Genome-Wide Analysis Reveals Genetic Structure and Selective Sweeps in Chinese Indigenous Pig Breeds

Xiong Tong
Guangdong Academy of Agricultural Sciences
Lianjie Hou
South China Agricultural University College of Animal Science
Weiming He
BGI
Chugang Mei
Northwest Agriculture and Forestry University
Bo Huang
South China Agricultural University College of Animal Science
Chi Zhang
BGI
Chingyuan Hu
University of Hawai'i at Manoa Libraries
Chong Wang (betty@scau.edu.cn)
South China Agricultural University College of Animal Science

Research article

Keywords: Chinese pigs, Genome sequencing, Population structure, Selection, Body size

Posted Date: May 9th, 2019

DOI: https://doi.org/10.21203/rs.2.9528/v1

Abstract

Background Chinese indigenous pigs exhibit considerable phenotypic diversity, but their population structure and the genetic basis of agriculturally important traits have not been explored. Results Here, we sequenced the whole genomes of 24 individual pigs representing 22 breeds distributed throughout China. For comparison with European and commercial breeds (one pig per breed), we integrated seven published pig genomes with our new genomes. Our results showed that pig domestication occurred at three places in Southeastern Asia, namely the Mekong region, the middle to downstream regions of the Yangtze River, and Tibetan highlands. Moreover, we demonstrated that classic morphological characteristics such as coat color are not consistent with genetic data. We found that genetic material from European pigs likely introgressed into five Chinese breeds. Two new subpopulations of domestic pigs have been identified in South and North China that encompass morphology-based criteria. The Southern Chinese subpopulation comprises the classical Southern China Type and part of the Central China Type, whereas the Northern Chinese subpopulation comprises the North China Type, the Lower Yangtze River Basin Type, the Southwest Type, the Plateau Type, and the remainder of the Central China Type. Eight haplotypes and two recombination sites were identified within a conserved 40.09 Mb linkage-disequilibrium block on the X chromosome. Potential selection and domestication signatures were identified, mainly influencing body size, along with adaptations to cold and hot temperature environments. Conclusions Our findings provide insights into the phylogeny of Chinese indigenous pig breeds, and will be of enormous benefit in identifying beneficial genes to develop superior pig breeds.

Background

Approximately 10,000 years ago, pigs *(Sus scrofa* L.) were independently domesticated in multiple Eurasian regions [1, 2]. China is a major center of early pig domestication [3] and therefore has numerous indigenous breeds that exhibit considerable phenotypic variation in response to both artificial and natural selection. Except for wild boars, Chinese indigenous pigs are historically classified into 48 breeds and split into six types (*South Chinese*, *North China*, *Lower Yangtze River Basin*, *Central China*, *Southwest*, and *Plateau*), based on geographic distribution, historical origin, and morphological characteristics [4]. Some molecular evidence suggests that this classification may be problematic, given the potential for admixture among different types. However, there was little evidence for this hypothesis because there are only a small number of studies that utilized molecular markers, including randomly amplified polymorphic DNA [5] and microsatellites [6-8].

With the development of genome sequencing and SNP chip technologies, the past decade has seen an increase in data on genome-wide variation. Indeed, comparative genomic analyses have identified genes involved in a wide variety of agriculturally-important traits, including coat color [9, 10], ear morphology [9], body size [11-13], meat yield [11], and disease resistance [11]. DNA-based techniques provide a good opportunity to clarify Chinese pig classification. Recent studies investigated only a few breeds with highly desirable agronomic traits, and they tended to focus on genetic mining [10, 13] and on identifying selective sweeps during domestication [14, 15]. Such research included genome-wide analyses of domestic breeds (e.g., Tibetan, [14]; Tongcheng, [10]; Enshi Black, [13]; Rongchang [16]) with a focus on tolerance to harsh environments, high fertility, and body size. Currently, too few Chinese pig breeds have been studied to provide a conclusive investigation of porcine evolution in China. Specific loci and genes underlying common phenotypic variation among Chinese domestic pig breeds have not yet been studied.

To address these deficiencies, we performed whole-genome resequencing of pigs representing 22 breeds distributed across different geographical areas in China. This new sequence data was integrated with publically-available sequence data from seven other pig breeds, including European ones. We uncovered population structures among Chinese indigenous pigs, along with selection and domestication signatures associated with body size and temperature-related adaptations.

Results And Discussion

Sequencing and variation identification

Twenty-four animals representing twenty-two pig breeds were individually resequenced (Additional file 1: Figure S1 and Table S1). The average effective sequencing depth was 17.54× and genomic coverage was 94.74% (Additional file 1: Figure S2 and Table S2), resulting in a high-quality resequencing resource for pigs.

To these data, we integrated publically-available genomic data from seven pigs of wild and commercial European and Chinese breeds (Additional file 1: Table S1). The combined dataset had 14.09 billion high-quality raw reads (1281.12 Gb raw bases, >90% Q30 bases) (Additional file 1: Figure S3).

A strict quality-filter pipeline resulted in 19 685 697 single-nucleotide polymorphisms (SNPs) from 31 pigs (Additional file 1: Table S4). Of these SNPs, 13 430 360 (68.22%) were in intergenic regions, 1 223 834 (6.22%) were 5-Kb upstream or downstream of gene regions, and 5 031 503 (25.56%) were within gene regions. The last group contained 46 618 non-synonymous (NS) and 53 028 synonymous (S) SNPs (Additional file 1: Figure S4), leading to a NS/S ratio (ω) of 0.88, which is higher than a previously reported ratio of 0.68 [14]. This result suggests that positive selection may be a more important factor in the evolution of domestic pigs in China than previously reported.

In addition, we identified 5 081 752 small-to-medium (1–20 bp) indels (Additional file 1: Table S5). As expected, most indels (3 486 145, 68.60%) occurred in intergenic regions; the remainder were either 5 Kb upstream or downstream of gene regions (352 227, 6.93%), or in gene regions (1 243 380, 24.47%). The Frameshift/Non-frameshift ratio was 2.24 (Additional file 1: Figure S5). Larger structural variations (SV, >45 bp) were detected using read-pair and read-depth methods. Across individuals, the SV count varied from 2881 to 49 939. Deletions and intra-chromosomal translocations were the two primary SV types identified in our samples (Additional file 1: Table S6).

Homozygous (Hom) and heterozygous (Het) SNPs were classified per individual. Homozygous SNPs were more common in all European pigs than in Chinese pigs, especially in two European wild boars that had Hom/Het SNP ratios of 1.627–4.460 (Additional file 1: Table S4). Furthermore, with the exception of the Large White (LW) pig, the Hom/Het indel ratio was consistent with SNP variations between European and Chinese pigs (Additional file 1: Table S5). These results suggest that population bottlenecks may be responsible for the reduced genetic diversity observed in European pigs compared with Chinese pigs [17].

Additionally, numerous specific alleles appear to have been fixed in European and Chinese populations after isolation.

