, and PRE0-SS families (**Fig. 1g**) (on average 90.75% of the medium-length SVs), especially for the L13 subfamily (**Fig. 1h**) (on average 36.15% of the medium-length SVs). This suggests that only a specific set of recently active SINE subfamilies was predominant in contributing to SVs (around 250 bp length) among diverse breeds during the recent evolution in pigs. We therefore further classified all SINEs into four categories of youngest (L13 subfamilies), younger (young SINE families except for L13 subfamilies, non-L13), older (non-young PRE families), and oldest (non-PRE families; ancient families) SINEs. (**Fig. 1i**, **Fig. S5**).

Widespread roles of young SINEs in gene regulatory networks

Previous studies proposed that TEs might be co-opted into regulatory sequences of genes *via* diverse epigenetic mechanisms[60-63]. To test this, we explored the impact of SINE subfamilies on the genome features regarding three-dimensional (3D) chromatin architecture (FR-AgENCODE[64]), chromatin accessibility, histone modifications, transcription factor binding sites (TFBS), and DNA methylation (**Fig. 2a**).

We observed a highly significant enrichment of all SINEs in the A compartments (active), whereas a depletion in the B compartments (inactive) (Wilcoxon test, *P*-values < 10^{-16}, **Fig. 2b, Fig. S6**). After the further separation of A/B compartments into topologically associating domains (TADs), as expected, we observed CTCF binding sites were significantly enriched in the boundary regions of TADs. SINEs showed a similar but slightly weaker trend of enrichment, while the young SINEs showed a higher but more variable enrichment than the old ones (**Fig. 2b, Fig. S7**).

Next, we explored the distribution of SINE families on the chromatin accessibility and nucleosome positioning near transcripts using the published ATAC-seq (14 tissues) and MNase-seq (five tissues) datasets, respectively[65-66]. TE enrichment in open and closed chromatin exhibited strong age-specific patterns, mainly reflecting that the youngest and younger SINE families were significantly depleted from open chromatin regions and enriched near the nucleosome (**Fig. 2c**). Whereas the older SINE families showed relatively high enrichment for open chromatin, particularly in the stomach, followed by adipose and cerebellum.

We further explored the relationship of SINE families with four active epigenetic marks (H3K4me1 - primed enhancers, H3K4me3 - enriched in transcriptionally active promoters, H3K27ac - distinguishes active enhancers from poised enhancers, and H3K36me3 - actively transcribed gene bodies) and two repressive marks (H3K9me3 - constitutively repressed genes and H3K27me3 - facultatively repressed genes). In **Fig. 2c**, we found that the majority of SINE families were underrepresented in all four active marks, consistent across tissues. Whereas SINE families, particularly for young SINEs, were highly enriched for H3K9me3

but not for H3K27me3. H3K27me3 has development-dependent repressed characteristics, while H3K9me3 indicates permanent repression[67]. In general, compared to older or oldest SINE groups, younger and youngest SINE groups (especially the youngest group) showed a much higher over-/under-representation for all types of histone modifications. In addition, we further investigated the enrichment of SINE families in 15 distinct chromatin states across 14 tissues (**Fig. 2d**, Supplementary Information)[68]. We observed that young SINE groups (youngest and younger) show lower enrichment in most of the functional chromatin states than old SINE groups (older and oldest), and the enrichment degree of SINE in chromatin states was roughly similar across 14 tissues. These enrichment characteristics can be further divided into four distinct enrichment patterns according to the enrichment degree of different SINE groups (**Fig. 3a)**. We observed that the oldest SINE's highly enriched pattern accounted for the vast majority (80%, 168 of 210 combinations of 14 tissues and 15 states), while the young SINE group showed the three enrichment patterns in remaining combinations (the enlarged inset). In two of them, only the youngest SINE group showed high enrichment for TssAHet (flanking active TSS without ATAC) and EnhAWk (weakly active enhancer)**.** In general, all SINE groups were significantly depleted from active promoters and enhancers, except for weak TSS and enhancers, and the young SINE groups showed a higher depletion than old ones. All these together indicated that young SINEs as new invaders might be silenced by histone modifications and DNA methylation, while the old SINEs might be tolerated by the pig genome.

It has been proposed that TEs carrying TFBS repertoire may indirectly contribute to the transcriptional regulation of genes[69-70]. We thus performed the motif enrichment analysis of SINEs to determine their possible contributions (**Fig. 3b**). In total, 31 known TFBS were predicted to have the binding motifs in at least one SINE family, 30 of which (96.8%) were found in old ones and 26 (83.9%) were related to the open chromatin, revealing that the recent exaptation of young SINEs into regulatory regions was relatively rare[71] and was repressed by less chromatin accessibility. The youngest SINE-specific TFBS related to the *ZNF148* gene has been proven to drive the formation of a muscle phenotype[72]. Besides, three members of the RFX transcription factor family were observed to be amplified in young and older SINEs and involved in the immune, reproductive, and nervous pathways[73]. For example, RFX1 and RFX3 were the candidate major histocompatibility complex (MHC) class II promoter binding proteins that were found to function as a *trans*-activator of the hepatitis B virus enhancer[74-75].

Given that TEs may play major roles in the regulation of gene expression by shaping the epigenetic modifications[76], we further assessed the epigenetic states of SINE families in terms of DNA methylation (MeDIP), the density of CG (CpG) sequence contexts, and AT:GC content (CpG islands) (**Fig. 3c**; Supplementary Information). The results showed that almost all SINE families exhibited a significant depletion of CpG islands, and the young SINE

families were more highly methylated than the old ones. Similarly, we observed that the average CG methylation levels within SINE bodies, especially for the young SINEs, were significantly higher than their flanking regions across 10 different tissues, which corresponded to the enrichment of SINE families in H3K9me3 (**Fig. 3d**). A previous study revealed that Piwi-interacting RNAs (piRNAs) played a major role in TE silencing via the ping-pong cycle in pig germline[77]. We further distinguished three classes of small non-coding RNAs to test the relationship between piRNA density and SINE families (**Fig. S8**, Supplementary Information). In contrast to siRNA and miRNA, piRNAs were significantly enriched for SINE-related sequences or sequence flanks, and there was a significant negative correlation between piRNA density and the age of SINE subfamily. Our result was in line with the findings in humans that young SINE families were more likely to be targeted by piRNAs[78], which can be regarded as the major reason leading to the high methylation levels of young SINE families, as we observed above.

Young SINE-derived transcriptome profiling in pigs

In addition to their interactions with epigenetic modifications, TEs can directly modify the transcription of host genes by remodeling new alternative splice events[79-80]. To test this, we analyzed the PacBio long-read isoform sequences (Iso-Seq) from 38 pig tissues[81-82] using a uniform pipeline to detect the transcripts of SINE-derived exonization and alternative splice sites. We estimated the contribution of TEs to gene expression across 52 tissues and 27 cell types by analyzing 3,570 published RNA-seq samples (**Fig. 4a,** Table S2) from the EBI database (https://www.ebi.ac.uk/).

After processing raw data by LoRDEC[83], we obtained 30,331,870 error-corrected Iso-seq reads with a mean length of 2,797 bp, of which 7.48% (2,267,973) was defined as TE-derived transcripts. Importantly, 68.81% (1,560,568) of TE-derived transcripts were recognized to be inserted by nearly full-length SINE (average coverage of 87.76%), indicating that SINE plays an important role in regulating gene expression.

Similar to a previous study[82], we next classified 337,746 young (younger and youngest) SINE-derived transcripts into four categories by comparing their genomic location with known transcripts in the currently available pig genome annotations (**Fig. S9**). Out of those, 1,028 young SINE-derived transcripts perfectly matched with 517 PCGs and 47 LncRNAs (Table S3), and 62,304 young SINE-derived transcripts potentially offered the novel alternative splice events for 8,103 PCGs and 405 LncRNAs (overlapping with at least one splice junction of a known transcript). The remaining young SINE-derived transcripts with no complete structural similarity with the currently available transcript annotation were classified as either 150,469 exon-covered (exonic overlap without any splice junction on the same or opposite strands) or 130,180 intronic transcripts (felled in a reference intron).

To understand the *cis*-functionality of young SINE-derived transcripts, we used the transcriptome data from 52 tissues and 27 cell types to quantify the abundance (RNA-seq counts) of PCGs and SINE-derived transcripts by Salmon tools[84]. A total of 3,138 protein-coding transcripts (3,112 PCGs) had significant associations with young SINE-derived transcripts at a Bonferroni significance threshold of 1.42×10^{-6} (0.05/35,135). Of those, 84 PCGs showed that the young SINE-derived exonization and their SINE-derived exons were involved in the expression of PCGs (Table S4). Interestingly, we found that the majority of SINE-derived exons (81.52%) were presented in mRNA 3′-untranslated regions (3′-UTRs) (**Fig. 4b**), suggesting that young SINEs can directly insert into the regulatory region to influence gene expression by the mechanism similar to Staufen-mediated decay (SMD)[85]. For example, we found that a full-length PRE0-SS (sus-specific SINE) was inserted in the 3′-UTR of the pig *PDK1* gene, which was in agreement with the previous report that the Alu and B1 (primate-specific and rodent-specific SINEs, respectively) regulated both human and mouse orthologs of *PDK1* by SMD[86]. This provides further supports for the convergent evolution of SINE-directed SMD. In contrast, a higher relative proportion of young SINEs was found in the CDS regions of exon-covered and intronic SINE-derived transcripts (10.03% and 11.15%, respectively) (**Fig. 4b**), suggesting that the selections against the young SINEs might be more relaxed in these two types of SINE-derived transcripts. However, it was consistent that most young SINEs do not directly participate in the protein translation but indirectly influence gene expression by affecting the UTRs of their derived transcripts[87].

We then retrieved young SINEs involved in SINE-derived transcripts and named them young-D SINEs. We observed that the average CG methylation levels of young-D SINEs that derived transcripts were significantly lower than the entire young SINEs in most tissues (**Fig. 4c**). Similarly, young-D SINEs were more significant enrichment in open chromatin and histone modifications than the entire young SINEs, especially the H3K4me3, providing more evidence that young-D SINEs and their derived transcripts were more likely active and functional across tissues (**Fig. 4d**).

Based on the normalized expression of PCGs by DESeq2[88], the *t*-SNE plots of unsupervised clustering of 3,570 samples clearly reflected tissue types (**Fig. 5a**). We further performed a co-expression network analysis of 14,403 PCGs using WGCNA R package[89] to explore the function of young SINE-derived transcripts across this wide range of tissues and cell types (**Fig. S10**). As a result, a total of 13,872 PCGs were grouped into 40 modules with the gene size ranging from 30 to 1,694 (**Fig. S11,** Table S5), and most of the modules showed high tissue specificity and likely play key roles in particular organ systems in pigs (**Fig. S12-S13**). The results were also supported by the gene-to-gene networks of topological clustering, which was performed using the Markov clustering (MCL) algorithm[90] (**Fig. S14**).

Importantly, 17.9% of co-expressed PCGs (2,744) were related to young SINE-derived transcripts and were mainly enriched in modules M38 (Trachea), M7 (Adipose), M18 (Fetal thymus), and M12 (Alveolar macrophages) (Z-score > 1), and the expression of all those modules showed high tissue specificity (**Fig. S15**).

In addition, as shown in **Figure 5b**, these young SINE-related PCGs were significantly enriched in the neural development, cellular metabolism, muscle development, and immune response, which may be responsible for the natural selection and domestication of modern pigs[91-93]. For instance, 186 PCGs were significantly enriched in the chemical synaptic transmission (GO:0007268), brain development (GO:0007420), and neuron projection morphogenesis (GO:0048812), which were mainly in the module M2 with a high expression in brain tissues (**Fig. S16**). Correspondingly, a total of 248 young-D SINEs, whose SINE-derived transcripts had significant associations with these 186 PCGs, were more significantly enriched in the active epigenetic marks (H3K4me1, H3K4me3, and H3K27ac) and depleted from the repressive mark (H3K27me3) at the nervous system (cerebellum, cortex, and hypothalamus) than other tissues, suggesting that young SINEs exhibited strong and concordant tissue specificity in both transcript expression and epigenetic regulation (**Fig. 5c**).

The roles of young SINEs in population admixture and local adaptation in pigs

TEs produce abundant raw materials for evolution in natural populations, and the burst of TEs was tightly related to significant evolutionary events such as the population admixture and local adaptation[23,94-95]. Young SINE polymorphisms represented the vast majority of TE polymorphisms in the pig genome[48,57,59,96]. Therefore, we developed a comprehensive map of polySINE from WGS of over 300 pigs, representing the majority of Eurasian pig breeds, to explore the roles of young SINEs and their derived genes in pig population admixture and local adaptation.

