TYPE Original Research
PUBLISHED 24 October 2022
DOI 10.3389/fgene.2022.1003069

OPEN ACCESS

EDITED BY
Lingyang Xu,
Institute of Animal Sciences (CAAS)
China

REVIEWED BY
Zhe Chao,
Hainan Academy of Agricultural Sciences, China
Zhe Zhang,
Zhejiang University, China

*CORRESPONDENCE
Andreia J. Amaral,
andreiaamaral@campus.ul.pt

†These authors share first authorship

SPECIALTY SECTION
This article was submitted to Livestock Genomics,
a section of the journal
Frontiers in Genetics

RECEIVED 25 July 2022
ACCEPTED 20 September 2022
PUBLISHED 24 October 2022

CITATION
Sá P, Santos D, Chiaia H, Leitão A, Cordeiro JM, Gama LT and Amaral AJ (2022), Lost pigs of Angola: Whole genome sequencing reveals unique regions of selection with emphasis on metabolism and feed efficiency.
Front. Genet. 13:1003069.
doi: 10.3389/fgene.2022.1003069

Lost pigs of Angola: Whole genome sequencing reveals unique regions of selection with emphasis on metabolism and feed efficiency

Pedro Sá[1,2†], Dulce Santos[1,2†], Hermenegildo Chiaia[3], Alexandre Leitão[1,2], José Moras Cordeiro[3], Luís T. Gama[1,2] and Andreia J. Amaral[1,2]*

[1]CIISA—Centro de Investigação Interdisciplinar em Sanidade Animal, Faculdade de Medicina Veterinária, Universidade de Lisboa, Lisboa, Portugal, [2]Laboratório Associado para a Ciência Animal e Veterinária (AL4AnimalS), Avenida da Universidade Técnica, Lisboa, Portugal, [3]Faculdade de Medicina Veterinária, Universidade José Eduardo dos Santos, Huambo, Angola

Angola, in the western coast of Africa, has been through dramatic social events that have led to the near-disappearance of native swine populations, and the recent introduction of European exotic breeds has also contributed to the erosion of this native swine repertoire. In an effort to investigate the genetic basis of native pigs in Angola (ANG) we have generated whole genomes from animals of a remote local pig population in Huambo province, which we have compared with 78 genomes of European and Asian pig breeds as well as European and Asian wild boars that are currently in public domain. Analyses of population structure showed that ANG pigs grouped within the European cluster and were clearly separated from Asian pig breeds. Pairwise F_{ST} ranged from 0.14 to 0.26, ANG pigs display lower levels of genetic differentiation towards European breeds. Finally, we have identified candidate regions for selection using a complementary approach based on various methods. All results suggest that selection towards feed efficiency and metabolism has occurred. Moreover, all analysis identified *CDKAL1* gene, which is related with insulin and cholesterol metabolism, as a candidate gene overlapping signatures of selection unique to ANG pigs. This study presents the first assessment of the genetic relationship between ANG pigs and other world breeds and uncovers selection signatures that may indicate adaptation features unique to this important genetic resource.

KEYWORDS

endangered pigs, signatures of selection, genomics, adaptation, metabolism

Introduction

Throughout the African continent human populations have relied on pigs as an important source of animal protein. Many of these animals survive in harsh conditions with poor availability of nutrients, water and in the presence of several highly infectious endemic pathogens. Nevertheless, the history of pigs in sub-Saharan Africa has been poorly studied. The domestication of pigs (*Sus scrofa*) since the Neolithic has produced a wide diversity of breeds distributed worldwide, that have undergone natural and artificial selection in different environments (Larson et al., 2007). It is known that Asian pig breeds derive from the domestication of Asian wild boar and that pigs have arrived in Europe along with humans coming from the Near East. However, the mitochondrial DNA signature of European pigs has been replaced by haplotypes associated with European wild boars (Frantz et al., 2019). It is hypothesized that African pigs derive from Near Eastern pigs that were introduced by land through Egypt (Blench, 2000). Later on, after the 15th century, it is known that the Portuguese introduced Iberian pigs along the seacoast in several trading posts. Until today, in these regions, the Portuguese word for "pig", *porco*, has influenced the local dialects. Through the study of linguistics, a few regions were proposed to harbor populations of native pigs in Africa, since the native word for pig does not derive from Portuguese, therefore, potentially deriving from the earlier introductions through Egypt. These regions are the ancient Egyptian, ancient Berber, Sennar populations, the West African Extension and the Angola Extension. From these, nowadays, due to cultural reasons, only the Angola Extension, which ranges from the Cameroon to Angola, harbors larger population of these pigs (Figure 1) (Blench, 2000).

More recently, a study based on variation of the *cytochrome b* gene (Amills et al., 2013) has shown that pigs from the west coast of Africa display European haplotypes whereas pigs from the East coast display haplotypes from the Far East, suggesting that ancient haplotypes have been replaced by European ones. A few studies have shown high levels of genetic diversity in several African pig populations (Ramírez et al., 2009; Swart et al., 2010). Nevertheless, these studies have not included samples from most of the regions of distribution of African native pigs, as described by Blench (2000), and are of limited scope given that they were

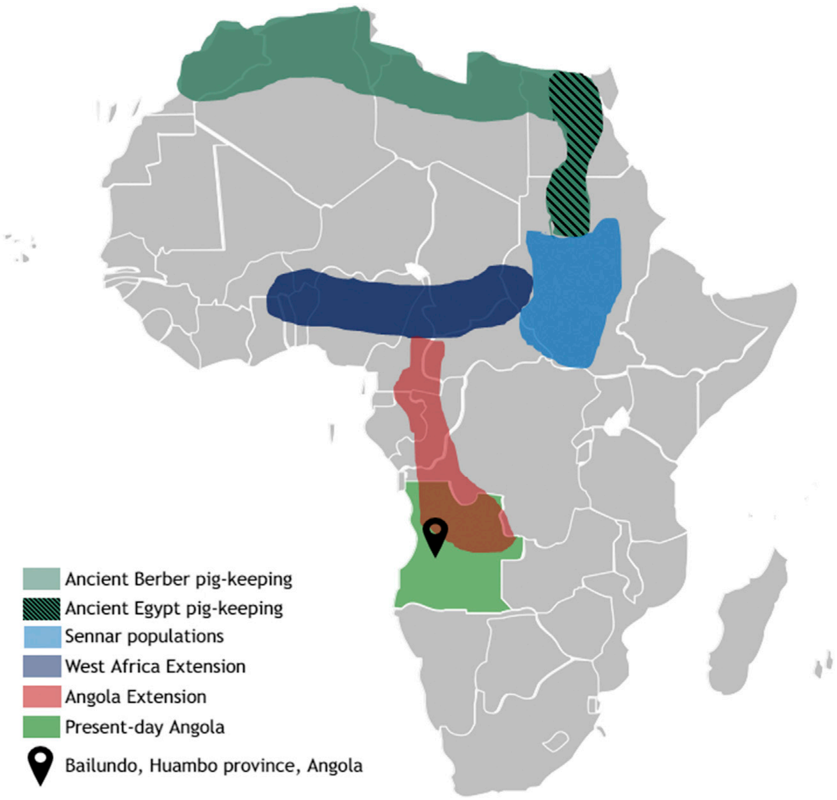
Ancient Berber pig-keeping
Ancient Egypt pig-keeping
Sennar populations
West Africa Extension
Angola Extension
Present-day Angola
Bailundo, Huambo province, Angola

FIGURE 1
Historical distribution of pig-keeping in Africa. Pig domestication was a very important practice for north African cultures (Berber and Egyptian), until the practice was eliminated with the spread of the Islamic culture. Despite this bottleneck in northern territories, reports exist describing domestication centers that kept existing in central Africa (Sennar region and the West Africa Extension), reaching the Angola extension—a strip stretching territory from the south of Cameroon, through Congo and down into Angola. Adapted from Blench, 2000 [3].

based on limited microsatellite, Y-chromosome and mitochondrial markers. An in-depth characterization of these populations at the genome level is thus required.

The region corresponding to the Angola extension of African pigs (Figure 1) is, still today, a remote area where pigs are bred in feral state and have never been characterized regarding their genetic variability. Their origins are still unclear, and some questions remain. Do they still have the genetic background that existed before the introduction of Iberian pigs? Have they been influenced by the recent introduction of other exotic breeds? In the local dialect, pigs are called “ongulu” or “olongulu,” a word that does not derive from the Portuguese language. The country has experienced dramatic social events that have led to a drastic decrease in the population of native pigs that are currently threatened. Their appearance is characterized by a variety of coat colors, with the black coat being the most common. The position of the ears is similar to Alentejano pigs, an Iberian strain from Portugal, and adult animals are smaller and thinner than cosmopolitan breeds and Iberian pigs.

In recent years, the knowledge obtained by studies of whole genome sequencing has enabled deciphering the origins and relatedness of many local breeds, including pigs. These studies have resulted in a remarkable rise in the availability of whole genome data for many breeds around the world (Groenen et al., 2012; Ramírez et al., 2014; Frantz et al., 2015), allowing to develop analysis at whole genome level for populations whose origins have not been explored. The goal of this study was to understand the origins of native pigs from Angola, based on a sample of individuals collected in the Bailundo municipality of Huambo province. Through the analysis of their genomes and by comparing these with the available genomes of 78 other pigs and wild boars from Europe and Asia, we were able to estimate introgression, relatedness and identify unique signatures of selection in this population.