Population structure and introgression

We constructed a non-rooted phylogenetic tree based on 9.2 million population SNPs (Fig. 1a) to understand the genetic relationships and structure among Chinese pigs with different geographical distributions. The estimated phylogeny revealed that the primary division was between European and Chinese pigs, consistent with previous studies [14, 17]. Our results lend further support to the hypothesis that pig domestication occurred independently in western Eurasia and East Asia. Moreover, Chinese domestic breeds split on geographical grounds, namely into South and North China (CnSouth and CnNorth) subpopulations. The former encompassed all individuals from the classical *South Chinese Type* and some of the *Central China Type*. The latter comprised the remainder of *Central China Type* and all those from the remaining four types (*North China*, *Lower Yangtze River Basin*, *Southwest*, and *Plateau*) (Fig. 1a). The genetic relationships among Chinese indigenous pig breeds were remarkably congruent with geographic distribution. Respectively, *South Chinese* and *Lower Yangtze River Basin Types* were closest to breeds Dahua Bai (DH; Xingning City, Guangdong Province, South China) and Jinhua (JH; Jinhua City, Zhejiang Province, Yangtze River lower reaches) (Fig. 1a and Additional file 1: Table S1). Notably, DH and JH are considered to be of the *Central China Type,* a consideration based on coat color phenotypes [4].

Principle component analysis (PCA) confirmed the phylogenetic analysis (Fig. 1b and Additional file 1: Table S7). Furthermore, a model-based clustering analysis with proportional contributions from five ancestral populations revealed the same subpopulations (CnNorth and CnSouth). Northern Chinese pigs could be further split into two subgroups (Fig. 1c): Subgroup 1 consisted of the *Lower Yangtze River Basin* and *North China* types, and Subgroup 2 comprises the *Southwest* and *Plateau* types. Structural features (Fig. 1c) and geographical distribution (Additional file 1: Figure S1) confirmed the three Southeast-Asian centers of pig domestication originally identified through

mitochondrial DNA. These centers are the Mekong region [18], middle and downstream regions of the Yangtze River [19, 20], and Tibetan highlands [18, 20]. Thus, our study provides evidence that the classical classification scheme [4, 21] requires updating with genetic information.

Our three analyses of population structure (phylogeny, PCA, and clustering analysis) (Fig. 1a-c) revealed that admixture likely took place in six Chinese indigenous breeds. Therefore, we employed haplotype sharing ratio to examine putative introgression across all pairs of four populations (South China, North China, Europe, and admixed, including domestic and wild pigs) corresponding to our model-based clusters (Fig. 1c). All autosomes from South China, North China, and Europe populations contained numerous discrete introgression fragments, indicating extensive gene flow had occurred under artificial or natural evolutionary processes. Multiple large and intensive regions on chromosomes 5, 14, 17, and 18 were introgressed from the European population into five Chinese breeds (Additional file 1: Figure S6a–d). Similar events have been reported for Min [22], Kele [22], and Zang/Tibetan [14] breeds.

We examined nucleotide variation (θπ and θw) to measure genetic diversity across three populations (wild pigs, European domestic pigs, and Chinese domestic pigs) and the two Chinese subpopulations (CnNorth and CnSouth). Tested populations were more genetically-diverse (θw/Kb: 2.01–2.80, θπ/Kb: 2.12–3.11; Additional file 1: Table S8 and Figure S7) than cattle breeds Angus and Holstein [23], θw/Kb and θπ/Kb: ~1.4), dogs [24]; θw/Kb: 0.61–1.28, θπ/Kb: 0.75–1.38), and giant pandas [25]; θw/Kb: 1.04–1.30, θπ/Kb: 1.13–1.37). European domestic pigs had the lowest θw/Kb (2.01) and θπ/Kb (2.12), consistent with their high linkage disequilibrium (LD) and F_{ST} values, as well as the maximum fluctuation of Tajima's D-values (Additional file 1: Table S8 and Figure S8–10). Our results support the hypotheses of expansion from a relatively small ancestral population [14, 17] and a large reduction of effective population size under intensive breeding [26]. Wild pigs had slightly higher nucleotide diversity than did Chinese domestic pigs (Additional file 1: Figure S7f). Within a short LD decayed distance (<30 Kb), wild pigs

had lower r^2 than Chinese pigs, but higher r^2 at long distances (≥30 Kb), suggesting narrow bottlenecks in wild-pig evolutionary history. Finally, CnNorth and CnSouth exhibited similar nucleotide diversity (θw/Kb: 2.68, θπ/Kb: 2.82) (Additional file 1: Table S8 and Figure S7d–f and Figure S8b).

Selective signatures from CnSouth and CnNorth

To reduce the impact of genetic admixture when analyzing selective footprints, admixed individuals MP, LWU, NX, and YH were removed. We then identified selective regions on autosomes and the X chromosome of CnSouth and CnNorth animals to evaluate the effects of local adaptation.

The nearly ten-fold greater length of genomic regions (note: windows with distances ≤50 Kb were merged into a single selected locus) indicated stronger selective pressure in CnNorth than in CnSouth (10.82 Mb versus 1.06 Mb), but the former had approximately five-fold fewer genes (167 versus 863 genes) (Fig. 2). This outcome probably reflects differences in selective pressure on the two subpopulations that caused variations in distribution, size, and gene content of genomic footprints. Our results lend further support to the native environment as a major force for rapid evolution, rather than mixture effects.

In CnSouth, 863 genes were under selection, most with functions linked to heat adaption. These include those associated with a specialized organelle, the lysosome, intestinal permeability (tight junction), central nervous system regulation (GABAergic synapse), as well as glycan biosynthesis and metabolism (glycosaminoglycan degradation) (Table 1 and Additional file 2: Table S9). Heat stress reduces intestinal blood flow because splanchnic-region blood is diverted to perfuse skin for heat dissipation [27, 28]. This decrease in blood flow can cause dysfunction of the intestinal epithelial tight junction barrier, resulting in inflammation from the passage of unwanted substances (e.g., endotoxins and pathogenic bacteria) into the internal environment [29, 30]. Thus, maintaining tight-junction function is crucial for proper intestinal health and metabolism, protecting CnSouth against heat stress-induced problems. Three selective genes (*AQP10*,

HspBP1 and *GABA*) , functional genes that protect cells against heat stress [31-33], were observed in CnSouth pigs.

The 167 strongly selected genes in CnNorth pigs were related mainly to vasopressin-regulated water reabsorption, bladder cancer, as well as gastric-acid secretion (Table 1), suggesting selection for extremely cold environments. Cold temperatures increase vasopressin secretion in the kidney, elevating water reabsorption while also boosting blood content and osmotic pressure to adequately supply blood to the heart and brain [34]. After water reabsorption, urinary concentration is very high, and therefore a potential risk factor for bladder cancer [35]. Moreover, cold exposure stimulates gastric acid secretion and causes stress ulcerations [36]. Thus, selection on genes related to regulating gastric acid secretion might prevent stomach damage. Moreover, we identified three important candidate genes (*AQP3*, *AQP7*, and *CA10*) for cold adaptation (Additional file 2: Table S9). Genes *AQP3* and *AQP7* are water and glycerol transporters in the skin, protecting against desiccation of the stratum corneum during cold stress [37]. *CA10* is associated with the molecular structure of hemoglobin and allows oxygen to reach respiring cells more easily [38].