To investigate whether the sequencing depth and polySINE detection tools had a prominent effect on the detection of polySINEs, we first benchmarked the polySINE detection tools that showed superior performances in previous human projects[13,97] under different sequencing depths (**Fig. S17-S25**). Based on the results of benchmarking, we customized the **P**ig **TE** **P**olymorphism (PigTEP) pipeline to maximize its performance in the current pig WGS datasets (**Fig. 6a**). We then applied it to identify polySINEs from 374 individuals with the uniform sequencing depth of 10× (**Fig. S26**, Table S6, average mapped bases: 27.18 GB and average mapping rates: 99.44%). The analyzed 374 individuals from 25 diverse populations (N ≥ 5) were further assigned into 10 major groups, i.e., PYGMY, ISEA, CHD, KOD, AWB, TWB, EUD, EWB, MINI, and COM (**Fig. 6b**).

In total, we identified 211,067 polySINEs (189,966 *Ref+* and 21,101 *Ref-*) in pig genome, 49.64% of which were located in non-intergenic regions. As expected, the vast majority of polySINEs were found at a low frequency (64.89% of polySINEs with < 5% minor allele frequencies, MAF) in the whole pig population (**Fig. S27**), but showed variable MAF among groups (**Fig. S28**). We found 85.58% of polySINEs were shared among groups, while only 30,441 polySINEs (PYGMY and ISEA accounted for 60.83% and 26.01%, respectively) were exclusively presented within a single group (**Fig. 6c**).

The principal component analysis (PCA) of samples using polySINE genotypes clearly reflected the four species of the Suidae (**Fig. 6d**). PC1 separated the *Porcula slavania* from *Sus*-species, while PC2 (19.41%) and PC3 (13.67%) showed the genetic separation between Asian and Western breeds (**Fig. 6e**). Interestingly, the Korean domestic pigs (KOD) separated from the Asian breeds and were closer to Western breeds than to Chinese breeds, suggesting the presence of gene flow and introgression from Western pig breeds to Korean domestic pigs and most likely mediated by humans.

The results were further supported by the TE-based phylogenetic tree and genetic admixture (**Fig. 6f and g**; **Fig. S29**), whose evolutionary relationships were basically consistent with previous studies on SNP-based genotypes[45,98]. The comparison of Chinese and European domestic pigs confirmed our previous findings on the TE-based introgression between Northern Chinese domestic pigs and European domestic pigs[57]. Specially, we found that Korean wild boars, unlike the Korean domestic pigs, clustered together with other Asian wild boars instead of European pigs.

To detect polySINEs associated with local adaptation, we calculated the pairwise Fst_i value for each polySINE between cluster i and the remaining clusters to measure its locus-specific divergence in allele frequencies. The polySINEs with extreme Fst_i (top 1%; n =337) were observed in functional regions (exonic, splicing, UTR5, UTR3, and upstream) of 330 PCGs (Table S7). While 77.94% of these PCGs existed in PYGMY (n = 223) and ISEA (n = 42), and the remaining 75 were likely associated with the breed-restricted phenotypes of domestic pigs (*Sus scrofa*) (**Fig. 7a**). For instance, a PRE1 insertion in the promoter of the *IGFBP7* gene, which was associated with tumor suppressor[99], was widespread in Chinese indigenous breeds rather than in commercial breeds[100]. The upstream (high signals in H3K4me1 and H3K27ac) of *FRZB* was inserted by a population-specific polySINE from Southern Chinese domestic pigs, which was associated with pig growth traits[101] (**Fig. 7b**).

Importantly, we found a fixed polySINE at the first exon of *RUNX3* in the Goettingen miniature pigs and MiniLEWE, which was also detected in Southern Chinese domestic pigs, particularly in the Luchuan pigs (**Fig. 7c**). The *RUNX3* gene was famous as a tumor suppressor in a human T-cell malignancy[102] and was a key part of the TGF-**β** induced signaling pathway[103]. In addition, 30 PCGs showed the extreme *Fst* in the long-term

399 laboratory-inbred Bama Xiang pigs[104]and were significantly enriched in the AMPK signaling
400 pathway (Corrected *P*-values = 0.00295, Table S8), and especially the *SLC3A2* (upstream)
401 and *SIRT1* (UTR3) genes had the polySINEs with a perfectly fixed frequency.
402 Nine of 75 candidate genes could be mapped to known QTX (Quantitative Trait
403 Loci/Gene/Nucleotide) data that were potentially associated with the phenotypic traits[105] (**Fig.**
404 **7d**). We observed that four PCGs related to laboratory-inbred Bama Xiang pigs were
405 associated with fat content and body weight, which was consistent with the direction of
406 selective breeding in laboratory[104]. Especially, the *LEP* gene was highly expressed in adipose
407 tissue and can produce a hormone called leptin, which was involved in the regulation of
408 appetite, fat storage, and body weight[106]. Overall, these findings demonstrate that polySINEs
409 can serve as a valuable source for studying genomic ancestry and local adaptive evolution in
410 pigs.

411 **Mapping young SINEs to the genetic associations of complex traits**

412 To explore the association of polySINEs with complex traits, we first collected a total of
413 4,072 trait-associated SNPs (T-SNPs) from 79 published GWAS studies of 97 complex traits
414 of economic value in pigs, including 18 reproduction, 22 production, 36 meat and carcass, six
415 health, and two exterior traits (**Fig. S30**). As shown in **Fig. 7e**, we identified 127
416 trait-associated polySINEs (T-polySINEs) that were in linkage disequilibrium (LD, $r^2 > 0.3$)
417 with T-SNPs among 296 domestic pigs (109 Asian and 187 European domestic pigs). In
418 particular, it was found that these T-polySINEs were more likely to be enriched in the
419 TxFlnkWk (Weak transcribed at gene), indicating that they have the potential for gene
420 regulation (**Fig. 8a**).
421 A total of 54 genes were affected by these T-polySINEs, which were associated with the
422 intramuscular fat composition and teat number. The majority of them showed specific
423 expression in certain tissues (Z-score > 2), particularly in the nervous system (plasmacytoid
424 dendritic cells, choroid plexus, hypothalamus, and brain), reproductive system (testis, oviduct,
425 and oocyte), and muscle satellite cells (**Fig. S31**, Table S9). Importantly, most of the intronic
426 T-polySINEs exhibited breed-specific MAF between Chinese and Western pigs, which was in
427 agreement with their differences in fatty acid content and teat number (**Fig. 8b**).
428 We identified a 320kb T-polySINE hotspots (chr14:112,965,840-113,285,513; $r^2 > 0.3$),
429 including six T-polySINEs and eight genes, which were significantly associated with
430 intramuscular fat composition (**Fig. S32**). Among these hotspot genes, C14H10orf76 (*ARMH3*,
431 $r^2 = 0.89$) and *GBF1* ($r^2 = 0.86$) have been reported to be essential for Golgi maintenance and
432 secretion[107]. The *ELOVL3* gene was a strong candidate gene for fatty acid composition[108-109].
433 A low-frequency polySINE was in the intron region of *ELOVL3*, while multiple T-polySINEs
434 within its upstream region of 15kb to 50kb. In particular, the T-polySINE (chr14:113,199,425)

near 27 kb upstream was at high frequency in Chinese domestic pigs, especially Southern Chinese domestic pigs (**Fig. 8c**).

Furthermore, there was a high level of pairwise LD (r^2 = 0.88) between the T-polySINE (chr8:109,447,835) in the intron of *ANK2* and the T-SNP that had a significant association with the fatty acid content of C14:0, C16:0, and C16:1n7 in backfat[110]. It was reported that ankyrin-B (AnkB) was a neuron-specific alternatively spliced variant of *ANK2* that was associated with obesity susceptibility in humans[111]. We found that the insertion of this T-polySINE was located in an LD block of 15kb ($r^2 > 0.5$, chr8:109,439,023-109,454,866, **Fig. S33**) and almost fixed in the Western domestic pig populations (**Fig. 8d**). *ANK2* gene was observed to be ubiquitously expressed in pig bodies and highly enhanced in the nervous system (**Fig. 8e**). We noticed that two predicted SINE-derived transcripts (supported by Iso-seq reads) overlapped with exons of *ANK2* had a significant correlation with the expression of *ANK2* (Table S10). The expression of *ANK2* was significantly upregulated in cultivars with high-fat deposition compared with those with low-fat deposition, such as Songliao black pigs vs. Landrace pigs[112], and fast-growing chickens vs. slow-growing chickens[113].

In addition, we found a high LD (r^2 = 0.75) between a T-SNP (chr7:97,606,621) and a T-polySINE (chr7:97,615,896) in the first intron of *VRTN* gene, which was significantly associated with teat number[114]. The *VRTN* was proposed as the most promising candidate gene to increase the number of thoracic vertebrae (ribs) in pigs[114], which was highly and specifically expressed in embryonic stem cells, embryo, and ovary (**Fig. 8f**), suggesting that *VRTN* functions at the early embryonic stage of pig development. We found that this T-polySINE showed the obvious difference in frequency between Chinese indigenous breeds and commercial breeds (**Fig. 8g**). Especially, a novel transcript (length = 2,191bp) derived by this T-polySINE and covered the first exon of *VRTN*, which was significantly correlated with the expression of *VRTN* (**Fig. 8h,** *P*-values = 3.03×10^{-201}, Table S10), and this transcript was supported by the RNA-seq exon coverage in NCBI annotation (**Fig. S34**). Of particular note, this region exhibited the open chromatin and enhancer signals (H3K4me1) while was facultatively repressed in most tissues (H3K27me3) (**Fig. 8i**). We clearly observed a significant decline in the repressed states at the stem cells and embryo-related tissues, which corresponded to the tissue-specific expression in *VRTN*, suggesting that this region was crucial for *VRTN*, and this T-polySINE was more likely to affect its expression.

Discussion

TEs are major, abundant, and polymorphic in the pig genome

In this study, we built a comprehensive atlas of TEs in the pig genome by using the newly developed PigTEDC pipeline, which combined the similarity-, structure-, and *de*

novo-based methods. Our results demonstrated that nearly a third (947 MB) of the pig genome was made up of TEs, and the majority of them were non-LTR retrotransposons (SINE and LINE). SINE was shorter and more complete than LINE. Similar to our previous findings[53], SINE, especially the PRE1-SS, PRE0-SS, and PRE1a families in SINE/tRNA, displayed the most recent ages and most polymorphic insertions. These polymorphic SINEs contributed nearly 90% of medium-length SVs among different assemblies, especially the L13 subfamilies, with 36.15% that were classified as the youngest SINEs in the pig genome.

The influence of SINEs on transcriptional networks are associated with their ages

Gene regulatory network is influenced by genomic components, chromatin accessibility, histone modifications, DNA methylation, and *cis*-regulatory elements (e.g., TFBS, promoters, and enhancers). TEs associated with unique chromosome features can contribute to gene regulatory networks in a variety of the above ways. This is the first time, to our knowledge, to use large-scale multi-omics data to fully explore the relationships between TEs and chromosome features in the pig genome. Our findings showed that SINEs were significantly enriched in the A compartment, and the enrichment of SINE in open or closed chromatin regions was associated with their ages. For example, young SINE families were frequently enriched in close chromatin-like nucleosomes but highly depleted from open chromatin.

As expected, SINEs were highly depleted from all active chromatin tags, and more signals of constitutive heterochromatin tags (H3K9me3 peaks) were observed on SINE. The exception was H3K27me3, which was associated with facultative suppressor genes and cannot make permanent silence on SINEs. Almost all signals of histone modifications in SINE showed a tendency to decay as the age of TE increased, which was in line with the contribution of SINE to TFBS and the distribution of DNA methylation on SINE. However, there is still evidence that some young SINEs enrichment in weak functional regions, such as the youngest SINEs families, were highly enriched in the regions of weak active enhancer at hypothalamus tissue (Fold >1.5). We speculate that the relationship between SINE and its host genome is a combination of both “arms race” and “co-evolution,” depending on how the symbiosis turned out. In the former case of parasitism, the young SINEs were more likely to be treated as new invaders that were constitutively silenced by histone modifications and DNA methylation of the host genome (e.g., PIWI -piRNA pathway during the TE in testis), while the old SINEs were mutated and gained new regulatory potential, and thus tolerated or even co-opted by the pig genome. In the latter case of mutualism, there might be rare cases where the SINEs were positively selected by nature thereby help the host genome better adapt to the local environment in the long run.

The non-coding RNAs derived from young SINEs may affect tissue-specific genes

The use of long-read isoform sequencing provided us a more complete characterization

of full-length transcripts, which made it possible to identify the young SINE-derived transcripts. Meanwhile, the Iso-seq reads we used here were collected from about 40 pig tissues, which ensured the investigation of abundance and tissue specificity of young SINE-derived transcripts.