Materials and methods

DNA extraction and preprocessing of WGS data

Whole-genome sequence (WGS) datasets with paired-end reads from European Nucleotide Archive (ENA) with at least 100 million reads were obtained (PRJEB9922, PRJEB1683, PRJNA320525 and PRJNA255085). Run accessions of samples used are shown in Supplementary Table S1. A total of 78 datasets were used, comprising data from 5 European domestic breeds (including the major cosmopolitan and Iberian pigs) and 1 Asian domestic breed, corresponding to 27 and 13 samples, respectively. For wild boars, a total of 38 samples were considered, of which 13 were Asian and 25 were European samples. One sample of *Sus verrucosus* (SV) was also included as an outgroup. Moreover, DNA was obtained from ear tissue collected from local pigs sampled (N = 4) in the municipality of Bailundo in Angola (Figure 1) using the phenol-chloroform extraction method. Following, sequencing libraries with paired-end reads (150bp) were generated from the obtained DNA. Sequencing was performed by outsourcing (Novogene company), using an Illumina NovaSeq 6000 sequencer. Illumina’s Casava V pipeline was used to remove reads i) with adapter sequences, ii) with unspecified bases (N) at more than 10% of the read length, and iii) with low quality (Qscore ≤5). Raw sequence data is available from the ENA database (accession number PRJEB49797). Read quality and adapter presence was evaluated with FastQC v.0.11.9 for all samples (N = 82 *Sus scrofa* + 1 outgroup *Sus verrucosus*) (Andrews, 2010). Adapters were trimmed using Flexbar v.3.4.0 (Dodt et al., 2012). Prinseq-lite v.0.20.4 (Schmieder and Edwards, 2011) was used to filter out reads outside a size range of 50–150 nt and an average phred score quality smaller than 20. Paired reads were mapped to the available reference genome (Sscrofa11.1) using BWA v.0.7.17 (Li and Durbin, 2009) mem command. Read groups corresponding to breed origin were added using “AddOrReplaceReadGroups” function of Picard v.2.23.4. Mapping files were sorted using SAMTools v.1.10 (Li et al., 2009) and PCR Duplicates were removed using “MarkDuplicates” function of Picard v.2.23.4.

Variant calling, filtering and effect prediction

Variant calling was performed for all samples using “SAMtools Variant Caller” v.1.0.6 (Li et al., 2009) following the pipeline in Supplementary Figure S1. Obtained variants were filtered using BCFtools v.1.10.2 (Narasimhan et al., 2016): i) removing small insertions and deletions (INDELs) and ii) filtering the selected variants based on phred scaled probability of false variant calling using varFilter function on default values except for *d* (minimum read depth) that was set to 10 and *a* (minimum number reads carrying the alternative allele) that was set to 3. Recently Lefouili & Nam (2022) have shown that this variant calling pipeline performs better for non-human data. Variant Effect Predictor v.105.0 (VEP) (McLaren et al., 2016) was used to determine effects of SNPs on genes and regulatory elements, transcripts and protein sequences in all samples (excluding SV). DAVID database (Huang et al., 2009) was used to assign gene ontology (GO) terms for biological processes to SNPs located in coding regions. A reduced representation of the obtained GO terms was generated with Revigo (Supek et al., 2011) using default settings.

Phylogenetic analysis

All identified SNPs as well as the subset of synonymous SNPs were used to generate multiple sequence alignment (MFA) sets respectively using VCF-kit v.0.2.6 (Cook and Andersen, 2017), which were further converted to phylip format using fasta-to-phylip.py (Davis-Richardson, 2019). PHYLIP v.3.695 software (Felsenstein, 2005) was used to generate a Neighbor-joining tree with bootstrap support as briefly described: "Seqboot" option was used to create 20 datasets for the MFA including all SNPs and 100 datasets for the MFA including synonymous SNPs which were used to estimate genetic distance matrices using the "Dnadist" option, neighbor-joining trees in Newick format were created for each matrix using option "neighbor", and a consensus tree was created using "Consense" option. FigTree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) was used for plotting the consensus tree. *Sus verrucosus* (SV) was used as outgroup.

Linkage disequilibrium decay

Linkage disequilibrium (LD) was estimated for each of the above populations as a function of genetic distance using PopLDdecay v.3.31 (Zhang et al., 2019) software considering a bin size for the mean r^2 for short and long distance of 10kb, breaks of 100 kb and a maximum distance of 250 kb and plotted using R statistical environment (Team, 2013).

Principal Component Analysis

SNP LD-pruning was performed using PLINK v.1.90 (Purcell et al., 2007) with the -indep-pairwise option considering a window size of 50kb, steps of 5 SNPs and an LD threshold of 0.5. The obtained LD-pruned SNP dataset was used to estimate the principal component matrix using PLINK (Purcell et al., 2007). R package AssocTests v.1.0-1 (Wang et al., 2020) was used to perform a Tracy-Widom test (Tracy and Widom, 1992; Tracy and Widom, 1994) in order to identify principal components with significant eigenvalues that represent a substantial proportion of the variance. Plotting was performed using Biovinci v.3.0.9. (BioTuring Inc., San Diego California USA).

Admixture analysis

Autosomal data for all samples were loaded from mapping files (.bam) and used to calculate genotype likelihood using ANGSD v.0.935 (Korneliussen et al., 2014), considering a SNP *p*-value threshold of 10^{-6}. Obtained genotype likelihood (GL) information was used to calculate admixture proportions using NGSAdmix v.33 (Skotte et al., 2013). Admixture proportions were estimated for k (number of ancestral populations) between 2 and 14. The obtained admixture proportions were plotted using R statistical environment (Team, 2013).

Fixation index and nucleotide diversity (θπ) cross analysis

Synonymous SNPs were selected to estimate pairwise breed F_{ST} (Weir and Cockerham, 1984). VCF format was converted to genepop format using script vcf2genepop.pl from the 2b-RAD pipeline (Wang et al., 2012), which was imported into R environment using adegenet package (Jombart and Ahmed, 2011), then F_{ST} was estimated using genet.dist and boot.ppfst functions from Hierfstat package (Goudet, 2005). Moreover, for the investigation of signatures of selection, the full set of identified SNPs was used to estimate F_{ST} between ANG and Iberian (IBN), Pietrain (PI), Large White (LW), Landrace (LR) and Duroc (DU), Meishan (MS), European (EWB) and Asian (AWB) Wild populations with VCFTools v.0.1.16 (Danecek et al., 2011) using the Weir & Cockerman method and considering a window size of 10 kb. Nucleotide diversity was estimated for each of the aforementioned populations using Nei and Li, 1979 (Nei and Li, 1979) method implemented in VCFTools (Danecek et al., 2011) and also considering a window size of 10 kb. In addition, the logarithm of the ratio of the nucleotide diversity was estimated:

$$\boldsymbol{\theta}_{\pi^{ratio}} = \boldsymbol{log}_{2}\frac{\boldsymbol{\theta}_{\pi^{population}}}{\boldsymbol{\theta}_{\pi^{ANG}}} \quad \text{population} = \{\text{IBN, LR, LW, PI, DU, MS, EWB, AWB}\} \tag{1}$$

Finally, genomic regions in the 5% right tail of F_{ST} and of $\theta\pi_{\text{ratio}}$ were selected ($\text{xpF}_{\text{ST}}/\theta\pi$). Using the wilcox.test function of R package stats v.3.6.2 (Team, 2013), a Mann-Whitney U test (Wilcoxon, 1945; Man and Whitney, 1947) was performed to examine the significance of the difference between the means of F_{ST} or $\theta\pi_{\text{ratio}}$ within the outlier regions and the whole genome. Genes in these outlier regions were identified using Ensembl's R package BiomaRt v.2.50.0 (Durinck et al., 2009). Finally, these gene lists were submitted to PigQTLdb (Hu and Reecy, 2007) (Release 44, Apr26, 2021) to identify QTLs overlapping these genes.

Signatures of selection with haplotype scans

BEAGLE v.5.2 (Browning and Browning, 2007; Browning et al., 2018) was used to generate phased genotypes for a total of 44 samples comprising data from the major European and Asian domestic populations, i.e. LR, LW, PI, DU, Meishan and IBN and ANG samples and considering a total of 24,809,344 markers. Phased genotypes were used to estimate Integrated Haplotype Scores (iHS) using the R package rehh v.3.2.1 (Gautier and Vitalis, 2012; Gautier et al., 2017). Outlier sweep regions were identified with overlapping windows, considering a window size of 10 kb and overlaps of 1 kb. As *p*iHS can be interpreted as a two sided –p-value associated with the null hypothesis of selective neutrality (Gautier et al., 2017) we have selected a highly conservative *p*-value threshold of 10^{-6} for the selection of candidate regions with a minimum number of two SNPs with *p*-values below the threshold. Moreover, a Cross-Population Extended Haplotype Homozygosity (xpEHH) analysis was performed using the R package rehh v.3.2.1 (Gautier and Vitalis, 2012; Gautier et al., 2017) considering ANG vs. European domestic pig populations, namely IBN, DU, PI, LW and LR, and MS to identify signatures of selection and to compare the ANG pigs with well-established domestic populations. Outlier sweep regions were identified with overlapping windows, considering windows of 10 kb and overlaps of 1 kb. Similarly, as before we have selected a highly conservative, *p*-value threshold of 10^{-4} and a minimum number of five SNPs. Gene annotation for the outlier regions was performed using Ensembl's R package BiomaRt v.2.50.0 (Durinck et al., 2009).

Results

Data quality control and variant calling

Raw sequence data generated from Bailundo pigs (ANG) ranged from 95 million to 126 million reads across samples, corresponding to an average size of 108 million reads per sample (Supplementary Table S2). Read filtering, retained over 99.80% of the reads indicating high-quality paired-end libraries, at least 99% of all reads were mapped against the reference genome and read mapping depth ranged from 9.80 to 12.12 (Supplementary Table S2). The same criterion was used for datasets publicly available. Results of data quality control and filtering of a total of 79 public datasets showed similar results as for ANG pigs data, except for samples from project PRJEB1683 (Groenen et al., 2012) which displays the lowest genome coverage and somewhat lower read depth (Supplementary Table S1).

