SCIENCE CHINA
Life Sciences

•RESEARCH PAPER•

May 2020 Vol.63 No.5: 750–763
https://doi.org/10.1007/s11427-019-9551-7

Building a sequence map of the pig pan-genome from multiple *de novo* assemblies and Hi-C data

Xiaomeng Tian[1†], Ran Li[1†], Weiwei Fu[1†], Yan Li[2†], Xihong Wang[1], Ming Li[1], Duo Du[1], Qianzi Tang[2], Yudong Cai[1], Yiming Long[1], Yue Zhao[1], Mingzhou Li[2*] & Yu Jiang[1*]

[1]*Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Yangling 712100, China;*
[2]*Institute of Animal Genetics and Breeding, College of Animal Science and Technology, Sichuan Agricultural University, Chengdu 611130, China*

Received January 7, 2019; accepted April 3, 2019; published online July 8, 2019

Pigs were domesticated independently in the Near East and China, indicating that a single reference genome from one individual is unable to represent the full spectrum of divergent sequences in pigs worldwide. Therefore, 12 *de novo* pig assemblies from Eurasia were compared in this study to identify the missing sequences from the reference genome. As a result, 72.5 Mb of non-redundant sequences (~3% of the genome) were found to be absent from the reference genome (Sscrofa11.1) and were defined as pan-sequences. Of the pan-sequences, 9.0 Mb were dominant in Chinese pigs, in contrast with their low frequency in European pigs. One sequence dominant in Chinese pigs contained the complete genic region of the tazarotene-induced gene 3 (*TIG3*) gene which is involved in fatty acid metabolism. Using flanking sequences and Hi-C based methods, 27.7% of the sequences could be anchored to the reference genome. The supplementation of these sequences could contribute to the accurate interpretation of the 3D chromatin structure. A web-based pan-genome database was further provided to serve as a primary resource for exploration of genetic diversity and promote pig breeding and biomedical research.

pan-genome, pig, reference genome, 3D chromatin structure, presence-absence variation

Citation: Tian, X., Li, R., Fu, W., Li, Y., Wang, X., Li, M., Du, D., Tang, Q., Cai, Y., Long, Y., et al. (2020). Building a sequence map of the pig pan-genome from multiple *de novo* assemblies and Hi-C data. Sci China Life Sci 63, 750–763. https://doi.org/10.1007/s11427-019-9551-7

INTRODUCTION

Sus scrofa (i.e., pig or swine) is of enormous agricultural significance and is an attractive biomedical model. Pigs were domesticated from wild boars independently in the Near East and China approximately 10,000 years ago following long-term gene flow from their local wild counterparts (Ai et al., 2015; Frantz et al., 2015; Groenen et al., 2012; Larson et al., 2005). The current European and Asian pigs have large phenotypic and genomic diversity as a result of geographical divergence and independent demographic events (Li et al., 2017). Therefore, neither the European nor the Asian pig genome is sufficient to capture the whole gene pool of pigs to represent the full range of genetic diversity. The current pig genome Sscrofa11.1 was derived from one European individual (Duroc breed). Although it represents one of the most continuous assemblies in mammals (Figure S1 in Supporting Information), aligning short reads to this single reference genome will lose many Asian variations probably associated with adaptions to various environmental conditions. Alternatively, comparisons of independently *de novo* assemblies and a reference genome sequence promise a more accurate understanding of genetic variations across breeds.

†Contributed equally to this work
*Corresponding authors (Yu Jiang, email: yu.jiang@nwafu.edu.cn; Mingzhou Li, email: mingzhou.li@sicau.edu.cn)

Recently, the pan-genome, the non-redundant collection of all genomic sequences of a species, has emerged as a fundamental resource for unlocking natural diversity in eukaryotes. Intraspecific comparisons in plants (e.g., soybean (Li et al., 2014), *Brassica oleracea* (Golicz et al., 2016), stiff brome (Gordon et al., 2017) and rice (Zhao et al., 2018)) and in animals (e.g., mosquito (Neafsey et al., 2015), macaque (Yan et al., 2011) and human (Li et al., 2010; Sherman et al., 2019)) have revealed surprisingly large amounts of variation within a species. To build a high-quality pan-genome, a number of individual genomes are required (Li et al., 2014; Monat et al., 2016; Wong et al., 2018), which remains an obstacle for most mammalian species. Our previous studies generated *de novo* assemblies of 10 geographically and phenotypically representative pig genomes from Eurasia (Li et al., 2017; Li et al., 2013). Together with the assembly of the Chinese Wuzhishan pig (Fang et al., 2012) and the current reference genome Sscrofa11.1, the availability of 12 pig genomes has provided an unprecedented opportunity to investigate their genetic differences at the individual, ethnic/breed or continental level.

Here, we carried out an in-depth comparison between 11 *de novo* assemblies and the reference genome using assembly-versus-assembly alignment. The final pan-genome comprised 39,744 newly added sequences (total length: 72.5 Mb). We further generated three trillion data from 12 Hi-C samples (Table S1 in Supporting Information). Using these data, the three-dimensional (3D) spatial structure of the pan-genome was depicted by revealing the characteristics of A/B compartment (generally euchromatic and heterochromatic regions, respectively) and topologically associating domain (TAD). We also build a pig pan-genome database (PIGPAN, http://animal.nwsuaf.edu.cn/code/index.php/panPig) to serve as a fundamental resource for uncovering variations within diverse pig breeds.

RESULTS

Initial characterization of pan-sequences in the pig genome

To construct the pig pan-genome, we first aligned 11 assemblies from 11 genetically distinct breeds (five from Europe, and six from China; Figure S2 and Table S2 in Supporting Information) against Sscrofa11.1 using BLASTN to generate the unaligned sequences (Figure 1A). We finally got 136.6 Mb of unaligned sequences from 11 assemblies. The length of the unaligned sequences in each Chinese pig was significantly longer than those in the European pigs (P<0.01) since the reference genome was from a European pig (Figure 1A). Notably, the Wuzhishan assembly, the only male-derived one in this study, contained a larger number of unaligned sequences compared with other assemblies, indicating it might encompass a large proportion of male-specific sequences. After removing redundant sequences and sequences shorter than 300 bp, we obtained 39,744 sequences with a total length of 72.5 Mb (Figure 1B, Supplementary Dataset 1 in Supporting Information), which were absent from Sscrofa11.1 and were defined as pan-sequences in this study (see Materials and methods). The content of repetitive elements (45.91%) and the GC (44.61%) content in these sequences were slightly higher than those of Sscrofa11.1 (45.02% and 41.53%, respectively) (Figure 1A). Except for those derived from the Wuzhishan assembly, the majority of the pan-sequences (74.7%) were shared in at least two assemblies (Figure 1B). Pan-sequences longer than 5 kb contributed 57% of the total length (Figure 1C).

To validate the authenticity of the pan-sequences, we first compared these sequences to 10 mammalian genomes and found that 67.5% of the pan-sequences had homologs in these genomes ($E<1\times10^{-5}$). Species of Cetacea had the greatest number of best hits in accordance with their close evolutionary relationship with pig (Figure 2A). To determine the potential presence of protein-coding genes within the pan-sequences, we used a translated BLAST search of pan-sequences against NCBI's nr database, which yielded an additional 1,322 contigs hitting a chordate protein with ⩾70% identity and an E-value less than 1×10^{-10}. Enrichment analysis revealed that these genes were mostly involved in olfactory transduction, metabolic pathways, immune function and fatty acid pathways (Table S3 in Supporting Information, corrected P-value ⩽0.01).