Characterization of a large-scale LD block in the X chromosome

Using SNP data, we identified a large-scale LD block (40.09 Mb, 44 595 487–84 684 295 bp) (Fig. 3) in the X chromosomes of all 31 pigs. This region had extremely low recombination (48 Mb segment, 44.0–91.5 Mb [15, 39]; and spanned the centromeric region (47.3–49.2 Mb). We observed three major haplotypes after selecting SNP markers with inter-marker distances of 3 Kb. Haplotype S was unique to domestic and wild pigs of southern China, whereas N was present in northern Chinese wild pigs, European domestic pigs, and European wild pigs. The third was a recombinant haplotype set (N-S-1 to N-S-6) found only in northern Chinese domestic pigs (Fig. 3). These LD patterns indicate that northern Chinese domestic pigs exhibit more haplotype diversity, and they corroborate findings of a 14 Mb X-linked sweep region [12, 15].

We then detected local breakdown spots in the recombinant haplotype set using all SNP markers from the LD block. We identified two intervals of recombination: spot 1 (left) was 46 219 219–46 419 569 bp and spot 2 (right) was 56 819 762–57 752 631 bp. Minimal distance between the two spots was a 10.40 Mb segment (46 419 569–56 819 762 bp) (Additional file 1: Figure S11), a highly conserved portion of haplotype N. This region was only in northern Chinese domestic pigs and is inside the 14 Mb X-linked sweep [15]. Overall, we found the eight haplotypes within the 40.09 Mb LD block, and a shorter conserved region (10.40 Mb) than described in previous reports [12, 15, 39]. This difference between studies is likely due to our use of high-density genetic markers from data with high sequencing depths, and from obtaining a greater number of Chinese pig breeds.

The 40.09 Mb LD block contained 189 annotated genes, 143 (75.66%) and 108 (57.14%) of which contained SNPs and nonsynonymous substitutions, respectively. KEGG analysis mapped these 189 genes onto the Shigellosis and Neurotrophin-signaling pathway (Additional file 2: Table S10 and Additional file 1: Table S-11). Of the 374 X-chromosome QTLs in the Pig Quantitative Trait Locus database (Pig QTLdb), we aligned 247 (66.04%) to the Wuzhishan pig genome. Furthermore, 47 X-chromosome QTLs overlapped with the 40.09 Mb LD block. Thirty-seven (37/47, 78.72%) and seven (7/47, 14.89%) QTLs belonged to Meat & Carcass Quality and Reproduction, respectively (Additional file 1: Table S12). Within the Meat & Carcass Quality QTLs, 26 (26/37, 70.27%) were related to fat traits (3 fat composition and 23 fatness QTLs), consistent with lipid-metabolism QTLs identified near the X-chromosome centromere [40]. Meanwhile, all overlapping Reproduction QTLs were assigned to the Reproductive organ, reflecting between-subpopulation (CnNorth, CnSouth, European) differences in reproductive characters. In autosomes, overlapping Reproduction QTLs containing selection regions differed greatly between CnSouth (95/1259, 7.55%) and CnNorth pigs (1/871, 0.11%) (Additional file 1: Table S13).

Across CnNorth and CnSouth pigs, we identified 4169 population-level indels in CDS regions of functional genes. After filtering out markers with $\chi^2 < 5$ for one group, 2711 indels remained. Six differed significantly between the two subpopulations, and five of these were distributed in three gene loci (ENSSSCG00000012830, HUWE1, and ITIH5L) on the

10.40 Mb conserved region (Additional file 1: Table S14). The first locus contained three indels that were matched against the InterPro database to reveal two specific cold-shock protein domains (IPR002059 and IPR011129). Variants of these genes in the CnNorth pigs were also found in northern Chinese wild pigs and European domestic and wild pigs.

We next selected the top 100 SVs out of 64 876 population-level SVs that exhibited significantly non-random distribution (χ^2 test with *FDR* correction, $P < 0.01$). Thirty-four of these SVs were located in the X chromosome (Additional file 2: Table S15), with 32 in the 10.40 Mb conserved region. The conserved region contained 63 annotated genes, and four (*EDA*, *HEPH*, *ARHGEF9*, and *HUWE1*) overlapped with six SVs that exhibited very high between-group differences ($P = 8.53 \times 10^{-4}$) (Additional file 1: Table S16). We identified two large loss-of-function deletion patterns (382 bp: 56,650,381–56,649,999, and 487 bp: 56,621,617–56,621,130, Additional file 1: Table S16) on *EDA* and found that they were fixed only in CnNorth pigs. The EDA signaling pathway is involved in ectodermal-organ (hair, teeth, and exocrine glands) development [41, 42], and EDA defects result in Tooth Agenesis [42]. Our findings are consistent with archaeological evidence of different tooth structural characters between CnNorth and CnSouth pigs [4].

Identification of candidate genes for body size

Our sample was split into small pigs (adult body length ≤100 cm, height ≤ 50 cm; N = 7) and large pigs (adult body length ≥ 120 cm, height ≥ 65 cm; N = 7), based on early phenotype characterization records [21] and our own measurements (Additional file 1: Table S3). We then identified 115 nonsynonymous substitutions, distributed in 95 gene regions, that varied in allele frequency between large versus small pigs (>80% in one group, approaching fixation; <20% in the other) (Additional file 1: Table S17). These nonsynonymous substitutions were putative candidate polymorphisms that resulted in size differences. Indeed, two genes (*LEPR* and *FANCC*) overlapping with nonsynonymous substitutions are reported as associated with body growth and development in some mammals [43, 44]. In humans, impaired *LEPR* function exerts a strong negative effect on

ponderal index at birth and height in adolescence [44]. Likewise, *FANCC* plays a major role in skeletal formation, and thus affects human height [45, 46].

We then analyzed differences (χ^2-test with *Bonferroni's* correction) in frequency of indels and SVs between large and small pigs, to understand their effects on body size. We found significant ($P < 0.05$) between-size-group differences for 10 indels and 20 SVs, located within 7 and 10 functional genes, respectively (Additional file 1: Tables S18 and S19). Among small pigs, we identified a 4 bp insertion in the third exon of *COL1A1*. COL1A1 is an α1(I) protein chain of type I collagen and a major structural component of bone. Nonfunctional COL1A1 markedly reduces skeletal mineral density and body height [47, 48]. We also found a 430 bp deletion in the third intron of the gene encoding propionyl CoA caboxylase α subunit (PCCA). A genetic defect in *PCCA* causes propionic acidemia, a condition that can lead to bone disease and growth failure [49].

Conclusions

In this population-genomic study on Chinese pig breeds, we performed whole-genome resequencing to generate a comprehensive catalog of genetic variants within a range of breeds spanning multiple geographical regions in China. We provided additional insights into the presence of three domestication centers (Mekong region, midstream and downstream of the Yangtze River, and Tibetan highlands) in Southeastern Asia, and modified the classical morphology-based categorizations using our new genetic data. We also identified five breeds with large and intensive introgressions from European pigs on chromosomes 5, 14, 17, and 18. Furthermore, in the southern and northern Chinese subpopulations, we noticed strong positive selection of candidate genes associated with adaption to hot and cold environments, respectively. Different geographical populations also contained numerous fine-scale structures of the X-chromosome 40.09 Mb LD block that were undergoing remarkably strong selection. Finally, we identified four candidate genes (*LEPR*, *FANCC*, *COL1A1*, and *PCCA*) influencing body size. Together, these results provide new insights into the phylogenetic relationships among Chinese indigenous pigs and contribute to identification of genes beneficial to breeding of superior pig breeds.

