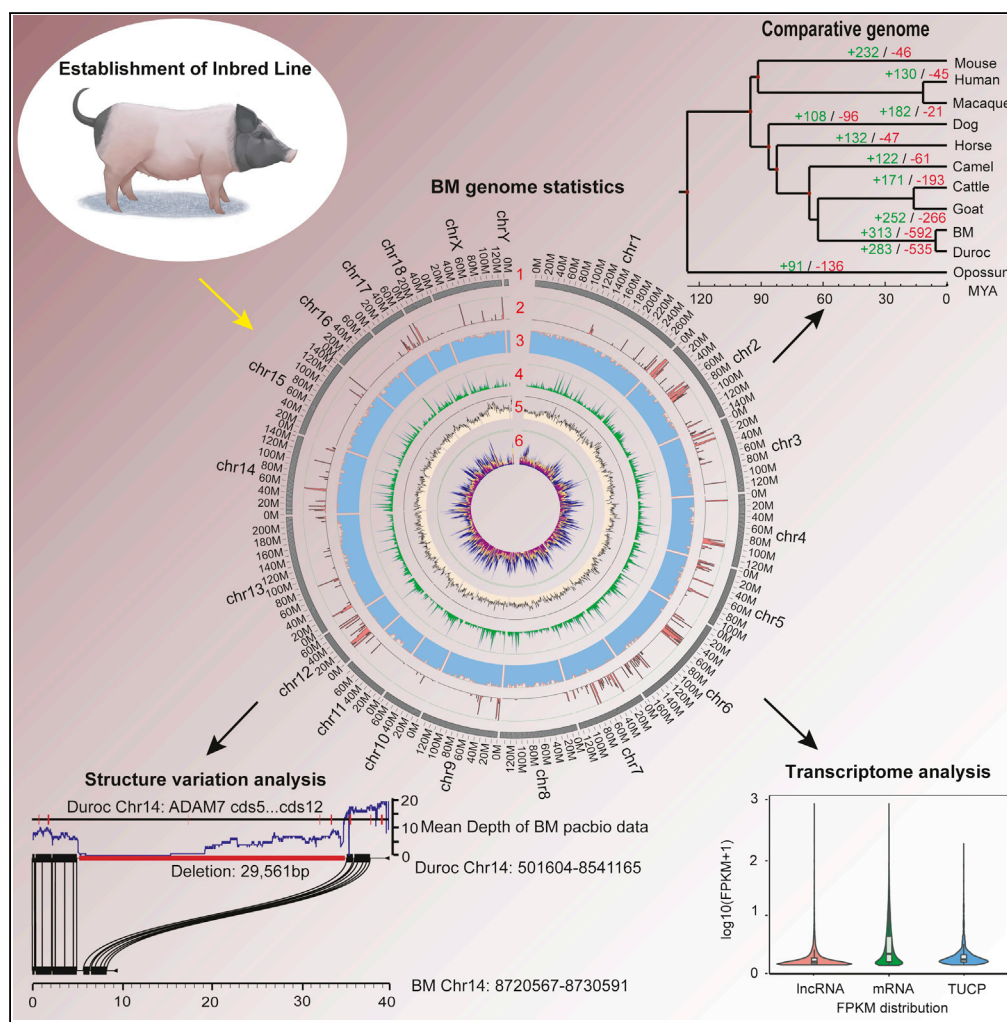


Article

Development and Genome Sequencing of a Laboratory-Inbred Miniature Pig Facilitates Study of Human Diabetic Disease



Li Zhang,
Yuemeng Huang,
Meng Wang, ...,
Kui Li, Ji-Feng Fei,
Ganqiu Lan

liangjing@gxu.edu.cn (J.L.)
chaoshi@qust.edu.cn (C.S.)
gqlan@gxu.edu.cn (G.L.)

HIGHLIGHTS

Bama miniature pig (BM) is one of the pig lines with the highest inbreeding coefficient

This atlas is a report on the chromosome-level genome assembly of miniature pig

Genomic analyses revealed genetic basis underlying BM's advantages to study diabetes

Some lncRNAs and mRNAs may be linked to resistance to diabetogenic environment

Zhang et al., iScience 19, 162–176
September 27, 2019 © 2019 The Author(s).
<https://doi.org/10.1016/j.isci.2019.07.025>



Article

Development and Genome Sequencing of a Laboratory-Inbred Miniature Pig Facilitates Study of Human Diabetic Disease

Li Zhang,^{1,10} Yuemeng Huang,^{1,2,10} Meng Wang,^{3,10} Yafen Guo,¹ Jing Liang,^{1,11,*} Xiurong Yang,¹ Wenjing Qi,¹ Yanjun Wu,¹ Jinglei Si,¹ Siran Zhu,¹ Zhe Li,³ Ruiqiang Li,³ Chao Shi,^{4,*} Shuo Wang,⁴ Qunjie Zhang,⁵ Zhonglin Tang,^{6,7} Lixian Wang,⁸ Kui Li,⁸ Ji-Feng Fei,⁹ and Ganqiu Lan^{1,*}

SUMMARY

Pig has been proved to be a valuable large animal model used for research on diabetic disease. However, their translational value is limited given their distinct anatomy and physiology. For the last 30 years, we have been developing a laboratory Asian miniature pig inbred line (Bama miniature pig [BM]) from the primitive Bama xiang pig via long-term selective inbreeding. Here, we assembled a BM reference genome at full chromosome-scale resolution with a total length of 2.49 Gb. Comparative and evolutionary genomic analyses identified numerous variations between the BM and commercial pig (Duroc), particularly those in the genetic loci associated with the features advantageous to diabetes studies. Resequencing analyses revealed many differentiated gene loci associated with inbreeding and other selective forces. These together with transcriptome analyses of diabetic pig models provide a comprehensive genetic basis for resistance to diabetogenic environment, especially related to energy metabolism.

INTRODUCTION

Pig (*Sus scrofa*) has served not only as one of the most economically important livestock but also as an important model organism used in many areas of medical research, including obesity, cardiovascular disease, endocrinology, diabetes, alcoholism, nephropathy, and organ transplantation, owing to parallels with humans in anatomy and physiology (Andersson, 2016; Ibrahim et al., 2006; Rocha and Plastow, 2006; Schook et al., 2005; Yan et al., 2018). There are over 730 distinct pig breeds worldwide, whose diverse phenotypes are shaped by the combined effects of local adaptation and artificial selection (Ai et al., 2015). However, the vast majority of pig breeds have been developed with a focus on economic benefits, rather than breeding an ideal laboratory animal, which directly resulted in almost non-existence of excellent inbred pig strains as model organisms used in biomedical research.

After several hundred years of intense artificial selection, current commercial pig breeds, represented by Duroc, have undergone drastic phenotypic changes and genetic adaptations that are economically important to the pig industry (e.g., reduction in feeding costs) and the consumer (e.g., higher production of lean meats) (Ai et al., 2015; Frantz et al., 2015; Rubin et al., 2012). In this context, a series of absolutely visible traits of current commercial pigs, such as large body size (adult individuals can reach 300–400 kg in weight), long life cycle, and weak inbreeding level (Table S1), have become obstacles in using pigs as biomedical animal models, especially in the studies of obesity and diabetes mellitus, because they result in high maintenance costs, specialized facility requirements, long experimental periods, and poor repeatability (Kleinert et al., 2018). Moreover, the different biomedical responses and performance of commercial pig breeds from those of wild boars, including severe resistance to “diabetogenic” (high-calorie and low-activity) environment (Gerstein and Waltman, 2006), also go against the construction of some pig models for human diseases. These defects mean the urgent need for a professionally experimental pig strain, which drove us to develop a laboratory Asian miniature pig inbred line—Bama miniature pig (BM) based on Bama xiang pig (BX), a primitive breed without artificial imprinting for commercial characters, whose many characteristics, such as small volume, early maturity, and long-term adaptation to inbreeding, are valuable in the construction of an ideal inbred laboratory pig line (Table S1), more than 30 years ago.

It is known that reference genome sequence is quite important to biomedical studies using pig model (Bains et al., 2016; Crawley et al., 1997). Although the genomes of some pig breeds had been published

¹College of Animal Science and Technology, Guangxi University, Nanning 530004, China

²College of Veterinary Medicine, Northwest A&F University, Yangling 712100, China

³Novogene Bioinformatics Institute, Beijing 100083, China

⁴Shandong Provincial Key Laboratory of Biochemical Engineering, College of Marine Science and Biological Engineering, Qingdao University of Science and Technology, Qingdao 266042, China

⁵Institution of Genomics and Bioinformatics, South China Agricultural University, Guangzhou 510642, China

⁶Research Centre for Animal Genome, Agricultural Genome Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China

⁷Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genome Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China

⁸Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing 100193, China

⁹Institute for Brain Research and Rehabilitation, South China Normal University, Guangzhou 510631, China

¹⁰These authors contributed equally

¹¹Lead Contact

*Correspondence: liangjing@gxu.edu.cn (J.L.), chaoshi@qust.edu.cn (C.S.), gqian@gxu.edu.cn (G.L.)
<https://doi.org/10.1016/j.isci.2019.07.025>



(Groenen et al., 2012; Li et al., 2017), the chromosome-level genomes with comprehensive annotation are still scarce resources to date. Moreover, most of the currently reported pig genomes are from commercial pig breeds, except, to our knowledge, only one highly fragmented draft genome sequence from an experimental inbreeding line, Wuzhishan miniature pig (Fang et al., 2012). Therefore, there is an inevitable quandary in most of the health studies involved in pig genome, that the Duroc reference genome (Sscrofa11.1, GenBank assembly accession: GCA_000003025.6) nearly becomes the only choice, no matter which kind of pig breed is selected. The presentation of BM high-quality reference genome can enrich the *Sus scrofa* genome database to effectively improve this dilemma, and likewise provide essential information needed to shed light on the genetic components of the BM phenotypes advantages to diabetic study, especially the relatively lower resistance to diabetic pressure, by comparative genomic analyses.

In this study, we have successfully inbred BMs to generation 19 (inbred line F19), which is, to our knowledge, the pig line with the highest inbreeding coefficient to date. Using combined technologies, we presented a chromosome-level genome sequence and the available annotation of highly inbred BM. Comparative analyses of BM and Duroc genomes revealed substantial genomic differences between them, as well as identified genetic basis underlying the BM's superior traits to study diabetic diseases. Resequencing analyses between BX and BM populations confirmed the leading inbreeding degree of BMs at the genome-wide level. Besides the positively selected genes (PSGs), selective sweep and transcriptome analyses also found some changes of energy metabolism systems, like phosphatidylinositol 3-kinase (PI3K)-Akt signaling pathway, related to diabetic resistance. This study provided not only an inbred miniature pig line to overcome these preexisting obstacles of using pig in diabetic researches but also a comprehensive molecular basis as reference to further optimize breeding of the experimental animal's phenotypes advantageous to diabetic diseases researches. Meanwhile, it systematically boosted the understanding of the mechanism of resistance to diabetogenic pressure from the genome level to the transcriptome level.

RESULTS

Development of BM Strain and Detection of Resistance to Diabetogenic Environment

To provide a dedicated laboratory pig strain for biomedical research, we have been developing a laboratory Asian miniature pig inbred line, the BM (Figure S1), from the original subtropical BX population (2 males and 14 females selected), which is native to south China (Table S1), for more than 30 years (since 1987) (Figure 1).

After 10 years (1987–1997) closed pure-bred breeding and directional selection, we had developed a closed colony (A strain) to generation 10. The F10 closed colony individuals' adult weight decreased about 10 kg (24-month-old F0: 52.78 ± 0.86 kg versus F10: 43.67 ± 0.77 kg). The ratio of individuals with uniform two-end-black coat color in closed colony increased from 70.44% (F0) to 94.73% (F10). Compared with F0 individuals (Table S1), the aggressive behaviors of F10 animals had disappeared almost completely. Subsequently, in 1997, establishment of inbred line (B strain) was initiated using four male and four female F10 closed colony (A strain) individuals (as inbred generation 0 [inbred line F0]), of which one couple has broken inbreeding bottleneck and has been bred for 19 generations (inbred line F19 maintaining all specific features of the F10 closed colony) so far.