Population structure of the pan-sequences in European and Chinese pigs

To explore whether these pan-sequences exhibited population-specific characteristics, we retrieved 87 publicly available pig resequencing data (with an average depth of 22×) from China and Europe and aligned them to the pan-genome (Sscrofa11.1 plus the pan-sequences) (Table S4 in Supporting Information). The presence of each sequence was defined by a whole-genome normalized read depth (NRD) of at least 0.2. Using the presence-absence information of these pan-sequences (Table S5 in Supporting Information), the 87 samples were clustered into three distinct groups corresponding to their geographical origin: southern China, northern China and Europe (Figure 2B, Table S4 in Supporting Information), which was consist with the previously reported genetic architecture of domestic pigs (Ai et al., 2015). We found that ~9.0 Mb of pan-sequences were dominant in Chinese pigs, as they showed much higher frequency in Chinese pigs than European pigs (Fisher's exact test, corrected P-value <0.05) (Table S6 in Supporting Information). Another ~2.8 Mb were dominant in European pigs. Among the sequences that dominated in Chinese pigs,

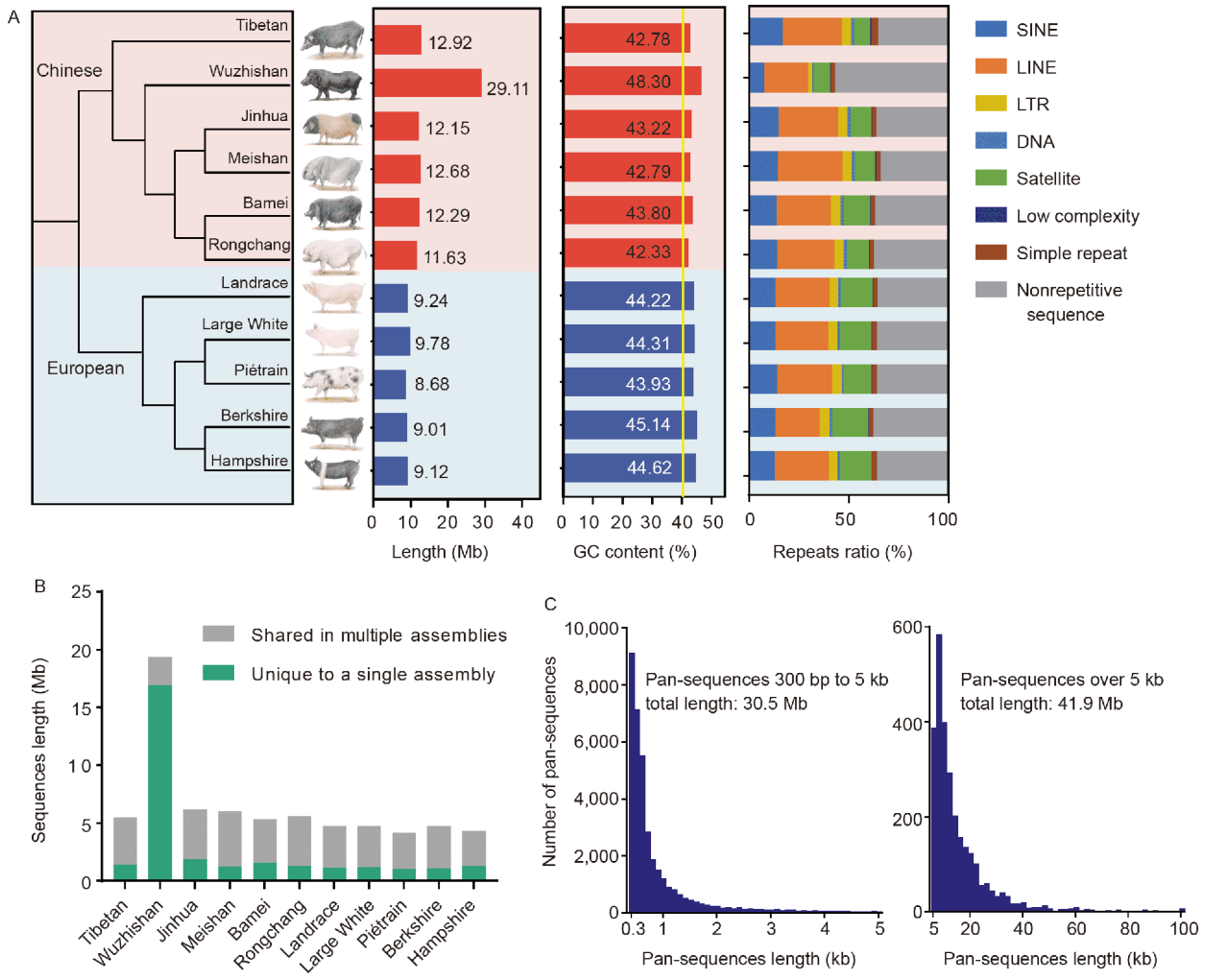

Figure 1 Construction of the pig pan-genome and the characterization of pan-sequences. A, Maximum likelihood phylogenetic tree, sequence length, GC content and repeat composition of missing sequences identified in each individual assembly of 11 breeds (left to right). B, The total sequence length and breed-specific sequence length of each breed for non-redundant pan-sequences. C, Length distribution of all pan-sequences. (Wuzhishan pigs had the largest number of sequences because this individual is the only male among all the 11 assemblies and the sequencing platform of this individual differed from that used for other samples.)

one sequence spanning 14.3 kb contained the complete genic region of tazarotene-induced gene 3 (*TIG3*) which encodes a tumor suppressor and an essential regulator of adipocyte lipolysis by releasing free fatty acids from glycerophospholipids (Uyama et al., 2012) (Figure 2C). We calculated the frequency of this gene in the 87 resequencing data, and found that *TIG3* was present in most of the Chinese pigs (62/71) but had low frequency in European pigs (1/16). Among the 10 assemblies whose RNA-seq data were available (Li et al., 2017), *TIG3* was found in all the Chinese pigs except the Tibetan pig whereas it was detected only in one European pig. Meanwhile, the RNA-seq data of these 10 pigs reflected that this gene showed considerable expression in subcutaneous adipose as well as other tissues for the pigs harbouring this gene (Figure 2D, Figure S3 in Supporting Information). We further examined the presence and protein structure of *TIG3* in mammals, chicken, alligator and zebrafish and found that this gene is under negative selection (Figures S4 and S5 in Supporting Information). The presence of *TIG3* in the outgroup suggests that its loss in European pigs could be due to different selection process as compared with Chinese pigs. In addition, its loss in human is associated with Poland Syndrome (Vaccari et al., 2014) characterized by fatty tissue atrophy (Christopoulos et al., 2018). Therefore, the presence/absence of this gene might be one of the genomic variations underlying physiological process difference among different groups of pigs.

Notably, we found that the 42 males had more pan-sequences (23,366 on average) than the 45 females (16,946 on average) (*t*-test, $P<0.01$). Therefore, we further examined the candidate male-specific sequences (from the male-derived Wuzhishan assembly) using the whole genome sequencing data (Table S4 in Supporting Information). In total, 10.4 Mb of male-specific sequences were identified by showing pre-