Methods

Samples

To clarify the genetic structure of Chinese pigs across different geographical locations, we selected individuals that represent all six Chinese indigenous breeds [4]: *South China* (n = 9), *North China* (n = 3), *Lower Yangtze River Basin* (n = 2), *Central China* (n = 3), *Southwest* (n = 1), *Plateau* (n = 2). We also included samples from southern and northern Chinese wild pigs (n = 4), as well as European wild and commercial pigs (n = 7) (Additional file 1: Figure S1 and Table S1). Altogether, data from 31 individual animals were used in this study: (i) 24 sampled from 22 breeds, which were handled by the South China Agricultural University, Guangzhou, People's Republic of China (Additional file 1: Table S2 and Figure S2) and (ii) seven (one pig per breed) downloaded from the Wageningen University Porcine Re-sequencing Phase 1 Project (http://www.ebi.ac.uk/ena/data/view/ERP001813) [12, 17] with the highest sequencing depths to supplement the breeds sampled here (Additional file 1: Table S2). Seven small pigs and seven large pigs were used to detect candidate genes for body size (Additional file 1: Table S3). Body size data were obtained for 14 pigs, 11 from the book *Animal genetic resources in China: pigs* [21], and three were measured according to the technical specifications for the registration of breeding pigs (NY/T 820-2004, 2004). A completed ARRIVE guidelines checklist is included in Checklist S1.

DNA isolation and genome sequencing

Genomic DNA was extracted from ear tissue of live collection using a phenol-chloroform-based method. For each sample, 1–15 μg of DNA was sheared into 200–800 bp fragments using the Covaris system (Life Technologies). Fragments were then treated according to the Illumina DNA-sample-preparation protocol. For library construction, fragments were end-repaired, A-tailed, ligated to paired-end adaptors, and PCR-amplified

with 500 bp inserts. Sequencing was performed to generate 100 bp paired-end reads on the HiSeq 2000 platform (Illumina), following the manufacturer's protocol.

Sequence alignment and genotype calling

Filtered reads were aligned to the Wuzhishan pig draft genome assembly (minipig_v1.0) [50] using the Burrows-Wheeler Aligner [51]. This genome was selected as the reference [6, 50] after considering the geographical distance and genetic divergence among the 31 breeds (Additional file 1: Table S1 and Figure S1).

Aligned bam files were sorted and indexed in Picard-tools version 1.117. Two GATK (Genome Analysis Toolkit version 2.4–9 [52] modules, RealignerTargetCreator and IndelRealigner, were used to realign the SNPs around indels in bam results. To obtain high-quality variants, additional GATK modules HaplotypeCaller and SAMtools [53] were used to call variants for each sample. Only concordance variants were selected, and SNPs were filtered with the parameter "QD < 2.0 || FS > 30.0 || MQ < 40.0 || DP<6 || DP>XXX || ReadPosRankSum < -8.0 || BaseQRankSum < -8," while indels were filtered with "QD < 2.0 || FS > 30.0 || ReadPosRankSum < -8.0." These variants were used to perform base quality score recalibration (BQSR), and resultant reads were applied calling population variants, done with the GATK HaplotypeCaller module using the parameter "--genotyping_mode DISCOVERY -stand_emit_conf 10 -stand_call_conf 30."

To detect structural variations, we followed an existing method [54], with some modifications. Reads were assembled into contigs and scaffolds using default parameters in SOAPdenovo. The assembled scaffold was mapped to the reference genome in BLAT [55], with the –fastmap option.

Criteria for determining the most well-aligned scaffold included coverage length in a given region and high contig support. Selected scaffolds and reference-genome regions with the highest alignment were extracted and aligned to each other in LASTZ (http://www.bx.psu.edu/miller_lab/). Unmapped scaffolds were further aligned against the

reference genome using BLASTn. Structural variations were extracted based on all aligned regions.

Phylogenetic and population genetic analyses

Genetic structure was inferred from high-density SNP data in FRAPPE [56], a program that applies maximum likelihood and expectation-maximization to estimate individual ancestry and admixture proportions. To explore individual convergence, we predefined the number of genetic clusters from K = 2 to K = 5. The maximum iteration of the expectation-maximization algorithm was set to 10,000.

A phylogenetic tree was generated from population-level SNPs in TreeBeST (http://treesoft.sourceforge.net/treebest.shtml), under the p-distances model. Population-level SNPs were then subjected to PCA in EIGENSOFT [57], and eigenvectors were obtained using the R function eigen.

To evaluate LD decay, Haploview [58] was used to calculate the squared correlation (r^2) between any two loci. Average r^2 was calculated for pairwise markers in a 5 Kb window and averaged across the whole genome.

SNP diversity and selective sweep analysis

Within-population, whole-genome SNP diversity was estimated with average pairwise divergence ($\theta\pi$) and Watterson's estimator (θw) [59]. These two parameters and Tajima's D [60] were estimated using sliding windows of different sizes (10 Kb, 100 Kb, and 500 Kb) with 50% overlap between adjacent windows. Parameters were calculated per window with an in-house PERL script. A representative window of 500 Kb was used to visualize whole-genome patterns. Next, population differentiation (F_{ST}) was calculated [61].

Regions and genes under artificial or natural selective sweeps should differ significantly in diversity across groups (with one being less diverse). We therefore calculated the ratio of genetic diversity between South and North groups ($\pi w/\pi c$) in 50-Kb sliding windows with a step size of 25 Kb. This ratio was used to identify regions with significantly fewer

polymorphisms per group. Windows containing the top 5% of πw/πc values were considered candidate regions of significantly lower diversity that likely experienced selection sweeps. If πw was <0.002, the window was removed from the analysis. Windows with distances ≤50 Kb were merged into a single selected locus.

Gene and QTL annotation

Pathway analyses of candidate genes under selective sweeps were performed using KEGG (https://www.genome.jp/kegg/pathway.html). Additionally, identified QTLs were functionally characterized using Pig QTLdb (https://www.animalgenome.org/cgibin/QTLdb/SS/index, Release 23, Apr 21, 2014), specifically with coordinate conversion of the Wuzhishan genome to the European-Duroc reference genome (Sscrofa10.2). Indels were matched to the InterPro database using EBI InterProScan (https://www.ebi.ac.uk/interpro/search/sequence-search).

Introgression analysis

Methods followed previous studies [62]. We applied a likelihood ratio test to study the ancestral contribution of groups to the genome of each sample. All putative introgressions between group pairs were examined. For every 100 Kb window containing at least 10 SNPs and when at least three comparisons were possible per group, we calculated the ratio of average sharing per variety with itself and with another group. Regions with an average sharing ratio of >0.8 were defined as introgressions. Shared introgression frequency was plotted and tabulated. Introgression length and number per sample were also tabulated. Regions of extensive haplotype sharing (≥90% shared SNPs) were considered introgression regions for each group pair.