Like other commonly used experimental animal inbred lines (Lilue et al., 2018), the BM has clear genetic background (without foreign gene flow influx), stable characters, and high homozygosity (inbreeding coefficient of inbred line F19: 0.9825) (Figure 1), which can ensure the reproducibility of conclusions of experiments based on this animal. Furthermore, it can be a valid substitute to overcome some shortcomings of the use of commercial pig breeds in biomedical studies, because of BM's observed phenotypic specificity. The BM was selected by inbreeding for a small adult body size (adult body weight: 40–50 kg) and short life cycle (<1.5 months for male sexual maturity), thereby lowering maintenance costs and reducing the duration of experiments (Table S1). In addition, breeding of the BM rigorously conforms to the standards of laboratory animal breeding husbandry, feeding a restricted food supply that provides ~70% of energy *ad libitum* of miniature pig (approximately equals to digestible energy requirement for maintenance) (Table S1), to just accomplish basic energy needs for essential life activities since the start of inbreeding. In short, the laboratory-inbred BMs harbor different physiological characters from those of commercial pig breeds, which make them more useful as disease models, especially for diet (high-fat and high-carbohydrate)-induced diabetes. We exerted diabetogenic pressures on BMs (BM-induced group) and Durocs (Duroc-induced group) by high-fat and high-carbohydrate diet and limited activity space, and the

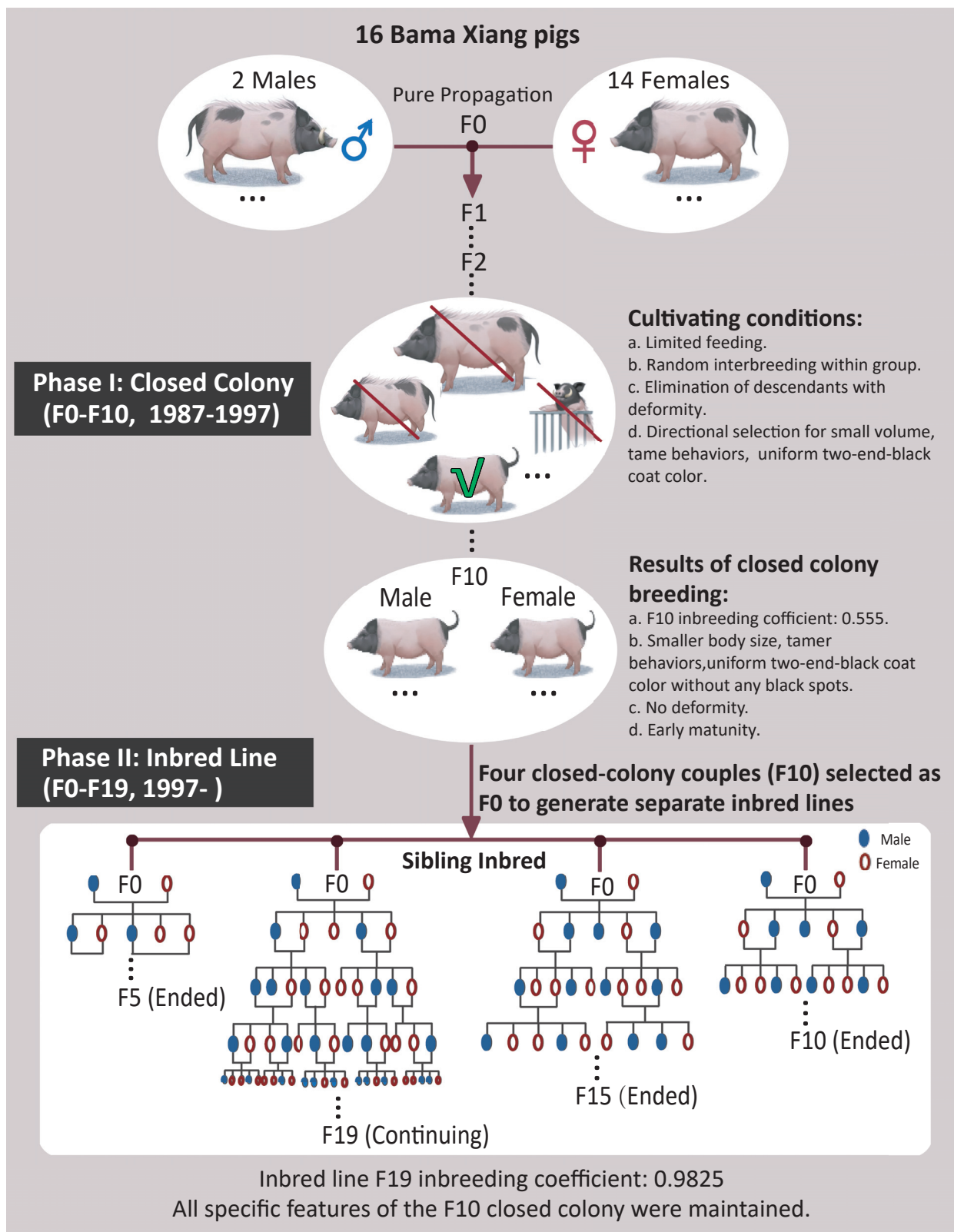


Figure 1. Establishment of Highly Inbred Laboratory BM from BX

Two male and fourteen female BX were introduced to breed a laboratory inbred line (BM) since 1987.

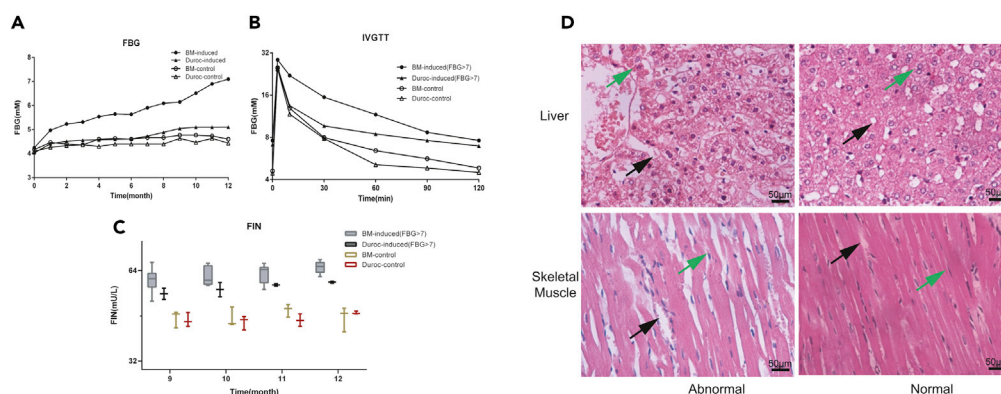


Figure 2. Detection of Resistance to Diabetogenic Environment

(A–D) Fifteen male BMs and fifteen male Durocs (6-month-old) were fed a high-fat and high-carbohydrate diet and had limited activity spaces for 12 months, and three male BMs and three male Durocs were selected to be fed with standard material as control groups (Table S2). Successful resistance of these pigs to diabetogenic environment was assessed by measuring changes in fasting blood glucose (FBG) at different times (A), intravenous glucose tolerance at month 12 (B), and fasting insulin (FINS) at different times (C) and by examining histopathological sections of the pancreases, kidney (Figure S2), liver, and skeletal muscle tissues after 12 months (D). (A) Changes in average FBG. Average FBG of BM-induced group rapidly increased after eighth month, in contrast to that of Duroc-induced group. (B) Intravenous glucose tolerance test (IVGTT) on individuals with FBG >126 mg/dL (7 mmol/L) conducted at month 12. The glucose clearance rate was reduced in BM-induced and Duroc-induced group individuals compared with the control group individuals. (C) Changes in average FINS. The FINS of individuals with FBG >7 mmol/L (10 BMs and 2 Durocs) were much higher than those of control group individuals (3 BMs and 3 Durocs). FINS levels are represented by box-and-whisker plots (with no box if $n < 4$). Boxes represent the interquartile range between the first and third quartiles and median (internal line), and whiskers denote the lowest and highest values, respectively. (D) Pathological sections of pig liver and skeletal muscle. In the liver, the liver cords exhibited a disordered arrangement, and many lipid droplets were observed in the abnormal liver tissue compared with the normal liver tissue in control group. Green and black arrows indicate liver cords and lipid droplets, respectively. In skeletal muscle, the abnormal tissues had fewer fibers, which were disordered, and significantly higher lipid content between fibers compared with normal tissues in control group. Green and black arrows indicate the muscle fibers and lipid droplets, respectively. Scale bars, 50 μ m.

resistance to “diabetogenic” environment was confirmed according to a series of parameters. The fact that 66.7% (10/15) of BMs and 13.3% (2/15) of Durocs have fasting blood glucose level >126 mg/dL (7 mmol/L), abnormal pathology, increased fasting insulin level, and decreased glucose disappearance rate at month 12 means that BMs’ resistance to “diabetogenic” environment is much lower than that of Durocs (Figures 2 and S2; Tables S2 and S3).

Sequencing, Assembly, and Annotation of the BM Reference Genome

To fill the gap of chromosome-level genome assembly of the miniature pig, we generated a reference genome assembly from a male BM. The BM has a diploid chromosome number of 38 (Figure S3) and an estimated genome size of 2.58 Gb (Figure S4; Table S4). To achieve high-quality genome assembly, we adopted a combination of sequencing methods including Illumina paired-end and mate-paired sequencing, 10 \times Genomics linked reads (>50 kb), Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing, Oxford Nanopore sequencing technology (ONT), and chromosome interaction mapping (Hi-C) sequencing. We developed a hybrid assembly pipeline to assemble this genome (Figure 3A) using a total of 932.04-Gb sequence data (equivalent to 361.25 genomic coverage). First, an initial draft genome was assembled by using 71.64-fold (184.83 Gb) 10 \times Genomics barcoded sequencing data and 152.87-fold (349.42 Gb) Illumina sequencing reads, with a scaffolds N50 size of 21.13 Mb (Tables S5 and S6). Second, PacBio long-reads (53 Gb; 20.54-fold), Oxford Nanopore sequences (26 Gb, 10.12-fold), and Hi-C sequence data (273.7 Gb; 106.09-fold) were used to upgrade the BM genome assembly and obtain a chromosome-scale genome with a contig N50 size of 1,010 kb, and total assembled length of 2.49 Gb, of which 97.49% was anchored to 20 chromosomes (18 autosomes and 2 sex chromosomes) ranging in length from 9,839,741 to 283,123,735 bp (Tables S6–S8). This new reference assembly has 5,723 gaps giving an estimated mean gap length of 2 kb (Figure 3B; Table 1). The high level of accuracy and completeness of BM genome assembly is demonstrated by the normal GC content (41.90% of genome), mapping of 97.69% of short sequencing reads, and Benchmarking Universal Single-Copy

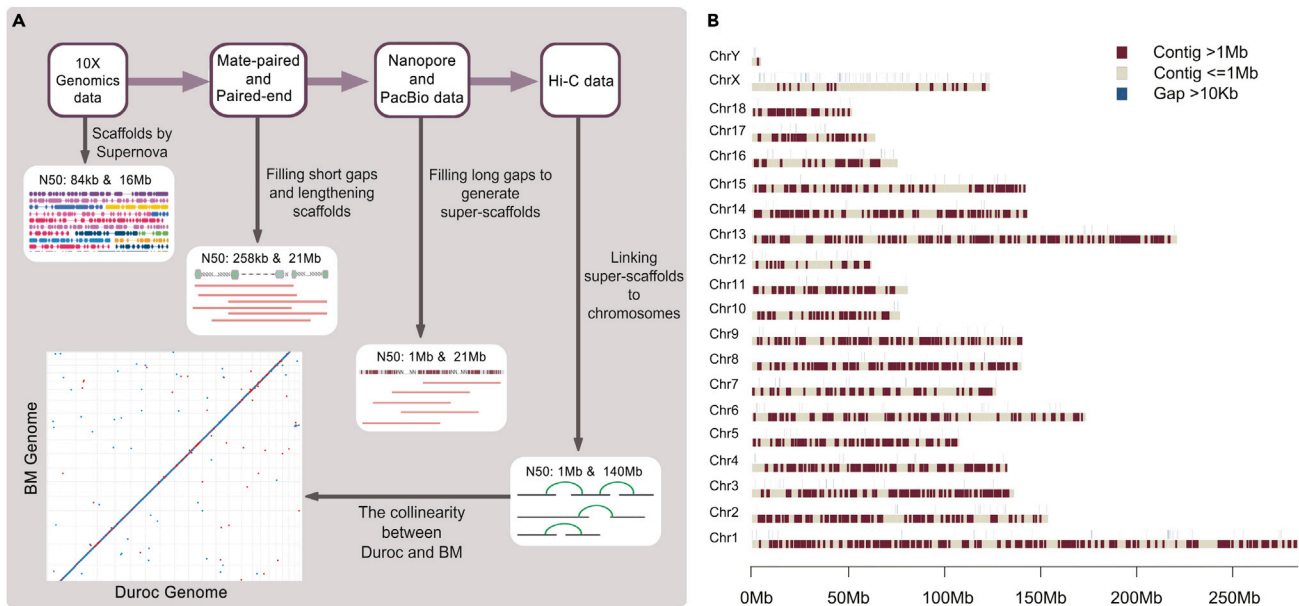


Figure 3. Genome Assembly

(A) Workflow for genome construction.

(B) Ideograms of BM reference chromosome-scale pseudomolecules. The upper track shows positions of all gaps in the pseudomolecules; few of them are longer than 10 kb. More than half of the assembly consists of contigs >1 Mb, which are shown as black red bars in the lower track.

Orthologs (BUSCO)-based completeness assessment (Figure S5; Tables S9–S12). The BM reference genome has 43-fold higher contiguity than the previously published short-read genome assembly of Wuzhishan pig (contig N50: 23.5 kb) (Table 1) (Fang et al., 2012). In short, these combined technologies produced one of the most continuous porcine *de novo* assemblies to date, with chromosome-length scaffolds and the shortest gap lengths.