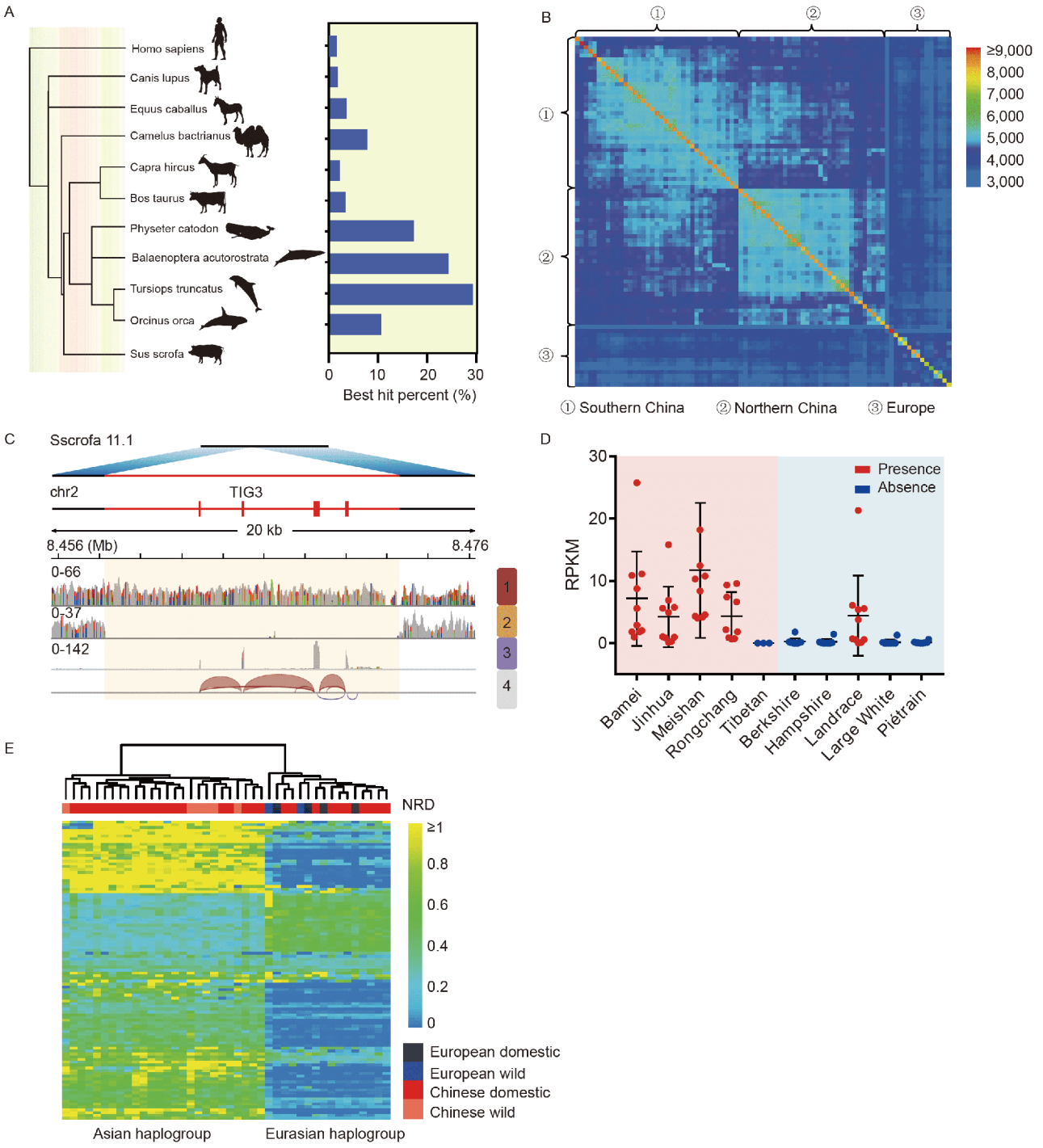

Figure 2 Pan-sequences validation and population-specific pattern. A, Homologue identification of pan-sequences in 10 mammalian genomes. Only the best hit was retained for each pan-sequence. B, An 87×87 matrix showing the number of shared pan-sequences among all the individuals by pairs. Each cell represents the number of shared pan-sequences by two individuals. See Table S3 in Supporting Information for the classification of each group. C, Genes contained in the pan-sequences. One pan-sequence of 14.3 kb harbours the complete genic region of *TIG3*. The four tracks at the bottom represent the reads mapping of whole-genome resequencing data of two samples (labelled "1" and "2") and the inferred exons as well as their splicing isoforms based on RNA-seq (labelled "3" and "4"). D, The expression of *TIG3* in 92 RNA-seq samples from 10 animals from China (light red background) and Europe (light blue background). The 10 animals corresponded to 10 of our 11 assemblies used in this study excluding the Wuzhishan assembly. E, The normalized read depth (NRD) of male-specific pan-sequences in each male. See Table S3 in Supporting Information for the classification of each group. (Only the sequences with the frequency ranging from 0.5 to 0.9 are shown.)

sence in more than half of the male individuals and absence in all females (Supplementary Dataset 2 in Supporting Information). Most of the male-specific sequences (10.3 Mb) were present in a high proportion of the males (≥90%). A small number of the male-specific sequences (171 kb) were found in 50%–90% of males and demonstrated a clear po-

pulation-specific pattern in the form of revealing two haplogroups: one Asian type confined to Asia and one Eurasian type distributed across Eurasia as has been reported in previous studies (Guirao-Rico et al., 2018) (Figure 2E).

Hi-C based analysis revealing the characteristics of pan-sequences regarding 3D structures and their potential function

Anchoring pan-sequences to the reference genome with refined positions will help to understand the functions of the pan-sequences. In order to maximally resolve the location of pan-sequences, we applied the common practice of using flanking sequences as well as a new strategy, Hi-C. The genomic regions that are close to each other tend to demonstrate frequent interactions which can be inferred from Hi-C data (Dong et al., 2017), thus providing us a new approach to anchor these pan-sequences to Sscrofa11.1. Here, we preformed 12 Hi-C experiments (~3 Tb data) in 10 individuals to anchor these pan-sequences to Sscrofa11.1 by inferring with their interactions with adjacent regions (Tables S1 and S7 in Supporting Information). Utilizing flanking sequences and the Hi-C method, 11,028 (27.7%) of the pan-sequences were anchored to Sscrofa11.1 (Table S8 in Supporting Information). By providing pan-sequences with refined positions, the genomic annotation was improved. For instance, a 18.6-kb pan-sequence containing six exons of zinc finger protein 622 (*ZNF622*) was lost in Sscrofa11.1 (Figure S6 in Supporting Information). *ZNF622* plays a role in embryonic development by activating the DNA-bound *MYBL2* transcription factor (Arumemi et al., 2013). The absence of the six exons resulted in an alternatively spliced isoform, as is evidenced in the RNA-seq data (Figure S6).

Hi-C also enabled us to understand the characteristics of the pan-genome's 3D structure (Figure 4A). Based on the positions resolved at 100-kb resolution, we found that pan-sequences were not enriched in the A compartment (euchromatic regions) or the B compartment (heterochromatic regions) (Figure 3B). At 20-kb resolution, we found that pan-sequences were enriched at TAD boundaries (Figure 3C). Notably, singe-nucleotide polymorphisms (SNPs) occurred more frequently at TAD boundary regions than at the TAD interior regions (Figure S7 in Supporting Information), indicating that the occurrence of pan-sequences could be associated with genomic variations.

We used one sample with high genome coverage (~300×) to identify promoter-enhancer interactions (PEIs) (Figure S8 in Supporting Information). In total, 47 of the pan-sequences contained potential enhancers, as they demonstrated interactions with known promoters (Table S9 in Supporting Information). The genes under the regulation of these enhancer-promoter interactions were significantly enriched in retinol metabolism, olfactory transduction, arachidonic acid metabolism and fatty acid degradation (Table S10 in Supporting Information). When the corresponding regions of low interaction were replaced with the pan-sequences, the interaction of pan-sequences with flanking sequences was revealed by more read contacts compared with the original weak interaction of the counterparts in the genome with the flanking sequences (Figure 3D). Therefore, our pan-sequences will help to accurately depict the 3D structures of the whole genome.

Improvement in read alignment and mapping efficacy using pan-sequences

Compared with Sscrofa11.1, the mapping rate of 87 re-sequencing data in the pan-genome was increased by 0.29%–0.43% (Figure 4A). Meanwhile, the mapping rate of pan-base (the Sscrofa11.1 proportion in the pan-genome) was decreased by approximately 1.43%, indicating that many reads had been adjusted to better positions in the pan-sequences accompanied by improved quality (Figure 4A and B). The adjustment of many reads from Sscrofa11.1 to pan-sequences improved SNP calling accuracy. An average of 41,729 heterozygous SNPs per sample were depressed when the read depth was adjusted to the whole-genome average level in these regions (Figure 4B–D, Figure S9 and Table S11 in Supporting Information). Meanwhile, ~12,888 novel SNPs per individual were recovered using the pan-genome thus providing enhanced resolution for genetic diversity studies.

The mapping quality and mapping rate of RNA-seq data were also improved based on the 92 samples of the 10 individuals as mentioned above (Figures S10 and S11 in Supporting Information). In total, 897 sequences containing 1,163 potential transcripts showed appreciable expression (FPKM≥1 in at least one sample) (Figure S12 in Supporting Information).