Our findings showed that the vast majority of young SINE-derived transcripts were non-coding RNAs that covered exons or felled in the introns. A total of 3,112 PCGs were found to be associated with young SINE-derived transcripts, and nearly 88% of them were enriched in co-expressed modules with high tissue specificity. The young SINEs that derived transcripts exhibited lower CG methylation levels and were more significant enrichment in open chromatin and histone modifications than whole young SINEs. Especially, some young SINEs exhibited strong and accordance tissue specificity in both transcript expression and epigenetic regulation. This was consistent with previous findings in other species[115-117], suggesting that SINE insertions may be a crucial component of genes and regulate tissue-specific expression of their target genes.

The detection of polySINE is significantly affected by detection tools and sequencing depth

PolySINEs belong to SVs that were more sensitive to sequencing depth than SNPs. In this study, to ensure the unbiased detection of polySINEs, we benchmarked the four polySINE detection tools under different sequencing depths. Our findings showed that MELT had robust performance in both *ref*+ and *ref*- detection (**Fig. S17-S18**). As expected, we found that the detected number of polySINEs increased significantly with the sequencing depth, especially from 5× to 10× that nearly doubled the average number of polySINEs (**Fig. S19**). Considering the sequencing depth in the current 838 publicly available whole-genome sequence datasets in pigs, we retained 374 individuals whose sequencing depth was greater than 10× and down-sampled their sequencing depth to ~10× through a strategy of randomly removing reads (average mapped bases: 27.17 GB and average mapping rates: 99.43%). Finally, the PigTEP pipeline was developed to identify both polySINEs and SNPs in individuals simultaneously. This pipeline will help other researchers to explore the role of polySINEs in pig genomic study and breeding.

TEs are non-negligible genetic markers for genetic diversity and complex traits

In pig genomic researches, the contribution of TEs to genetic diversity and complex traits was underestimated, even though SINEs can be more active when the pig is under selective pressure. The lack of a comprehensive map of polySINEs based on the large-scale re-sequencing dataset has limited our understanding of SINEs in pig population genetics.

Here, we genotyped and analyzed 211,067 polySINE loci in 374 individuals across 25 pig populations and found that polySINE loci were highly variable in polySINE allele

543 frequencies among populations. The genetic relationships of samples based on polySINEs
544 were consistent with previous studies based on SNP genotypes[118], revealing that polySINE
545 loci were informative in studying population genetics. Especially, we observed that ten major
546 clusters representing 25 pig populations well corresponded to the geographic differentiation.
547 These polySINEs with high pairwise *Fst* value were useful resources to understand local
548 adaptation in domestic pigs. Our findings confirmed the results of previous studies (e.g.,
549 *IGFBP7*) and provided novel candidate genes that are potential to contribute to economically
550 complex traits in pigs.

551 The genome-wide association studies (GWAS) based on SNPs have discovered
552 thousands of QTLs of important economic traits in pigs, but most of these loci have not been
553 functionally characterized. One possible reason is that what really affects phenotypic changes
554 is not SNPs but SVs (MEIs) that were in linkage disequilibrium (LD) with them. In this study,
555 127 polySINEs were found to be LD with significant GWAS SNPs of complex traits, and
556 nearly a third of them showed high tissue specificity in terms of expression. Importantly, a
557 part of these polySINEs has been found to have the ability to derive novel functional
558 transcripts (in H3K4me1 and H3K27me3), as the exon-covered transcript was found in the
559 upstream of the *VRTN* gene. Future researches will be required to functionally validate
560 whether and how these polySINEs affect complex traits by regulating their target genes in
561 particular tissues (e.g., *VRTN* gene in embryonic stem cells).

References

1. Ricci M, Peona V, Guichard E, Taccioli C, Boattini A. Transposable Elements Activity is Positively Related to Rate of Speciation in Mammals. *J Mol Evol* **86**, 303-310 (2018).
2. Platt RN, 2nd, Blanco-Berdugo L, Ray DA. Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. *Genome Biol Evol* **8**, 403-410 (2016).
3. Sahebi M*, et al.* Contribution of transposable elements in the plant's genome. *Gene* **665**, 155-166 (2018).
4. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691-703 (2009).
5. Levin HL, Moran JV. Dynamic interactions between transposable elements and their hosts. *Nat Rev Genet* **12**, 615-627 (2011).
6. Streva VA, Faber ZJ, Deininger PL. LINE-1 and Alu retrotransposition exhibit clonal variation. *Mob DNA* **4**, 16 (2013).
7. Flynn JM*, et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A* **117**, 9451-9457 (2020).
8. Beck CR*, et al.* LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159-1170 (2010).
9. Ade C, Roy-Engel AM, Deininger PL. Alu elements: an intrinsic source of human genome instability. *Curr Opin Virol* **3**, 639-645 (2013).
10. Wicker T*, et al.* A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**, 973-982 (2007).
11. Biemont C. A brief history of the status of transposable elements: from junk DNA to major players in evolution. *Genetics* **186**, 1085-1093 (2010).
12. Arkhipova IR. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob DNA* **8**, 19 (2017).
13. Hoen DR*, et al.* A call for benchmarking transposable element annotation methods. *Mob DNA* **6**, 13 (2015).
14. Goerner-Potvin P, Bourque G. Computational tools to unmask transposable elements. *Nat Rev Genet* **19**, 688-704 (2018).
15. Xing J, Witherspoon DJ, Jorde LB. Mobile element biology: new possibilities with high-throughput sequencing. *Trends Genet* **29**, 280-289 (2013).
16. Sotero-Caio CG, Platt RN, 2nd, Suh A, Ray DA. Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biol Evol* **9**, 161-177 (2017).
17. Chenais B, Caruso A, Hiard S, Casse N. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene* **509**, 7-15 (2012).
18. Zeh DW, Zeh JA, Ishida Y. Transposable elements and an epigenetic basis for punctuated equilibria. *Bioessays* **31**, 715-726 (2009).
19. Bourque G*, et al.* Ten things you should know about transposable elements. *Genome Biol* **19**, 199 (2018).
20. Stewart C*, et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* **7**, e1002236 (2011).
21. Manthey JD, Moyle RG, Boissinot S. Multiple and Independent Phases of Transposable Element Amplification in the Genomes of Piciformes (Woodpeckers and Allies). *Genome Biol*

Evol **10**, 1445-1456 (2018).

22. Schmitz J. SINEs as driving forces in genome evolution. *Genome Dyn* **7**, 92-107 (2012).
23. Carpentier MC*, et al.* Retrotranspositional landscape of Asian rice revealed by 3000 genomes. *Nat Commun* **10**, 24 (2019).
24. Marmoset Genome S, Analysis C. The common marmoset genome provides insight into primate biology and evolution. *Nat Genet* **46**, 850-857 (2014).
25. Rishishwar L, Tellez Villa CE, Jordan IK. Transposable element polymorphisms recapitulate human evolution. *Mob DNA* **6**, 21 (2015).
26. Rishishwar L, Wang L, Wang J, Yi SV, Lachance J, Jordan IK. Evidence for positive selection on recent human transposable element insertions. *Gene* **675**, 69-79 (2018).
27. Van't Hof AE*, et al.* The industrial melanism mutation in British peppered moths is a transposable element. *Nature* **534**, 102-105 (2016).
28. Mao H*, et al.* A transposable element in a NAC gene is associated with drought tolerance in maize seedlings. *Nat Commun* **6**, 8326 (2015).
29. Feschotte C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**, 397-405 (2008).
30. Attig J*, et al.* Splicing repression allows the gradual emergence of new Alu-exons in primate evolution. *Elife* **5**, (2016).
31. Hirsch CD, Springer NM. Transposable element influences on gene expression in plants. *Biochim Biophys Acta Gene Regul Mech* **1860**, 157-165 (2017).
32. Trizzino M*, et al.* Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res* **27**, 1623-1633 (2017).
33. Wang J*, et al.* MIR retrotransposon sequences provide insulators to the human genome. *Proc Natl Acad Sci U S A* **112**, E4428-4437 (2015).
34. Sundaram V*, et al.* Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res* **24**, 1963-1976 (2014).
35. Funkhouser SA, Steibel JP, Bates RO, Raney NE, Schenk D, Ernst CW. Evidence for transcriptome-wide RNA editing among Sus scrofa PRE-1 SINE elements. *BMC Genomics* **18**, 360 (2017).
36. Jeck WR*, et al.* Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19**, 141-157 (2013).
37. Roberts JT*, et al.* Continuing analysis of microRNA origins: Formation from transposable element insertions and noncoding RNA mutations. *Mob Genet Elements* **3**, e27755 (2013).
38. Chishima T, Iwakiri J, Hamada M. Identification of Transposable Elements Contributing to Tissue-Specific Expression of Long Non-Coding RNAs. *Genes (Basel)* **9**, (2018).
39. Fedoroff NV. Presidential address. Transposable elements, epigenetics, and genome evolution. *Science* **338**, 758-767 (2012).
40. Rey O, Danchin E, Mirouze M, Loot C, Blanchet S. Adaptation to Global Change: A Transposable Element-Epigenetics Perspective. *Trends Ecol Evol* **31**, 514-526 (2016).
41. Lanciano S, Mirouze M. Transposable elements: all mobile, all different, some stress responsive, some adaptive? *Curr Opin Genet Dev* **49**, 106-114 (2018).
42. Horvath V, Merenciano M, Gonzalez J. Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response. *Trends Genet* **33**, 832-841 (2017).
43. Niederhuth CE*, et al.* Widespread natural variation of DNA methylation within angiosperms.

Genome Biol **17**, 194 (2016).

44. Lerat E, Casacuberta J, Chaparro C, Vieira C. On the Importance to Acknowledge Transposable Elements in Epigenomic Analyses. *Genes (Basel)* **10**, (2019).
45. Groenen MA, *et al.* Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**, 393-398 (2012).
46. Cooper DK, *et al.* The pathobiology of pig-to-primate xenotransplantation: a historical review. *Xenotransplantation* **23**, 83-105 (2016).
47. Ekser B, Markmann JF, Tector AJ. Current status of pig liver xenotransplantation. *Int J Surg* **23**, 240-246 (2015).
48. Fang X, *et al.* The sequence and analysis of a Chinese pig genome. *Gigascience* **1**, 16 (2012).
49. Li M, *et al.* Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nat Genet* **45**, 1431-1438 (2013).
50. Li M, *et al.* Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple de novo assemblies. *Genome Res* **27**, 865-874 (2017).
51. Groenen MAM, *et al.* Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**, 393-398 (2012).
52. Vamathevan JJ, *et al.* Minipig and beagle animal model genomes aid species selection in pharmaceutical discovery and development. *Toxicol Appl Pharmacol* **270**, 149-157 (2013).
53. Zhao P, *et al.* PRE-1 Revealed Previous Unknown Introgression Events in Eurasian Boars during the Middle Pleistocene. *Genome Biol Evol* **12**, 493 (2020).
54. Chalopin D, Naville M, Plard F, Galiana D, Volff JN. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biol Evol* **7**, 567-580 (2015).
55. Groenen MAM. A decade of pig genome sequencing: a window on pig domestication and evolution. *Genet Sel Evol* **48**, (2016).
56. Gingerich PD. Environment and evolution through the Paleocene-Eocene thermal maximum. *Trends Ecol Evol* **21**, 246-253 (2006).
57. Zhao P, *et al.* PRE-1 Revealed Previous Unknown Introgression Events in Eurasian Boars during the Middle Pleistocene. *Genome Biol Evol* **12**, 1751-1764 (2020).
58. Zhou R, *et al.* Developmental stage-specific A-to-I editing pattern in the postnatal pineal gland of pigs (Sus scrofa). *J Anim Sci Biotechnol* **11**, 90 (2020).
59. Chen C, *et al.* SINE jumping contributes to large-scale polymorphisms in the pig genomes. *Mob DNA* **12**, 17 (2021).
60. Jordan KW, He F, de Soto MF, Akhunova A, Akhunov E. Differential chromatin accessibility landscape reveals structural and functional features of the allopolyploid wheat chromosomes. *Genome Biol* **21**, 176 (2020).
61. Vrljicak P, *et al.* Analysis of chromatin accessibility in decidualizing human endometrial stromal cells. *FASEB J* **32**, 2467-2477 (2018).
62. Benoit M, *et al.* Environmental and epigenetic regulation of Rider retrotransposons in tomato. *PLoS Genet* **15**, e1008370 (2019).
63. Cournac A, Koszul R, Mozziconacci J. The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. *Nucleic Acids Res* **44**, 245-255 (2016).
64. Foissac S, Djebali S, Munyard K, Vialaneix N, Giuffra E. Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biology* **17**, 108 (2019).