To aid genome annotation, we also sequenced the transcriptomes of 10 tissues (brain, liver, heart, spleen, lung, kidney, pancreas, stomach, skeletal muscle, and adipose) from the BM. A total of 21,334 protein-encoding genes were annotated using both *de novo* and homologous-based predictions (Figures S6, S7, and S9; Tables S13 and S14). Moreover, it was found that the BM genome is composed of 37.32% repetitive elements, fewer than that (40.55%) of Duroc genome, and encodes functionally important noncoding RNAs (Figures S8 and S9; Tables S15–S17).

Comparison of the BM genome with the human, and three common experimental animal (macaque, mouse, and dog), genomes unveiled three gene families, including *ARF1* and *IGHD*, shared between the BM and human genomes but absent in macaque, mouse, and dog genomes (Figure S10). These genes may play roles in Alzheimer disease, pituitary dwarfism, and growth failure (from database “DisGeNET”). The presence of these genes in the BM potentially facilitates research on the above-mentioned diseases using this animal model. Moreover, BM has fewer unique genes compared with the Duroc (1,303 versus 1,531) (Figure S10), and the genes specific to BM were significantly enriched in the “steroid hormone biosynthesis” Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway ($p = 0.00908$), which is associated with sex hormone secretion, male testicles development, and rapid maturation of sperm.

Porcine endogenous retroviruses (PERVs) are ancient viral sequences integrated into the pig genome and transmitted vertically to the offspring, which has made them difficult to eliminate in the field of organ transplantation (Niu et al., 2017). BM has significantly fewer PERV gene copies (including the *gag*, *pol*, and *env* genes), with a total of 45 copies of potential virus-derived genes in the BM genome assembly, compared with 171 copies in the Duroc genome (Table S18). This reduction in PERV genes included complete loss of the PERV-A and PERV-C *env* genes, and the latter was further validated by general *env* genotyping (Figure S11). Fewer PERVs in BMs alleviate the requirement of knocking out PERV-C genes for pig-to-human

	BM Genome	Wuzhishan Pig Genome
Sequencing technology	10X Genomics, Illumina, PacBio, Nanopore and Hi-C	Illumina
Sequence coverage (X)	361.25	126
Assembly level	Chromosome	Scaffold
Genome size (Gb)	2.49	2.64
Contig N50 (bp)	1,010,657	23,535
Scaffold N50 (bp)	140,438,739	5,432,118
Total assembly gap length (bp)	11,642,258	–
Mean gap length (bp)	2034.2928	–

Table 1. Comparison of the Quality between the BM and Published Inbred Miniature Pig Genome Assemblies

xenotransplantation, especially for the organ transplantation of the tissues that diabetic diseases mainly target on, like liver, heart, kidney, and pancreas.

Structural Variants between the BM and Duroc Genomes

Although BM and Duroc share a high degree of chromosomal collinearity (Figure S12), we found that these two genomes still have numerous blocks of DNA sequence variations. To identify the large-block sequence variations between BM and Duroc, we conducted a comprehensive survey of structural variants (SVs), including genome-wide deletions, duplications, insertions, and inversions, in the BM by alignment of the BM genome with the Duroc reference genome assembly (Sscrofa 11.1). We focused on identifying these SVs >50 bp because of their severe effect on gene function. Our genome-wide alignment identified 59,373 SVs in the BM genome, most of which (98%) were located in intergenic and intronic regions (37,080 [62.45%] within intergenic regions and 21,092 [35.52%] within intronic regions) and few (622 or 0.01%) were located within exons (Table S19), indicating that few of the SVs affected the coding sequences of genes, with most located in noncoding regions.

During inbreeding, BM maintained the early sexual maturity of BX (BM/BX can produce mature sperm for insemination at the age of 45 days) circumventing the inherently long life cycle of the commercial pig (180–240 days for Duroc). We found that some of the SVs between BM and Duroc are located within the exonic regions of genes related to male sperm maturation (Figure 4; Table S20). The functions of these genes cover many aspects of sperm development, including formation of the oviductal sperm reservoir in the pig (*AQN1*) (Dostalova et al., 1994; Sanz et al., 1992); protective effect on boar sperm functionality (*PSP1* and *PSP2*) (Garcia et al., 2006) (Figure 4A); epididymosomes and sperm plasma membrane development (*ADAM7*) (Oh et al., 2009) (Figure 4B); processes in meiosis, germ cell apoptosis, and male infertility (*HSP70-2*) (Dix et al., 1996, 1997); idiopathic male infertility (*UBE2B*) (Suryavathi et al., 2008); and regulation of meiotic pachytene progression during spermatogenesis (*OVOL1*) (Li et al., 2005) (Table S21).

SVs identified influence genes involved in metabolic disorders. Three genes, *AHNAK*, *ADGRF5/GPR116*, and *ATP10D*, reported to regulate obesity (Ramdas et al., 2015), were affected by both deletion and duplication events within exonic regions (Table S21). Knocking out the *AHNAK* gene in mice results in protection from diet-induced obesity, *ADGRF5/GPR116* affects insulin sensitivity via modulation of adipose function (Nie et al., 2012), and *ATP10D* is involved in endoplasmic reticulum-to-Golgi ceramide processing and regulation of obesity (Kengia et al., 2013). These SVs might be correlated with that BMs are more unbearable to diabetogenic pressure compared with Durocs (Figures 2 and S2; Tables S2 and S3).

Evolutionary Status and Small Body Size of BM

It is necessary to determine the medically applied scope of laboratory animals by dissecting the genetic relationship between them and humans. To detect the exact phylogenetic position of BM and Duroc, we constructed a highly resolved phylogenomic tree. BM and Duroc clustered within one clade as expected, but the branch length of Duroc is longer than that of BM (Figure 5A), suggesting that the

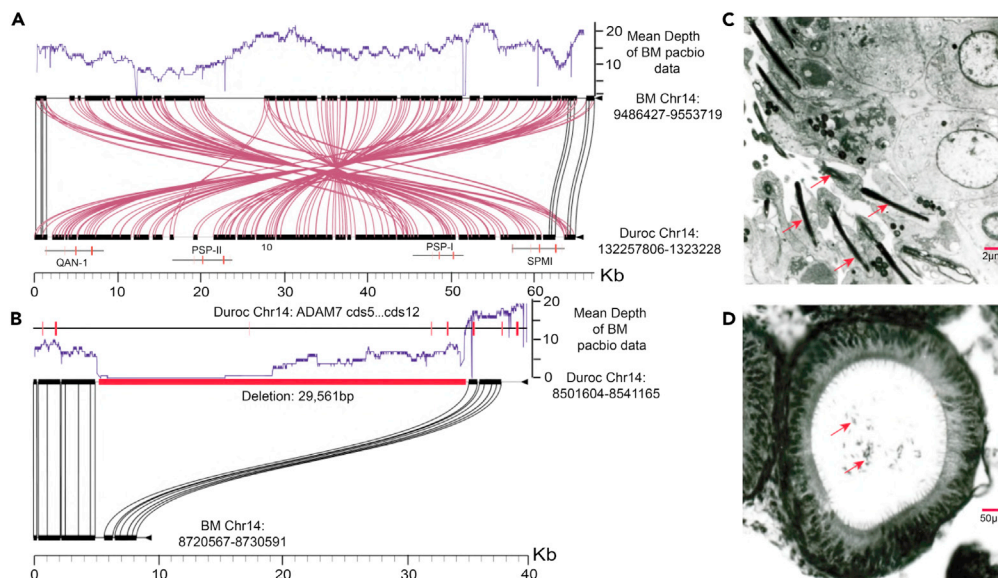


Figure 4. Genome Structural Variation Related to Sperm Maturity between BM and Duroc

(A) An ~67-kb inversion on chromosome 14 (9486427–9553719) in the BM genome (annotated red lines) revealed by PacBio sequencing.
 (B) A large-scale deletion (~29kb) resulted in removal of the putative coding exons of *ADAM7* from the BM genome.
 (C) Microstructures of BM testes (30 days of age). The red arrows indicate spermiogenesis of spermatids. Scale bar, 2 μm .
 (D) Section of BM epididymis (48 days of age). The red arrow indicates the appearance of mature sperm. Scale bar, 50 μm .

commercial pigs have undergone rapid evolution, which gives rise to less genetic divergences between humans and BM than between humans and Duroc.

Among the 11,368 single-copy orthologous genes, the number of orthologous genes, which are closer to counterparts of humans than those of mouse, is higher in BM than that in Duroc (8,547 in BM versus 8,240 in Duroc) (Figure S13). The specific genes, more similar to genes of humans than mice, of BM relative to Duroc were associated with a wide range of physiological processes. These genes were significantly ($p < 0.01$, *t* test) enriched in a major energy metabolic KEGG pathway, “insulin secretion” pathway, which is tightly intertwined with diabetes mellitus and other comorbidities (such as atherosclerosis). We further surveyed gene sets pertinent to eight common human diseases (obesity, Curtasu et al., 2019; type 2 diabetes mellitus [T2DM], Okitsu et al., 2004; nonalcoholic fatty liver disease, Yamada et al., 2017; atherosclerosis, Natarajan et al., 2002; Parkinson disease, Danielsen et al., 2000; Huntington disease, Yan et al., 2018; Alzheimer disease, Holm et al., 2016; and amyotrophic lateral sclerosis, Chieppa et al., 2014) suitably studied by pig model and found that these human-encoded genes are better conserved in the BM than in Duroc and mouse (Figure S14). Taken together, these genome-level analyses illuminated better similarities in physiological genetic background between BM and humans than between Duroc and humans, suggesting that the BM may be more appropriate for analyses of some common diabetic diseases.

Next, we focused on insights into gene family evolution, during which expansion or contraction of gene families have contributed to the phenotypic evolution of animal (Kim et al., 2016; Nowoshilow et al., 2018). To study whether expanded and contracted gene families are responsible for the biological phenotypic changes in BMs relative to commercial pigs, we calculated the number of gene families that diverged along different branches with marked changes (expansion or contraction) (Figure 5A). BM displayed relatively less events of gene family expansion (283 versus 313) and contraction (535 versus 592) compared with Duroc.

Compared with commercial pigs, small body size is one of the most visible characteristics of BM, probably as an ecological response called out by the need of relatively large heat dissipation area formed through body volume loss to a long-term warm-temperature environment in low latitudes (Figure 5B) like other species (Forster et al., 2012; Sheridan and Bickford, 2011). Our analysis identified that the contraction of the *LILRA* and *LILRB* subfamilies (from 11 and 4 copies in Duroc to 5 and 2 copies in BM, respectively),

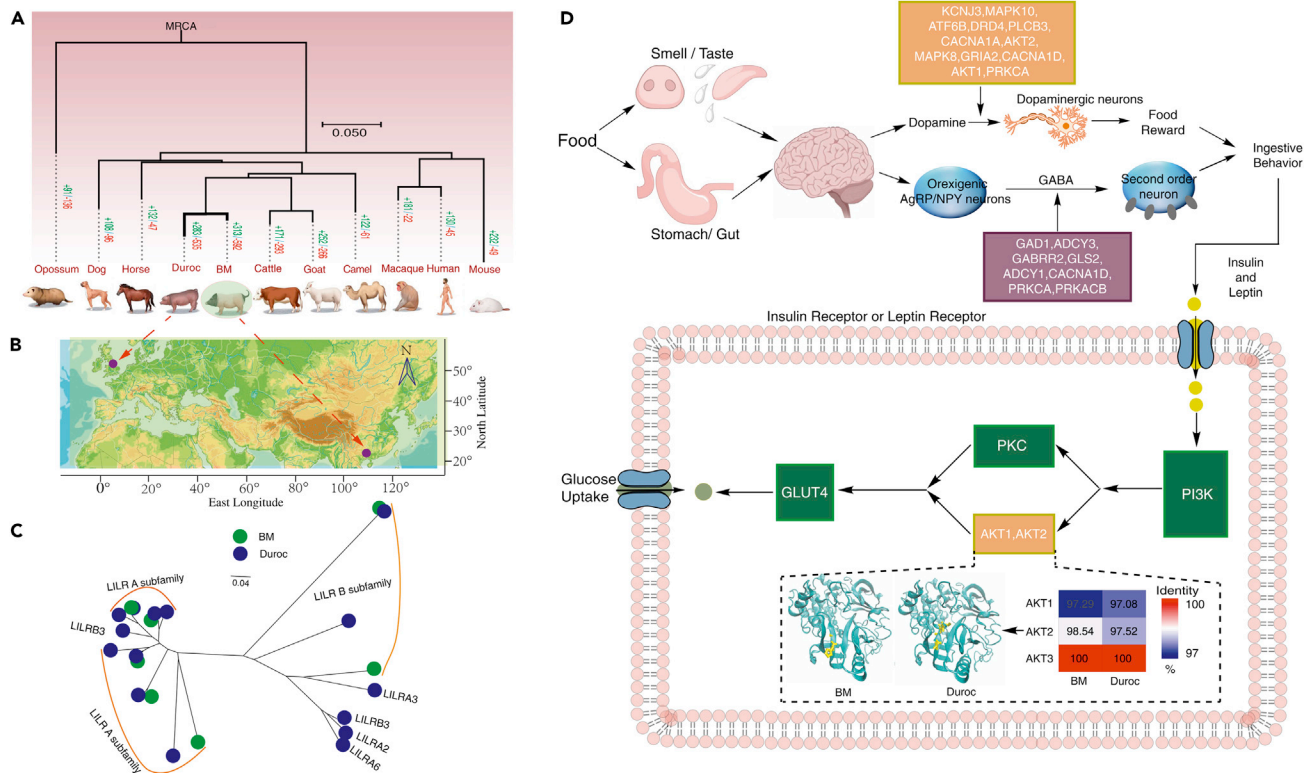


Figure 5. Comparative and Evolutionary Genomic Analysis and Regulation in Terms of Difference of Resistance to Diabetogenic Environment between BM and Duroc

(A) Phylogenomic tree showing expansion (green, +) and contraction (red, -) of gene families in BM and Duroc, and other nine Eutherian species. MRCA, most recent common ancestor.