Pig pan-genome database

To facilitate the use of the pig pan-genome by the scientific community, a pig pan-genome database (PIGPAN) was constructed. PIGPAN is a comprehensive repository of integrated genomics, transcriptomics and regulatory data. The system diagram is shown in Figure 5A. In our local UCSC Genome Browser server (Gbrowse), 17 tracks were released against the pig pan-genome (Figure 5B). Users can search the database using a gene symbol or chromosome location to obtain results in terms of four aspects: (i) the reference genome and pan-sequence annotations (Supplementary Dataset 3 in Supporting Information), (ii) the gene expression in 10 corresponding tissues, seven types of regulation signals and the conserved elements of a 20-way mammalian alignment, (iii) the chromosomal locations of pan-sequences, and

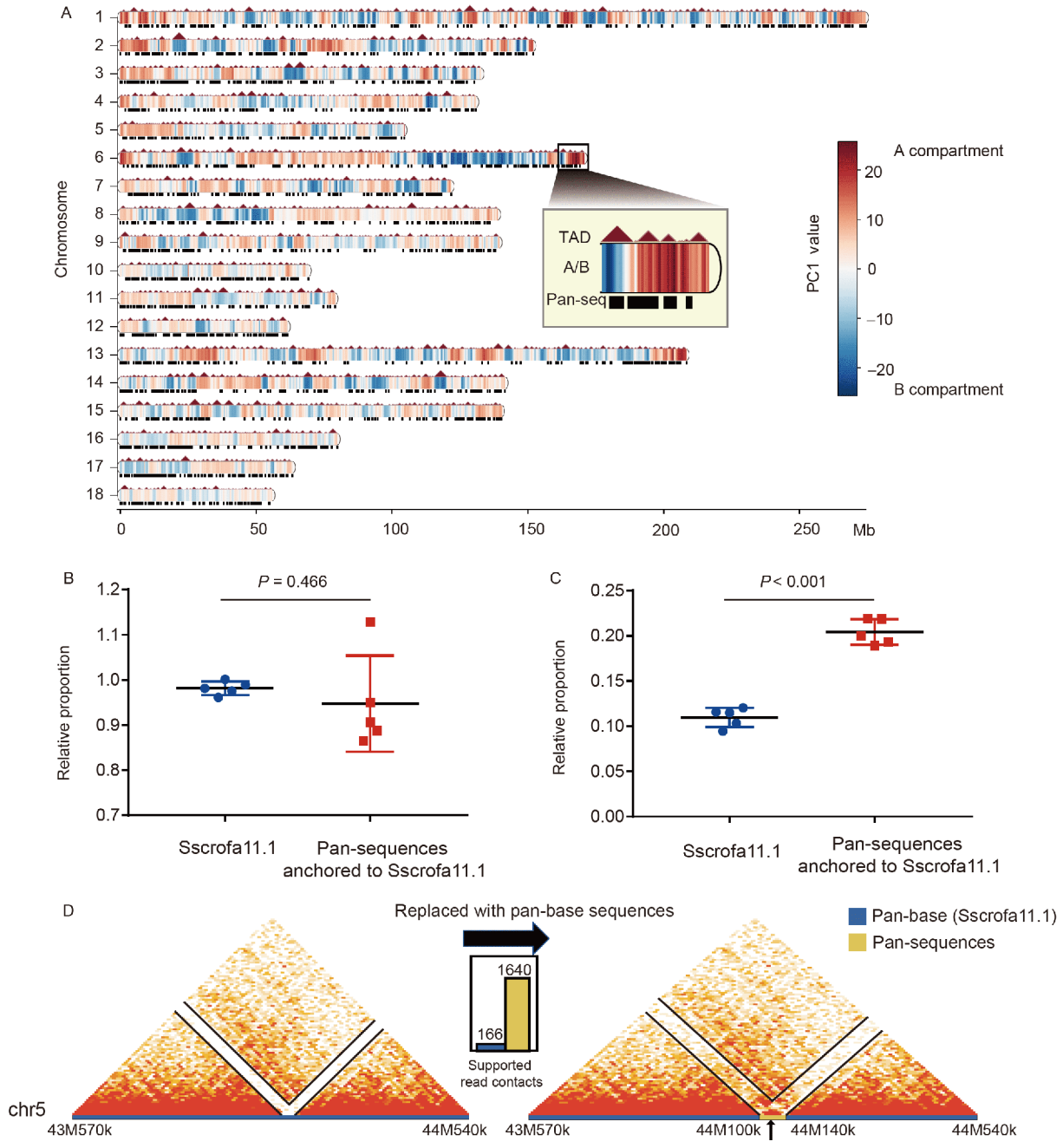

Figure 3 The 3D spatial structure of the pan-genome. A, The distributions of the A/B compartment, TAD and anchored pan-sequences. B, The relative length-proportion of the A compartment over the B compartment in the pig genome (left) and the relative length-proportion of pan-sequences located in the A compartment over those located in the B compartment (right). C, The relative length-proportion of TAD boundary regions over TAD interior regions in Sscrofa11.1 (left) and the relative length-proportion of pan-sequences located in TAD boundary regions over TAD interior regions (right). D, An example of improving a 3D spatial structure after replacing the weakly interacting sequences with the non-reference pan-sequences. The interaction of pan-sequences with flanking sequences was supported by more read contacts than the original interaction of the counterparts in the genome with the flanking sequences.

(iv) the haploid copy number of 87 pigs (Supplementary Dataset 4 in Supporting Information). We also provided basic search functions to retrieve basic gene information, GO annotation and KEGG pathways. Here, we present one case using PIGPAN showing the copy number difference of the *KIT* gene between European and Chinese pigs (Figure 5C). Moreover, users can download data from http://animal.nwsuaf.edu.cn/panPig/download.php. As the functions and associated traits of more genes in the pig genome are determined in the future, our browser will be updated regularly to meet the various needs of the scientific community.

DISCUSSION

In this study, we utilized 12 independent *de novo* assemblies

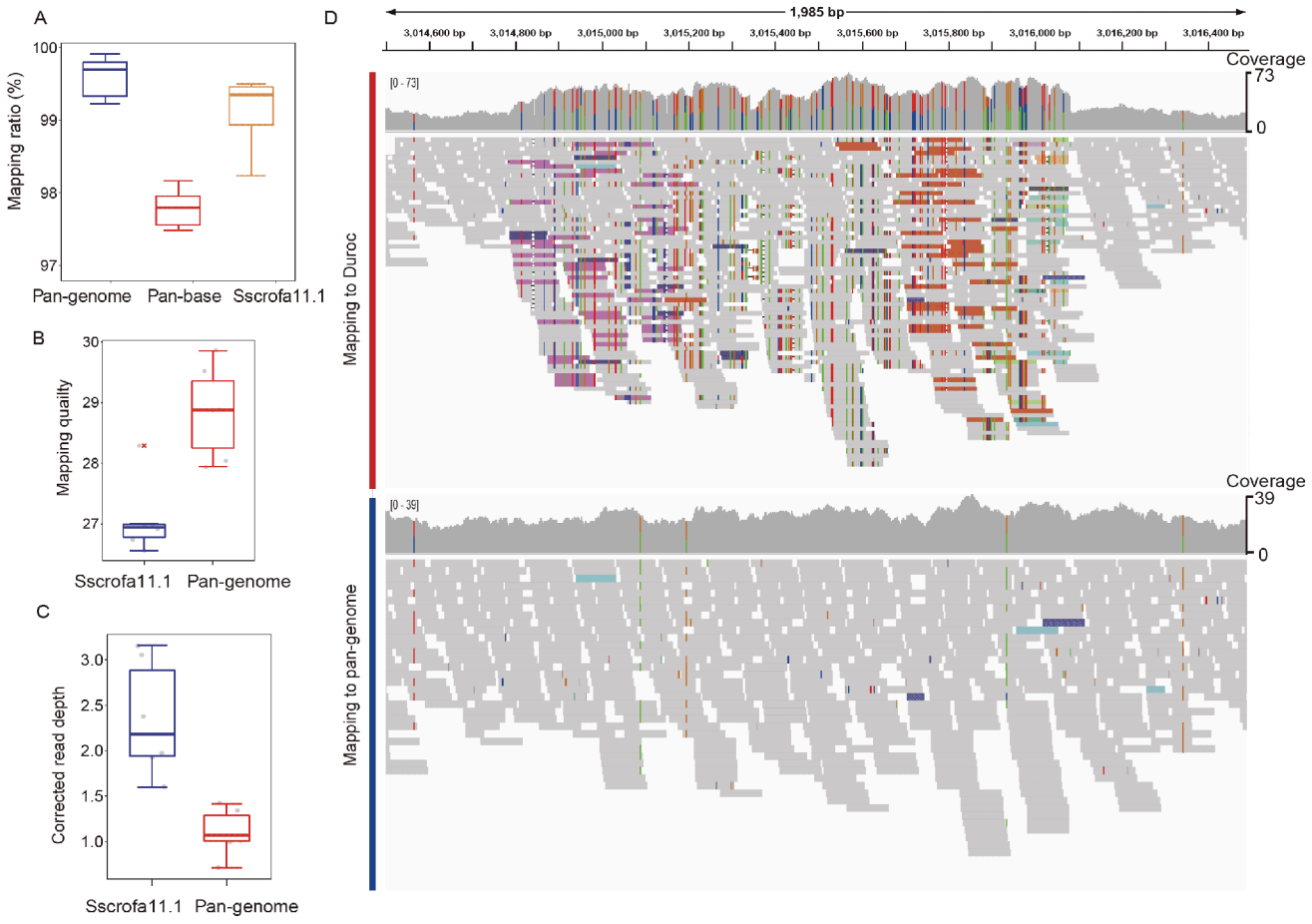

Figure 4 Improvements of genomic analyses by using the pan-genome. A, Comparison of the mapping ratio of resequencing data using the pan-genome versus Sscrofa11.1. B, Comparison of read-mapping quality using the pan-genome versus Sscrofa11.1. C, Comparison of corrected read-mapping depth using the pan-genome versus Sscrofa11.1. D, Improved read mapping using the pan-genome versus Sscrofa11.1 as viewed with IGV.