Abbreviations

SNP: Single-nucleotide polymorphism; SV: Structural variation; NS SNPs: Non-synonymous SNPs; S SNPs: Synonymous SNPs; Hom SNPs: Homozygous SNPs; Het SNPs: heterozygous SNPs; CnSouth: South China subpopulation; CnNorth: North China subpopulation; DH: Dahua Bai pig; JH: Jinhua pig; MP: Min pig; LWU: Laiwu pig; NX: Ningxiang pig; YH:

Yuedong Hei pig; PCA: Principle component analysis; LD: linkage disequilibrium; KEGG: Kyoto encyclopedia of genes and genomes; QTLdb: Quantitative Trait Locus database

Declarations

Ethics approval and consent to participate

All animals used in this study were reared and euthanized with the approval of the College of Animal Science, South China Agricultural University. All experiments were performed in accordance with relevant guidelines and regulations of ‘the instructive notions with respect to caring for laboratory animals’ issued by the Ministry of Science and Technology of the People's Republic of China.

Consent for publication

Not applicable

Availability of data and material

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Key Foundation for Basic and Application Research in Higher Education of Guangdong, China (2017KZDXM009); the Team Project of Guangdong Agricultural Department, China (2017LM2148); and the Provincial Agricultural Science Innovation and Promotion Project in 2018 (2018LM2150). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

C.W. and X.T. conceived and designed the experiments. W-M.H. and C.Z. performed variation identification and population analyses. W-M.H., X.T., C.Z., and B.H. contributed to computational analyses. X.T. and L-J.H. collected samples and prepared them for sequencing. C.W., C-G.M., and C-Y.H. provided suggestions and helped check the manuscript. X.T. wrote the manuscript, and C-G.M. helped revise the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Dr Jilong Liu for his helpful suggestions and Editage (www.editage.cn) for English language editing.

References

1. Larson G, Dobney K, Albarella U, Fang M, Matisoo-Smith E, Robins J, Lowden S, Finlayson H, Brand T, Willerslev E: Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science* 2005, 307(5715):1618-1621.

2. Larson G, Liu R, Zhao X, Yuan J, Fuller D, Barton L, Dobney K, Fan Q, Gu Z, Liu X-H: Patterns of East Asian pig domestication, migration, and turnover revealed by modern and ancient DNA. *Proceedings of the National Academy of Sciences* 2010, 107(17):7686-7691.

3. Cucchi T, Hulme-Beaman A, Yuan J, Dobney K: Early Neolithic pig domestication at Jiahu, Henan Province, China: clues from molar shape analyses using geometric morphometric approaches. *Journal of Archaeological Science* 2011, 38(1):11-22.

4. Zhang ZG, Li B, Chen X: Pig breeds in China. *Shanghai Scientific and Technical Publisher, Shanghai* 1986.

5. Yongfu H, Yaping Z: Study on random amplified polymorphic DNA of four local pig breeds in Sichuan Province [China]. *Journal of Sichuan Agricultural University (China)* 1997.

6. Zhang G-X, Wang Z-G, Sun F-Z, Chen W-S, Yang G-Y, Guo S-J, Li Y-J, Zhao X-L, Zhang Y, Sun J: Genetic diversity of microsatellite loci in fifty-six Chinese native pig breeds. *Yi Chuan Xue Bao* 2003, 30(3):225-233.

7. Fang M, Hu X, Jiang T, Braunschweig M, Hu L, Du Z, Feng J, Zhang Q, Wu C, Li N: The phylogeny of Chinese indigenous pig breeds inferred from microsatellite markers. *Animal genetics* 2005, 36(1):7-13.

8. Yang S-L, Wang Z-G, Liu B, Zhang G-X, Zhao S-H, Yu M, Fan B, Li M-H, Xiong T-A, Li K: Genetic variation and relationships of eighteen Chinese indigenous pig breeds. *Genetics Selection Evolution* 2003, 35(7):657.

9. Wilkinson S, Lu ZH, Megens H-J, Archibald AL, Haley C, Jackson IJ, Groenen MA, Crooijmans RP, Ogden R, Wiener P: Signatures of diversifying selection in European pig breeds. *PLoS genetics* 2013, 9(4):e1003453.

10. Wang C, Wang H, Zhang Y, Tang Z, Li K, Liu B: Genome-wide analysis reveals artificial selection on coat colour and reproductive traits in Chinese domestic pigs. *Molecular ecology resources* 2015, 15(2):414-424.

11. Li M, Tian S, Yeung CK, Meng X, Tang Q, Niu L, Wang X, Jin L, Ma J, Long K: Whole-genome sequencing of Berkshire (European native pig) provides insights into its origin and domestication. *Scientific reports* 2014, 4:4678.

12. Rubin C-J, Megens H-J, Barrio AM, Maqbool K, Sayyab S, Schwochow D, Wang C, Carlborg Ö, Jern P, Jørgensen CB: Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences* 2012, 109(48):19529-19536.

13. Fu Y, Li C, Tang Q, Tian S, Jin L, Chen J, Li M, Li C: Genomic analysis reveals selection in Chinese native black pig. *Scientific reports* 2016, 6:36354.

14. Li M, Tian S, Jin L, Zhou G, Li Y, Zhang Y, Wang T, Yeung CK, Chen L, Ma J: Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nature genetics* 2013, 45(12):1431-1438.

15. Ai H, Fang X, Yang B, Huang Z, Chen H, Mao L, Zhang F, Zhang L, Cui L, He W: Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nature genetics* 2015, 47(3):217-225.

16. Lei C, Shilin T, Long J, Zongyi G, Dan Z, Lan J, Tiandong C, Qianzi T, Siqing C, Liang ZHANG TZ: Genome-wide analysis reveals selection for Chinese Rongchang pigs. *Frontiers of Agricultural Science and Engineering* 2017, 4(3):319-326.

17. Groenen MA, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens H-J: Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 2012, 491(7424):393-398.

18. Wu G-S, Yao Y-G, Qu K-X, Ding Z-L, Li H, Palanichamy MG, Duan Z-Y, Li N, Chen Y-S, Zhang Y-P: Population phylogenomic analysis of mitochondrial DNA in wild boars and domestic pigs revealed multiple domestication events in East Asia. *Genome biology* 2007, 8(11):R245.

19. Jin L, Zhang M, Ma J, Zhang J, Zhou C, Liu Y, Wang T, Jiang A-a, Chen L, Wang J: Mitochondrial DNA evidence indicates the local origin of domestic pigs in the upstream region of the Yangtze River. *PloS one* 2012, 7(12):e51649.

20. Yang S, Zhang H, Mao H, Yan D, Lu S, Lian L, Zhao G, Yan Y, Deng W, Shi X: The local origin of the Tibetan pig and additional insights into the origin of Asian pigs. *PloS one* 2011, 6(12):e28215.

21. Wang L, Wang A, Wang L, Li K, Yang G, He R, Qian L, Xu N, Huang R, Peng Z: Animal genetic resources in China: pigs. In.: Beijing: China Agricultural Press; 2011.

22. Ai H, Huang L, Ren J: Genetic diversity, linkage disequilibrium and selection signatures in Chinese and Western pigs revealed by genome-wide SNP markers. *PloS one* 2013, 8(2):e56001.