694 65. Yumei L*, et al.* Human exonization through differential nucleosome occupancy. *Proceedings*
695 *of the National Academy of Sciences*, 201802561- (2018).
696 66. Kern C*, et al.* Functional annotations of three domestic animal genomes provide vital
697 resources for comparative and agricultural research. *Nat Commun* **12**, 1821 (2021).
698 67. Saksouk N, Simboeck E, Dejardin J. Constitutive heterochromatin formation and transcription
699 in mammals. *Epigenetics Chromatin* **8**, 3 (2015).
700 68. Pan Z*, et al.* Pig genome functional annotation enhances the biological interpretation of
701 complex traits and human disease. *Nat Commun* **12**, 5848 (2021).
702 69. Luis V-CJ, Vivien H, Laura A, Josefa G. Diverse families of transposable elements affect the
703 transcriptional regulation of stress-response genes in Drosophila melanogaster. *Nucleic acids*
704 *research*, 6842 (2019).
705 70. Jordi M, Fatima M, Jordi P, Casacuberta JM. Plant lineage-specific amplification of
706 transcription factor binding motifs by Miniature Inverted-repeat Transposable Elements
707 (MITEs). *Genome Biology & Evolution*, 5 (2018).
708 71. Corinne, Simonti, Mihaela, Pavlicev, John, Capra. Transposable Element Exaptation into
709 Regulatory Regions Is Rare, Influenced by Evolutionary Age, and Subject to Pleiotropic
710 Constraints. *Molecular Biology & Evolution*, (2017).
711 72. Bakke J*, et al.* Transcription factor ZNF148 is a negative regulator of human muscle
712 differentiation. *Scientific Reports* **7**, 8138 (2017).
713 73. Piasecki BP, Burghoorn J, Swoboda P. Regulatory Factor X (RFX)-mediated transcriptional
714 rewiring of ciliary genes in animals. *Proceedings of the National Academy of Sciences of the*
715 *United States of America* **107**, (2010).
716 74. Reith W*, et al.* Cloning of the major histocompatibility complex class II promoter binding
717 protein affected in a hereditary defect in class II gene regulation. *Proceedings of the National*
718 *Academy of Sciences of the United States of America* **86**, 4200-4204 (1989).
719 75. Siegrist CA*, et al.* RFX1 is identical to enhancer factor C and functions as a transactivator of
720 the hepatitis B virus enhancer. *Molecular and Cellular Biology* **13**, 6375-6384 (1993).
721 76. Noshay JM, Anderson SN, Zhou P, Ji L, Springer NM. Monitoring the interplay between
722 transposable element families and DNA methylation in maize. *PLOS Genetics* **15**, e1008291-
723 (2019).
724 77. Gebert D, Ketting RF, Zischler H, Rosenkranz D. piRNAs from Pig Testis Provide Evidence
725 for a Conserved Role of the Piwi Pathway in Post-Transcriptional Gene Regulation in
726 Mammals. *Plos One* **10**, (2015).
727 78. Ha H, Song J, Wang S, Kapusta A, Feschotte… C. A comprehensive analysis of piRNAs from
728 adult human testis and their relationship with genes and mobile elements. *Bmc Genomics* **15**,
729 545 (2014).
730 79. Jang HS*, et al.* Transposable elements drive widespread expression of oncogenes in human
731 cancers. *Nat Genet* **51**, 611-617 (2019).
732 80. Moschetti R, Palazzo A, Lorusso P, Viggiano L, Marsano RM. "What You Need, Baby, I Got
733 It": Transposable Elements as Suppliers of Cis-Operating Sequences in Drosophila.
734 *Biology-Basel* **9**, (2020).
735 81. Li Y*, et al.* A survey of transcriptome complexity in Sus scrofa using single-molecule
736 long-read sequencing. *DNA Res* **25**, 421-437 (2018).
737 82. Beiki H*, et al.* Improved annotation of the domestic pig genome through integration of

Iso-Seq and RNA-seq data. *BMC Genomics* **20**, 344 (2019).

83. Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction. *Bioinformatics* **30**, 3506-3514 (2014).

84. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419 (2017).

85. Brogna S, Wen J. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat Struct Mol Biol* **16**, 107-113 (2009).

86. Lucas BA*, et al.* Evidence for convergent evolution of SINE-directed Staufen-mediated mRNA decay. *Proc Natl Acad Sci U S A* **115**, 968-973 (2018).

87. Tajaddod M*, et al.* Transcriptome-wide effects of inverted SINEs on gene expression and their impact on RNA polymerase II activity. *Genome Biology* **17**, 220 (2016).

88. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

89. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

90. van Dongen S, Abreu-Goodger C. Using MCL to extract clusters from networks. *Methods Mol Biol* **804**, 281-295 (2012).

91. Zhu Y*, et al.* Signatures of Selection and Interspecies Introgression in the Genome of Chinese Domestic Pigs. *Genome Biol Evol* **9**, 2592-2603 (2017).

92. Bosse M*, et al.* Artificial selection on introduced Asian haplotypes shaped the genetic architecture in European commercial pigs. *Proc Biol Sci* **282**, 20152019 (2015).

93. Chen M, Wang J, Wang Y, Wu Y, Fu J, Liu JF. Genome-wide detection of selection signatures in Chinese indigenous Laiwu pigs revealed candidate genes regulating fat deposition in muscle. *BMC Genet* **19**, 31 (2018).

94. Chen J*, et al.* Tracking the origin of two genetic components associated with transposable element bursts in domesticated rice. *Nat Commun* **10**, 641 (2019).

95. Li ZW*, et al.* Transposable Elements Contribute to the Adaptation of Arabidopsis thaliana. *Genome Biol Evol* **10**, 2140-2150 (2018).

96. Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiol Spectr* **3**, MDNA3-0061-2014 (2015).

97. Wang L, Jordan IK. Transposable element activity, genome regulation and human health. *Curr Opin Genet Dev* **49**, 25-33 (2018).

98. Ai H*, et al.* Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nat Genet* **47**, 217-225 (2015).

99. Godina C*, et al.* Prognostic impact of tumor-specific insulin-like growth factor binding protein 7 (IGFBP7) levels in breast cancer: A prospective cohort study. *Carcinogenesis*, (2021).

100. Wu Q*, et al.* Linkage disequilibrium and functional analysis of PRE1 insertion together with SNPs in the promoter region of IGFBP7 gene in different pig breeds. *J Appl Genet* **59**, 231-241 (2018).

101. Wang Z*, et al.* Single nucleotide polymorphism scanning and expression of the FRZB gene in pig populations. *Gene* **543**, 198-203 (2014).

102. Zhou WN*, et al.* RUNX3 plays a tumor suppressor role by inhibiting cell migration, invasion and angiogenesis in oral squamous cell carcinoma. *Oncol Rep* **38**, 2378-2386 (2017).

103. Fainaru O, *et al.* Runx3 regulates mouse TGF-beta-mediated dendritic cell function and its absence results in airway inflammation. *EMBO J* **23**, 969-979 (2004).

104. Zhang L, *et al.* Development and Genome Sequencing of a Laboratory-Inbred Miniature Pig Facilitates Study of Human Diabetic Disease. *iScience* **19**, 162-176 (2019).

105. Fu Y, *et al.* A gene prioritization method based on a swine multi-omics knowledgebase and a deep learning model. *Commun Biol* **3**, 502 (2020).

106. Barb CR, Hausman GJ, Czaja K. Leptin: a metabolic signal affecting central regulation of reproduction in the pig. *Domest Anim Endocrinol* **29**, 186-192 (2005).

107. Chan CJ, *et al.* BioID Performed on Golgi Enriched Fractions Identify C10orf76 as a GBF1 Binding Protein Essential for Golgi Maintenance and Secretion. *Mol Cell Proteomics* **18**, 2285-2297 (2019).

108. Viterbo VS, Lopez BIM, Kang H, Kim H, Song CW, Seo KS. Genome wide association study of fatty acid composition in Duroc swine. *Asian-Australas J Anim Sci* **31**, 1127-1133 (2018).

109. Westerberg R, *et al.* ELOVL3 is an important component for early onset of lipid recruitment in brown adipose tissue. *J Biol Chem* **281**, 4958-4968 (2006).

110. Crespo-Piazuelo D, *et al.* Identification of strong candidate genes for backfat and intramuscular fatty acid composition in three crosses based on the Iberian pig. *Sci Rep* **10**, 13962 (2020).

111. Lorenzo DN, Bennett V. Cell-autonomous adiposity through increased cell surface GLUT4 due to ankyrin-B deficiency. *Proc Natl Acad Sci U S A* **114**, 12743-12748 (2017).

112. Xing K, *et al.* Comparative adipose transcriptome analysis digs out genes related to fat deposition in two pig breeds. *Sci Rep* **9**, 12925 (2019).

113. Claire D'Andre H, *et al.* Identification and characterization of genes that control fat deposition in chickens. *J Anim Sci Biotechnol* **4**, 43 (2013).

114. van Son M, *et al.* A QTL for Number of Teats Shows Breed Specific Effects on Number of Vertebrae in Pigs: Bridging the Gap Between Molecular and Quantitative Genetics. *Front Genet* **10**, 272 (2019).

115. Miao B, Fu S, Lyu C, Gontarz P, Wang T, Zhang B. Tissue-specific usage of transposable element-derived promoters in mouse development. *Genome Biol* **21**, 255 (2020).

116. Smith CEL, Alexandraki A, Cordery SF, Parmar R, Bonthron DT, Valleley EMA. A tissue-specific promoter derived from a SINE retrotransposon drives biallelic expression of PLAGL1 in human lymphocytes. *PLoS One* **12**, e0185678 (2017).

117. Jjingo D, Conley AB, Wang J, Marino-Ramirez L, Lunyak VV, Jordan IK. Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. *Mob DNA* **5**, 14 (2014).

118. Zhao P, *et al.* Evidence of evolutionary history and selective sweeps in the genome of Meishan pig reveals its genetic and phenotypic characterization. *Gigascience* **7**, (2018).

821 **FIGURE LEGENDS**

822 **Figure 1. Transposable elements annotation and SINE classification in the pig genome. a**
823 A schematic of the Pig Transposable Element Detection and Classification (PigTEDC)
824 pipeline. It is composed of three TE detection approaches, using the similarity-, structure-, and
825 *de novo*-based algorithms. **b** The proportion of TEs from different superfamilies in the count.
826 **c** The proportion of TEs from different superfamilies in length. **d** Classification of pig TE
827 superfamilies and families (≥ 3,000 TE copies in each family). **e** Sequence divergence
828 distribution for TE superfamilies (upper panel) and families (bottom panel) in the pig genome.
829 Sequence divergence distributions are plotted in bins of 0.01 increments. **f** Phylogenetic tree
830 and sequence divergence distribution for SINE families in the pig genome. On the right panel,
831 the x-axis represents the divergence, and the y-axis represents the counts of the SINE families.
832 **g** Boxplots displays the proportion of genomic SVs formed by different SINE families. **h**
833 Boxplots displays the proportion of genomic SVs formed by different SINE new subfamilies.
834 See Table S1 for their definitions. **i** Classification of pig SINE families based on their ages.

835 **Figure 2. Distribution of SINE on pig genome and functional regions. a** The five types of
836 genomic features used in this study included 3D chromatin architecture, chromatin
837 accessibility, histone modifications, DNA methylation, and transcription factor binding sites.
838 **b** The distribution of SINE families between 3D chromatin architectures (Compartments A vs.
839 B), as well as near topologically associating domain (TAD). **c** The reads density distributions
840 of chromatin accessibility and histone modifications near transcripts across four different
841 SINE groups. **d** Boxplots displays the enrichment of four SINE groups in 15 distinct
842 chromatin states across 14 tissues.

843 **Figure 3. Enrichment of SINE in functional elements and methylation modification. a**
844 Hierarchical clustering of enrichment patterns in 15 chromatin states for four SINE groups
845 across 14 tissues (left panel). The Heatmap of three distinct enrichment patterns for the high
846 enrichment in the young SINE group (right panel). **b** A heatmap for the enrichment of
847 transcription factor binding motifs in SINE families, chromatin accessibility, and histone

848 modifications. **c** The signal density of MeDIP-seq and CpG island within different SINE
849 families. **d** Boxplots displays the DNA methylation levels on different SINE families. L and
850 R represent the upstream and downstream directions of the SINE body. E.g., L250 represents
851 0 to 250 bp, and L500 represents 250 to 500 bp window upstream of SINE.