(Fang et al., 2012; Li et al., 2017; Li et al., 2013) and a large amount of whole-genome resequencing data to build a sequence map of the pig pan-genome. The *de novo* assemblies in the present study cover a wide range of diverse breeds across Eurasia and thus ensure a comprehensive discovery of the missing pan-sequences. These pan-sequences as well as the accompanying genomic variation and expression information will be a valuable resource for fully depicting porcine phenotypic and genomic diversity.

The importance of pan-genomes has been widely accepted in the field of plant genomics (Golicz et al., 2016; Hirsch et al., 2014; Schatz et al., 2014; Sun et al., 2017; Zhao et al., 2018). The high genomic plasticity of plants can result in the complete gain/loss of a large number of genes within a species (Golicz et al., 2016; Gordon et al., 2017; Zhao et al., 2018). In contrast, animal genomes are much more conserved and have longer genes with complex splicing events, which means that generally, only intergenic or fragmented genic regions are involved in the gain/loss of genomic sequences in animals. Nonetheless, this difference does not mean that animal pan-genomes are less important. Instead, the pan-sequences could represent an important type of structure variations that contribute to the phenotypic variations. This study shows that ~9.0 Mb of the pan-sequences are dominant in Chinese pigs. Among them, one sequence contains *TIG3*, an essential regulator of adipocyte lipolysis whose expression can result in accumulation of free fatty acids (Uyama et al., 2012). This finding indicates that the pan-sequences should be taken into consideration when exploring the genetic mechanisms underlying differences in growth rate and fat deposition between European and Chinese pigs. Furthermore, our research suggests that these pan-sequences may act as enhancers of some genes that regulate metabolic activity in different breeds. We also found a large number of SNPs residing in the pan-sequences, which can lead to an accurate assessment of true variations, thereby providing enhanced resolution of the genetic diversity of different pig populations. The enriched genomic sequence repertoire can help in identifying causal mutations that were previously unrecognized by linkage, association and copy-number-variation studies.

In conclusion, our study has shown that the pan-genome,

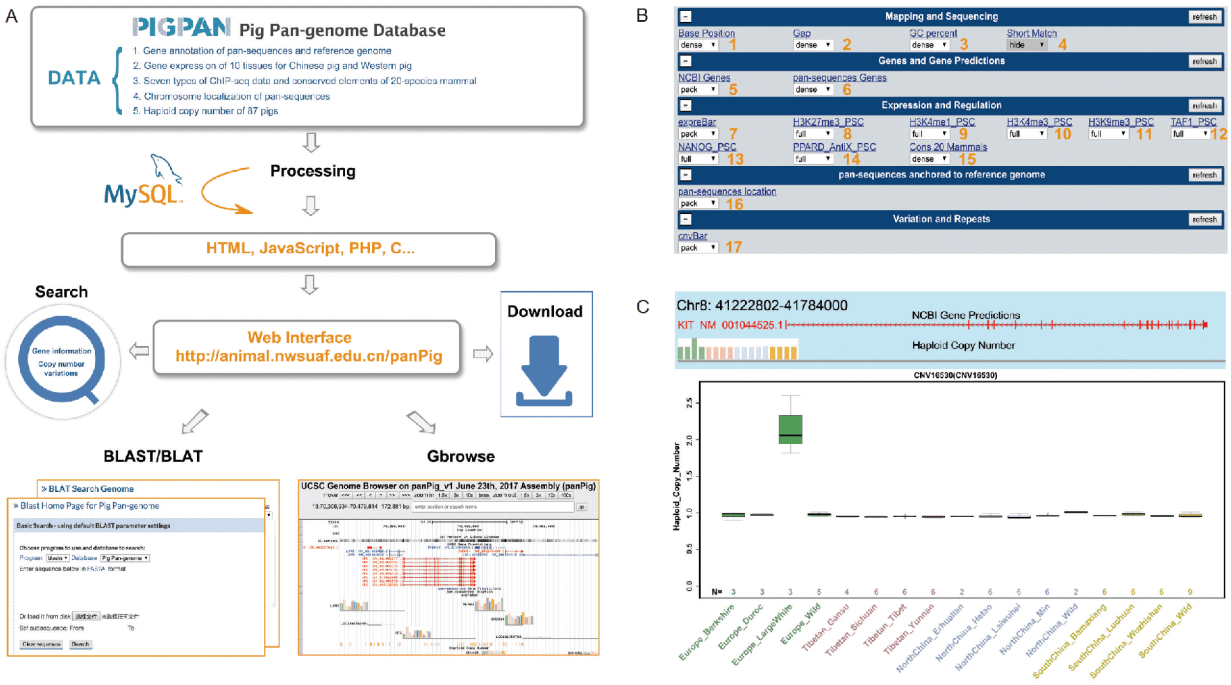

Figure 5 The processing pipeline used to construct the PIGPAN database. PIGPAN integrated genomics, transcriptomics and regulatory data. Users can search for a gene symbol or a genomic region to obtain results in the form of an interactive table and graph. A, The system diagram of PIGPAN. B, The 17 tracks released against the pig pan-genome in our local UCSC Genome Browser server. C, One case showing the copy number difference of the *KIT* gene between European and Chinese pigs by using PIGPAN.

when used as a reference, can ensure a more comprehensive repertoire of genomic variations and can facilitate downstream genomic, transcriptomic and even 3D genome analyses. Therefore, we highlight the transition from the current reference genome to the pan-genome.

MATERIALS AND METHODS

Construction of the pan-genome

We downloaded the publicly available pig genome assemblies of 10 female and one male individuals from 11 diverse breeds (five originated in Europe and six originated in China) (Figure S2 and Table S2 in Supporting Information) (Fang et al., 2012; Li et al., 2017; Li et al., 2013). To identify the sequences which could not align to the reference genome, we split the 11 assemblies by gap region and iteratively aligned them to the reference pig genome assembly (Sscrofa11.1) using BLASTN (Camacho et al., 2009). Sscrofa11.1 was masked by WindowMasker (Morgulis et al., 2006) before alignment to speed up the alignment process. The sequences with <90% identity and ≥300 bp in length were retained. We then used BLASTN to generate a non-redundant collection of unaligned sequences identified from all the assemblies. For the unaligned sequences from each assembly, we first aligned them to the rest of unaligned sequences from the other assemblies. We then removed those sequences which had a hit (≥90% identity and ≥90% coverage) with the targets and were shorter than the targets. After that, the remaining unaligned sequences from each assembly were merged into one FASTA file and were defined as pan-sequences.

Determining the characteristics of the pan-sequences

To explore whether the pan-sequences had homologous regions across species and were potentially functional, we aligned these sequences to 10 mammalian reference genomes (i.e., *Homo sapiens*, *Camelus bactrianus*, *Equus caballus*, *Canis lupus familiaris*, *Capra hircus*, *Bos Taurus*, *Orcinus orca*, *Physeter catodon*, *Balaenoptera acutorostrata scammoni*, *Tursiops truncates*) to search for any matches ($E<1\times10^{-5}$) using BLASTN (Camacho et al., 2009). Only the best hit remained for each query.

To validate the authenticity of these pan-sequences and identify assembly-specific sequences, we aligned all of them to each of the 11 *de novo* pig assemblies to search for any matches (≥90% coverage and ≥95% identity) using BLASTN (Camacho et al., 2009). If the sequence of an assembly did not have a high similarity with other assemblies, this sequence was considered an assembly-specific sequence.