Variant calling analysis in ANG pigs allowed to identify more than eight million SNPs in autosomes with an average frequency of 1 SNP per 0.23 kb, of which slightly more than 900 thousand are unique to this population in comparison with European and Asian domestic and wild populations. Most identified single nucleotide variations were located in intronic regions and have unknown functional significance (Supplementary Table S3). Among the exonic SNPs, approximately 21 thousand were annotated as missense, stop-and-start gain and loss. A large portion of these SNPs is related with tissue remodeling biological processes (Supplementary Figure S2).

Phylogenetic analysis and population structure

We have further performed a phylogenetic analysis of all identified variants. The result shows a clear separation between the European and Asian populations (Figure 2). ANG pigs are placed closer to European domestic in a separate subclade, between the clades formed by a few Iberian samples and the clade in which are placed the European domestic breeds and the Duroc. This analysis was also performed including only synonymous SNPs (130 K SNPs), and results are very similar (Supplementary Figure S3).

Following, pairwise F_{ST} was estimated in order to assess the levels of genetic differentiation between ANG pigs, worldwide breeds and wild boars. The lowest levels of genetic differentiation were observed while comparing ANG with Landrace or Large White (F_{ST} = 0.16) or with European Wild Boars (F_{ST} = 0.15). When comparing ANG pigs with other European domestic breeds the values showed slightly higher levels for the relationship with Iberian (F_{ST} = 0.21), Pietrain (F_{ST} = 0.19) and Duroc (F_{ST} = 0.21). In contrast, the levels of genetic differentiation were observed to be higher between ANG and Asian *Sus* populations, namely when compared to Meishan (F_{ST} = 0.25) or with Asian Wild Boar (F_{ST} = 0.23) (Figure 3A). All these values were significant at a threshold *p*-value <0.05 (Figure 3B).

The Principal Component Analysis (PCA) allowed identifying three components that cumulatively represent 38.20% of the total variation of genotypes. In Figure 4 are shown the two principal components that explained the highest level of the genetic variance (PC1 = 16.27%, PC2 = 14.05%). PC2 axis clustered pig breeds according to their geographic origin, namely Europe and Asia with the exception of Duroc breed. ANG pigs were clustered together with the European domestic populations. PC1 axis separates European wild boar and some Iberian samples from the remaining European pig breeds.

Finally, to investigate the proportion of breed composition in ANG pigs genetic background, we performed an admixture-based clustering analysis. The analysis was performed for different levels of K ranging between 2 and 14 (Figure 5; Supplementary Figure S4). The best likelihood value was obtained when considering K = 2 ancestral populations, which represents the number of the two large geographic regions of Europe and Asia. Then, as K increases we may observe that the different breeds start to display different proportions of breed composition. ANG pigs were clearly differentiated from all other breeds and wild boars at K = 9, which is the number of populations included in the study.

The pattern of linkage disequilibrium decay across genomic distances is shown in Figure 6. Wild boars of Asia and of Europe display lower LD levels in comparison with domestic *Sus scrofa*. Among domestic swine, the Meishan displays lower levels of LD in comparison with European pig breeds, with the exception of Large

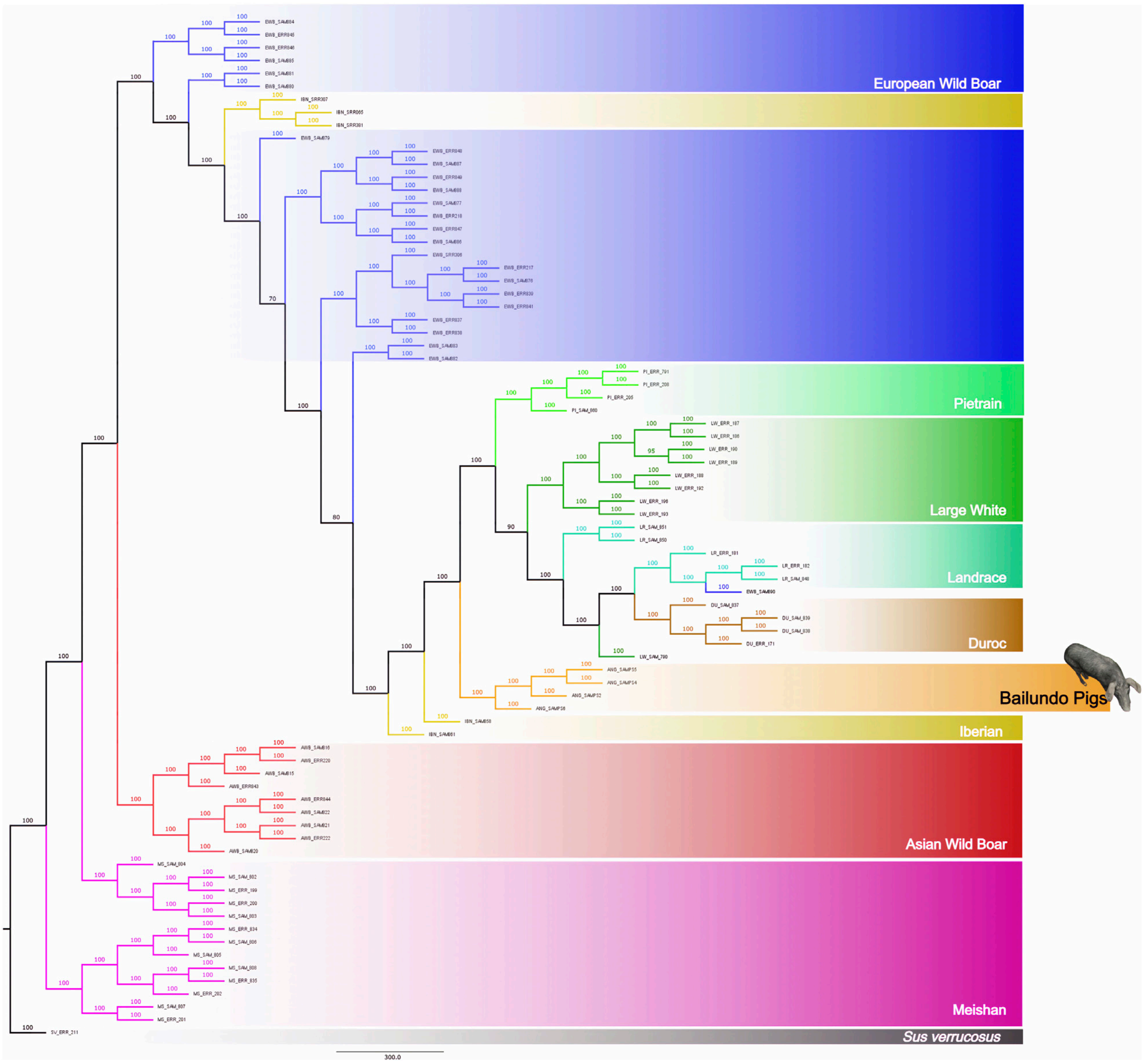

FIGURE 2
Phylogenetic analysis using autosomal SNPs of Bailundo pigs (ANG) and of European and Asian Sus populations. ANG pigs (orange) were clustered closer to European domestic populations, i.e., LW (dark green), PI (light green), LR (cyan) and DU (brown). Two IBN (yellow) pigs were clustered closer to ANG cluster, while the remaining IBN pigs were clustered among EWB (blue). A distant cluster was also formed comprising Asian populations, namely AWB (red) and MS (pink). Bootstrap support is shown in each branch.

White. When compared to European domestic populations, ANG pigs showed high levels of LD surpassed only by Pietrain and Duroc pigs.

Signatures of selection in ANG pigs

The detection of selection signatures in the genomes of ANG pigs was performed using the integrated haplotype scores method (iHS). iHS allows to measure the amount of extended haplotype homozygosity (EHH) at a given SNP along the ancestral allele relative to the derived allele, allowing to detect selective sweeps positively selected and that have not yet reached fixation (Sabeti et al., 2002; Sabteti et al., 2007). Considering sliding windows of 10 kb and overlaps of 1kb, and using a p-value threshold of 10^{-6}, 25 candidate regions representing a total of 475 Kb were selected which harbor the SNPs at the top 0.003% of the iHS empirical distribution. These are distributed along the genome (Figure 7). The candidate regions contain a total of 75 outlier SNPs, of which

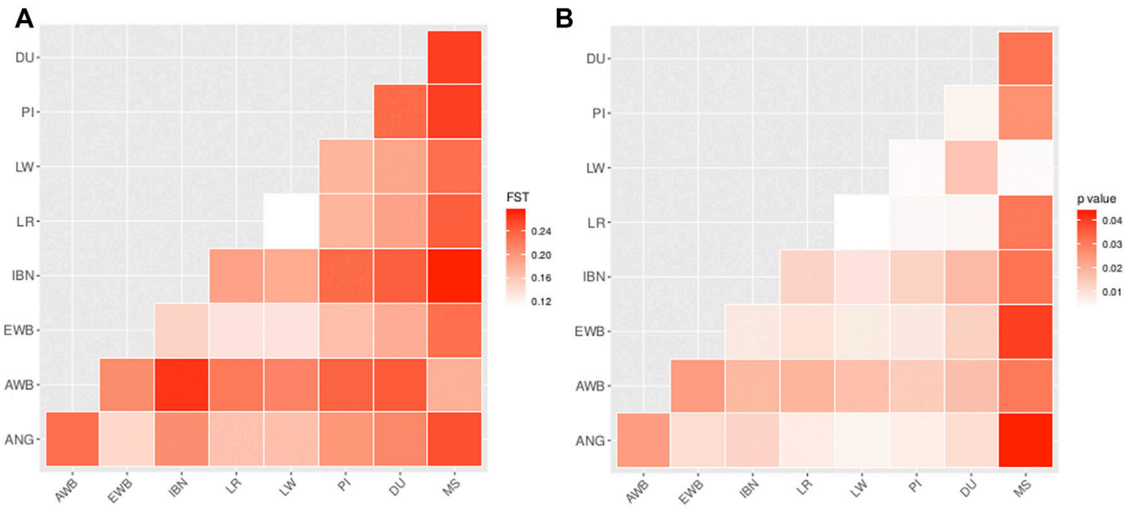

FIGURE 3
Pairwise F_{ST} estimations in all studied pig populations. **(A)** Pairwise F_{ST} between the Bailundo pigs (ANG) and the major domestic and wild populations of Eurasia. IBN—Iberian; LR—Landrace; LW—Large White; DU—Duroc; PI—Pietrain; EWB—European Wild Boar; MS—Meishan; AWB—Asian Wild Boar. **(B)** Pairwise F_{ST} corresponding p-values with 100 bootstraps.