(B) Geographic origin of BMs and Durocs. The BMs originated from low latitudes but the Durocs from high latitudes.

(C) An unrooted neighbor-joining tree constructed using *LILR* family genes identified in BM and Duroc.

(D) Control from CNS to intracellular glucose balance. When mammals eat, the CNS responds to the hormonal and sensory signals produced by the food to regulate ingestive behaviors (upper part). The box in purple indicates the PSGs of BM, and the other box in light yellow indicates the PSGs of Duroc involved in this process. After food intake, the PI3K-AKT pathway plays an important role in the transport of glucose from blood to inside the cell to regulate glucose homeostasis (lower part). The light yellow box indicates the *AKT1* and *AKT2* PSGs of Duroc, and the Duroc-specific amino acid substitutions have resulted in variation in the protein structure of *AKT2* (the dashed box) (Figure S18). Specifically, compared with the BM and human proteins, substitutions in amino acids 13 and 14 in *AKT2* of Duroc have affected the protein's β -fold (shown in left bottom). In contrast to the absence of amino acids 15 and 16 in *AKT2* of BM and human, the α -helix structure has disappeared from the structure of *AKT2* of Duroc (shown in left bottom). Colors in the heatmap (the dashed box) represent the degree of sequence identity of the AKT proteins at the amino acid level between BMs/Durocs and humans (shown in right bottom).

belonging to the *LILR* gene family, which is associated with bone development, can result in pycnodysostosis characterized by osteosclerosis, short stature, clavicular dysplasia, and skull deformities in humans (Song et al., 2017) (Figure 5C). This may explain why BM has a relatively short body length with only 19–20 thoracic and lumbar vertebrae, less than that in commercial pigs (21–23 thoracic and lumbar vertebrae in Duroc) and low body height with a dramatically shorter fibula than that in Duroc (Figure S15; Table S1).

Bidirectional Selection in BM and Duroc

To look for rapidly evolving genes that underlie different adaptive traits between BM and Duroc under divergent selective conditions, we identified 789 and 990 PSGs in BM and Duroc by estimating ω values (nonsynonymous/synonymous rate ratio [Ka/Ks]), respectively (Tables S22 and S23). Genes related to organ development and morphology in Duroc appear to have undergone rapid evolution (correct $p < 0.05$; Figures S16 and S17; Table S23). Many PSGs of Duroc were significantly enriched in the categories “focal adhesion” (17 PSGs), “Hippo signaling pathway-fly” (5 PSGs), and “extracellular matrix-receptor interaction” (8 PSGs), all of which play important roles in tissue and organ morphogenesis (Hynes, 2009) (Figure S16; Table S23). This information coincides with the findings that BMs exhibit an organ weight that is more

comparable with that in humans than that in commercial pigs (Table S24), supporting that BMs provide better donors, including liver (Shah et al., 2016), heart (Mohiuddin et al., 2016), kidney (Higginbotham et al., 2015), spleen (van der Windt et al., 2009), and lung (Kubicki et al., 2015), for xenotransplantation.

Notably, although both these PSGs in BM and Duroc are over-represented in the candidate gene set related to energy homeostasis processes ("PI3K-Akt signaling pathway" and "glycerolipid metabolism" KEGG pathways shared in both BM and Duroc [correct $p < 0.05$]; "AGE-RAGE signaling pathway in diabetic complications" KEGG pathway unique to Duroc [correct $p < 0.05$]), the specific gene contents within these sets differ drastically between the two pigs (Figure S16; Tables S22 and S23). Rapidly evolving energy metabolism genes in BM are involved mainly in adipose deposition (such as *AGPAT2*, *AGPAT4*, *AWAT1*, and *FABP6*) and the growth and development of cardiac and skeletal muscles (such as *FGFR2*, *FGFR4*, and *IGFBP4*), reflecting a need to enhance the efficiency of biomass production under the nutrient-restricted feeding condition. Conversely, energy metabolism PSGs in Duroc are involved mainly in diabetic diseases, including eight PSGs playing important roles in resistance to diabetes (*GAPT4*, *ZNF608*, and *BBS2*) (Nishimura et al., 2001; Speliotes et al., 2010), atherosclerosis (*SERPINE1*) (Koch et al., 2010), insulin secretion (*UCP2*) (Bordone et al., 2006), and energy expenditure (*GPAM*, *PRKCA*, and *NCOA3*) (Han et al., 2017; Yu et al., 2017).

We additionally found different PSGs potentially involved in the neurocircuitry control of peripheral metabolism between BM and Duroc. Unlike the enrichment of the "dopaminergic synapse (12 PSGs)" KEGG category for Duroc PSGs (correct $p < 0.05$), the PSGs of BM were significantly enriched in KEGG pathway related to the functional regulation of the central nervous system (CNS), "GABAergic synapse (8 PSGs)" (correct $p < 0.05$) (Figures 5D and S16; Table S22). BMs, in contrast to commercial pigs feeding *ad libitum*, have undergone long-term limited feeding (over their entire breeding history of more than 30 years) without any flavor supplementation. As a result, BMs, similar to humans under restrained eating condition (Laessle et al., 1989), consumed ~2.46-fold energy beyond their need-based requirements when they were implemented free food intake (Table S1), which was supported by a large number of distinct PSGs involved in synapse in the CNS between two breeds. Duroc PSGs are enriched in "dopaminergic synapse" affecting "food reward" mechanisms (overconsumption of rewarding palatable foods, often in quantities exceeding energetic needs) (Clemmensen et al., 2017). The BM PSGs over-represented in "GABAergic synapse," which release an inhibitory projection (GABA) onto MC4R (Kleinriders et al., 2009; Myers and Olson, 2012), are involved in promoting food intake and hyperphagia, even when the basic energy needs are met. These two genetic changes in the CNS may interact to affect the eating behavior of BMs different from Durocs by increasing food intake and reducing sensitivity to dietary excess (Figure 5D). Thus, with free access to enough nutrients despite less appealing food, BMs are still willing to consume excess nutrients beyond their basic physiological needs due to decreased satiation (Table S1). After food digestion in the stomach or gut, cellular signaling pathways involved in energy metabolism, such as PI3k-AKT pathway, play a key role in the peripheral glucose homeostatic loop. Insulin receptor and insulin-like growth factor receptor 1 promote the phosphorylation of receptor tyrosine residues (pY), leading to the recruitment and phosphorylation of the insulin receptor substrate (IRS). These recruit PI3K, which activates AKT by targeting its pleckstrin homology domain indirectly to control glucose transporter 4 (GLUT4) translocation to the plasma membrane and thus cellular uptake of glucose (Hribal et al., 2002; Kleinriders et al., 2009; Manning and Toker, 2017; Myers and Olson, 2012). In contrast to BMs, Durocs have undergone positive selection for *AKT1* and *AKT2* (Figures 5D and S18); especially the *AKT2*-encoded protein has multiple alterations in pleckstrin homology domain, which likely affects cellular uptake of glucose and regulation of blood glucose levels. The differential positive selection of genes involved in the CNS and cellular energy metabolism may be responsible for the lower resistance to "diabetogenic" environment in BM compared with Duroc.

Effects of Inbreeding and Directional Selection on BM

Inbreeding of the BM from BX population was also accompanied by the selection of some features (e.g., smaller body size, tamer behavior) favorable for use as an animal model (Figure 1). To detect genomic footprints left by selection, we conducted population resequencing analyses to measure genome-wide variants between BMs and BXs and found over 16 million single nucleotide polymorphisms (SNPs) in these two populations (9,169,662 in BXs versus 7,051,076 in BMs).

Among all SNPs detected in the BMs and BXs, we found 6,040,262 (58.9%) unique heterozygous variations in BXs, six-fold more than the number in BMs (928,628 or 9.1%), and the number of fixed homozygous variations was significantly smaller in BXs (167,992, 1.6%) than in BMs (3,161,040, 30.9%) (Figure 6A), confirming that long-term inbreeding has dramatically decreased the degree of heterozygosity in BMs. Chromosome-wide

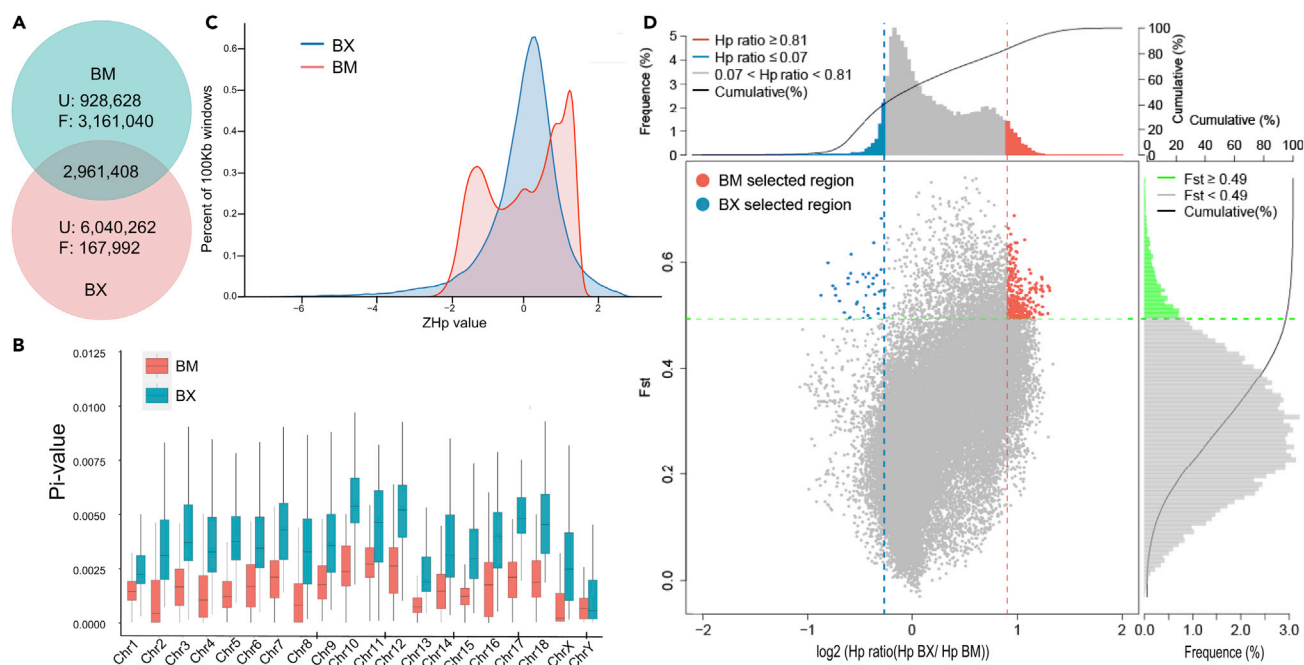


Figure 6. Single Nucleotide Divergence and Genomic Regions with Strong Selective Sweep Signals in BMs and BXs Revealed by Population Resequencing

(A) Classification of single nucleotide variation between BMs and BXs. The ~17 million single nucleotide differences between the two Bama pig lines were classified into three subclasses. The overlapping regions represent heterozygous variations shared between BMs and BXs. U, unique heterozygous variations evident in each species; F, the number of fixed homozygous variations in each species.

(B) Population diversity (P_i) values between BMs ($n = 50$) and BXs ($n = 50$). Data are represented by box-and-whisker plots. Boxes represent the interquartile range between the first and third quartiles and median (internal line). Whiskers denote the lowest and highest values within 1.5 times the range of the first and third quartiles, respectively.

(C) Z score of heterozygosity (ZHp) patterns in BMs and BXs within 100-kb windows across the genome.

(D) Distribution of Hp ratios and F_{ST} values. Data points located to the left and right of the left and right vertical dashed lines, respectively, and above the horizontal dashed line were identified as selected sweep regions for BM and BXs, respectively.

comparison of nucleotide diversity (P_i , π) for fixed and unique variants revealed fewer variants in BMs than in BXs on all autosomes and X chromosome (Figure 6B). These indicate high genetic similarity or homology among individual BMs and confirmed that the BM inbred line has already become genetically stable, which was also further verified by genetic structural analyses of microsatellite data (Figures S19 and S20; Table S25).