852 **Figure 4. Young SINE-derived transcriptome landscape. a** Overview of RNA-seq libraries
853 in 3570 samples across 52 tissues and 27 types of cells. **b** The bar plot indicates the
854 proportion of functional regions affected by SINE across four different categories of
855 SINE-derived transcripts. **c** Boxplots display CG methylation levels on young SINE families.
856 Young-D group represents the SINE families that derived the young SINE-derived transcripts.
857 The Younger and Youngest groups represent all the younger and youngest SINEs in the entire
858 genome, respectively. **d** Boxplots displays the reads density of chromatin accessibility and
859 histone modifications on Young-D, Younger, and Youngest groups.

860 **Figure 5. Functional enrichment of young SINE-derived transcripts. a** The tSNE plots
861 display the expression differentiation among different tissues and cells. **b** Top 20 results of
862 functional enrichment analysis for young SINE-derived genes. **c** The bar plot indicates the
863 enrichment of 248 young-D SINEs in chromatin states and histone modifications across
864 different tissues.

865 **Figure 6. Young SINE-derived genetic diversity of pigs. a** The Pig TE polymorphism
866 pipeline. The pipeline was constructed to identify both polySINEs and SNPs for each
867 individual simultaneously. **b** Overviews of whole-genome re-sequencings in 374 individuals.
868 **c** Venn plot represents the distribution of polySINEs among different populations. **d** PCA plot
869 displays the genetic relationship based on polySINEs among 374 individuals. **e** PCA plot
870 displays the genetic relationship based on polySINEs among 364 individuals from the modern
871 pigs (*Sus scrofa*). **f** Phylogenetic tree based on polySINEs for 374 individuals. **h** Population
872 structure based on polySINEs for 374 individuals when K was 3 and 10.

873 **Figure 7. Potential candidate genes for young SINE-derived local adaptation. a** The
874 scatter diagram displays the 75 PCGs that are possibly associated with local adaptation. The

x-axis represents the *Fst,* and the y-axis represents population frequency. **b** Chromatin accessibility and histone modifications for *FRZB* and their ploySINE. **c** Bar plot displays the population frequency of the polySINEs in the first exon of *RUNX3*. **d** Overviews of nine candidate genes under local adaptation. Bar charts indicate the population frequency of candidate polySINEs and the average TPM values of their corresponding candidate genes across tissues. **e** The scatter diagram displays the linkage disequilibrium between T-SNPs and polySINEs. The x-axis represents the chromosome, and the y-axis represents the r2 values.

Figure 8. Mapping young SINEs to the complex traits. a Bar plot displays the enrichment of T-ploySINEs in different chromatin states. **b** Heatmap displays the frequency of T-polySINEs among different pig populations. The darker red color represents a higher population frequency for T-polySINEs. **c** Population frequency of T-polySINEs in *ELOVL3* gene among different pig populations. **d** Population frequency of T-polySINEs in *ANK2* gene among different pig populations. **e** The expression of the *ANK2* gene at the top 10 tissues sorted by gene expression. **f** The expression of the *VRTN* gene at the top 10 tissues sorted by gene expression. **g** Population frequency of T-polySINEs in *VRTN* gene among different pig populations. **h** The *VRTN* gene structure and the neighboring SINE-derived transcript. **i** Chromatin accessibility and histone modifications for the upstream of *VRTN* gene. H3K27me3 signals for the upstream of *VRTN* gene.

Acknowledgements

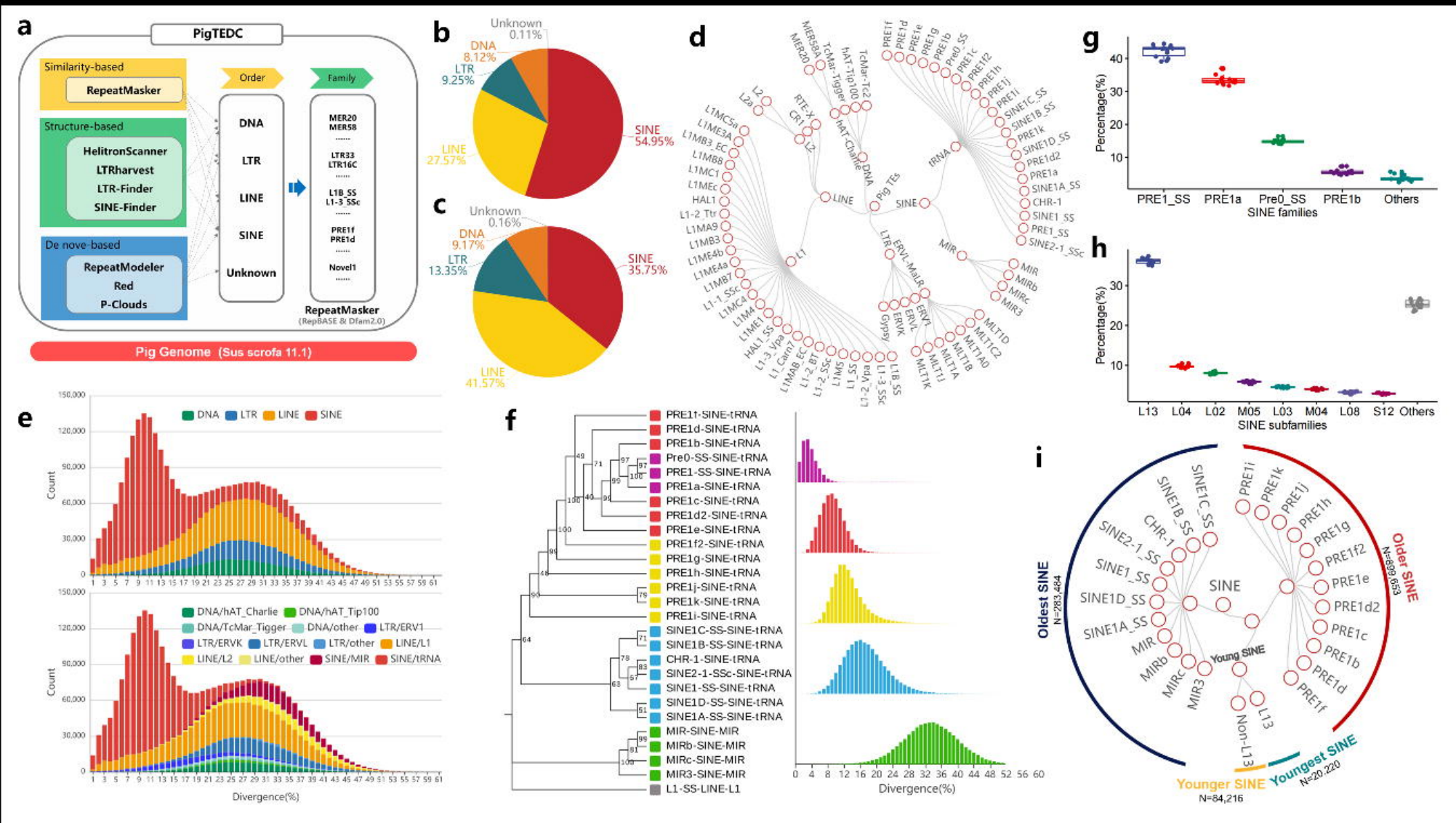
This work is financially supported by the Sanya Yazhou Bay Science and Technology City (202002007), the National Natural Science Foundation of China (31941007), and the AFRI grant numbers 2019-67015-29321 and 2021-67015-33409 from the USDA National Institute of Food and Agriculture (NIFA). We thank Sylvain Foissac from Université de Toulouse and other members associated with the FRAGENCODE project for their contributions to data of chromatin accessibility used in this manuscript.