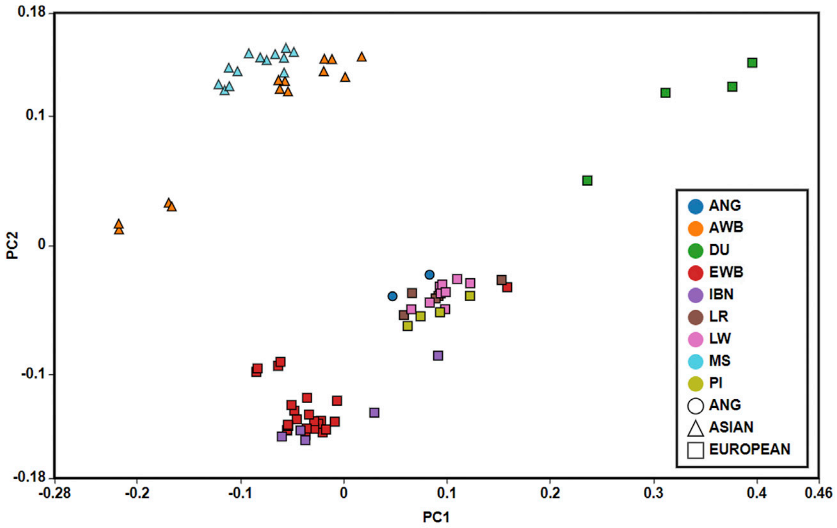

FIGURE 4
Principal Component Analysis. Eigenvectors for Principal Component Analysis show that, considering PC1 and PC2, two major clusters are formed: a cluster with Asian samples (triangle) and a mainly European cluster (square). ANG samples (circle) were clustered with European samples. IBN—Iberian; LR—Landrace; LW—Large White; DU—Duroc; PI—Pietrain; EWB—European Wild Boar; MS—Meishan; AWB—Asian Wild Boar.

37 are located within 15 genes (Supplementary Table S4). Five of these 15 genes have a total of 108 associated GO terms that were reduced to 10 parental GO terms. These GO terms are related to "sulfur compound metabolic process," "detection of stimulus," "neuron-neuron synaptic transmission," "amide transport," "regulation of hormone levels," "activation of adenylate cyclase activity," "regulation of purine nucleotide biosynthetic process," "negative regulation of response to external stimulus," "adult behavior" and "behavior." The GO term for sulfur compound metabolic process (GO:0006790) is associated with a higher number of child GO terms (Supplementary Table S5). One QTL overlaps with one of these genes (*HK2*, QTL:194695

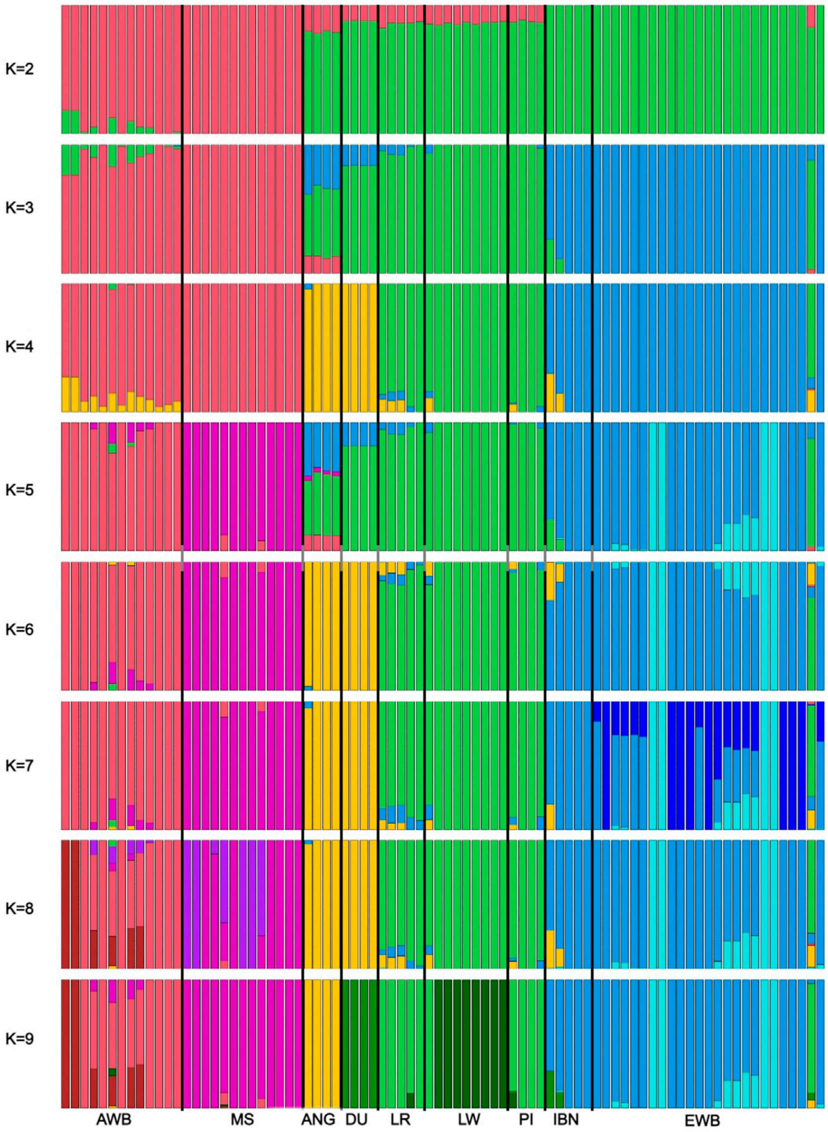

FIGURE 5
Admixture-based clustering considering k between 2 and 9. For each individual on the x-axis, the amount of shared genetic material is shown on the y-axis. The genetic structure of ANG samples was compared with European and Asian samples when k between 2 and 9 were forced for the five main populations: Bailundo pigs (ANG), Asian wild boar (AWB) and Meishan samples (MS), European wild boar (EWB) and European domestic, including Large White (LW), Duroc (DU), Landrace (LR), Pietrain (PI) and Iberian (IBN) pigs.

(Drag et al., 2019), "Fat androstenone level"). Moreover, two sets of genomic regions exhibited very strong signals and included a high number of SNPs and were highlighted on chromosomes 7 and 8 (Supplementary Table S4). In chromosome 7, 12 SNPs (p-value $<10^{-6}$) were identified that are located in the *CDKAL1* gene and 9 SNPs (p-value $<10^{-6}$) were identified in *SLC2A9* gene, on chromosome 8. The iHS analysis was performed for each of the other domestic pig breeds (Supplementary Table S6) and results showed that *CDKAL1* and *SLC2A9* genes only overlap signatures of selection that are unique to ANG pigs.

Detection of selection signatures by comparison of ANG pigs with European and Meishan pig breeds

The cross-analysis aimed to detect regions of divergent selection, i.e., regions in the genome that display simultaneously unusual levels (95% outlier selection) of genetic differentiation F_{ST} and of reduced nucleotide diversity ($\log_2$ θπ ratio) between pairs of breeds. The results (Table 1) indicate that the outlier identified regions

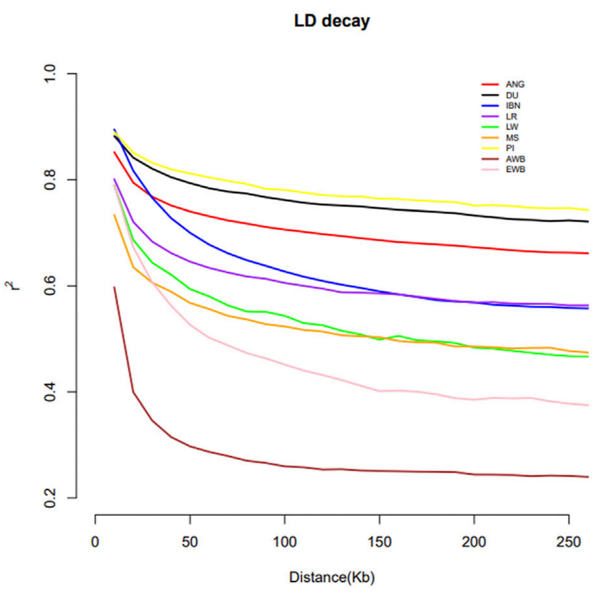

FIGURE 6
Linkage disequilibrium pattern of ANG pigs, European and Asia pig breeds and of their wild counterparts. Bailundo pigs (ANG), Asian wild boar (AWB) and Meishan samples (MS), European wild boar (EWB) and European domestic, including Large White (LW), Duroc (DU), Landrace (LR), Pietrain (PI) and Iberian (IBN) pigs.

exhibit significant differences in comparison with the genomic background. The candidate regions identified by the comparison of ANG and IBN displayed lower extent (outlier region size), when considering the total number of regions, and higher degree of genetic differentiation, in contrast with the candidate regions identified by the comparison of ANG and European commercial populations, i.e., Landrace and Large White, which reported the lowest threshold for F_{ST} and θπ ratio. Regarding the functional impact, more than 50% of regions with strong sweep signals were found in non-coding regions (Supplementary Figure S5). In the comparison between of ANG and IBN pigs, it should be kept in mind that both breeds are adapted to poor environmental conditions in terms of water and food availability, and genes overlapping in candidate regions have been reported to be associated with three major groups of traits: feed efficiency and meat related features, body features and immune related features (Supplementary Table S7).