Besides, a genome-wide screen to determine the Z score of heterozygosity (Hp) of BMs showed a skewed distribution and low average value (0.365), in contrast to the normal distribution and higher average value (0.414) in BXs (Figure 6C). This suggests that BM is already a non-natural population after the 30-year inbreeding period, which is coincident to the differences of some biological characters of BMs from BXs. To reveal the genomic basis underlying these biological differences, we further calculated the fixation index (F_{ST}), a measure of genetic differentiation, and identified convincing selective sweep regions using significant high F_{ST} values and low/high Hp ratios as cutoff values in sliding windows of 100 kb with a 50-kb step size along the entire genome, in BMs and BXs (Figure 6D). Even under such a stringent criterion, we found eight times as many genomic regions with strong selective sweep signals in BMs (271) as in BXs (41), quantitatively suggesting that the selection in BMs is more powerful than that in BXs in shaping the genome, resulting in rapid changes in the phenotypic or behavioral traits of BMs (Figure 6D).

We subsequently extracted the annotated protein-coding genes from these regions (282 genes in BMs and 36 genes in BXs) to examine the precise correlation between selection and the altered physiological traits of BMs relative to BXs (Table S26). One of the greatest challenges during the over 30-year period of BM breeding was overcoming inbreeding bottleneck (Charlesworth and Willis, 2009). Our inbreeding practices resulted in increased genetic homozygosity, and individuals carrying homologous deleterious recessive alleles are at risk of reduced survival and fertility, the severity of which intensifies with each generation (only a few individuals

from one of the four pair of original parents were bred to the 19th generation; Figure 1). Our population-level analysis revealed clearly selective footprints on inbreeding depression for six autosomal recessive disease-related genes involved in autosomal recessive cone dystrophy (*CACNA2D4*) (Wycisk et al., 2006), autosomal-dominant familial Meniere disease (*DTNA*) (Requena et al., 2015), early coronary disease and metabolic risk (*LRP6*) (Mani et al., 2007), autosomal recessive autism spectrum disease (*MYO1A*) (Talebi et al., 2018), limb defects associated with epicanthus inversus syndrome and Möbius syndrome (*SOX14*) (Hargrave et al., 2000), and autosomal dominant retinitis pigmentosa (*SPP2*) (Liu et al., 2015). Thus, selective elimination of recessive deleterious mutations within these genes concurs with our inbreeding practice of weeding out the individuals that lack fitness traits, allowing the healthiest individuals to continue breeding.

Selection of individuals with a small body size has been a major focus during BM inbreeding. We found four genes related to bone development, showing evidence of a strong signature of selective sweeps. *ATP6V1H* regulates the growth and differentiation of bone marrow stromal cells (Zhang et al., 2017). *CHMP5* controls bone turnover rates by decreasing nuclear factor- κ B activity in osteoclasts (Greenblatt et al., 2015). *GPR55* is a putative cannabinoid receptor that regulates osteoclast function and bone mass (Whyte et al., 2009). *UHMK1*, as a bone mineral density-related protein, regulates osteoblasts and osteoclasts (Choi et al., 2016). These genes encoding factors associated with bone development may explain the smaller body size of the BM compared with the BX (Table S1).

A visible trait of BM is its uniform two-end-black fur color. Its coat color differs from that of the BX, in that it lacks the black spots of different sizes on the shoulders, back, and waist characteristic of BXs (Table S1). The *SNAI2* gene, deletion of which results in human piebaldism characterized by congenital patches of skin and hair from which melanocytes are completely absent (Sanchez-Martin et al., 2003), exhibits strong selective sweep signals. We infer this is why the black spots disappeared from BM skin (Table S1).

Another key artificially selected trait in BM inbreeding was tamer behavior for adaptation to a captive environment, as opposed to the anxiety-associated aggressive behavior of the primitive population. Four BM genes predominantly involved in anxious behaviors in humans or animal models were identified in the regions with strong selective sweep. *GPR55* receptor agonists and antagonists modulate anxiety-related behaviors in rats (Rahimi et al., 2015). *ASIC1* is highly expressed in patients with a panic disorder characterized by unexpected, recurrent panic attacks, associated with a fear of dying and worry about possible future attacks or other behavioral changes as a consequence of the attacks (Gugliandolo et al., 2016). *RXRG* is associated with mouse anxiety and human bipolar disorders (Alliey-Rodriguez et al., 2011; Ashbrook et al., 2015). *CRY1* directly influences cognitive function and anxiety-related behaviors (De Bundel et al., 2013). In short, BMs have broken the inbred bottleneck and some phenotypes of them have changed, which are consistent with the altered genomic regions resulting from long-term intense selection.

Difference in Resistance to Diabetogenic Environment between BM and Duroc

It is speculated that the divergent feeding conditions contribute to different evolution of energy metabolism system between BMs and commercial pigs, so as to weaken diabetogenic pressure endurance of BMs relative to Durocs (Figures 2 and S2; Tables S2 and S3). To investigate the molecular mechanism underlying difference of tolerance to diabetogenic pressures between these two breeds, we first conducted selective sweep analysis based on F_{ST} values and H_p ratios to dissect selection in BM and Duroc for adaptation to the divergent feeding conditions (Figure S21). Population-level analysis revealed that the energy metabolism systems of BM and Duroc were under distinct selections. The BM energy metabolic genes embedded in selected regions belong mainly to categories that are related to energy deposition (Gene Ontology [GO] term: "lipid storage," "positive regulation of lipid storage," "regulation of lipid storage," $p < 0.05$) and diabetic disease (KEGG pathway: "Maturity onset diabetes of the young," correct $p < 0.05$) (Figures S22 and S23). Conversely, we identified that five Duroc lipid-related genes were overrepresented in "response to lipid" and "cellular response to lipid" GO term ($p < 0.05$) (Figure S23). These distinct selective sweep events related to energy metabolism are coincident with feeding difference between BM and Duroc, which potentially facilitates the BM's predisposition of diabetes.

Currently, long noncoding RNAs (lncRNAs), their target genes, and diabetes have drawn increasing attention among researchers (Knoll et al., 2015). As an important post-transcriptional pathogenesis of diabetes, lncRNAs and their associated orchestrated networks are implicated in mediating complex pathological mechanisms of diabetes (Kato et al., 2016; Liu et al., 2014). To delineate the influence of lncRNAs and

mRNAs on the different resistance between BM and Duroc to diabetogenic environment, we next generated transcriptome sequencing by using two key glycogen-metabolizing tissues—liver and skeletal muscle of BMs (BM-induced group) and Durocs (Duroc-induced group) in “diabetogenic” environment (Figures 2 and S2; Tables S2 and S3), respectively. After RNA sequencing (RNA-seq) and a series of lncRNA identification steps, 5,186 and 6,552 lncRNAs were identified in liver and skeletal muscle, respectively, and subjected to subsequent analyses (Figure S24). As is characteristic of lncRNAs, our identified lncRNAs had fewer exons, a shorter average length, and lower expression levels compared with mRNAs (Figure S25). We found a total of 258 (20 upregulated and 238 downregulated) and 227 (26 upregulated and 201 downregulated) lncRNA transcripts and 477 (279 upregulated and 198 downregulated) and 277 (140 upregulated and 137 downregulated) mRNA transcripts differentially expressed in liver and skeletal muscle, respectively, between the BM-induced and Duroc-induced groups (Figure S26).

GO analysis of the potential targets of differentially expressed lncRNAs revealed diabetes-associated terms among the top five significantly enriched terms ($p < 0.01$) (Tables S27 and S28), such as “insulin-like growth factor binding” (GO: 0005520) in liver and “glycerol-3-phosphate and alditol phosphate metabolic process” (GO: 0006072 and GO: 0052646) in skeletal muscle. We also found a total of nine and eight significantly enriched KEGG pathways among the potential targets of the differentially expressed lncRNAs in liver and skeletal muscle, respectively ($p < 0.05$) (Tables S29 and S30), of which the AMPK and PI3K-Akt signaling pathways were related to diabetes. Next, GO and KEGG analyses of the significantly dysregulated mRNAs in liver and skeletal muscle also indicated significant enrichment of transcripts related to the AMPK signaling pathway (Tables S31–S34).

Furthermore, we selectively analyzed the lncRNAs and their target genes that (1) were both significantly differentially expressed between BM-induced and Duroc-induced groups and (2) should be associated with diabetes based on functional enrichment. We detected a total of six pairs fulfilling these criteria involving six lncRNAs and two target genes. In the liver, we found that XLOC_006422 and XLOC_026958 were correlated (*trans*-acting) with *PGC-1 α* and XLOC_044402 with *PEPCK*. In skeletal muscle, XLOC_003015 and XLOC_023027 were correlated with *PEPCK* through *trans* activity, and XLOC_026564 was correlated with *PGC-1 α* . Among these, the target genes *PGC-1 α* and *PEPCK* have been directly linked to T2DM previously (Samuel et al., 2009; Sawada et al., 2014; Soccio et al., 2015), and their expression was significantly higher in BM-induced liver and skeletal muscle than in Duroc-induced tissues ($p < 0.05$; Figure S26). Furthermore, KEGG pathway analysis showed that *PGC-1 α* and *PEPCK* were both enriched in the “AMPK signaling pathway” ($p < 0.05$) related to energy metabolism. The selected lncRNAs and target gene expression in RNA-seq were validated by qRT-PCR analysis (Figure S27; Table S35). Given these results, we suspect that these lncRNAs probably participate in the regulation of resistance to “diabetogenic” environment by influencing the expression of diabetes-related genes such as *PGC-1 α* and *PEPCK*, although the underlying mechanisms require additional investigation.

DISCUSSION

The establishment of BM line with high inbreeding level is an important action, which can strengthen the laboratory pig-resource gene pool and enrich the pig diversity, as well as directly compensate the disadvantages of using pig in diabetes study. The chromosome-level reference genome sequence of BM reported here is a high-quality miniature pig genome that offers the needed information to expedite current efforts in developing BM as an ideal experimental animal to reveal the genetic basis of diabetes. Comparing the genome sequence of BM with that of commercial Duroc provided insights into the distinct evolutionary scenarios, especially in the CNS and cellular energy metabolism, which leads to the different resistance to diabetogenic pressure between them, occurring under inbred selection for experimental use and artificial selection for commercialization. Besides, genomic comparison of these two pig breeds also identified many genetic loci related to body size and sexual maturation, which may not only promote the miniaturization and prematurity of experimental animals but also provide possible selection markers for the breeding of commercial pigs. The whole-genome resequencing analysis between BX and BM revealed genomic loci that have been under selection during BM inbreeding, maximizing the scientific value of these BM populations as a reference for improving the inbreeding practices of other inbred strains. The dissection of different resistance to diabetogenic environment between BM and Duroc by selective sweep, transcriptome, and comparative genome analyses help to improve current understanding of different diabetes susceptibility in humans. In conclusion, the data in this study provide a valuable resource and tool for functional genomic studies on BM as well as increases use of the BM as an animal model in broader field, particularly human diabetes.

Limitations of the Study

In this study, we developed a laboratory miniature pig inbred line with advantages for translational medicine, and presented the chromosome-scale BM genome. Although the genomic analyses revealed that BMs strongly resemble human beings, it also demonstrated that attention should be paid to the interspecific differences when selecting BMs for use as human disease models. Besides, functional research should be performed to further validate the candidate genes involved in susceptibility to diabetes.

METHODS

All methods can be found in the accompanying [Transparent Methods supplemental file](#).

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at <https://doi.org/10.1016/j.isci.2019.07.025>.

ACKNOWLEDGMENTS

We thank Professor. Wen Wang (Chinese Academy of Sciences), Dr. Wensheng lin (University of Minnesota), and Dr. Anna V. Kukekova (University of Illinois at Urbana) for critical review of the manuscript, Mr. Yajie Wang for the design of all figures. This work was supported by grants from the National Natural Science Foundation of China (81860150), Special Project on Innovation Driven Development of Guangxi (Guike-AA17204029), Science and Technology Funds of the Chairman of the Autonomous Region (16449-10), National Modern Agricultural Industrial Technology System (nycytxgxcxd-15-01), Science and Technology Major Special Project of Guangxi (Guike-AA17292002), and Guangxi Natural Science Foundation Program (2017GXNSFBA198157 and 2018GXNSFAA294038).