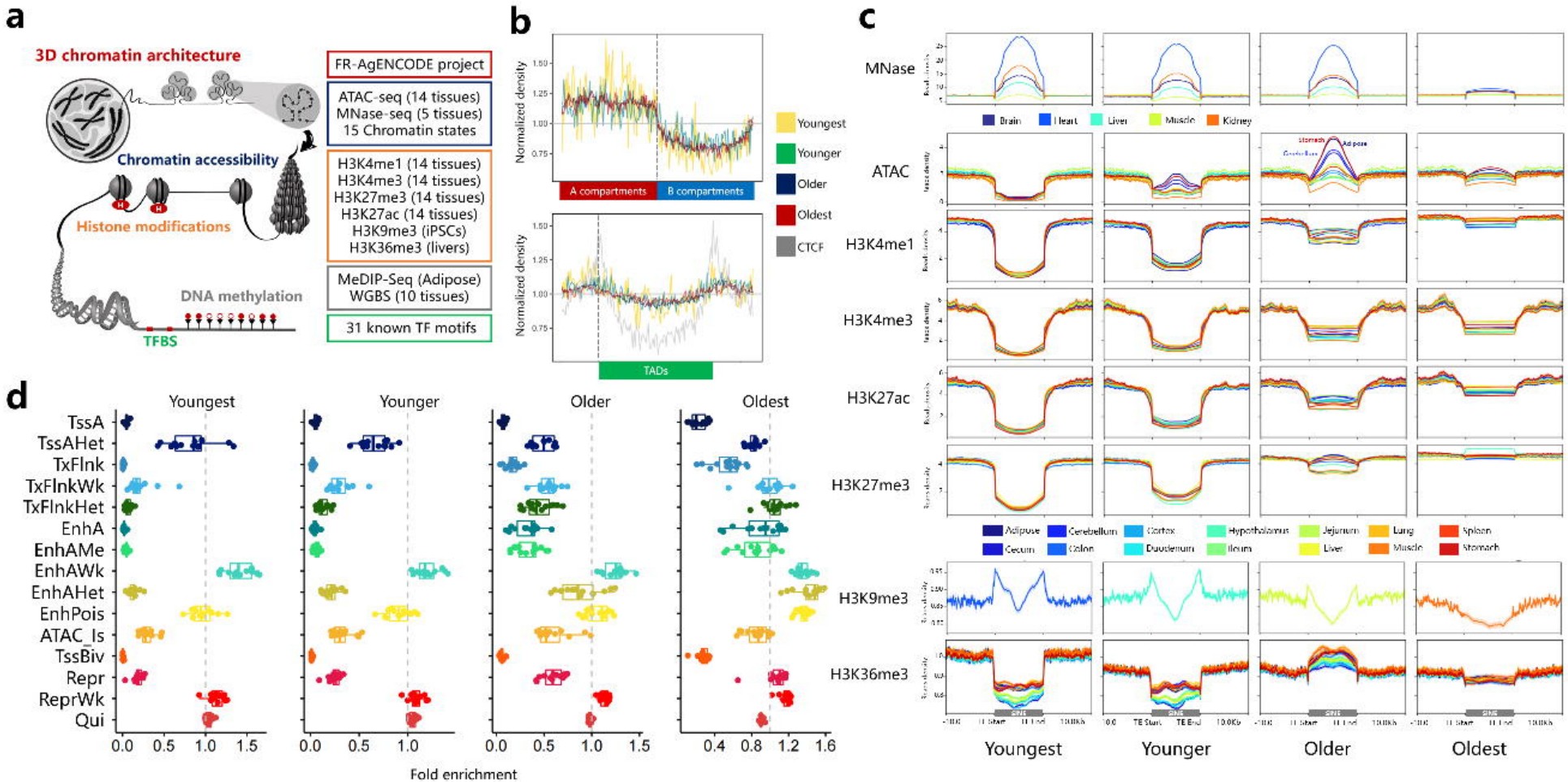
Author's contribution

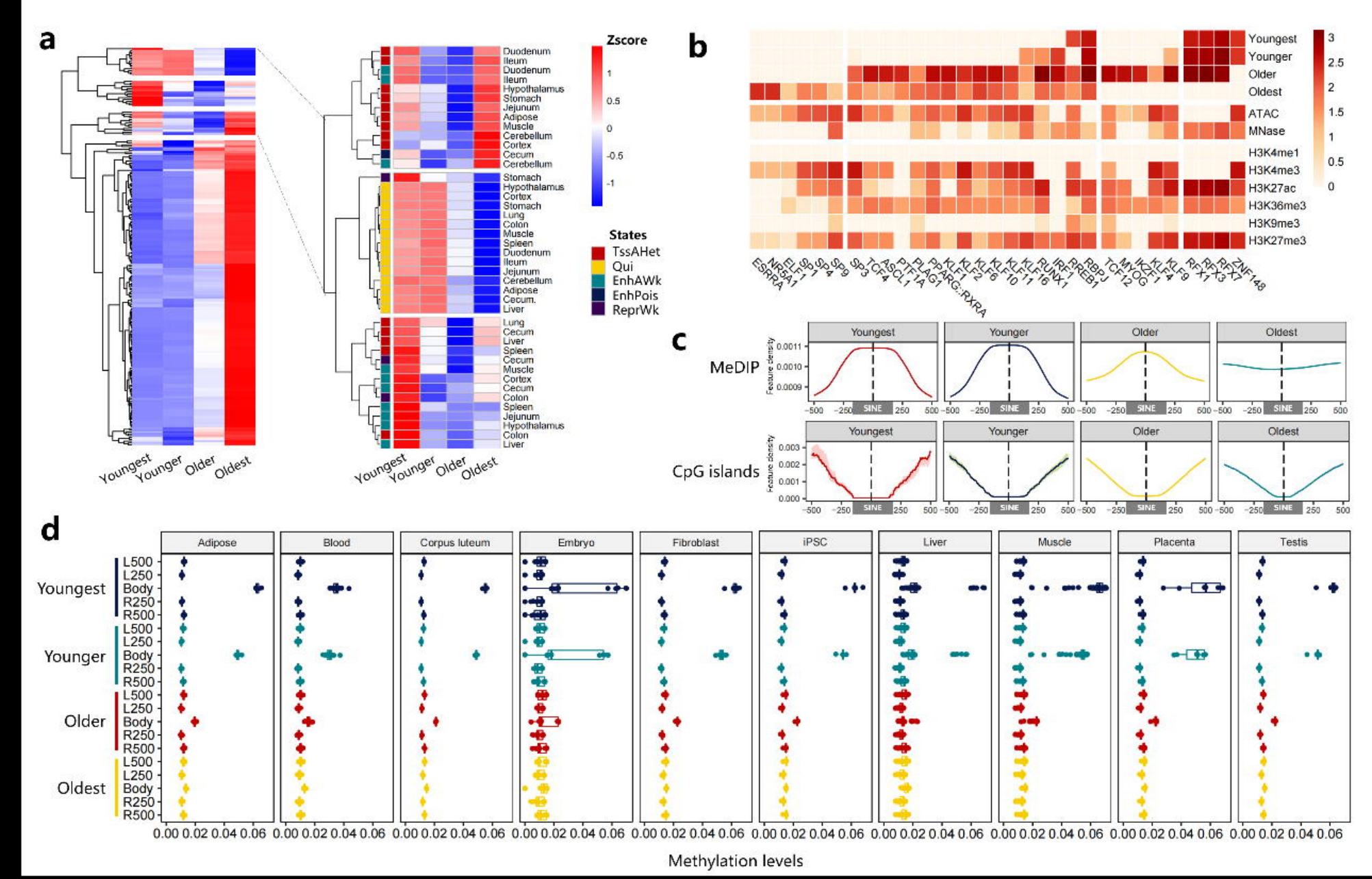
P.Z., Z.W., L.F., and G.L. conceived and designed the experiments. P.Z. designed analytical strategy and performed all analysis process. Y.G. assisted the DNA methylation analysis. Z.P. assisted the Histone modifications analysis. L.L. and X.L. assisted the bioinformatics analysis process. L.G., H.Z., X.H., and L.Q. collected and prepared for sequencing data. P.Z., G.L., L.F., Z.W., H.Z., and D.Y. wrote and revised the paper. All authors read and approved the final manuscript.

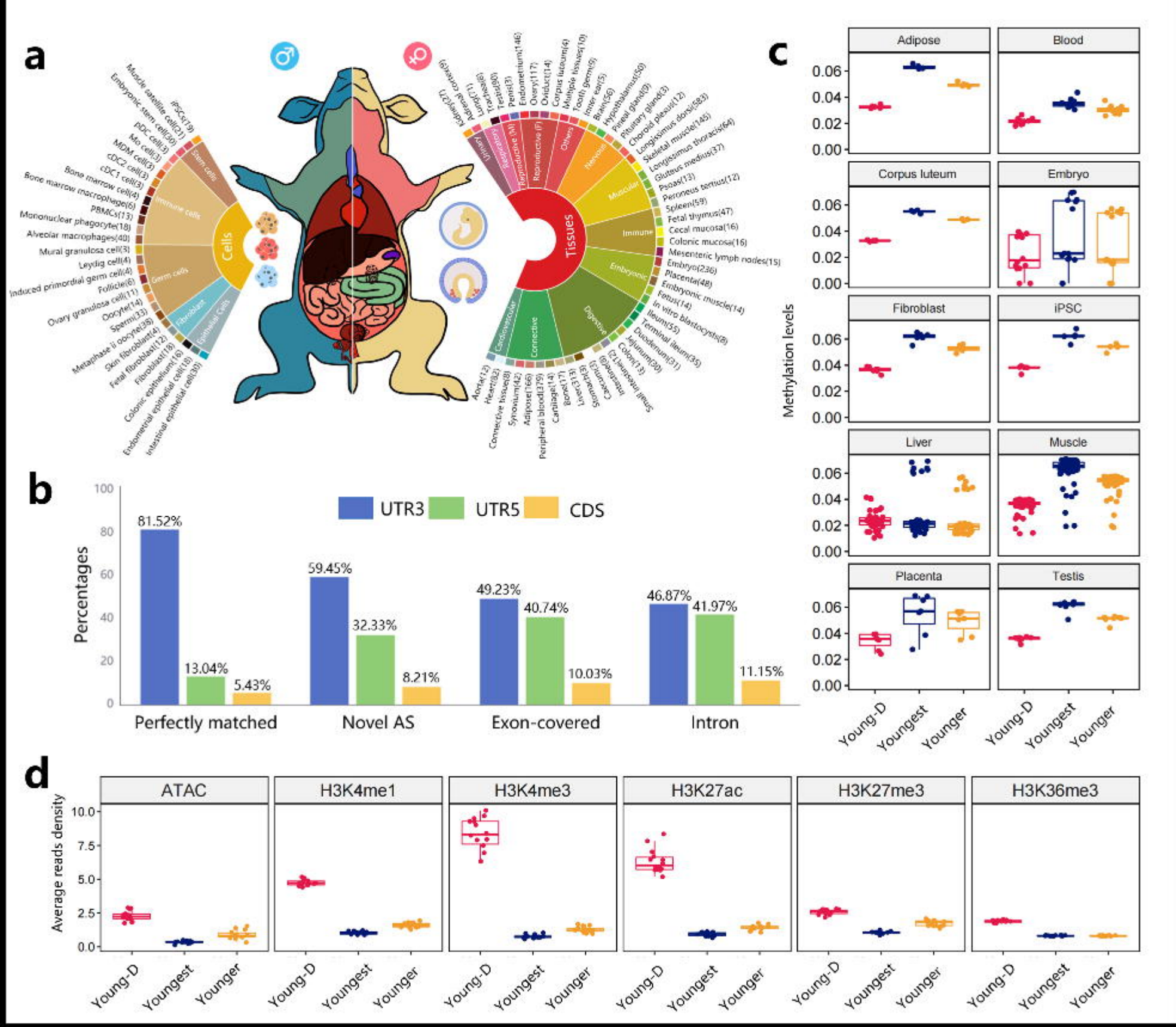
Competing interests

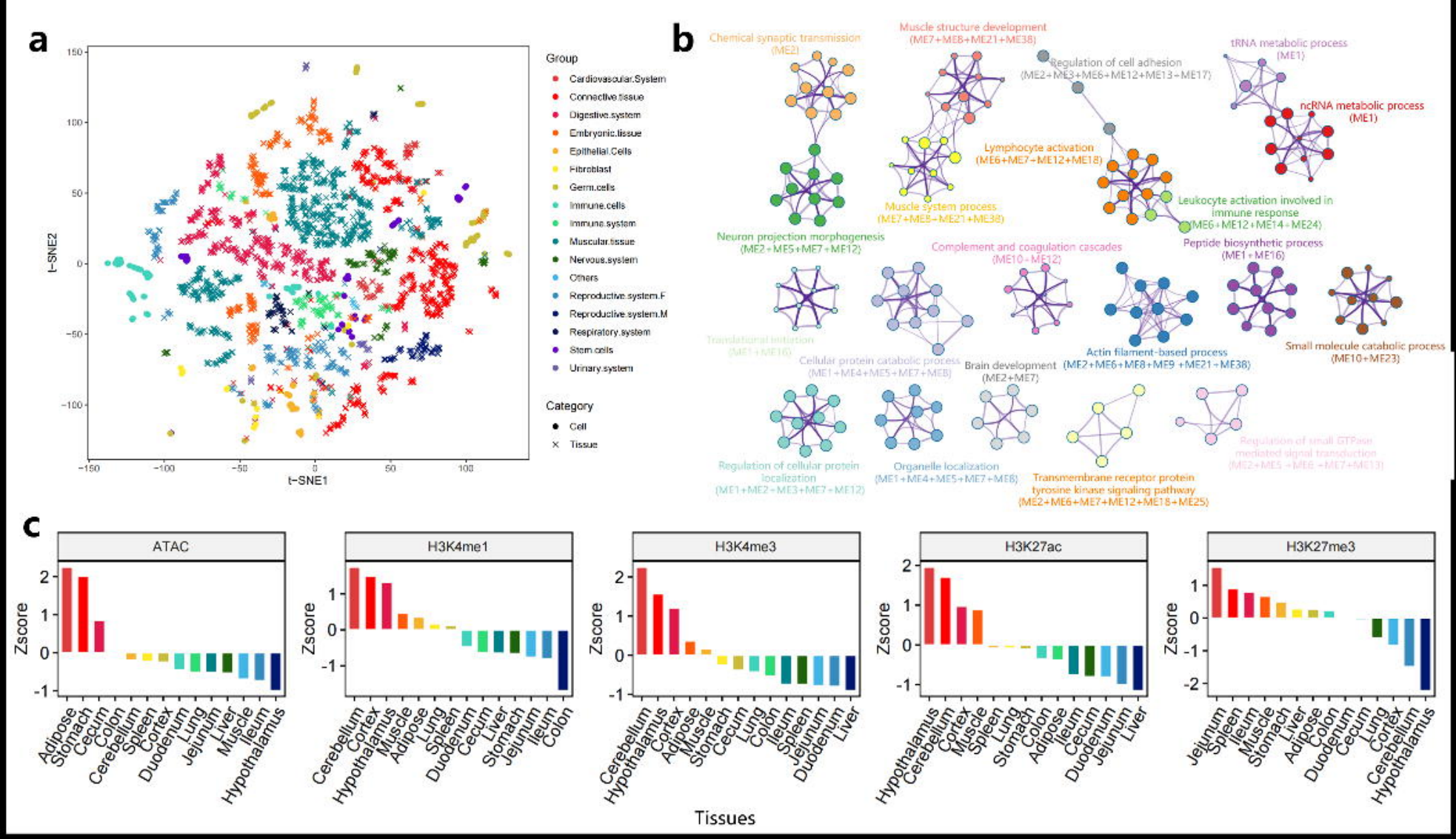
The authors declare no competing interests.