The identification of signatures of selection by investigating the Extended Haplotype Homozygosity (xpEHH) also allowed identifying strong selective sweeps between ANG pigs and other world breeds. Considering ANG vs. IBN, the extent of candidate regions overlapping signatures of selection was 1,277 Mb, harboring a total of 1,125 outlier SNPs (that were at least at the top 0.007% of the XP-EHH empirical distribution) and overlapping a total of 28 genes spread across the genome (Table 2). When considering ANG *versus* each of the major European commercial populations, i.e., LR, LW and PI, similar results were obtained, with a lower extent in the genome and corresponding to a lower number of genes. Finally, to complete the assessment of European domestic populations, the xpEHH analysis between ANG vs. DU revealed a total of 134 kb outlier regions overlapping three genes. The analysis between ANG and MS revealed a total of 264 kb outlier regions, overlapping three candidate genes. Similarly as observed for the outlier regions identified using the xpF_{ST}/θπ method, more than 50% of candidate regions overlap with non-coding regions (Supplementary Figure S6). The analysis of QTLs overlapping candidate genes when comparing ANG vs. IBN pigs allowed to identify two major groups of traits: feed efficiency and features related with meat and body traits (Supplementary Table S8).

Finally, the total list of candidate genes obtained using the xpEHH method and the xpF_{ST}/θπ were compared for ANG vs. IBN, LR, LW and PI to identify common genes among the pairwise comparisons of those breeds. In Figure 8A it can be observed that 29 genes appear as candidates of selection in all comparisons when the xpF_{ST}/θπ method is used. In contrast, no candidate genes could be identified in all comparisons through the xpEHH method (Figure 8B). Finally, we have investigated which were the overlapping QTLs for candidate genes identified using both methods xpEHH and xpF_{ST}/θπ. As shown in Table 3, several genes had no associated QTLs, but QTLs associated with backfat and feed intake were identified for *DOCK5* and *DLGAP2* genes, respectively. Of note we observe that *CDKAL1* gene overlaps outlier regions when ANG pigs are compared with the other European pig breeds.

Discussion

Remote regions of Angola are part of what the historians have described as the “Angola extension” (Blench, 2000), one of the territories of native pigs that derived from pigs introduced by land through Egypt. In this study we used whole genome sequencing to investigate the origins of native pigs from Angola, their relatedness with other world breeds and to detect signatures of selection within these pigs and by comparison with other pig breeds. The whole genome sequencing data that was generated displayed an average depth of 10x and a mapping rate of at least 97.5%, similar to previous studies in other pig breeds (Groenen et al., 2012; Frantz et al., 2015) and which are within the range of the optimal value for SNP calling in pigs (Jiang et al., 2019). The variant calling analysis allowed to identify more than eight million SNPs in autosomes for the ANG pigs, resulting in an average of 1 SNP per 0.23 kb, crucial for the identification of signatures of selection (Ma et al., 2015).

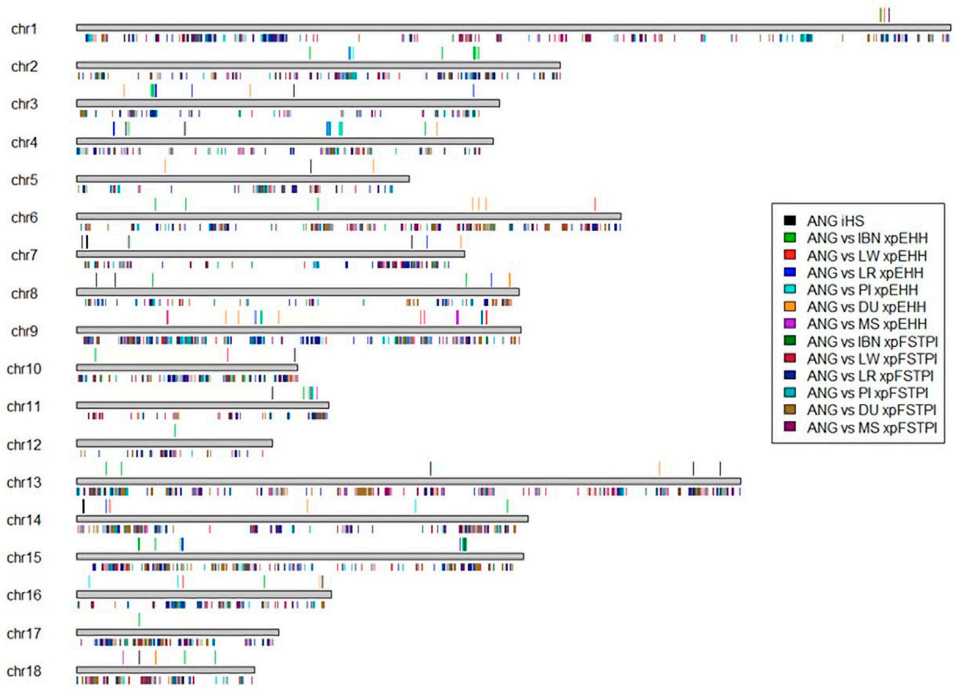

FIGURE 7
Summary representation of candidate sweeps found using the nucleotide diversity/fixation index cross analysis (xpFST/θπ) and haplotype scans (iHS and xpEHH). Three different statistics were used to identify regions under positive selection: iHS for within population and xpFST/θπ and xpEHH for between-populations analysis. All candidate regions identified with these tests are summarized in this plot. Horizontal grey bold lines represent chromosomes. Colored vertical lines represent positions of candidate selective sweeps. Colored lines above chromosomes represent candidate regions identified using haplotype-based methods (iHS and xpEHH). Colored lines below chromosomes represent candidate regions identified using nucleotide diversity/fixation index cross analysis (xp$_{FST}$/θ_π).

TABLE 1 Fixation Index and nucleotide diversity thresholds for the selection of outlier regions.

Breed pairs	F_{ST}[a]		θ_π ratio[a]		Size of outlier regions (Mb)	Number of outlier regions
	Threshold	Mann-Whitney test (W_{FST})	Threshold	Mann-Whitney test ($W_{\theta\pi}$)		
ANG vs. IBN	0.43	1685185 p-value < 2.2e^{-16}	2.94	1336999 p-value < 2.2e^{-16}	3.73	373
ANG vs. LR	0.35	9249564 p-value < 2.2e^{-16}	2.74	7489717 p-value < 2.2e^{-16}	18.65	1865
ANG vs. LW	0.33	5625687 p-value < 2.2e^{-16}	1.76	4757995 p-value < 2.2e^{-16}	11.53	1153
ANG vs. PI	0.48	7171221 p-value < 2.2e^{-16}	1.89	5278266, p-value < 2.2e^{-16}	16.50	1650
ANG vs. DU	0.49	5967174, p-value < 2.2e^{-16}	2.76	4406428 p-value < 2.2e^{-16}	11.79	1179
ANG vs. MS	0.61	4072772 p-value < 2.2e^{-16}	2.49	6019531 p-value < 2.2e^{-16}	12.30	1230
ANG vs. EWB	0.33	2143168 p-value < 2.2e^{-16}	1.81	1852636 p-value < 2.2e^{-16}	3.95	395
ANG vs. AWB	0.53	5911630 p-value < 2.2e^{-16}	2.59	9101593 p-value < 2.2e^{-16}	17.85	1785

[a]The selection of 95% regions with the highest values for F_{ST} and θ_π ratio was conducted for each cross analysis. The threshold values for each metric were registered. A Mann-Whitney U test was conducted to compare the selected outlier regions with the whole genome for each metric.

Phylogenetic analysis and population structure

The investigation of the relatedness of ANG pigs with other world breeds placed these within the European clade, within a differentiated sub-clade suggesting the existence of a common ancestry that may have derived from the older populations of African pigs. This analysis shows that Iberian samples are clustered scattered, closer to samples of EWB or closer to ANG pigs. The close relationship of Iberian pigs with

TABLE 2 Selected outlier regions exhibiting strong positive selection in cross-population Extent Haplotype Homozygosity (xpEHH) analysis.