AUTHOR CONTRIBUTIONS

L.Z., Y.H., C.S., J.L., and G.L. conceived the study and designed major scientific objectives. L.Z., J.L., and G.L. coordinated the whole project. L.Z., Y.H., Y.W., J.S., J.L., X.Y., Y.G., and G.L. participated in the inbred strain establishment. Y.W., J.S., S.Z., and W.Q. prepared materials for genome and transcriptome sequencing. L.Z., Y.H., and M.W. conducted genome and transcriptome analysis. L.Z., M.W., and Q.Z. finished charts. Y.H., Y.W., J.S., S.Z., and W.Q. participated in PCR and microsatellite analysis. L.Z., Y.H., M.W., J. L., and C.S. did most of the writing. L.Z., Y.H., C.S., S.W., Z.L., Z.T., L.W., K.L., R.L., J.-F.F., and G.L. did the writing as well as review and editing. J.L., and G.L. conducted funding acquisition.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 18, 2019

Revised: June 11, 2019

Accepted: July 13, 2019

Published: September 27, 2019

REFERENCES

- Ai, H.S., Fang, X.D., Yang, B., Huang, Z.Y., Chen, H., Mao, L.K., Zhang, F., Zhang, L., Cui, L.L., He, W.M., et al. (2015). Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nat. Genet.* **47**, 217–225.
- Alliey-Rodriguez, N., Zhang, D.D., Badner, J.A., Lahey, B.B., Zhang, X.T., Dinwiddie, S., Romanos, B., Pleny, N., Liu, C.Y., and Gershon, E.S. (2011). Genome-wide association study of personality traits in bipolar patients. *Psychiatr. Genet.* **21**, 190–194.
- Andersson, L. (2016). Domestic animals as models for biomedical research. *Ups. J. Med. Sci.* **121**, 1–11.
- Ashbrook, D.G., Williams, R.W., Lu, L., and Hager, R. (2015). A cross-species genetic analysis identifies candidate genes for mouse anxiety and human bipolar disorder. *Front Behav. Neurosci.* **9**, 171.
- Bains, R.S., Cater, H.L., Sillito, R.R., Chartsias, A., Sneddon, D., Concas, D., Keskkivali-Bond, P., Lukins, T.C., Wells, S., Arozena, A.A., et al. (2016). Analysis of individual mouse activity in group housed animals of different inbred strains using a novel automated home cage analysis system. *Front. Behav. Neurosci.* **10**, 106.
- Bordone, L., Motta, M.C., Picard, F., Robinson, A., Jhala, U.S., Apfeld, J., McDonagh, T., Lemieux, M., McBurney, M., Szilvasi, A., et al. (2006). Sirt1 regulates insulin secretion by repressing UCP2 in pancreatic beta cells. *PLoS Biol.* **4**, e31.
- Charlesworth, D., and Willis, J.H. (2009). The genetics of inbreeding depression. *Nat. Rev. Genet.* **10**, 783–796.
- Chieppa, M.N., Perota, A., Corona, C., Grindatto, A., Lagutina, I., Vallino Costassa, E., Lazzari, G., Colleoni, S., Duchi, R., Lucchini, F., et al. (2014). Modeling amyotrophic lateral sclerosis in hSOD1 transgenic swine. *Neurodegener. Dis.* **13**, 246–254.
- Choi, H.J., Park, H., Zhang, L., Kim, J.H., Kim, Y.A., Yang, J.Y., Pei, Y.F., Tian, Q., Shen, H., Hwang, J.Y., et al. (2016). Genome-wide association study in East Asians suggests UHMK1 as a novel bone mineral density susceptibility gene. *Bone* **91**, 113–121.

- Clemmensen, C., Muller, T.D., Woods, S.C., Berthoud, H.R., Seeley, R.J., and Tschöp, M.H. (2017). Gut-brain cross-talk in metabolic control. *Cell* 168, 758–774.
- Crawley, J.N., Belknap, J.K., Collins, A., Crabbe, J.C., Frankel, W., Henderson, N., Hitzemann, R.J., Maxson, S.C., Miner, L.L., Silva, A.J., et al. (1997). Behavioral phenotypes of inbred mouse strains: implications and recommendations for molecular studies. *Psychopharmacology* 132, 107–124.
- Curtasu, M.V., Knudsen, K.E.B., Callesen, H., Purup, S., Stagsted, J., and Hedemann, M.S. (2019). Obesity development in a miniature yucatan pig model: a multi-compartmental metabolomics study on cloned and normal pigs fed restricted or Ad libitum high-energy diets. *J. Proteome Res.* 18, 30–47.
- Danielsen, E.H., Cumming, P., Andersen, F., Bender, D., Brevig, T., Falborg, L., Gee, A., Gillings, N.M., Hansen, S.B., Hermansen, F., et al. (2000). The DaNeX study of embryonic mesencephalic, dopaminergic tissue grafted to a minipig model of Parkinson's disease: preliminary findings of effect of MPTP poisoning on striatal dopaminergic markers. *Cell Transplant.* 9, 247–259.
- De Bundel, D., Gangarossa, G., Biever, A., Bonnefont, X., and Valjent, E. (2013). Cognitive dysfunction, elevated anxiety, and reduced cocaine response in circadian clock-deficient cryptochrome knockout mice. *Front. Behav. Neurosci.* 7, 152.
- Dix, D.J., Allen, J.W., Collins, B.W., Mori, C., Nakamura, N., PoormanAllen, P., Goulding, E.H., and Eddy, E.M. (1996). Targeted gene disruption of Hsp70-2 results in failed meiosis, germ cell apoptosis, and male infertility. *Proc. Natl. Acad. Sci. U S A* 93, 3264–3268.
- Dix, D.J., Allen, J.W., Collins, B.W., PoormanAllen, P., Mori, C., Blizard, D.R., Brown, P.R., Goulding, E.H., Strong, B.D., and Eddy, E.M. (1997). HSP70-2 is required for desynapsis of synaptonemal complexes during meiotic prophase in juvenile and adult mouse spermatocytes. *Development* 124, 4595–4603.
- Dostalova, Z., Calvete, J.J., Sanz, L., and Topfer-Petersen, E. (1994). Quantitation of boar spermadhesins in accessory sex gland fluids and on the surface of epididymal, ejaculated and capacitated spermatozoa. *Biochim. Biophys. Acta* 1200, 48–54.
- Fang, X.D., Mu, Y.L., Huang, Z.Y., Li, Y., Han, L.J., Zhang, Y.F., Feng, Y., Chen, Y.X., Jiang, X.T., Zhao, W., et al. (2012). The sequence and analysis of a Chinese pig genome. *Gigascience* 1, 16.
- Forster, J., Hirst, A.G., and Atkinson, D. (2012). Warming-induced reductions in body size are greater in aquatic than terrestrial species. *Proc. Natl. Acad. Sci. U S A* 109, 19310–19314.
- Frantz, L.A.F., Schraiber, J.G., Madsen, O., Megens, H.J., Cagan, A., Bosse, M., Paudel, Y., Crooijmans, R.P.M.A., Larson, G., and Groenen, M.A.M. (2015). Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. *Nat. Genet.* 47, 1141–1148.
- Garcia, E.M., Vazquez, J.M., Calvete, J.J., Sanz, L., Caballero, I., Parrilla, I., Gil, M.A., Roca, J., and Martinez, E.A. (2006). Dissecting the protective effect of the seminal plasma spermadhesin PSP-I/PSP-II on boar sperm functionality. *J. Androl.* 27, 434–443.
- Gerstein, H.C., and Waltman, L. (2006). Why don't pigs get diabetes? Explanations for variations in diabetes susceptibility in human populations living in a diabetogenic environment. *CMAJ* 174, 25–26.
- Greenblatt, M.B., Park, K.H., Oh, H., Kim, J.M., Shin, D.Y., Lee, J.M., Lee, J.W., Singh, A., Lee, K.Y., Hu, D., et al. (2015). CHMP5 controls bone turnover rates by dampening NF-kappa B activity in osteoclasts. *J. Exp. Med.* 212, 1283–1301.
- Groenen, M.A.M., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.J., et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491, 393–398.
- Gugliandolo, A., Gangemi, C., Caccamo, D., Curro, M., Pandolfo, G., Quattrone, D., Crucitti, M., Zoccali, R.A., Bruno, A., and Muscatello, M.R.A. (2016). The RS685012 polymorphism of ACCN2, the human ortholog of murine acid-sensing ion channel (ASIC1) gene, is highly represented in patients with panic disorder. *Neuromol. Med.* 18, 91–98.
- Han, H., Gu, S., Chu, W., Sun, W., Wei, W., Dang, X., Tian, Y., Liu, K., and Chen, J. (2017). miR-17-5p regulates differential expression of NCOA3 in pig intramuscular and subcutaneous adipose tissue. *Lipids* 52, 939–949.
- Hargrave, M., James, K., Nield, K., Toomes, C., Georgas, K., Sullivan, T., Verzijl, H.T., Oley, C.A., Little, M., De Jonghe, P., et al. (2000). Fine mapping of the neurally expressed gene SOX14 to human 3q23, relative to three congenital diseases. *Hum. Genet.* 106, 432–439.
- Higginbotham, L., Mathews, D., Breeden, C.A., Song, M., Farris, A.B., 3rd, Larsen, C.P., Ford, M.L., Lutz, A.J., Tector, M., Newell, K.A., et al. (2015). Pre-transplant antibody screening and anti-CD154 costimulation blockade promote long-term xenograft survival in a pig-to-primate kidney transplant model. *Xenotransplantation* 22, 221–230.
- Holm, I.E., Alstrup, A.K., and Luo, Y. (2016). Genetically modified pig models for neurodegenerative disorders. *J. Pathol.* 238, 267–287.
- Hribal, M.L., Oriente, F., and Accili, D. (2002). Mouse models of insulin resistance. *Am. J. Physiol. Endocrinol. Metab.* 282, E977–E981.
- Hynes, R.O. (2009). The extracellular matrix: not just pretty fibrils. *Science* 326, 1216–1219.
- Ibrahim, Z., Busch, J., Awwad, M., Wagner, R., Wells, K., and Cooper, D.K. (2006). Selected physiologic compatibilities and incompatibilities between human and porcine organ systems. *Xenotransplantation* 13, 488–499.
- Kato, M., Wang, M., Chen, Z., Bhatt, K., Oh, H.J., Lanting, L., Deshpande, S., Jia, Y., Lai, J.Y.C., O'Connor, C.L., et al. (2016). An endoplasmic reticulum stress-regulated lncRNA hosting a microRNA megacluster induces early features of diabetic nephropathy. *Nat. Commun.* 7, 12864.
- Kengia, J.T., Ko, K.C., Ikeda, S., Hiraishi, A., Mieno-Naka, M., Arai, T., Sato, N., Muramatsu, M., and Sawabe, M. (2013). A gene variant in the Atp10d gene associates with atherosclerotic indices in Japanese elderly population. *Atherosclerosis* 231, 158–162.
- Kim, S., Cho, Y.S., Kim, H.M., Chung, O., Kim, H., Jho, S., Seomun, H., Kim, J., Bang, W.Y., Kim, C., et al. (2016). Comparison of carnivore, omnivore, and herbivore mammalian genomes with a new leopard assembly. *Genome Biol.* 17, 211.
- Kleinert, M., Clemmensen, C., Hofmann, S.M., Moore, M.C., Renner, S., Woods, S.C., Huypens, P., Beckers, J., de Angelis, M.H., Schurmann, A., et al. (2018). Animal models of obesity and diabetes mellitus. *Nat. Rev. Endocrinol.* 14, 140–162.
- Kleinridders, A., Konner, A.C., and Bruning, J.C. (2009). CNS-targets in control of energy and glucose homeostasis. *Curr. Opin. Pharmacol.* 9, 794–804.
- Knoll, M., Lodish, H.F., and Sun, L. (2015). Long non-coding RNAs as regulators of the endocrine system. *Nat. Rev. Endocrinol.* 11, 151–160.
- Koch, W., Schrepf, M., Erl, A., Mueller, J.C., Hoppmann, P., Schomig, A., and Kastrati, A. (2010). 4G/5G polymorphism and haplotypes of SERPINE1 in atherosclerotic diseases of coronary arteries. *Thromb. Haemost.* 103, 1170–1180.
- Kubicki, N., Laird, C., Burdorf, L., Pierson, R.N., 3rd, and Azimzadeh, A.M. (2015). Current status of pig lung xenotransplantation. *Int. J. Surg.* 23, 247–254.
- Laessle, R.G., Tuschl, R.J., Kotthaus, B.C., and Pirke, K.M. (1989). Behavioral and biological correlates of dietary restraint in normal life. *Appetite* 12, 83–94.
- Li, B.A., Nair, M., Mackay, D.R., Bilanchone, V., Hu, M., Fallahi, M., Song, H.Q., Dai, Q., Cohen, P.E., and Dai, X. (2005). Ovol1 regulates melotic pachytene progression during spermatogenesis by repressing Id2 expression. *Development* 132, 1463–1473.
- Li, M.Z., Chen, L., Tian, S.L., Lin, Y., Tang, Q.Z., Zhou, X.M., Li, D.Y., Yeung, C.K.L., Che, T.D., Jin, L., et al. (2017). Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple de novo assemblies. *Genome Res.* 27, 865–874.
- Lilue, J., Doran, A.G., Fiddes, I.T., Abrudan, M., Armstrong, J., Bennett, R., Chow, W., Collins, J., Collins, S., Czechanski, A., et al. (2018). Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nat. Genet.* 50, 1574–1583.
- Liu, J.Y., Yao, J., Li, X.M., Song, Y.C., Wang, X.Q., Li, Y.J., Yan, B., and Jiang, Q. (2014). Pathogenic role of lncRNA-MALAT1 in endothelial cell dysfunction in diabetes mellitus. *Cell Death Dis.* 5, e1506.
- Liu, Y., Chen, X., Xu, Q., Gao, X., Tam, P.O., Zhao, K., Zhang, X., Chen, L.J., Jia, W., Zhao, Q., et al. (2015). SPP2 mutations cause autosomal dominant Retinitis Pigmentosa. *Sci. Rep.* 5, 14867.
- Mani, A., Radhakrishnan, J., Wang, H., Mani, A., Mani, M.A., Nelson-Williams, C., Carew, K.S., Mane, S., Najmabadi, H., Wu, D., et al. (2007).