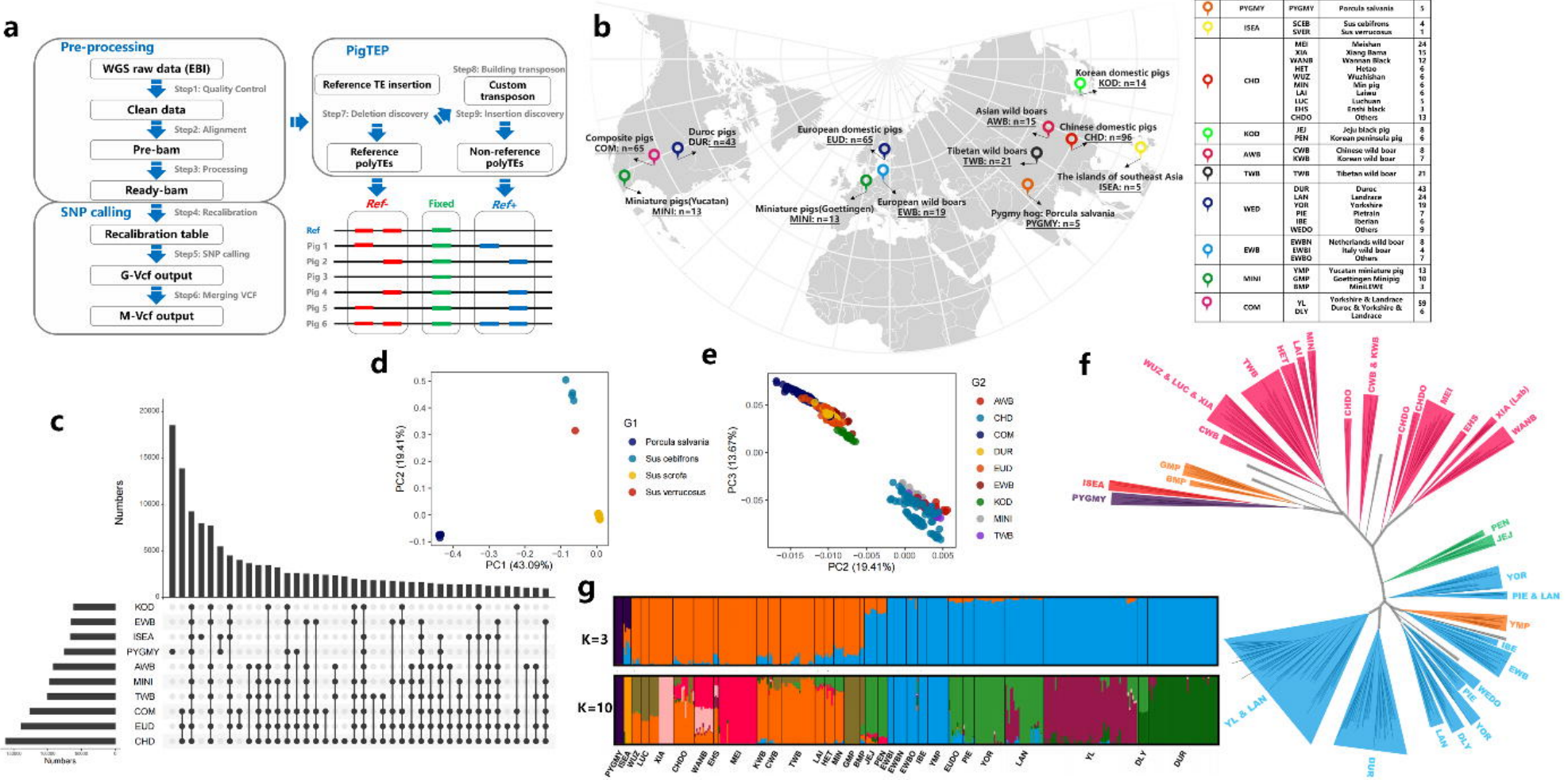

a
PigTEDC
Similarity-based
RepeatMasker
Structure-based
HelitronScanner
LTRharvest
LTR-Finder
SINE-Finder
De nove-based
RepeatModeler
Red
P-Clouds
Order
DNA
LTR
LINE
SINE
Unknown
Family
MER20
MER58
LTR33
LTR16C
L1B_SS
L1-3_SSc
PRE1f
PRE1d
Novel1
RepeatMasker
(RepBASE & Dfam2.0)
Pig Genome (Sus scrofa 11.1)
b
Unknown 0.11%
DNA 8.12%
LTR 9.25%
SINE 54.95%
LINE 27.57%
c
Unknown 0.16%
DNA 9.17%
LTR 13.35%
SINE 35.75%
LINE 41.57%
d
L2
L2a
L1MC5a
L1ME3A
L1MB3_EC
L1MB8
L1MC1
L1MEc
HAL1
L1-2_Ttr
L1MA9
L1MB3
L1ME4b
L1ME4a
L1MB7
L1-1_SSc
L1MC4
L1M4
L1ME1
HAL1_SS
L1-3_Vpa
L1_Carn7
L1MA8_EC
L1-2_BT
L1-2_SSc
L1M5
L1_SS
L1-2_Vpa
L1-3_SSc
L1B_SS
L1
MER58a
MER20
TcMar-Tigger
hAT-Tip100
TcMar-Tc2
RTE-X
CR1
L2
hAT-Charlie
DNA
LINE
Pig TEs
SINE
PRE1f
PRE1d
PRE1e
PRE1g
PRE1b
Pre0_SS
PRE1c
PRE1f2
PRE1h
PRE1i
SINE1C_SS
SINE1B_SS
PRE1k
SINE1D_SS
PRE1d2
PRE1a
SINE1A_SS
CHR-1
SINE1_SS
PRE1_SS
SINE2-1_SSc
tRNA
LTR
ERVL-MaLR
ERV1
ERVK
Gypsy
MIR
MIR
MIRb
MIRc
MIR3
MLT1D
MLT1C2
MLT1A0
MLT1B
MLT1A
MLT1J
MLT1K
g
Percentage(%)
PRE1_SS
PRE1a
Pre0_SS
PRE1b
Others
SINE families
h
Percentage(%)
L13
L04
L02
M05
L03
M04
L08
S12
Others
SINE subfamilies
e
DNA
LTR
LINE
SINE
Count
DNA/hAT_Charlie
DNA/hAT_Tip100
DNA/TcMar_Tigger
DNA/other
LTR/ERV1
LTR/ERVK
LTR/ERVL
LTR/other
LINE/L1
LINE/L2
LINE/other
SINE/MIR
SINE/tRNA
Count
Divergence(%)
f
PRE1f-SINE-tRNA
PRE1d-SINE-tRNA
PRE1b-SINE-tRNA
Pre0-SS-SINE-tRNA
PRE1-SS-SINE-tRNA
PRE1a-SINE-tRNA
PRE1c-SINE-tRNA
PRE1d2-SINE-tRNA
PRE1e-SINE-tRNA
PRE1f2-SINE-tRNA
PRE1g-SINE-tRNA
PRE1h-SINE-tRNA
PRE1j-SINE-tRNA
PRE1k-SINE-tRNA
PRE1i-SINE-tRNA
SINE1C-SS-SINE-tRNA
SINE1B-SS-SINE-tRNA
CHR-1-SINE-tRNA
SINE2-1-SSc-SINE-tRNA
SINE1-SS-SINE-tRNA
SINE1D-SS-SINE-tRNA
SINE1A-SS-SINE-tRNA
MIR-SINE-MIR
MIRb-SINE-MIR
MIRc-SINE-MIR
MIR3-SINE-MIR
L1-SS-LINE-L1
Divergence(%)
i
Oldest SINE
N=283,484
Older SINE
N=899,653
Young SINE
Youngest SINE
N=20,220
Younger SINE
N=84,216
SINE
SINE1C_SS
SINE1B_SS
CHR-1
SINE2-1_SS
SINE1_SS
SINE1D_SS
SINE1A_SS
MIR
MIRb
MIRc
MIR3
PRE1i
PRE1k
PRE1j
PRE1h
PRE1g
PRE1f2
PRE1e
PRE1d2
PRE1c
PRE1b
PRE1d
PRE1f
L13
Non-L13

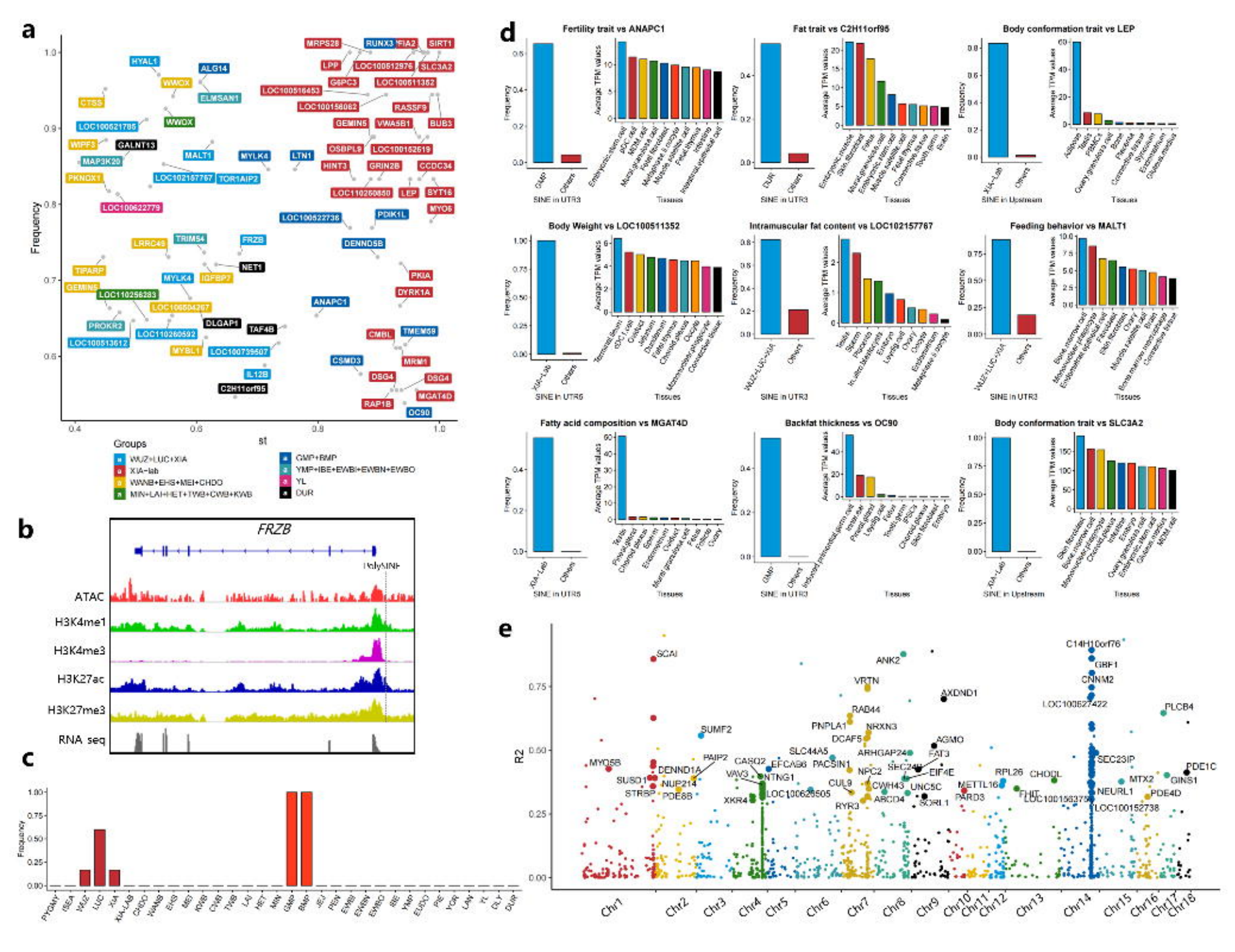

a
3D chromatin architecture
Chromatin accessibility
Histone modifications
DNA methylation
TFBS
FR-AgENCODE project
ATAC-seq (14 tissues)
MNase-seq (5 tissues)
15 Chromatin states
H3K4me1 (14 tissues)
H3K4me3 (14 tissues)
H3K27me3 (14 tissues)
H3K27ac (14 tissues)
H3K9me3 (iPSCs)
H3K36me3 (livers)
MeDIP-Seq (Adipose)
WGBS (10 tissues)
31 known TF motifs
b
Normalized density
A compartments
B compartments
TADs
Youngest
Younger
Older
Oldest
CTCF
c
MNase
ATAC
H3K4me1
H3K4me3
H3K27ac
H3K27me3
H3K9me3
H3K36me3
Brain
Heart
Liver
Muscle
Kidney
Adipose
Cerebellum
Cortex
Hypothalamus
Jejunum
Lung
Spleen
Cecum
Colon
Duodenum
Ileum
Liver
Muscle
Stomach
Youngest
Younger
Older
Oldest
d
Youngest
Younger
Older
Oldest
TssA
TssAHet
TxFlnk
TxFlnkWk
TxFlnkHet
EnhA
EnhAMe
EnhAWk
EnhAHet
EnhPois
ATAC_Is
TssBiv
Repr
ReprWk
Qui
Fold enrichment