Breed pairs	Number of outlier regions	Regions total extent (Kb)	Total SNP count	Outlier SNP count (p-value > 10^{-4})	Gene count
ANG vs. IBN	55	1,277	7,137	1125	28
ANG vs. LR	19	391	2,623	204	9
ANG vs. LW	16	329	2031	214	10
ANG vs. PI	23	457	2,529	246	10
ANG vs. DU	6	134	892	83	3
ANG vs. MS	13	264	1025	173	3

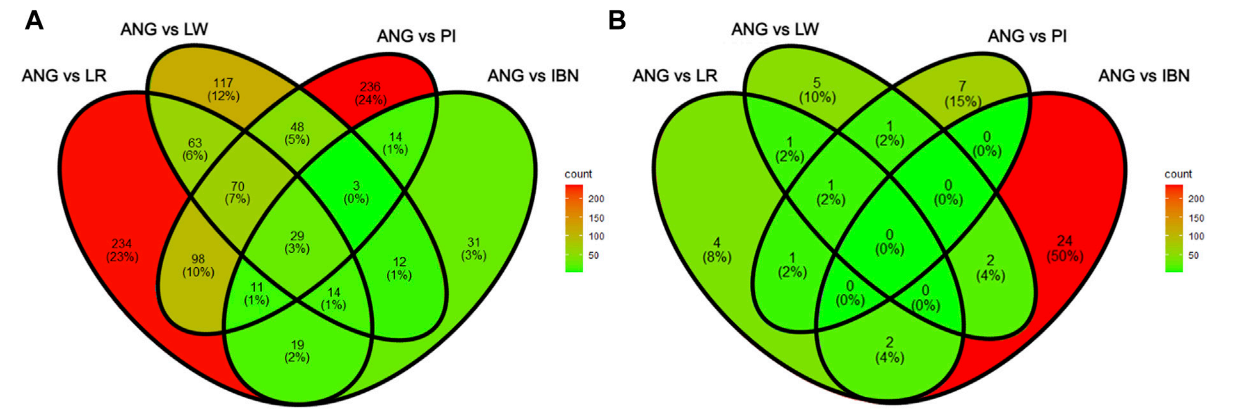

FIGURE 8
Venn diagrams of shared genes identified when comparing ANG with European domestic pigs with xp$_{FST}$/θ_{π} and xpEHH analyses. **(A)** Number of shared candidate genes between the four pairwise comparisons obtained using xp$_{FST}$/θ_{π}. **(B)** Number of shared candidate genes between the four pairwise comparisons obtained using xpEHH.

TABLE 3 Quantitative trait *loci* related to common genes found in xpF_{ST}/θ_{π} and xpEHH analysis.

Gene	QTL ID	Description	Breed comparison	Ref
NRCAM	—	—	ANG vs. IBN, ANG vs. LW	—
SNX24	—	—	ANG vs. IBN	—
RBFOX1	—	—	ANG vs. IBN	—
U6	—	—	ANG vs. IBN	—
SNTG1	—	—	ANG vs. PI, ANG vs. LR	—
CDKAL1	—	—	ANG vs. PI, ANG LW, ANG vs. DU	—
DOCK5	31226	Backfat between 3rd and 4th last ribs	ANG vs. LR	Fowler et al. (2013)
DLGAP2	22424	Feed intake per feeding	ANG vs. LR	Do et al. (2013)

European wild boars has been reported (Ramírez et al., 2015). Introgression events between these Iberian populations and their wild counterparts may be pointed out as a potential explanation for this observation considering that these pigs are traditionally bred at least for a few months rearing in acorn fields (Gama et al., 2013; Herrero-Medrano et al., 2013). In terms of genetic differentiation the analysis of pairwise F_{ST} revealed a lower distance between ANG and EWB, followed by LR and LW. This result suggests that ANG pigs are closer to European pigs breeds and clearly differentiated from Asian pigs. This result is further supported by the PCA analysis that clearly places ANG pigs among European pig breeds. Regarding the Duroc, the placement of samples reflects a gradient due to the multiple genetic influences that affected the origin of this breed (Edea et al., 2014; Ramírez et al., 2015; Traspov et al., 2016; Shu-qi et al., 2021). In spite of the relationship of ANG pigs with European breeds, they display a unique genetic signature that differentiates them from other breeds, as shown in the admixture analysis. Also, the admixture analysis showed that Duroc and ANG pigs shared a common genetic background, supporting the hypothesis of an African origin for the Duroc breed, as reported by Porter (1993) who indicated that the Duroc breed established by Wisconsin breeders descended from multiple genetic sources, including Berkshire, Iberian, Tamworth and Red Guinea Hog.

Results obtained regarding the patterns of LD decay were similar as previous studies in which Meishan displays lower levels in comparison with European pig breeds (Larson et al., 2005; Amaral et al., 2008; Badke et al., 2012; Diao et al., 2019; Muñoz et al., 2019). The high level of LD observed in ANG pigs might result from the reduction of the population size that these free-ranging pigs have suffered, due to the difficult sociological events during the last ~40 years in Angola (Amills et al., 2013). This reduction in population size may have directly increased inbreeding (Lindblad-Toh et al., 2005).

Detection of selection

For the analysis of signatures of selection it is important to take into account the power of the data generated in order to provide an adequate strategy of analysis. It has been shown that the power of xpF_{ST}/θπ depends on sample size, requiring a large sample size in order to achieve high power (Ma et al., 2015). Power is also highly influenced by other factors such as SNP density, in which the use of WGS increases power to 80% accuracy (Ma et al., 2015). Regarding other methods such as iHS and xpEHH, these are not so affected by sample size and require SNP densities at the level of 50K SNP arrays, in order to achieve high power in the detection of selection, revealing accuracy levels above 90% in the case of WGS data even with a small sample size (Ma et al., 2015). In this study, the population analyzed was ANG pigs which is extremely difficult to sample. Nevertheless the small sample size is similar as for other populations in similar studies (Frantz et al., 2015; Tong et al., 2020) that have used WGS and the obtained SNP density (1/0.23 Kb) allows to achieve a power that is expected to be at least 80% for xpF_{ST}/θπ analysis and more than 90% using as iHS and xpEHH. Therefore a complementary approach for the identification of selection signatures was followed and three methods were used. The iHS method, which allows to identify signatures of strong sweeps within a breed, and the xpEHH and xpF_{ST}/θπ methods, both allowing to detect signatures of strong selection sweeps between breeds. The use of these methods allows identifying signatures of selection that might have occurred in different time scales. The iHS and xpEHH are aimed to detect more recent events of selection whereas xpF_{ST}/θπ detects events of selection that have occurred in ancient times. Overall, the results obtained are consistent with other studies which report the identification of a large number of candidate regions with a short extent using F_{ST} or θπ or xpF_{ST}/θπ, and report a reduced number of candidate regions of larger extent using iHS or xpEHH methods (Li et al., 2014; Ma et al., 2015; Chen et al., 2016). The identification of genes with associated QTLs for feed efficiency under positive selection suggests that this population is under local adaptation to a harsh environment, with limited food availability. This is similar to other studies assessing adaptation of breeds in different species; in sheep, positive selection signatures in genes related to resistance to infection, bone formation and fat deposition were identified in several landraces across Africa, Asia and Europe (Li et al., 2020). Selection for body shape has been reported in pigs as an important character for adaptation of different breeds (Li et al., 2014; Bovo et al., 2020) which results in different composition of tissues and ability to manage short availability of water and food in some environments. The *CDKAL1* gene is located within candidate outlier regions unique in ANG pigs and identified by the three methods of analysis used, using both within breed and between breed methods and therefore should be further explored in the future. Previous studies have shown that *CDKAL1* gene overlaps with identified QTLs associated with cholesterol metabolism, homeostasis, transport and regulation in humans (Cheon et al., 2018) and pigs (Bovo et al., 2019). Polymorphisms in the *CDKAL1* gene are reported to be associated with type 2 *diabetes mellitus* in several human populations (Hu et al., 2009; Han et al., 2010; Naser et al., 2021; Amin et al., 2022; Ghosh et al., 2022), and the *CDKAL1* gene is involved in the metabolism of insulin that affects glucose mediated mechanisms, crucial for feed-efficiency and other growth and adaptation traits (Santos et al., 2020).

This study generated the first whole genome sequencing data of a native pig population from Angola that is highly threatened. Our results have provided novel insights regarding its origins and also how it may have influenced other world breeds, namely the Duroc. Our results suggest the existence of a unique genetic background and provide the identification of several genes related with the adaptation to conditions of drought and to an environment scarce in nutrient availability. These results may provide novel opportunities for conservation and for the genetic improvement of these pigs as well as for other worldwide pig breeds, especially in a context of climate change.

Data availability statement

The data presented in the study are deposited in the European Nucleotide Archive repository (https://www.ebi.ac.uk/ena), accession number PRJEB49797.

Ethics statement

Ethical review and approval was not required for the animal study because ear tissue samples of Angolan pigs were collected by qualified veterinarians through their routine practice in the framework of Angolan Law 4/04 (13th August 2004) for the identification and health control of animals in Angola.

Author contributions

AJA, LTG, JC, and AL conceptualized the project. AJA managed and organized the research project. JC and HC performed sampling. AJA and PS performed the bioinformatics analyses and LTG provided insights regarding analysis. DS performed molecular biology assays. AL, JC, HC, LTG, and DS provided insightful suggestions and comments on the manuscript. AJA and PS wrote the manuscript.

Funding

This work was supported by FCT- Fundação para a Ciência e Tecnologia IP (UIDB/00276/2020; LA/P/0059/2020 - AL4AnimalS; 2021.02058.CEECIND).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1003069/full#supplementary-material

References

Amaral, A. J., Megens, H.-J., Crooijmans, R. P. M. A., Heuven, H. C. M., and Groenen, M. A. M. (2008). Linkage disequilibrium decay and haplotype block structure in the pig. *Genetics* 179, 569–579. doi:10.1534/GENETICS.107.084277

Amills, M., Ramírez, O., Galman-Omitogun, O., and Clop, A. (2013). Domestic pigs in Africa. *Afr. Archaeol. Rev.* 30, 73–82. doi:10.1007/S10437-012-9111-2

Amin, U. S. M., Parvez, N., Rahman, T. A., Hasan, M. R., Das, K. C., Jahan, S., et al. (2022). CDKAL1 gene rs7756992 A/G and rs7754840 G/C polymorphisms are associated with gestational diabetes mellitus in a sample of Bangladeshi population: Implication for future T2DM prophylaxis. *Diabetol. Metab. Syndr.* 14, 18. doi:10.1186/s13098-021-00782-w

Andrews, S. (2010). FastQC A quality control tool for high throughput sequence data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Badke, Y. M., Bates, R. O., Ernst, C. W., Schwab, C., and Steibel, J. P. (2012). Estimation of linkage disequilibrium in four US pig breeds. *BMC Genomics* 131 (13), 24–10. doi:10.1186/1471-2164-13-24

Blench, R. M. (2000). "A history of pigs in Africa," in *The origins and development of African livestock*. Editors R. M. Blench and K. C. McDonald (London: UCL press), 355–367.