- LRP6 mutation in a family with early coronary disease and metabolic risk factors. *Science* 315, 1278–1282.
- Manning, B.D., and Toker, A. (2017). AKT/PKB signaling: navigating the network. *Cell* 169, 381–405.
- Mohiuddin, M.M., Singh, A.K., Corcoran, P.C., Thomas, M.L., Clark, T., Lewis, B.G., Hoyt, R.F., Eckhaus, M., Pierson, R.N., Belli, A.J., et al. (2016). Chimeric 2C10R4 anti-CD40 antibody therapy is critical for long-term survival of GTKO.hCD46.hTBM pig-to-primate cardiac xenograft. *Nat. Commun.* 7, 11138.
- Myers, M.G., Jr., and Olson, D.P. (2012). Central nervous system control of metabolism. *Nature* 491, 357–363.
- Natarajan, R., Gerrity, R.G., Gu, J.L., Lanting, L., Thomas, L., and Nadler, J.L. (2002). Role of 12-lipoxygenase and oxidant stress in hyperglycaemia-induced acceleration of atherosclerosis in a diabetic pig model. *Diabetologia* 45, 125–133.
- Nie, T., Hui, X.Y., Gao, X.F., Li, K., Lin, W.H., Xiang, X.L., Ding, M.X., Kuang, Y., Xu, A.M., Fei, J., et al. (2012). Adipose tissue deletion of Gpr116 impairs insulin sensitivity through modulation of adipose function. *FEBS Lett.* 586, 3618–3625.
- Nishimura, D.Y., Searby, C.C., Carmi, R., Elbedour, K., Van Maldergem, L., Fulton, A.B., Lam, B.L., Powell, B.R., Swiderski, R.E., Bugge, K.E., et al. (2001). Positional cloning of a novel gene on chromosome 16q causing Bardet-Biedl syndrome (BBS2). *Hum. Mol. Genet.* 10, 865–874.
- Niu, D., Wei, H.J., Lin, L., George, H., Wang, T., Lee, H., Zhao, H.Y., Wang, Y., Kan, Y.N., Shrock, E., et al. (2017). Inactivation of porcine endogenous retrovirus in pigs using CRISPR-Cas9. *Science* 357, 1303–1307.
- Nowoshilow, S., Schloissnig, S., Fei, J.F., Dahl, A., Pang, A.W.C., Pippel, M., Winkler, S., Hastie, A.R., Young, G., Roscito, J.G., et al. (2018). The axolotl genome and the evolution of key tissue formation regulators (vol 554, pg 50, 2018). *Nature* 559, E2.
- Oh, J.S., Han, C., and Cho, C. (2009). ADAM7 is associated with epididymosomes and integrated into sperm plasma membrane. *Mol. Cells* 28, 441–446.
- Okitsu, T., Kobayashi, N., Jun, H.S., Shin, S., Kim, S.J., Han, J., Kwon, H., Sakaguchi, M., Totsugawa, T., Kohara, M., et al. (2004). Transplantation of reversibly immortalized insulin-secreting human hepatocytes controls diabetes in pancreatectomized pigs. *Diabetes* 53, 105–112.
- Rahimi, A., Moghaddam, A.H., and Roobakhsh, A. (2015). Central administration of GPR55 receptor agonist and antagonist modulates anxiety-related behaviors in rats. *Fund Clin. Pharmacol.* 29, 185–190.
- Ramdas, M., Harel, C., Armoni, M., and Karnieli, E. (2015). AHNK KO mice are protected from diet-induced obesity but are glucose intolerant. *Horm. Metab. Res.* 47, 265–272.
- Requena, T., Cabrera, S., Martin-Sierra, C., Price, S.D., Lysakowski, A., and Lopez-Escamez, J.A. (2015). Identification of two novel mutations in FAM136A and DTNA genes in autosomal-dominant familial Meniere's disease. *Hum. Mol. Genet.* 24, 1119–1126.
- Rocha, D., and Plastow, G. (2006). Commercial pigs: an untapped resource for human obesity research? *Drug Discov. Today* 11, 475–477.
- Rubin, C.J., Megens, H.J., Barrio, A.M., Maqbool, K., Sayyab, S., Schwochow, D., Wang, C., Carlborg, O., Jern, P., Jorgensen, C.B., et al. (2012). Strong signatures of selection in the domestic pig genome. *Proc. Natl. Acad. Sci. U S A* 109, 19529–19536.
- Samuel, V.T., Beddow, S.A., Iwasaki, T., Zhang, X.M., Chu, X., Still, C.D., Gerhard, G.S., and Shulman, G.I. (2009). Fasting hyperglycemia is not associated with increased expression of PEPCCK or G6Pc in patients with Type 2 diabetes. *Proc. Natl. Acad. Sci. U S A* 106, 12121–12126.
- Sanchez-Martin, M.S., Perez-Losada, J., Rodriguez-Garcia, A., Gonzalez-Sanchez, B., Korf, B.R., Kuster, W., Moss, C., Spritz, R.A., and Sanchez-Garcia, I. (2003). Deletion of the SLUG (SNAI2) gene results in human piebaldism. *Am. J. Med. Genet. A* 122A, 125–132.
- Sanz, L., Calvete, J.J., Jonakova, V., and Topfer-Petersen, E. (1992). Boar spermadhesin AQN-1 and AWN are sperm-associated acrosin inhibitor acceptor proteins. *FEBS Lett.* 300, 63–66.
- Sawada, N., Jiang, A., Takizawa, F., Safdar, A., Manika, A., Tesmenitsky, Y., Kang, K.T., Bischoff, J., Kalwa, H., Sartoretto, J.L., et al. (2014). Endothelial PGC-1 alpha mediates vascular dysfunction in diabetes. *Cell Metab.* 19, 246–258.
- Schook, L., Beattie, C., Beever, J., Donovan, S., Jamison, R., Zuckermann, F., Niemi, S., Rothschild, M., Rutherford, M., and Smith, D. (2005). Swine in biomedical research: creating the building blocks of animal models. *Anim. Biotechnol.* 16, 183–190.
- Shah, J.A., Navarro-Alvarez, N., DeFazio, M., Rosales, I.A., Elias, N., Yeh, H., Colvin, R.B., Cosimi, A.B., Markmann, J.F., Hertl, M., et al. (2016). A bridge to somewhere: 25-day survival after pig-to-baboon liver xenotransplantation. *Ann. Surg.* 263, 1069–1071.
- Sheridan, J.A., and Bickford, D. (2011). Shrinking body size as an ecological response to climate change. *Nat. Clim. Chang.* 1, 401–406.
- Soccio, R.E., Chen, E.R., Rajapurkar, S.R., Safabakhsh, P., Marinis, J.M., Dispirito, J.R., Emmett, M.J., Briggs, E.R., Fang, B., Everett, L.J., et al. (2015). Genetic variation determines PPAR gamma function and anti-diabetic drug response in vivo. *Cell* 162, 33–44.
- Song, H.K., Sohn, Y.B., Choi, Y.J., Chung, Y.S., and Jang, J.H. (2017). A case report of pycnodysostosis with atypical femur fracture diagnosed by next-generation sequencing of candidate genes. *Medicine (Baltimore)* 96, e6367.
- Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Lango Allen, H., Lindgren, C.M., Luan, J., Magi, R., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* 42, 937–948.
- Suryavathi, V., Khattry, A., Gopal, K., Rani, D.S., Panneerdoss, S., Gupta, N.J., Chakravarty, B., Deenadayal, M., Singh, L., and Thangaraj, K. (2008). Novel variants in UBE2B gene and idiopathic male infertility. *J. Androl.* 29, 564–571.
- Talebi, F., Ghanbari Mardasi, F., Mohammadi Asl, J., Tizno, S., and Najafvand Zadeh, M. (2018). Identification of novel PTPRQ and MYO1A mutations in an Iranian pedigree with autosomal recessive hearing loss. *Cell J.* 20, 127–131.
- van der Windt, D.J., Bottino, R., Casu, A., Campanile, N., Smetanka, C., He, J., Murase, N., Hara, H., Ball, S., Loveland, B.E., et al. (2009). Long-term controlled normoglycemia in diabetic non-human primates after transplantation with hCD46 transgenic porcine islets. *Am. J. Transplant.* 9, 2716–2726.
- Whyte, L.S., Ryberg, E., Sims, N.A., Ridge, S.A., Mackie, K., Greasley, P.J., Ross, R.A., and Rogers, M.J. (2009). The putative cannabinoid receptor GPR55 affects osteoclast function in vitro and bone mass in vivo. *Proc. Natl. Acad. Sci. U S A* 106, 16511–16516.
- Wycisk, K.A., Zeitz, C., Feil, S., Wittmer, M., Forster, U., Neidhardt, J., Wissinger, B., Zrenner, E., Wilke, R., Kohl, S., et al. (2006). Mutation in the auxiliary calcium-channel subunit CACNA2D4 causes autosomal recessive cone dystrophy. *Am. J. Hum. Genet.* 79, 973–977.
- Yamada, S., Kawaguchi, H., Yamada, T., Guo, X., Matsuo, K., Hamada, T., Miura, N., Tasaki, T., and Tanimoto, A. (2017). Cholic acid enhances visceral adiposity, atherosclerosis and nonalcoholic fatty liver disease in microminipigs. *J. Atheroscler. Thromb.* 24, 1150–1166.
- Yan, S., Tu, Z.C., Liu, Z.M., Fan, N.N., Yang, H.M., Yang, S., Yang, W.L., Zhao, Y., Ouyang, Z., Lai, C.D., et al. (2018). A huntingtin knockin pig model recapitulates features of selective neurodegeneration in huntington's disease. *Cell* 173, 989–1002.e13.
- Yu, H., Zhao, Z., Yu, X., Li, J., Lu, C., and Yang, R. (2017). Bovine lipid metabolism related gene GPAM: molecular characterization, function identification, and association analysis with fat deposition traits. *Gene* 609, 9–18.
- Zhang, Y.H., Huang, H.G., Zhao, G.X., Yokoyama, T., Vega, H., Huang, Y., Sood, R., Bishop, K., Maduro, V., Accardi, J., et al. (2017). ATP6V1H deficiency impairs bone development through activation of MMP9 and MMP13. *PLoS Genet.* 13, e1006481.

ISCI, Volume 19

Supplemental Information

Development and Genome Sequencing of a Laboratory-Inbred Miniature Pig Facilitates Study of Human Diabetic Disease

Li Zhang, Yuemeng Huang, Meng Wang, Yafen Guo, Jing Liang, Xiurong Yang, Wenjing Qi, Yanjun Wu, Jinglei Si, Siran Zhu, Zhe Li, Ruiqiang Li, Chao Shi, Shuo Wang, Qunjie Zhang, Zhonglin Tang, Lixian Wang, Kui Li, Ji-Feng Fei, and Ganqiu Lan

Supplementary Figures



Figure S1. Appearance of Bama miniature pig (BM), Related to Figure 1.