23. Consortium BH: Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 2009, 324(5926):528-532.

24. Gou X, Wang Z, Li N, Qiu F, Xu Z, Yan D, Yang S, Jia J, Kong X, Wei Z: Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. *Genome research* 2014, 24(8):1308-1315.

25. Zhao S, Zheng P, Dong S, Zhan X, Wu Q, Guo X, Hu Y, He W, Zhang S, Fan W: Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nature Genetics* 2013, 45(1):67-71.

26. Bosse M, Megens H-J, Frantz LA, Madsen O, Larson G, Paudel Y, Duijvesteijn N, Harlizius B, Hagemeijer Y, Crooijmans RP: Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. *Nature communications* 2014, 5.

27. Travis S, Menzies I: Intestinal permeability: functional assessment and significance. *Clinical science* 1992, 82(5):471-488.

28. Lambert G: Stress-induced gastrointestinal barrier dysfunction and its inflammatory effects. *Journal of animal science* 2009, 87(14_suppl):E101-E108.

29. Dokladny K, Moseley PL, Ma TY: Physiologically relevant increase in temperature causes an increase in intestinal epithelial tight junction permeability. *American Journal of Physiology-Gastrointestinal and Liver Physiology* 2006, 290(2):G204-G212.

30. Lambert GP: Intestinal barrier dysfunction, endotoxemia, and gastrointestinal symptoms: the 'canary in the coal mine'during exercise-heat stress? In: *Thermoregulation and Human Performance.* vol. 53: Karger Publishers; 2008: 61-73.

31. Mobasheri A, Shakibaei M, Marples D: Immunohistochemical localization of aquaporin 10 in the apical membranes of the human ileum: a potential pathway for luminal water and small solute

absorption. *Histochemistry and cell biology* 2004, 121(6):463-471.

32. Nylandsted J, Gyrd-Hansen M, Danielewicz A, Fehrenbacher N, Lademann U, Høyer-Hansen M, Weber E, Multhoff G, Rohde M, Jäättelä M: Heat shock protein 70 promotes cell survival by inhibiting lysosomal membrane permeabilization. *Journal of Experimental Medicine* 2004, 200(4):425-435.

33. Shekhar A, Johnson PL, Sajdyk TJ, Fitz SD, Keim SR, Kelley PE, Gehlert DR, DiMicco JA: Angiotensin-II is a putative neurotransmitter in lactate-induced panic-like responses in rats with disruption of GABAergic inhibition in the dorsomedial hypothalamus. *Journal of Neuroscience* 2006, 26(36):9205-9215.

34. Bankir L, Bouby N, Ritz E: Vasopressin: a novel target for the prevention and retardation of kidney disease? *Nature Reviews Nephrology* 2013, 9(4):223-239.

35. Braver DJ, Modan M, Chêtrit A, Lusky A, Braf Z: Drinking, micturition habits, and urine concentration as potential risk factors in urinary bladder cancer. *Journal of the National Cancer Institute* 1987, 78(3):437-440.

36. Niida H, Takeuchi K, Ueshima K, Okabe S: Vagally mediated acid hypersecretion and lesion formation in anesthetized rat under hypothermic conditions. *Digestive diseases and sciences* 1991, 36(4):441-448.

37. Garcia N, Gondran C, Menon G, Mur L, Oberto G, Guerif Y, Dal Farra C, Domloge N: Impact of AQP3 inducer treatment on cultured human keratinocytes, ex vivo human skin and volunteers. *International journal of cosmetic science* 2011, 33(5):432-442.

38. Campbell KL, Roberts JE, Watson LN, Stetefeld J, Sloan AM, Signore AV, Howatt JW, Tame JR, Rohland N, Shen T-J: Substitutions in woolly mammoth hemoglobin confer biochemical properties adaptive for cold tolerance. *Nature genetics* 2010, 42(6):536-540.

39. Ma J, Iannuccelli N, Duan Y, Huang W, Guo B, Riquet J, Huang L, Milan D: Recombinational landscape of porcine X chromosome and individual variation in female meiotic recombination associated with haplotypes of Chinese pigs. *BMC genomics* 2010, 11(1):159.

40. Ma J, Gilbert H, Iannuccelli N, Duan Y, Guo B, Huang W, Ma H, Riquet J, Bidanel J-P, Huang L: Fine mapping of fatness QTL on porcine chromosome X and analyses of three positional candidate genes. *BMC genetics* 2013, 14(1):46.

41. Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T: A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Human molecular genetics* 2007, 17(6):835-843.

42. Pantalacci S, Chaumot A, Benoît G, Sadier A, Delsuc F, Douzery EJ, Laudet V: Conserved features and evolutionary shifts of the EDA signaling pathway involved in vertebrate skin appendage development. *Molecular Biology and Evolution* 2008, 25(5):912-928.

43. do Carmo JM, da Silva AA, Cai Z, Lin S, Dubinion JH, Hall JE: Control of blood pressure, appetite, and glucose by leptin in mice lacking leptin receptors in proopiomelanocortin neurons. *Hypertension* 2011, 57(5):918-926.

44. Labayen I, Ruiz JR, Moreno LA, Ortega FB, Beghin L, DeHenauw S, Benito PJ, Diaz LE, Ferrari M, Moschonis G: The effect of ponderal index at birth on the relationships between common LEP and LEPR polymorphisms and adiposity in adolescents. *Obesity* 2011, 19(10):2038-2045.

45. Kemper KE, Visscher PM, Goddard ME: Genetic architecture of body size in mammals. *Genome biology* 2012, 13(4):244.

46. Allen HL, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S: Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 2010, 467(7317):832.

47. Suuriniemi M, Kovanen V, Mahonen A, Alén M, Wang Q, Lyytikäinen A, Cheng S: COL1A1 Sp1 polymorphism associates with bone density in early puberty. *Bone* 2006, 39(3):591-597.

48. Pochampally R, Horwitz E, DiGirolamo C, Stokes D, Prockop D: Correction of a mineralization defect by overexpression of a wild-type cDNA for COL1A1 in marrow stromal cells (MSCs) from a patient with osteogenesis imperfecta: a strategy for rescuing mutations that produce dominant-negative protein defects. *Gene therapy* 2005, 12(14):1119.

49. Van Gosen L: Organic acidemias: a methylmalonic and propionic focus. *Journal of Pediatric Nursing: Nursing Care of Children and Families* 2008, 23(3):225-233.

50. Fang X, Mou Y, Huang Z, Li Y, Han L, Zhang Y, Feng Y, Chen Y, Jiang X, Zhao W: The sequence and analysis of a Chinese pig genome. *GigaScience* 2012, 1(1):16.

51. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009, 25(14):1754-1760.

52. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010, 20(9):1297-1303.

53. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: The sequence alignment/map format and SAMtools. *Bioinformatics* 2009, 25(16):2078-2079.

54. Li Y, Zheng H, Luo R, Wu H, Zhu H, Li R, Cao H, Wu B, Huang S, Shao H: Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nature biotechnology* 2011, 29(8):723.