Bovo, S., Mazzoni, G., Bertolini, F., Schiavo, G., Galimberti, G., Gallo, M., et al. (2019). Genome-wide association studies for 30 haematological and blood clinical-biochemical traits in Large White pigs reveal genomic regions affecting intermediate phenotypes. *Sci. Rep.* 91 (9), 7003–7017. doi:10.1038/s41598-019-43297-1

Bovo, S., Ribani, A., Muñoz, M., Alves, E., Araujo, J. P., Bozzi, R., et al. (2020). Whole-genome sequencing of European autochthonous and commercial pig breeds allows the detection of signatures of selection for adaptation of genetic resources to different breeding and production systems. *Genet. Sel. Evol.* 52, 33. doi:10.1186/s12711-020-00553-7

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 103, 338–348. doi:10.1016/J.AJHG.2018.07.015

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097. doi:10.1086/521987

Chalkias, H., Jonas, E., Andersson, L., Jacobson, M., de Koning, D., N, L., et al. (2017). Identification of novel candidate genes for the inverted teat defect in sows using a genome-wide marker panel. *J. Appl. Genet.* 58, 249–259. doi:10.1007/S13353-016-0382-1

Chen, M., Pan, D., Ren, H., Fu, J., Li, J., Su, G., et al. (2016). Identification of selective sweeps reveals divergent selection between Chinese Holstein and Simmental cattle populations. *Genet. Sel. Evol.* 48, 76. doi:10.1186/s12711-016-0254-5

Cheon, E. J., Cha, D. H., Cho, S. K., Noh, H. M., Park, S., Kang, S. M., et al. (2018). Novel association between CDKAL1 and cholesterol efflux capacity: Replication after GWAS-based discovery. *Atherosclerosis* 273, 21–27. doi:10.1016/J.ATHEROSCLEROSIS.2018.04.011

Cook, D., and Andersen, E. (2017). VCF-Kit: Assorted utilities for the variant call format. *Bioinformatics* 33, 1581–1582. doi:10.1093/bioinformatics/btx011

Crespo-Piazuelo, D., Criado-Mesas, L., Revilla, M., Castelló, A., Noguera, J. L., Fernández, A. I., et al. (2020). Identification of strong candidate genes for backfat and intramuscular fatty acid composition in three crosses based on the Iberian pig. *Sci. Rep.* 10, 13962. doi:10.1038/S41598-020-70894-2

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi:10.1093/BIOINFORMATICS/BTR330

Davis-Richardson, A. (2019). BioHAcks. Available at: https://github.com/audy/bioinformatics-hacks

Diao, S., Huang, S., Xu, Z., Ye, S., Yuan, X., Chen, Z., et al. (2019). Genetic diversity of indigenous pigs from south China area revealed by SNP array. *Animals.* 9, E361. doi:10.3390/ANI9060361

Do, D. N., Strathe, A. B., Ostersen, T., Jensen, J., Mark, T., and Kadarmideen, H. N. (2013). Genome-wide association study reveals genetic architecture of eating behavior in pigs and its implications for humans obesity by comparative mapping. *PLoS One* 8, e71509. doi:10.1371/JOURNAL.PONE.0071509

Dodt, M., Roehr, J., Ahmed, R., and Dieterich, C. (2012). FLEXBAR—flexible barcode and adapter processing for next-generation sequencing platforms. *Biol. (Basel)* 1, 895–905. doi:10.3390/biology1030895

Drag, M. H., Kogelman, L. J. A., Maribo, H., Meinert, L., Thomsen, P. D., and Kadarmideen, H. N. (2019). Characterization of eqtls associated with androstenone by rna sequencing in porcine testis. *Physiol. Genomics* 51, 488–499. doi:10.1152/PHYSIOLGENOMICS.00125.2018

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* 4, 1184–1191. doi:10.1038/NPROT.2009.97

Edea, Z., Kim, S.-W., Lee, K.-T., Kim, T. H., and Kim, K.-S. (2014). Genetic structure of and evidence for admixture between western and Korean native pig breeds revealed by single nucleotide polymorphisms. *Asian-Australas. J. Anim. Sci.* 27, 1263–1269. doi:10.5713/AJAS.2014.14096

Felsenstein, J. (2005). Phylogeny inference package. Felsenstein, Joseph, Dep. Genome Sci. , Univ. of Washington, Seattle. Available at: http://evolution.genetics.washington.edu/phylip.html.

Fowler, K., R, P.-W., J, B., Ej, C., Cp, R., Na, A., et al. (2013). Genome wide analysis reveals single nucleotide polymorphisms associated with fatness and putative novel copy number variants in three pig breeds. *BMC Genomics* 14, 784. doi:10.1186/1471-2164-14-784

Frantz, L. A. F., Haile, J., Lin, A. T., Scheu, A., Geörg, C., Benecke, N., et al. (2019). Ancient pigs reveal a near-complete genomic turnover following their introduction to Europe. *Proc. Natl. Acad. Sci. U. S. A.* 116, 17231–17238. doi:10.1073/pnas.1901169116

Frantz, L. A. F., Schraiber, J. G., Madsen, O., Megens, H. J., Cagan, A., Bosse, M., et al. (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. doi:10.1038/ng.3394

Gama, L. T., Martínez, A. M., Carolino, I., Landi, V., Delgado, J. V, Vicente, A. A., Vicente, A. A., et al. (2013). Genetic structure, relationships and admixture with wild relatives in native pig breeds from Iberia and its islands. *Genet. Sel. Evol.* 45, 18. doi:10.1186/1297-9686-45-18

Gautier, M., Klassmann, A., and Vitalis, R. (2017). Rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol. Ecol. Resour.* 17, 78–90. doi:10.1111/1755-0998.12634

Gautier, M., and Vitalis, R. (2012). rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* 28, 1176–1177. doi:10.1093/BIOINFORMATICS/BTS115

Ghosh, C., Das, N., Saha, S., Kundu, T., Sircar, D., and Roy, P. (2022). Involvement of Cdkal1 in the etiology of type 2 diabetes mellitus and microvascular diabetic complications: A review. *J. Diabetes Metab. Disord.* 21, 991–1001. doi:10.1007/s40200-021-00953-6

Goudet, J. (2005). HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* 2, 184–186. doi:10.1111/j.1471-8286.2004.00828.x

Groenen, M. A. M., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491, 393–398. doi:10.1038/nature11622

Han, X., Luo, Y., Ren, Q., Zhang, X., Wang, F., Sun, X., et al. (2010). Implication of genetic variants near SLC30A8, HHEX, CDKAL1, CDKN2A/B, IGF2BP2, FTO, TCF2, KCNQ1, and WFS1 in Type 2 Diabetes in a Chinese population. *BMC Med. Genet.* 111 11, 81–89. doi:10.1186/1471-2350-11-81

Herrero-Medrano, J. M., Megens, H.-J., Groenen, M. A., Ramis, G., Bosse, M., Pérez-Enciso, M., et al. (2013). Conservation genomic analysis of domestic and wild pig populations from the Iberian Peninsula. *BMC Genet.* 141 14, 106–113. doi:10.1186/1471-2156-14-106

Hu, C., Zhang, R., Wang, C., Wang, J., Ma, X., Lu, J., et al. (2009). PPARG, KCNJ11, CDKAL1, cdkn2a-CDKN2B, IDE-KIF11-HHEX, IGF2BP2 and SLC30A8 are associated with type 2 diabetes in a Chinese population. *PLoS One* 4, e7643. doi:10.1371/JOURNAL.PONE.0007643

Hu, Z. L., and Reecy, J. M. (2007). Animal QTLdb: Beyond a repository - a public platform for QTL comparisons and integration with diverse types of structural genomic information. *Mamm. Genome* 18, 1–4. doi:10.1007/s00335-006-0105-8

Huang, D., Sherman, B., and Lempick, R. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nat. Protoc.* 4, 44–57. doi:10.1038/nprot.2008.211

Jiang, Y., Jiang, Y., Wang, S., Zhang, Q., and Ding, X. (2019). Optimal sequencing depth design for whole genome re-sequencing in pigs. *BMC Bioinforma.* 20, 556. doi:10.1186/S12859-019-3164-Z

Jombart, T., and Ahmed, I. (2011). Adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* 27, 3070–3071. doi:10.1093/bioinformatics/btr521

Keel, B., Snelling, W., Lindholm-Perry, A., Oliver, W., Kuehn, L., and Rohrer, G. (2020). Using SNP weights derived from gene expression modules to improve GWAS power for feed efficiency in pigs. *Front. Genet.* 10, 1339. doi:10.3389/FGENE.2019.01339

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinforma.* 15, 356. doi:10.1186/S12859-014-0356-4

Larson, G., Albarella, U., Dobney, K., Rowley-Conwy, P., Schibler, J., Tresset, A., et al. (2007). Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Proc. Natl. Acad. Sci. U. S. A.* 104, 15276–15281. doi:10.1073/PNAS.0703411104