Calling presence/absence for the pan-sequences

We downloaded the whole genome resequencing data for 71

Chinese pigs and 16 European pigs for population analysis of pan-sequences. The sequences data were retrieved from NCBI under the Bioproject PRJNA213179, PRJNA281548, PRJNA309108 and PRJEB9922 (Ai et al., 2015; Frantz et al., 2015; Jeong et al., 2015; Li et al., 2017) (Table S4 in Supporting Information). After alignment using BWA (version 0.7.15-r1140) (Li and Durbin, 2009) with default parameters, we used CNVcaller (Wang et al., 2017) to calculate the whole genome normalized read depth (NRD) of each sequence. The presence and absence of each pan-sequence were then determined by NRD. We further compared the frequency distribution of pan-sequences in Chinese pigs and European pigs using Fisher's exact test followed by multiple testing corrections, and found that there were ~9.0 Mb pan-sequences with higher frequency in Chinese pigs and ~2.8 Mb pan-sequences with higher frequency in European pigs (corrected $P<0.05$).

ChIP-seq short-read alignment and peak calling

To confirm the content of regulatory elements in pan-sequences, we downloaded seven ChIP-seq data from NCBI Bioproject PRJNA152995, including H3K27me3, H3K4me1, H3K4me3, H3K9me3, NANOG, PPARD AntiX and TAF1 signals (Xiao et al., 2012). Sequencing reads were aligned to the pig pan-genome using BWA (version 0.7.17-r1188) (Li and Durbin, 2009) with default parameters. Low-quality and multiple-mapping reads were removed using SAMtools (Li et al., 2009) with option "-q 20". Enriched regions (or peaks) were called ($P<1\times10^{-5}$; no filtering on fold enrichment or false discovery rate (FDR) correction) using MACS (version 2.1.1) (Zhang et al., 2008) with total DNA input as control.

Identification of male-specific sequences

There were 42 males and 45 females in our whole genome resequencing data (Table S4 in Supporting Information). We compared the NRD between females (NRD<0.1, sample size=45) and males (0.2<NRD<0.7, sample size=42) to identify the putative male-specific pan-sequences. Thus, we identified 1,638 male-specific scaffolds which were present in most male individuals (frequency ≥50%) but absent in females (frequency=0) with a combined length of 10,432,972 bp (Supplementary Dataset 2 in Supporting Information).

Gene annotation and functional enrichment analysis

Homology-based and *de novo* prediction were used to annotate protein-coding genes. For homology-based prediction, pan-sequences were aligned onto the repeat-masked assembly using TBLASTN (Camacho et al., 2009) with an *E*-value cutoff of 1×10^{-5}. Aligned sequences as well as corresponding query proteins were then filtered and passed to GeneWise to search for accurately spliced alignments (Doerks et al., 2002). For *de novo* prediction, GenScan (Burge and Karlin, 1998), Augustus (Stanke et al., 2006), and geneid (Blanco et al., 2007) were then used to predict genes.

Annotated genes of novel sequences were analysed for Kyoto Encyclopedia of Genes and Genomes (KEGG) terms and pathway enrichment using KOBAS (Xie et al., 2011).

SNP calling

To verify whether using the pan-genome as reference could improve SNP calling efficacy, we randomly selected six pig samples (ranging from 10 to 30× coverage) (Table S11 in Supporting Information) and mapped their clean reads to the pan-genome and Sscrofa11.1 for comparison. Duplicate reads were removed using Picard Tools. Then, the Genome Analysis Toolkit (GATK, version 3.6) (McKenna et al., 2010) was used to detect SNPs. The following criteria were applied to all SNPs: (1) variant confidence/quality by depth (QD) >2; (2) RMS mapping quality (MQ) >30.0.

RNA-seq analysis

The 92 strand-specific RNA-seq data (7–10 tissue libraries for each of 10 individuals) were downloaded from the NCBI database (Bioproject: PRJNA311523) (Li et al., 2017). All reads were mapped to the pan-genome by HISAT2 (Kim et al., 2015). Transcripts including novel splice variants were assembled using StringTie version 1.2.2 (Pertea et al., 2015), and the FPKM (Fragments Per Kilobase per Million mapped reads) values for these transcripts and genes in each sample were determined using Ballgown (Frazee et al., 2015). Finally, transcripts with FPKM ≥1 in at least one sample were retained.

Materials for Hi-C experiment

Liver samples labelled BH-33, BH-34, BH-35, and BH-36 were collected from four 2-year-old female Bama minipigs. The liver sample labelled F2 was collected from a 90-day-old female foetus of Bama minipig. Ear skin fibroblasts DB-2 and DB-3 were established by using two 12-day-old female Large White pigs. Ear skin fibroblasts XYZ were established by using a 2-year-old female Wild Boar. Embryonic fibroblasts RC-7 and RC-8 were established by using two 40-day-old female foetuses of Chinese Rong Chang pigs. Mature adipocytes DB-2-Y and DB-3-Y were derived from pre-adipocytes which were established by using the same pigs as for ear skin fibroblasts DB-2 and DB-3, by inducing adipogenic differentiation.

All of the fibroblasts were grown in DMEM Dulbecco's Modified Eagle Medium (DMEM, 11995-065, Gibco) containing 10% foetal bovine serum (FBS, 10099-141, Gibco) and 1× penicillin/streptomycin (P/S, 15140-122, Gibco), incubated at 37°C in 5% CO_2.

Pre-adipocytes were cultured in 10% FBS/DMEM-F12 (11330-032, Gibco) with 1× P/S until confluence and induced to differentiation as previously described. Briefly, two days' post-confluence, cells were exposed to differentiation medium containing 0.5 mmol L^{-1} isobutylmethylxanthine (I5879, Sigma), 1 μmol L^{-1} dexamethasone (D2915, Sigma), 850 nmol L^{-1} insulin (I6634, Sigma), 1 μmol L^{-1} rosiglitazone (R2408, Sigma) and 10% FBS for three days. At the end of day 3, the differentiation medium was replaced with maintenance medium with only 850 nmol L^{-1} insulin, 1 μmol L^{-1} rosiglitazone and 10% FBS, and was replenished every other day. After the differentiation process, at least 90% of the cells had accumulated lipid droplets at day 15, and were used as mature adipocytes (DB-2-Y and DB-3-Y).

Hi-C experiment

Hi-C experiments on cells were performed according to the previously published Hi-C protocol with some minor modifications (Lieberman-Aiden et al., 2009). Briefly, 25 million (M) cells were resuspended in 45 mL serum free DMEM, and 37% formaldehyde was added to obtain a final concentration of 2% for chromatin cross-linking. Cells were incubated at room temperature (20−25°C) for 5 min. Then glycine was added to obtain a final concentration of 0.25 mol L^{-1} to quench the formaldehyde. The mixture was incubated at room temperature for 5 min, and subsequently on ice for at least 15 min. Fixed cells were lysed using a Dounce homogenizer in the presence of cold lysis buffer (10 mmol L^{-1} Tris-HCl, pH 8.0, 10 mmol L^{-1} NaCl, 0.2% IGEPAL CA-630, and 1× protease inhibitor solution). Chromatin digestion (restriction enzyme *Hin*d III), labelling, and ligation steps were performed according to the original protocol (Lieberman-Aiden et al., 2009). After deproteinization, removal of biotinylated free-ends, and DNA purification, Hi-C libraries were controlled for quality and sequenced on an Illumina HiSeq X Ten sequencer (paired-end sequencing with 150 bp in read length).

Hi-C experiments on liver tissue were performed as previously described using the *Mbo* I restriction enzyme (Rao et al., 2014), with minor modifications pertaining to handling flash frozen primary tissues (Leung et al., 2015). Briefly, 0.5 g flash-frozen liver tissue was pulverized in liquid nitrogen. Then, the tissue was cross-linked by adding 37% formaldehyde to a final concentration of 4% and incubated at room temperature for 30 min. Glycine was added to obtain a final concentration of 0.25 mol L^{-1} to quench the formaldehyde. The mixture was incubated at room temperature for 5 min, and subsequently on ice for at least 15 min. Cross-linked liver cells were filtered through 70-μm and 40-μm nylon cell strainers and spun down to collect the liver cells. Approximately 25 mg liver cell precipitate was used for Hi-C library preparation. The Hi-C library preparation procedure was performed as previously described using the *Mbo* I restriction enzyme (Rao et al., 2014).