55. Kent WJ: BLAT—the BLAST-like alignment tool. *Genome research* 2002, 12(4):656-664.

56. Tang H, Peng J, Wang P, Risch NJ: Estimation of individual admixture: analytical and study design considerations. *Genetic epidemiology* 2005, 28(4):289-301.

57. Patterson N, Price AL, Reich D: Population structure and eigenanalysis. *PLoS Genet* 2006, 2(12):e190.

58. Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005, 21(2):263-265.

59. Watterson G: On the number of segregating sites in genetical models without recombination. *Theoretical population biology* 1975, 7(2):256-276.

60. Tajima F: Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989, 123(3):585-595.

61. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 2002, 12(12):1805-1814.

62. McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, Zeller G, Clark RM, Hoen DR, Bureau TE: Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences* 2009, 106(30):12273-12278.

Tables

Table 1. Functional gene categories enriched for genes affected by selection from CnSouth population and CnNorth population

Functional category	Pathway ID	*P* values*	Gene count
CnSouth population			
Lysosome	ko04142	0.007	15 (181)
Basal cell carcinoma	ko05217	0.007	8 (70)
Melanogenesis	ko04916	0.013	12 (143)
Tight junction	ko04530	0.023	25 (404)
Glycosaminoglycan degradation	ko00531	0.028	4 (29)
GABAergic synapse	ko04727	0.034	10 (128)
CnNorth population			
Bladder cancer	ko05219	0.002	4 (69)
Vasopressin-regulated water reabsorption	ko04962	0.003	4 (73)
Prion diseases	ko05020	0.006	4 (94)
Thiamine metabolism	ko00730	0.031	1 (4)
Gastric acid secretion	ko04971	0.038	4 (162)

*: *P* values represent *FDR*-adjusted *P* values.

Additional File Legend

Additional file 1: This file includes Figures S1 to S11, Tables S1- S8, Tables S11- S14, and Tables S16- S19. (DOCX 48710 kb)

Additional file 2: This file includes Tables S9, S10, and S15. (XLSX 77kb)

Checklist S1: A completed ARRIVE guidelines checklist for reporting animal data in this manuscript.(PDF 1122kb)

Figures

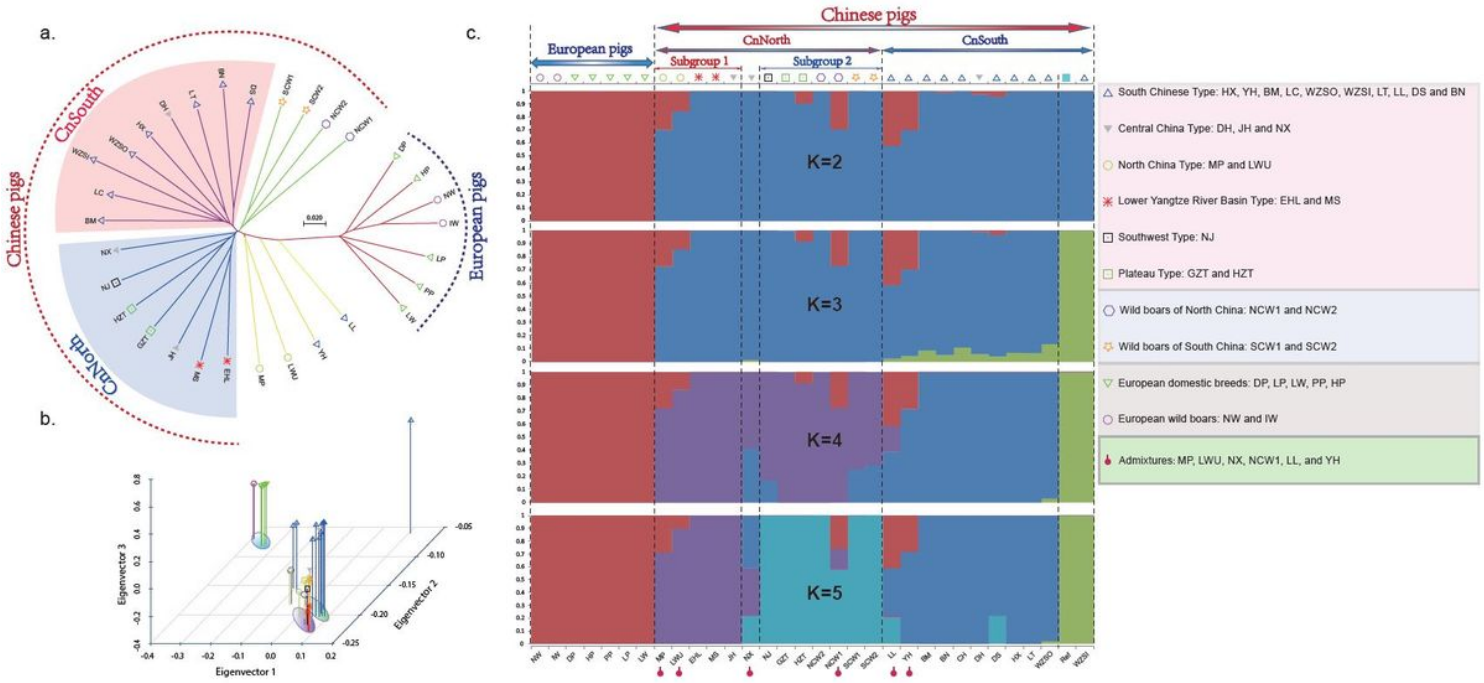

Figure 1

Population structure of wild and domestic pigs from different geographical regions. a Neighbor-joining tree of all pigs based on the 9.2 million population SNPs. The scale bar denotes p distance between individuals. b Three-way PCA plot of all individuals. c Genetic structure analysis of samples using FRAPPE, with changing ancestral populations from K=2 to K=5.

Population	Selected windows	Length	Genes
CnSouth	1,508	1.06 Mb	863
CnNorth	285	10.82 Mb	167