Larson, G., Dobney, K., Albarella, U., Fang, M., Matisoo-Smith, E., Robins, J., et al. (2005). Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science* 307, 1618–1621. doi:10.1126/SCIENCE.1106927

Lefouili, M., and Nam, K. (2022). The evaluation of Bcftools mpileup and GATK HaplotypeCaller for variant calling in non-human species. *Sci. Rep.* 12, 11331. doi:10.1038/s41598-022-15563-2

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Li, M., Tian, S., Yeung, C. K. L., Meng, X., Tang, Q., Niu, L., et al. (2014). Whole-genome sequencing of Berkshire (European native pig) provides insights into its origin and domestication. *Sci. Rep.* 4, 4678–4687. doi:10.1038/srep04678

Li, X., Yang, J., Shen, M., Xie, X.-L., Liu, G.-J., Xu, Y.-X., et al. (2020). Whole-genome resequencing of wild and domestic sheep identifies genes associated with morphological and agronomic traits. *Nat. Commun.* 111 11, 2815–2816. doi:10.1038/s41467-020-16485-1

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., et al. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438, 803–819. doi:10.1038/nature04338

Liu, Q., Yue, J., Niu, N., Liu, X., Yan, H., Zhao, F., et al. (2020). Genome-wide association analysis identified BMPR1A as a novel candidate gene affecting the number of thoracic vertebrae in a large white × minzhu intercross pig population. *Animals.* 10, 21866–E2212. doi:10.3390/ANI10112186

Liu, X., Wang, L. G., Luo, W. Z., Li, Y., Liang, J., Yan, H., et al. (2014). Genome-wide SNP scan in a porcine Large White×Minzhu intercross population reveals a locus influencing muscle mass on chromosome 2. *Anim. Sci. J.* 85, 969–975. doi:10.1111/ASJ.12230

Ma, Y., Ding, X., Qanbari, S., Weigend, S., Zhang, Q., and Simianer, H. (2015). Properties of different selection signature statistics and a new strategy for combining them. *Heredity* 115, 426–436. doi:10.1038/HDY.2015.42

Man, H. B., and Whitney, D. R. (1947). On a test whether one of two random variables is stochatiscally larger than the other. *Ann. Math. Stat.* 18, 50–60. doi:10.1214/aoms/1177730491

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., et al. (2016). The Ensembl variant effect predictor. *Genome Biol.* 171 17, 122–214. doi:10.1186/S13059-016-0974-4

Muñoz, M., Bozzi, R., García-Casco, J., Núñez, Y., Ribani, A., Franci, O., et al. (2019). Genomic diversity, linkage disequilibrium and selection signatures in European local pig breeds assessed with a high density SNP chip. *Sci. Rep.* 91 9, 13546–13614. doi:10.1038/s41598-019-49830-6

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., and Durbin, R. (2016). BCFtools/RoH: A hidden markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* 32, 1749–1751. doi:10.1093/bioinformatics/btw044

Naser, F. H., Fadheel, H. K., Hussain, M. K., Algenabi, A. H. A., Mohammad, H. J., Kaftan, A. N., et al. (2021). Association of CDKAL1 gene polymorphisms with type 2 diabetes mellitus in a sample of Iraqi population. *Gene Rep.* 25, 101371. doi:10.1016/j.genrep.2021.101371

Nei, M., and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* 76, 5269–5273. doi:10.1073/PNAS.76.10.5269

Niu, N., Wang, H., Shi, G., Liu, X., Liu, H., Liu, Q., et al. (2021). Genome scanning reveals novel candidate genes for vertebral and teat number in the Beijing Black Pig. *Anim. Genet.* 52, 734–738. doi:10.1111/AGE.13111

Nonneman, D., Lents, C., Rohrer, G., Rempel, L., and Vallet, J. (2014). Genome-wide association with delayed puberty in swine. *Anim. Genet.* 45, 130–132. doi:10.1111/AGE.12087

Park, H., Han, S., Lee, J., and Cho, I. (2017). Rapid Communication: High-resolution quantitative trait loci analysis identifies LTBP2 encoding latent transforming growth factor beta binding protein 2 associated with thoracic vertebrae number in a large F2 intercross between Landrace and Korean native pigs. *J. Anim. Sci.* 95, 1957–1962. doi:10.2527/JAS.2017.1390

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi:10.1086/519795

Ramírez, O., Burgos-Paz, W., Casas, E., Ballester, M., Bianco, E., Olalde, I., et al. (2014). Genome data from a sixteenth century pig illuminate modern breed relationships. *Heredity* 1142 114, 175–184. doi:10.1038/hdy.2014.81

Ramírez, O., Burgos-Paz, W., Casas, E., Ballester, M., Bianco, E., Olalde, I., et al. (2015). Genome data from a sixteenth century pig illuminate modern breed relationships. *Heredity* 114, 175–184. doi:10.1038/HDY.2014.81

Ramírez, O., Ojeda, A., Tomàs, A., Gallardo, D., Huang, L. S., Folch, J. M., et al. (2009). Integrating Y-chromosome, mitochondrial, and autosomal data to analyze the origin of pig breeds. *Mol. Biol. Evol.* 26, 2061–2072. doi:10.1093/MOLBEV/MSP118

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 4497164 449, 913–918. doi:10.1038/nature06250

Sabeti, P., Reich, D., Higgins, J., Levine, H., Richter, D., Schaffner, S., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837. doi:10.1038/NATURE01140

Santos, M. C. F. dos, Anderson, C. P., Neschen, S., Zumbrennen-Bullough, K. B., Romney, S. J., Kahle-Stephan, M., et al. (2020). Irp2 regulates insulin production through iron-mediated Cdkal1-catalyzed tRNA modification. *Nat. Commun.* 111 11, 296–316. doi:10.1038/s41467-019-14004-5

Sato, S., Uemoto, Y., Kikuchi, T., Egawa, S., Kohira, K., Saito, T., et al. (2017). Genome-wide association studies reveal additional related loci for fatty acid composition in a Duroc pig multigenerational population. *Anim. Sci. J.* 88, 1482–1490. doi:10.1111/ASJ.12793

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi:10.1093/bioinformatics/btr026

Shu-qi, D., Zhi-ting, X., Shao-pan, Y., Shu-wen, H., Jin-yan, T., Xiao-long, Y., et al. (2021). Exploring the genetic features and signatures of selection in South China indigenous pigs. *J. Integr. Agric.* 2021, 1359–1371. doi:10.1016/S2095-3119(20)63260-9

Skotte, L., Korneliussen, T. S., and Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics* 195, 693–702. doi:10.1534/GENETICS.113.154138

Supek, F., Bošnjak, M., õ Kunca, S., and õ Muc, S. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6, 21800. doi:10.1371/journal.pone.0021800

Swart, H., Kotze, A., Olivier, P. A. S., and Grobler, J. P. (2010). Microsatellite-based characterization of Southern African domestic pigs (*Sus scrofa* domestica). *S. Afr. J. Anim. Sci.* 40, 121–132. doi:10.4314/sajas.v40i2.57280

Team, R. C (2013). R: A language and environment for statistical computing. Available at: http://www.R-project.org/.

Tong, X., Hou, L., He, W., Chugang, M., Huang, B., Zhang, C., et al. (2020). Whole genome sequence analysis reveals genetic structure and X-chromosome haplotype structure in indigenous Chinese pigs. *Sci. Rep.* 10, 9433. doi:10.1038/s41598-020-66061-2

Tracy, C. A., and Widom, H. (1992). Level-Spacing distributions and the airy kernel. Available at: http://arxiv.org/abs/hep-th/9210074 (Accessed September 22, 2021).

Tracy, C. A., and Widom, H. (1994). Level-spacing distributions and the Airy kernel. *Commun. Math. Phys.* 159, 151–174. doi:10.1007/bf02100489

Traspov, A., Deng, W., Kostyunina, O., Ji, J., Shatokhin, K., Lugovoy, S., et al. (2016). Population structure and genome characterization of local pig breeds in Russia, Belorussia, Kazakhstan and Ukraine. *Genet. Sel. Evol.* 481 48, 16–19. doi:10.1186/S12711-016-0196-Y

Wang, L., Zhang, W., and Li, Q. (2020). Assoctests: An R package for genetic association studies. *J. Stat. Softw.* 94, 1–26. doi:10.18637/jss.v094.i05

Wang, S., Meyer, E., McKay, J. K., and Matz, M. V. (2012). 2b-RAD: A simple and flexible method for genome-wide genotyping. *Nat. Methods* 9, 808–810. doi:10.1038/nmeth.2023

Weir, B. S., and Cockerham, C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358–1370. doi:10.1111/j.1558-5646.1984.tb05657.x

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biom. Bull.* 1, 80–83. doi:10.2307/3001968

Wu, Z. C., Liu, Y., Zhao, Q. H., Zhu, S. P., Huo, Y. J., Zhu, G. Q., et al. (2015). Association between polymorphisms in exons 4 and 10 of the BPI gene and immune indices in Sutai pigs. *Genet. Mol. Res.* 14, 6048–6058. doi:10.4238/2015.JUNE.8.2

Zhang, C., Dong, S. S., Xu, J. Y., He, W. M., and Yang, T. L. (2019). PopLDdecay: A fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* 35, 1786–1788. doi:10.1093/bioinformatics/bty875

Zhuang, Z., Li, S., Ding, R., Yang, M., Zheng, E., Yang, H., et al. (2019). Meta-analysis of genome-wide association studies for loin muscle area and loin muscle depth in two Duroc pig populations. *PLoS One* 14, e0218263. doi:10.1371/JOURNAL.PONE.0218263