Hi-C reads mapping, filtering, and generation of contact matrices

Pre-processing paired-end sequencing data, reads mapping as well as filtering of mapped di-tags were performed using the Juicer pipeline (v.1.8.9) (Durand et al., 2016). Briefly, short reads were mapped to the pan-genome using BWA (version 0.7.15-r1140) (Li and Durbin, 2009). Reads of low mapping quality were filtered using Juicer with default parameters, discarding the invalid self-ligated and un-ligated fragments, as well as PCR artefacts. Filtered di-tags were further processed with Juicer command line tools to bin di-tags (10 kb bins) and to normalize matrices with KR normalization (Knight and Ruiz, 2013). Valid Hi-C read pairs should harbour more intrachromosomal (cis) interactions than interchromosomal (trans) ones (Table S7 in Supporting Information). To improve resolution, we combined the Hi-C data from the same tissue of the same pig breed after we randomly extracted 20 Gb data for correlation coefficient testing. We combined Hi-C data from DB-2 and DB-3 (Pearson's r=0.99); RC-7 and RC-8 (Pearson's r=0.99); DB-2-Y and DB-3-Y (Pearson's r=0.96) (Figure S13 in Supporting Information). After combining samples, all processes were done on all the data. Normalized interaction matrices were generated at two resolutions, low (100 kb) and high (20 kb).

Identification of compartment A and B

Identification of compartment A/B was performed as previously described using the 100-kb interaction matrix (Lieberman-Aiden et al., 2009). Principal component analysis (PCA) was performed to generate the first principal component (PC1) vectors of each chromosome, and Spearman's correlation between PC1 and genomic characteristics (gene density and GC content) were then calculated. The GC content (%) for each bin (100-kb bin sizes) was calculated using SeqKit (v.0.8.0) (Shen et al., 2016). Gene density (number of genes per bin) was calculated based on the number of promoters (from –2,000 to +500 bp relative to the transcription start site (TSS)) located within (i.e., more than 50% of the region overlapped) each bin. Compartment A and B were determined by the PC1 values. Bins with positive Spearman's correlation between PC1 values and genomic features were assigned as compartment A, otherwise B.

Identification of topologically associating domains (TADs) and topological boundaries

Higher-resolution TAD calls were generated following the previously described procedure by using the directionality index (DI) metric (Dixon et al., 2012). DI was calculated using raw interaction counts between 20-kb bins to capture the observed upstream or downstream interaction bias of genomic regions. A hidden Markov model (HMM) was then used to predict the states of DI for final TAD generation. The same criterion of 400 kb (distance between two adjacent TADs) was used to distinguish unorganized chromatin from topological boundaries. That is, the topological boundaries were less than 400 kb and unorganized chromatin was larger than 400 kb.

Locating pan-sequences on Sscrofa11.1 based on Hi-C

We normalized all Hi-C matrices on the same scale by KR normalization (Knight and Ruiz, 2013), ensuring that any differences between Hi-C were not attributable to variation in sequence length. The maximum 100-kb bin of each pan-sequence interaction (interaction intensity ≥5) was collected as a potential location of pan-sequences. Starting with the filtered 100-kb resolution bin of pan-sequences, we got the higher-resolution interval of 20 kb by taking the maximum 20-kb bin with each 100-kb bin.

Identification of putative promoter and enhancer interactions

We kept the interactions identified by PHYCHIC (Ron et al., 2017) with FDR <0.01 as high confidence interactions and used them to identify promoter-enhancer interactions (PEI). A promoter segment was defined as a region from −2,000 to +500 bp relative to the transcription start site (TSS). When at least half of a promoter segment was in either one of the two bins involved in a chromatin interaction, this interaction was defined as a putative promoter interaction.

The bins which were distal (at least 40 kb upstream or downstream) to the promoter and demonstrate a stronger interaction with the promoter than any other regions were defined as the enhancer interacting with the corresponding promoter. This interaction of the two bins corresponding to the promoter and enhancer was defined as a potential PEI. If our pan-sequences were located in a bin harbouring an enhancer of a PEI, the pan-sequences could be potentially involved in the regulatory functions of the enhancer. If the pan-sequences further demonstrate interactions with the promoter of the same PEI, the involvement of the pan-sequences in the regulatory functions of the enhancer was regarded as highly confident and the pan-sequences could be potential enhancers themselves.

The pig pan-genome web server

The web interface of PIGPAN was built by combining the Apache web server, PHP, HTML, JavaScript and a relational MySQL database. Users can use all online resources without preregistration. Our browser can be divided into two parts: frontend and backend interfaces. The frontend consists of a home page, a download page and several search pages. The MySQL relational database server stores 16 tables containing gap information, GC percentage, seven regulatory signals of potential stem cells (H3K27me3, H3K4me1, H3K4me3, H3K9me3, NANOG, PPARD AntiX, TAF1), conserved elements of 20 mammalian species, haploid copy number of 87 pigs, gene expression, location of pan-sequences and gene annotation. The appropriate index was built on the corresponding retrieval columns of the table. When a user submits an entry, the backend will respond quickly to execute an SQL statement. PHP and JavaScript manage the data analysis processes and display the results. Moreover, we have introduced web-based software such as BLAST (Camacho et al., 2009), BLAT (Kent, 2002) and Gbrowse (Casper et al., 2017). Accordingly, users can query data with rapid visualization in Gbrowse or enter a query sequence to search for homologous regions in the genome. PIGPAN was tested in all major modern Internet browsers, including Firefox, Chrome, Internet Explorer, Safari and Opera. Therefore, PIGPAN is a robust and easy-to-use website to facilitate the search for and visualization of results for pig pan-genome analyses.

Data availability

The sequencing reads of each sequencing library have been deposited at NCBI for Hi-C data (Project ID: PRJNA482496). The assembly of the pig pan-genome and subsequent analysis results are available from our PIGPAN website (http://animal.nwsuaf.edu.cn/code/index.php/panPig). All other data supporting the findings of this study are available in the article and its supplementary information files are available from the corresponding author on request.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *This work was supported by the National Natural Science Foundation of China (31822052 and 31572381) to Y.J and the Science & Technology Support Program of Sichuan (2016NYZ0042 and 2017NZDZX0002) to M.Z.L. We thank the High Performance Computing platform of Northwest A&F University for their assistance with the computing.*

References

Ai, H., Fang, X., Yang, B., Huang, Z., Chen, H., Mao, L., Zhang, F., Zhang, L., Cui, L., He, W., et al. (2015). Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome