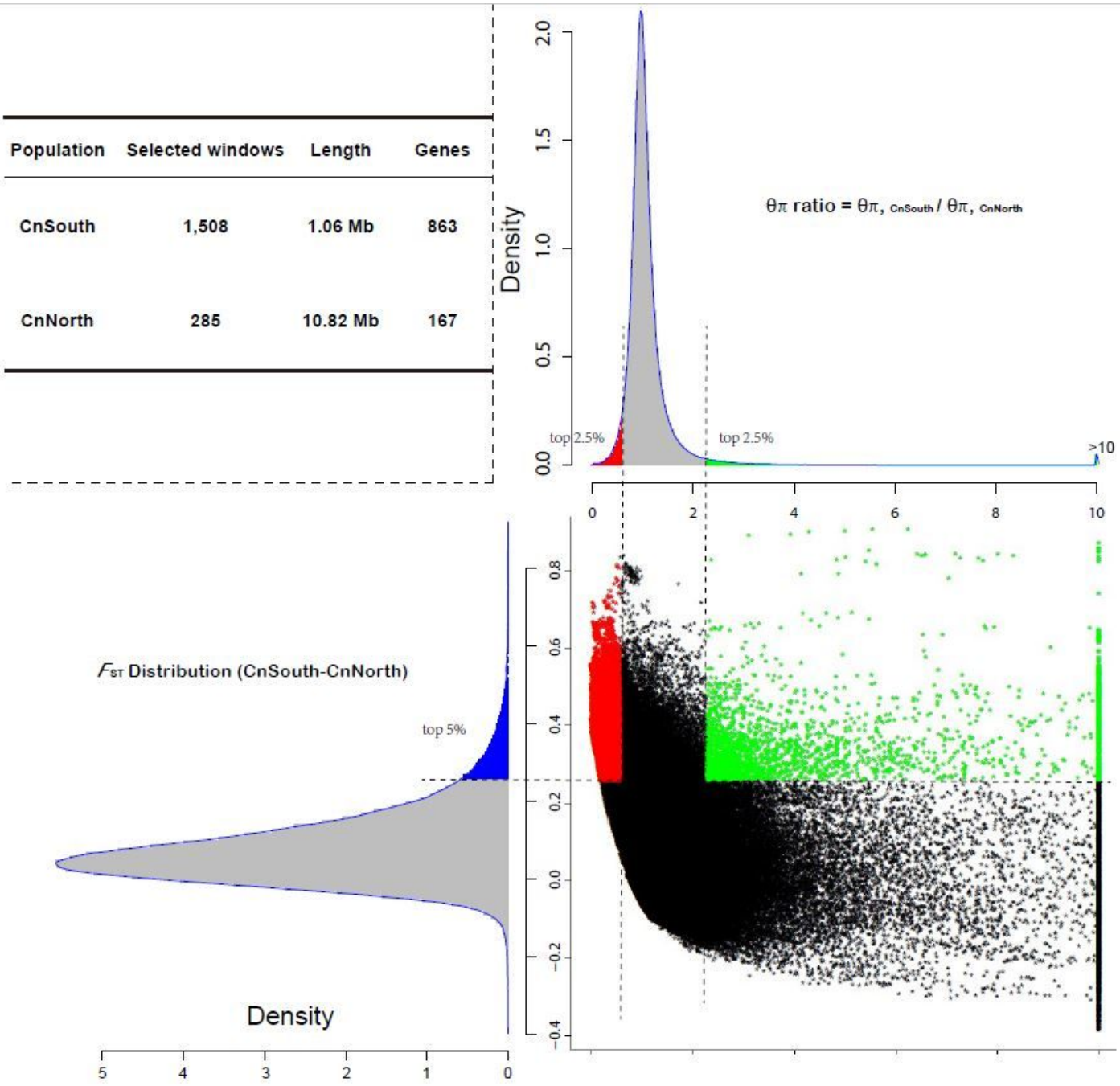

Figure 2

Distribution of θπ ratios and Fst values and selected regions from CnSouth and CnNorth. Red and green data points represent the selected windows for CnSouth and CnNorth, respectively.

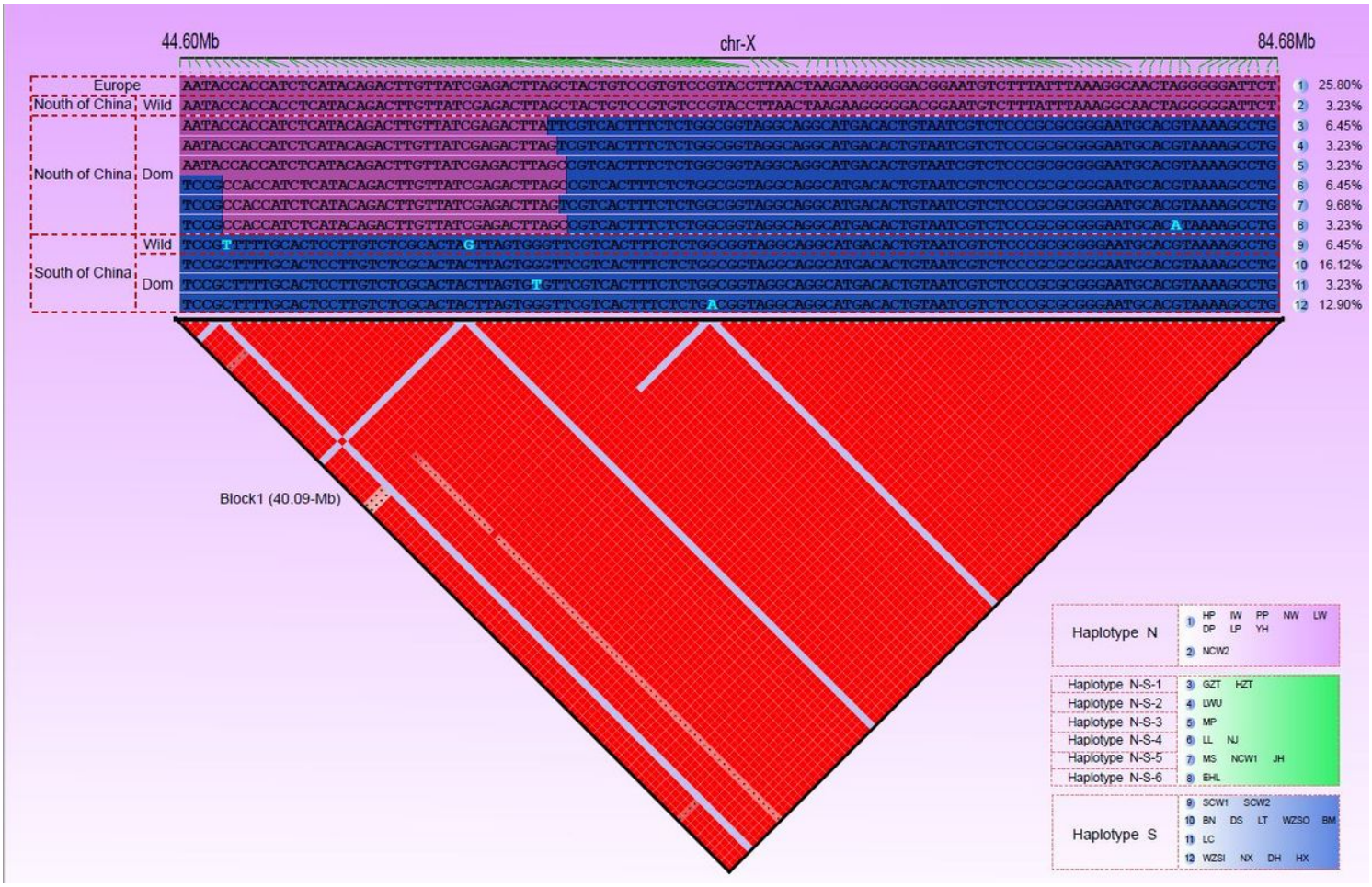

Figure 3

Haplotype pattern of LD block region on the chromosome X (1 SNP/0.3 Mb). Haplotype S is identified in South China pigs (domestic pigs and boars) (purple regions). Haplotype N is identified in European pigs (domestic pigs and boars) and wild boar of South China (blue regions). Six derived Haplotypes (Haplotype N-S-1-6) are ident¬¬¬ified in domestic pigs of North China (purple and blue regions).

Supplementary Files

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

- Additionalfile1.docx
- Additionalfile2.xlsx
- ChecklistS1.pdf