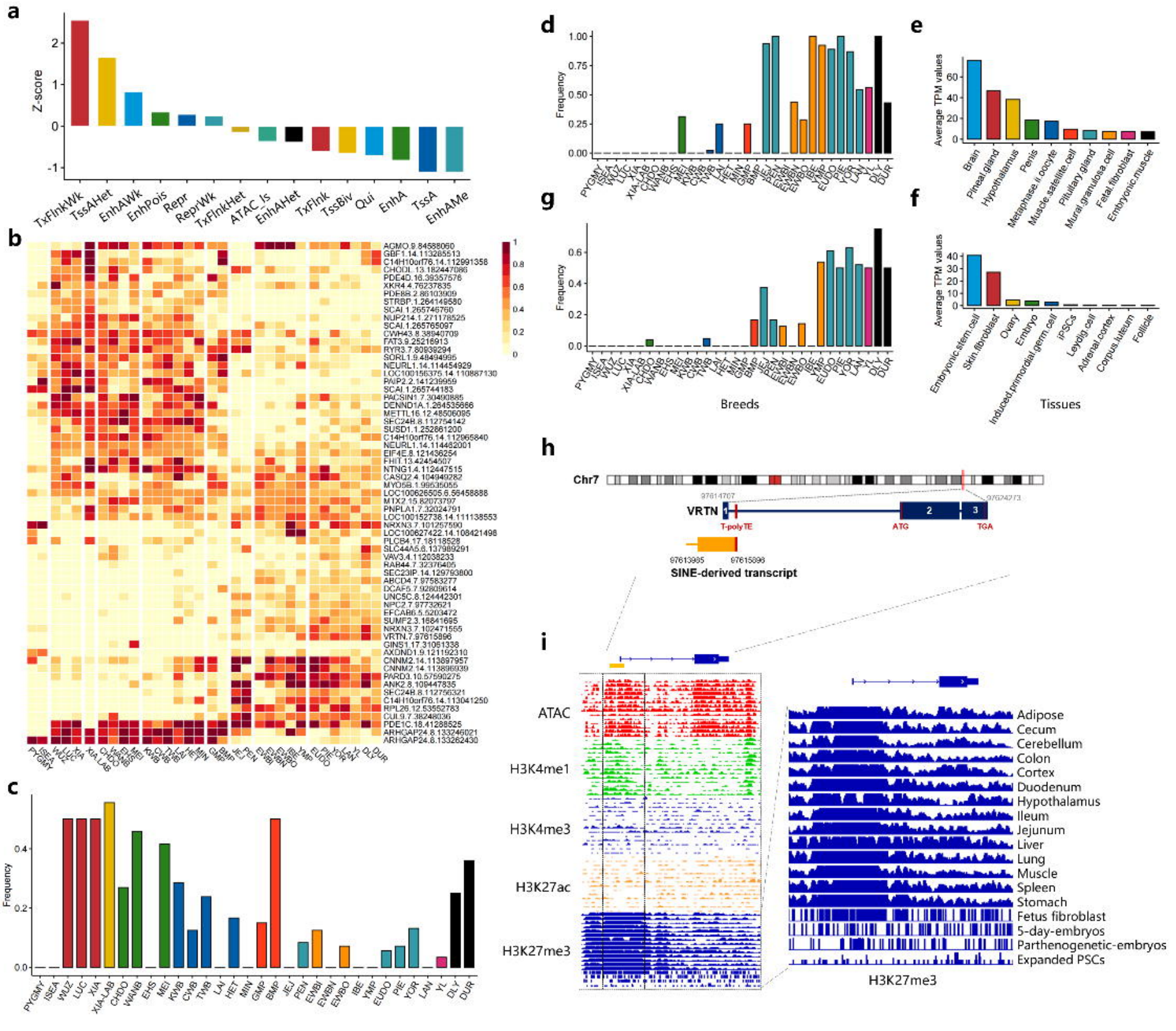

a
Zscore
1
0.5
0
-0.5
-1
States
TssAHet
Qui
EnhAWk
EnhPois
ReprWk
Youngest
Younger
Older
Oldest
b
Youngest
Younger
Older
Oldest
ATAC
MNase
H3K4me1
H3K4me3
H3K27ac
H3K36me3
H3K9me3
H3K27me3
3
2.5
2
1.5
1
0.5
0
ESRRA
NR5A1
ELF1
SP1
SP4
SP9
SP3
TCF4
ASCL1
PTF1A
PLAG1
PPARG::RXRA
KLF1
KLF2
KLF6
KLF10
KLF11
KLF16
RUNX1
IRF1
RREB1
RBPJ
TCF12
MYOG
KLF1
KLF4
KLF9
RFX1
RFX3
RFX7
ZNF148
c
MeDIP
CpG islands
Youngest
Younger
Older
Oldest
Feature density
SINE
d
Adipose
Blood
Corpus luteum
Embryo
Fibroblast
iPSC
Liver
Muscle
Placenta
Testis
Youngest
Younger
Older
Oldest
L500
L250
Body
R250
R500
0.00 0.02 0.04 0.06
Methylation levels


a
Cells
Stem cells
Immune cells
Germ cells
Fibroblast
Epithelial Cells
Tissues
Reproductive (F)
Reproductive (M)
Respiratory
Urinary
Others
Nervous
Muscular
Immune
Embryonic
Digestive
Connective
Cardiovascular
b
UTR3
UTR5
CDS
Percentages
81.52%
13.04%
5.43%
59.45%
32.33%
8.21%
49.23%
40.74%
10.03%
46.87%
41.97%
11.15%
Perfectly matched
Novel AS
Exon-covered
Intron
c
Methylation levels
Adipose
Blood
Corpus luteum
Embryo
Fibroblast
iPSC
Liver
Muscle
Placenta
Testis
Young-D
Youngest
Younger
d
Average reads density
ATAC
H3K4me1
H3K4me3
H3K27ac
H3K27me3
H3K36me3


a
t-SNE2
t-SNE1
Group
Cardiovascular.System
Connective.tissue
Digestive.system
Embryonic.tissue
Epithelial.Cells
Fibroblast
Germ.cells
Immune.cells
Immune.system
Muscular.tissue
Nervous.system
Others
Reproductive.system.F
Reproductive.system.M
Respiratory.system
Stem.cells
Urinary.system
Category
Cell
Tissue
b
Chemical synaptic transmission (ME2)
Muscle structure development (ME7+ME8+ME21+ME38)
Regulation of cell adhesion (ME2+ME3+ME6+ME12+ME13+ME17)
tRNA metabolic process (ME1)
ncRNA metabolic process (ME1)
Lymphocyte activation (ME6+ME7+ME12+ME18)
Muscle system process (ME7+ME8+ME21+ME38)
Leukocyte activation involved in immune response (ME6+ME12+ME14+ME24)
Neuron projection morphogenesis (ME2+ME5+ME7+ME12)
Complement and coagulation cascades (ME10+ME12)
Peptide biosynthetic process (ME1+ME16)
Translational initiation (ME1+ME16)
Cellular protein catabolic process (ME1+ME4+ME5+ME7+ME8)
Brain development (ME2+ME7)
Actin filament-based process (ME2+ME6+ME8+ME9 +ME21+ME38)
Small molecule catabolic process (ME10+ME23)
Regulation of cellular protein localization (ME1+ME2+ME3+ME7+ME12)
Organelle localization (ME1+ME4+ME5+ME7+ME8)
Transmembrane receptor protein tyrosine kinase signaling pathway (ME2+ME6+ME7+ME12+ME18+ME25)
Regulation of small GTPase mediated signal transduction (ME2+ME5+ME6+ME7+ME13)
c
ATAC
H3K4me1
H3K4me3
H3K27ac
H3K27me3
Zscore
Adipose
Stomach
Cecum
Colon
Cerebellum
Spleen
Cortex
Duodenum
Lung
Jejunum
Liver
Muscle
Ileum
Hypothalamus
Tissues


a
Pre-processing
WGS raw data (EBI)
Step1: Quality Control
Clean data
Step2: Alignment
Pre-bam
Step3: Processing
Ready-bam
SNP calling
Step4: Recalibration
Recalibration table
Step5: SNP calling
G-Vcf output
Step6: Merging VCF
M-Vcf output
PigTEP
Reference TE insertion
Step8: Building transposon
Custom transposon
Step7: Deletion discovery
Step9: Insertion discovery
Reference polyTEs
Non-reference polyTEs
Ref-
Fixed
Ref+
Ref
Pig 1
Pig 2
Pig 3
Pig 4
Pig 5
Pig 6
b
Korean domestic pigs
KOD: n=14
Asian wild boars
AWB: n=15
Chinese domestic pigs
CHD: n=96
Composite pigs
COM: n=65
Duroc pigs
DUR: n=43
European domestic pigs
EUD: n=65
Tibetan wild boars
TWB: n=21
The islands of southeast Asia
ISEA: n=5
Miniature pigs(Yucatan)
MINI: n=13
Miniature pigs(Goettingen)
MINI: n=13
European wild boars
EWB: n=19
Pygmy hog: Porcula salvania
PYGMY: n=5
Group Short Full N
PYGMY PYGMY Porcula salvania 5
ISEA SCEB Sus cebifrons 4
SVER Sus verrucosus 1
CHD MEI Meishan 24
XIA Xiang Bama 15
WANB Wannan Black 12
HET Hetao 6
WUZ Wuzhishan 6
MIN Min pig 6
LAI Laiwu 6
LUC Luchuan 5
EHS Ehshi black 3
CHDO Others 13
KOD JEJ Jeju black pig 8
PEN Korean peninsula pig 6
AWB CWB Chinese wild boar 8
KWB Korean wild boar 7
TWB TWB Tibetan wild boar 21
WED DUR Duroc 43
LAN Landrace 24
YOR Yorkshire 19
PIE Pietrain 7
IBE Iberian 6
WEDO Others 9
EWB EWBN Netherlands wild boar 8
EWBI Italy wild boar 4
EWBO Others 7
MINI YMP Yucatan miniature pig 13
GMP Goettingen Minipig 10
BMP MiniLEWE 3
COM YL Yorkshire & Landrace 59
DLY Duroc & Yorkshire & Landrace 6
c
Numbers
KOD
EWB
ISEA
PYGMY
AWB
MINI
TWB
COM
EUD
CHD
d
PC2 (19.41%)
PC1 (43.09%)
G1
Porcula salvania
Sus cebifrons
Sus scrofa
Sus verrucosus
e
PC3 (13.67%)
PC2 (19.41%)
G2
AWB
CHD
COM
DUR
EUD
EWB
KOD
MINI
TWB
f
WUZ & LUC & XIA
CWB
TWB
HET
LAI
MIN
CWB & KWB
CHDO
MEI
EHS
XIA (Lab)
WANB
GMP
BMP
ISEA
PYGMY
PEN
JEJ
YOR
PIE & LAN
YMP
IBE
EWB
WEDO
PIE
YOR
LAN
DLY
DUR
YL & LAN
g
K=3
K=10
PYGMY ISEA WUZ LUC XIA CHDO WANB EHS MEI KWB CWB TWB LAI HET MIN GMP BMP JEJ PEN EWBI EWBN EWBO IBE YMP EUDO PIE YOR LAN YL DLY DUR


a
Frequency
st
Groups
WUZ+LUC+XIA
XIA-lab
WANB+EHS+MEI+CHDO
MIN+LAI+HET+TWB+CWB+KWB
GMP+BMP
YMP+IBE+EWBI+EWBN+EWBO
YL
DUR
b
FRZB
IsoSINE
ATAC
H3K4me1
H3K4me3
H3K27ac
H3K27me3
RNA seq
c
Frequency
PYGMY
ISEA
WUZ
LUC
XIA
XIA-LAB
CHDO
WANB
EHS
MEI
KWB
CWB
TWB
LAI
HET
MIN
GMP
BMP
IEJ
PEN
EWBI
EWBN
EWBO
IBE
YMP
EUDO
PE
YOR
LAN
YL
OLY
DUR
d
Fertility trait vs ANAPC1
Fat trait vs C2H11orf95
Body conformation trait vs LEP
Body Weight vs LOC100511352
Intramuscular fat content vs LOC102157767
Feeding behavior vs MALT1
Fatty acid composition vs MGAT4D
Backfat thickness vs OC90
Body conformation trait vs SLC3A2
Frequency
Average TPM values
SINE in UTR3
SINE in Upstream
Tissues
e
R2
SCAI
ANK2
VRTN
AXDND1
RAD44
SUMF2
PNPLA1
NRXN3
DCAF5
AGMO
SLC44A5
ARHGAP24
FAT3
PAIP2
CASQ2
EFCAB6
PACSIN1
SEC24B
EIF4E
MYO5B
DENND1A
VAV3
NTNG1
NPC2
SEC23IP
SUSD1
NUP214
CUL9
CWH43
UNC5C
RPL26
CHODL
PDE1C
STRBP
PDE6B
XKR4
LOC100628505
ARCD4
SORL1
NETTL16
PARD3
FHIT
LOC100152738
NEURL1
MTX2
GINS1
PDE4D
C14H10orf76
GBF1
CNNM2
LOC100627422
PLCB4
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18


a
Z-score
TxFlnkWk
TssAHet
EnhAWk
EnhPois
Repr
ReprWk
TxFlnkHet
ATAC_Is
EnhAHet
TxFlnk
TssBiv
Qui
EnhA
TssA
EnhAMe
b
c
Frequency
d
Frequency
Breeds
e
Average TPM values
Brain
Pineal gland
Hypothalamus
Penis
Metaphase II oocyte
Muscle satellite cell
Pituitary gland
Mural granulosa cell
Fetal fibroblast
Embryonic muscle
f
Average TPM values
Embryonic stem cell
Skin fibroblast
Ovary
Embryo
Induced primordial germ cell
iPSCs
Leydig cell
Adrenal cortex
Corpus luteum
Follicle
Tissues
g
Frequency
Breeds
h
Chr7
97614707
97624773
VRTN
T-polyTE
ATG
TGA
97613985
97615096
SINE-derived transcript
i
ATAC
H3K4me1
H3K4me3
H3K27ac
H3K27me3
Adipose
Cecum
Cerebellum
Colon
Cortex
Duodenum
Hypothalamus
Ileum
Jejunum
Liver
Lung
Muscle
Spleen
Stomach
Fetus fibroblast
5-day-embryos
Parthenogenetic-embryos
Expanded PSCs
H3K27me3