sequencing. Nat Genet 47, 217–225.
Arumemi, F., Bayles, I., Paul, J., and Milcarek, C. (2013). Shared and discrete interacting partners of ELL1 and ELL2 by yeast two-hybrid assay. ABB 04, 774–780.
Blanco, E., Parra, G., and Guigo, R. (2007). Using geneid to identify genes. Curr Protoc Bioinformatics Chapter 4, Unit 4.3.
Burge, C.B., and Karlin, S. (1998). Finding the genes in genomic DNA. Curr Opin Struct Biol 8, 346–354.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC BioInf 10, 421.
Casper, J., Zweig, A.S., Villarreal, C., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Karolchik, D., et al. (2017)OUP accepted manuscript. Nucleic Acids Res .
Christopoulos, A., Ligoudistianou, C., Bethanis, P., and Gazouli, M. (2018). Successful use of adipose-derived mesenchymal stem cells to correct a male breast affected by Poland Syndrome: a case report. J Surg Case Rep 2018(7), rjy151.
Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380.
Doerks, T., Copley, R.R., Schultz, J., Ponting, C.P., and Bork, P. (2002). Systematic identification of novel protein domain families associated with nuclear functions. Genome Res 12, 47–56.
Dong, P., Tu, X., Chu, P.Y., Lü, P., Zhu, N., Grierson, D., Du, B., Li, P., and Zhong, S. (2017). 3D chromatin architecture of large plant genomes determined by local A/B compartments. Mol Plant 10, 1497–1509.
Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst 3, 95–98.
Fang, X., Mou, Y., Huang, Z., Li, Y., Han, L., Zhang, Y., Feng, Y., Chen, Y., Jiang, X., Zhao, W., et al. (2012). The sequence and analysis of a Chinese pig genome. Gigascience 1, 16.
Frantz, L.A.F., Schraiber, J.G., Madsen, O., Megens, H.J., Cagan, A., Bosse, M., Paudel, Y., Crooijmans, R.P.M.A., Larson, G., and Groenen, M.A.M. (2015). Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. Nat Genet 47, 1141–1148.
Frazee, A.C., Pertea, G., Jaffe, A.E., Langmead, B., Salzberg, S.L., and Leek, J.T. (2015). Ballgown bridges the gap between transcriptome assembly and expression analysis. Nat Biotechnol 33, 243–246.
Golicz, A.A., Bayer, P.E., Barker, G.C., Edger, P.P., Kim, H.R., Martinez, P. A., Chan, C.K.K., Severn-Ellis, A., McCombie, W.R., Parkin, I.A.P., et al. (2016). The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat Commun 7, 13390.
Gordon, S.P., Contreras-Moreira, B., Woods, D.P., Des Marais, D.L., Burgess, D., Shu, S., Stritt, C., Roulin, A.C., Schackwitz, W., Tyler, L., et al. (2017). Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. Nat Commun 8, 2184.
Groenen, M.A.M., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.J., et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491, 393–398.
Guirao-Rico, S., Ramirez, O., Ojeda, A., Amills, M., and Ramos-Onsins, S. E. (2018). Porcine Y-chromosome variation is consistent with the occurrence of paternal gene flow from non-Asian to Asian populations. Heredity 120, 63–76.
Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M.A., Barry, K., et al. (2014). Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26, 121–135.
Jeong, H., Song, K.D., Seo, M., Caetano-Anollés, K., Kim, J., Kwak, W., Oh, J.D., Kim, E.S., Jeong, D.K., Cho, S., et al. (2015). Exploring evidence of positive selection reveals genetic basis of meat quality traits in Berkshire pigs through whole genome sequencing. BMC Genet 16, 104.
Kent, W.J. (2002). BLAT—The BLAST-like alignment tool. Genome Res 12, 656–664.
Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12, 357–360.
Knight, P.A., and Ruiz, D. (2013). A fast algorithm for matrix balancing. IMA J Numer Anal 33, 1029–1047.
Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33, 1870–1874.
Larson, G., Dobney, K., Albarella, U., Fang, M., Matisoo-Smith, E., Robins, J., Lowden, S., Finlayson, H., Brand, T., Willerslev, E., et al. (2005). Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. Science 307, 1618–1621.
Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A.Y., Yen, C.A., Lin, S., Lin, Y., Qiu, Y., et al. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature 518, 350–354.
Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.
Li, M., Chen, L., Tian, S., Lin, Y., Tang, Q., Zhou, X., Li, D., Yeung, C.K. L., Che, T., Jin, L., et al. (2017). Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple *de novo* assemblies. Genome Res 27, 865–874.
Li, M., Tian, S., Jin, L., Zhou, G., Li, Y., Zhang, Y., Wang, T., Yeung, C.K. L., Chen, L., Ma, J., et al. (2013). Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. Nat Genet 45, 1431–1438.
Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., Qian, W., Ren, Y., Tian, G., Li, J., et al. (2010). Building the sequence map of the human pan-genome. Nat Biotechnol 28, 57–63.
Li, Y., Zhou, G., Ma, J., Jiang, W., Jin, L., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng, L., et al. (2014). *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol 32, 1045–1052.
Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289–293.
McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297–1303.
Monat, C., Pera, B., Ndjiondjop, M.N., Sow, M., Tranchant-Dubreuil, C., Bastianelli, L., Ghesquière, A., and Sabot, F. (2016). *de novo* assemblies of three *Oryza glaberrima* accessions provide first insights about pan-genome of African rices. Genome Biol Evol evw253.
Morgulis, A., Gertz, E.M., Schäffer, A.A., and Agarwala, R. (2006). WindowMasker: window-based masker for sequenced genomes. Bioinformatics 22, 134–141.
Neafsey, D.E., Waterhouse, R.M., Abai, M.R., Aganezov, S.S., Alekseyev, M.A., Allen, J.E., Amon, J., Arcà, B., Arensburger, P., Artemov, G., et al. (2015). Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes. Science 347, 1258522–43.
Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33, 290–295.
Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680.
Ron, G., Globerson, Y., Moran, D., and Kaplan, T. (2017). Promoter-

enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. Nat Commun 8, 2237.
Schatz, M.C., Maron, L.G., Stein, J.C., Hernandez Wences, A., Gurtowski, J., Biggers, E., Lee, H., Kramer, M., Antoniou, E., Ghiban, E., et al. (2014). Whole genome *de novo* assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. Genome Biol 15, 506.
Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS ONE 11, e0163962.
Sherman, R.M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M.P., Chavan, S., Vergara, C., Ortega, V.E., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet 51, 30–35.
Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34, W435–W439.
Sun, C., Hu, Z., Zheng, T., Lu, K., Zhao, Y., Wang, W., Shi, J., Wang, C., Lu, J., Zhang, D., et al. (2017). RPAN: rice pan-genome browser for ~3000 rice genomes. Nucleic Acids Res 45, 597–605.
Uyama, T., Ichi, I., Kono, N., Inoue, A., Tsuboi, K., Jin, X.H., Araki, N., Aoki, J., Arai, H., and Ueda, N. (2012). Regulation of peroxisomal lipid metabolism by catalytic activity of tumor suppressor H-rev107. J Biol Chem 287, 2706–2718.
Vaccari, C.M., Romanini, M.V., Musante, I., Tassano, E., Gimelli, S., Divizia, M.T., Torre, M., Morovic, C.G., Lerone, M., Ravazzolo, R., et al. (2014). De novo deletion of chromosome 11q12.3 in monozygotic twins affected by Poland Syndrome. BMC Med Genet 15, 63.
Wang, X., Zheng, Z., Cai, Y., Chen, T., Li, C., Fu, W., and Jiang, Y. (2017). CNVcaller: highly efficient and widely applicable software for detecting copy number variations in large populations. GigaScience 6.
Wong, K.H.Y., Levy-Sakin, M., and Kwok, P.Y. (2018). *De novo* human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. Nat Commun 9, 3040.
Xiao, S., Xie, D., Cao, X., Yu, P., Xing, X., Chen, C.C., Musselman, M., Xie, M., West, F.D., Lewin, H.A., et al. (2012). Comparative epigenomic annotation of regulatory DNA. Cell 149, 1381–1392.
Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., Kong, L., Gao, G., Li, C.Y., and Wei, L. (2011). KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 39, W316–W322.
Yan, G., Zhang, G., Fang, X., Zhang, Y., Li, C., Ling, F., Cooper, D.N., Li, Q., Li, Y., van Gool, A.J., et al. (2011). Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. Nat Biotechnol 29, 1019–1023.
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B. E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol 9, R137.
Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., et al. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet 50, 278–284.

SUPPORTING INFORMATION

Figure S1 Comparison of contig N50 among pig, human and other animal reference genomes.

Figure S2 Geographic distributions of the original pig breeds collected in this study.

Figure S3 The expression of TIG3 in subcutaneous adipose tissue (light red background) and other tissues (light blue background) of pigs harboring this gene.

Figure S4 Protein alignment of TIG3 in mammals (pig, dog, panda and human), chicken, alligator and zebrafish.

Figure S5 Selection test for TIG3 in pig and other species.

Figure S6 One pan-sequence covers partial genic regions of ZNF622, representing a new splicing event.

Figure S7 SNP density in TAD boundary (blue) and TAD internal (red) region in five samples digested by MboI enzyme.

Figure S8 Schematic diagram showing our strategy in identifying potential putative enhancer in pan-sequences.

Figure S9 An IGV view of Illumina reads at two example regions in chr2 and chr1.

Figure S10 Comparison of RNA-seq read mapping quality using the pan-genome versus Sscrofa11.1.

Figure S11 Comparison of RNA-seq read mapping rate using the pan-genome versus Sscrofa11.1.

Figure S12 Transcriptional potential of the pan-sequences.

Figure S13 Correlation coefficient of Hi-C data from different samples.

Table S1 Sample information of Hi-C data

Table S2 Detailed statistics of assemblies used in pig-pangenome construction

Table S3 Enriched KEGG functional classes among genes that annotated in pan-sequences

Table S4 Summary statistics of whole genome resequencing data in this study

Table S5 The presence and absence of pan-sequences in 87 resequencing samples

Table S6 The frequency distribution of population-specific pan-sequences in Chinese pigs and European pigs

Table S7 Summary statistics of the Hi-C data used in our experiment

Table S8 The anchored location of pan-sequences to Sscrofa11.1 by flanking sequences based and Hi-C based methods

Table S9 List of pan-sequences which shown interaction with a known promoter identified using Hi-C analysis

Table S10 Enriched KEGG functional classes among genes that might be regulated by the putative enhancers of pan-sequences

Table S11 Summary of adjusted SNPs after addition of pan-sequences

Supplementary Dataset 1 The sequences of pig pan-genome

Supplementary Dataset 2 The male-specific pan-sequences

Supplementary Dataset 3 The annotation of pan-sequences

Supplementary Dataset 4 The copy number variation dataset of pig pan-genome

The supporting information is available online at http://life.scichina.com and https://link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.