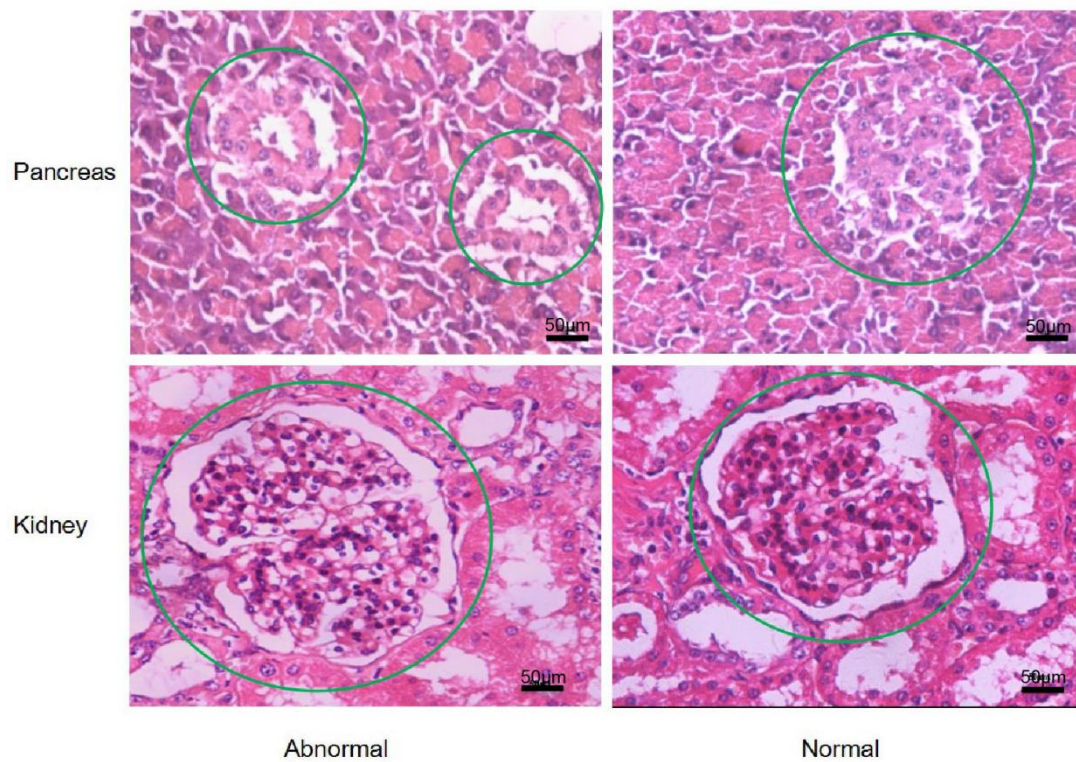


Figure S2. Pathological sections of pig pancreas and kidney, Related to Figure 2.

In the abnormal pancreas tissues, islets were atrophied with significantly decreased beta-cells, cell volume was increased and partial cells were vacuolar degeneration, compared with normal pancreas tissues in control group. In the abnormal kidney tissues, some glomerular size were enlarged with partial cell volume increased, and some cells were vacuolar degeneration, compared with normal kidney tissues in control group. For all panels, scale bars represent 50 μm.

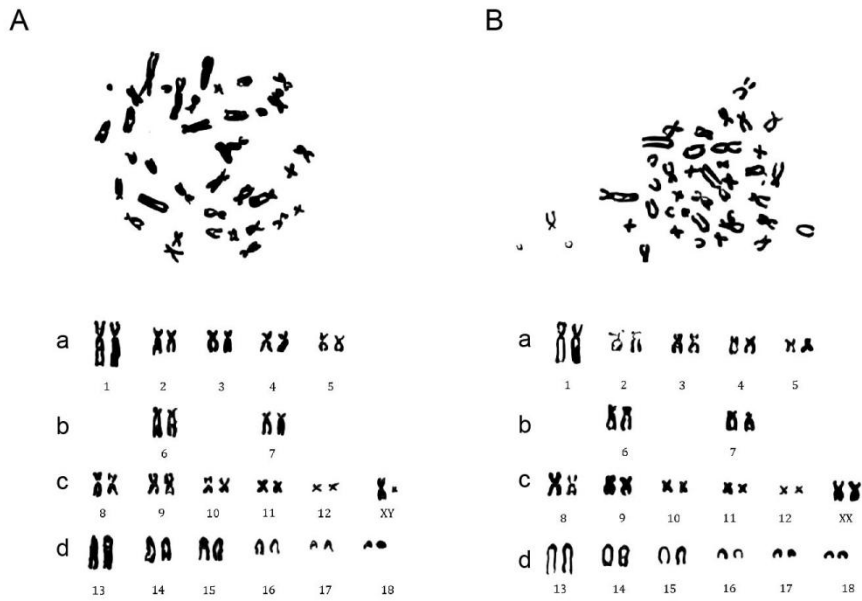


Figure S3. Chromosome karyotype of Bama miniature pig, Related to Figure 3.

(A) Male and (B) Female. The examination of karyotype of Bama miniature pig by means of peripheral blood lymphocytes culture showed that the diploid chromosomes number was 38, 18 pairs of autosomes and one pair of sex chromosomes in both males (XY) and females (XX). The chromosomes were divided into four groups of a, b, c and d according to the standard of Reading Congress. The karyotype of the autosomes was $10sm+4st+10m+12t$. The X chromosome was a metacentric chromosome whose length was between the 8th and 9th chromosome, while the Y chromosome was the smallest metacentric chromosome.

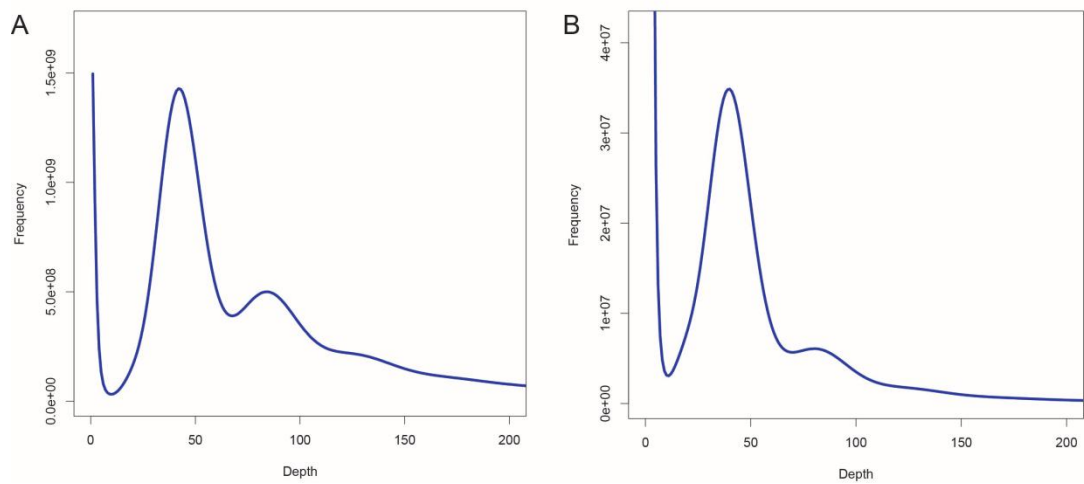


Figure S4. Distribution of 17-mer frequency, Related to Figure 3 and Table 1.

(A) 17-mer number frequency and (B) 17-mer type frequency. In total 120.46 Gb of high-quality short-insert reads (250 and 500 bp) were used to generate the 17-mer depth distribution curve frequency information.

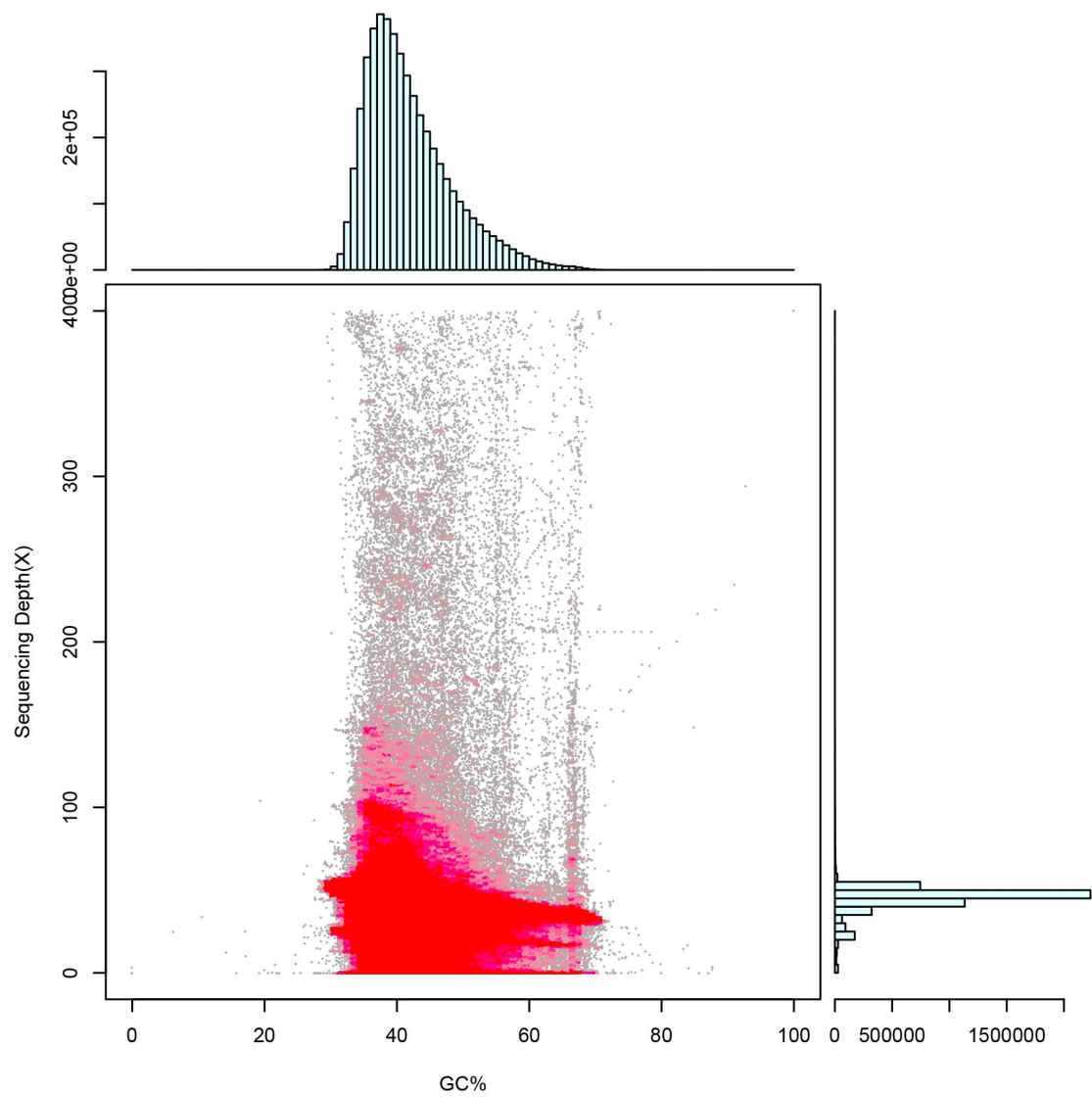


Figure S5. GC content against the sequencing depth of BM genome, Related to Figure 3 and Table 1.

We used 10 kb nonoverlapping sliding windows along the assembled sequence to calculate GC content and average sequencing depth using short reads. The x-axis represents the GC content, and the y-axis indicates the average sequencing depth.

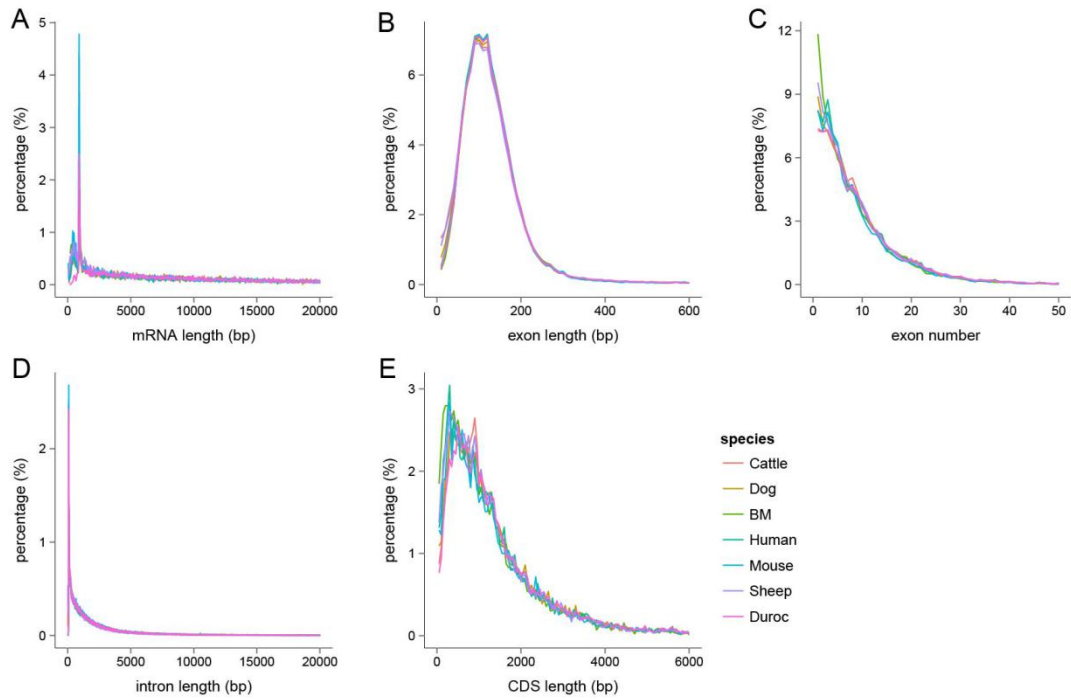


Figure S6. Comparison of gene parameters among the BM and eight other mammalian genomes, Related to Figure 3.

(A) mRNA length, (B) exon length, (C) exon number, (D) intron length, and (E) CDS length. The similar gene parameters between the BM and other mammals indicate the high quality gene structure annotation of the BM genome.

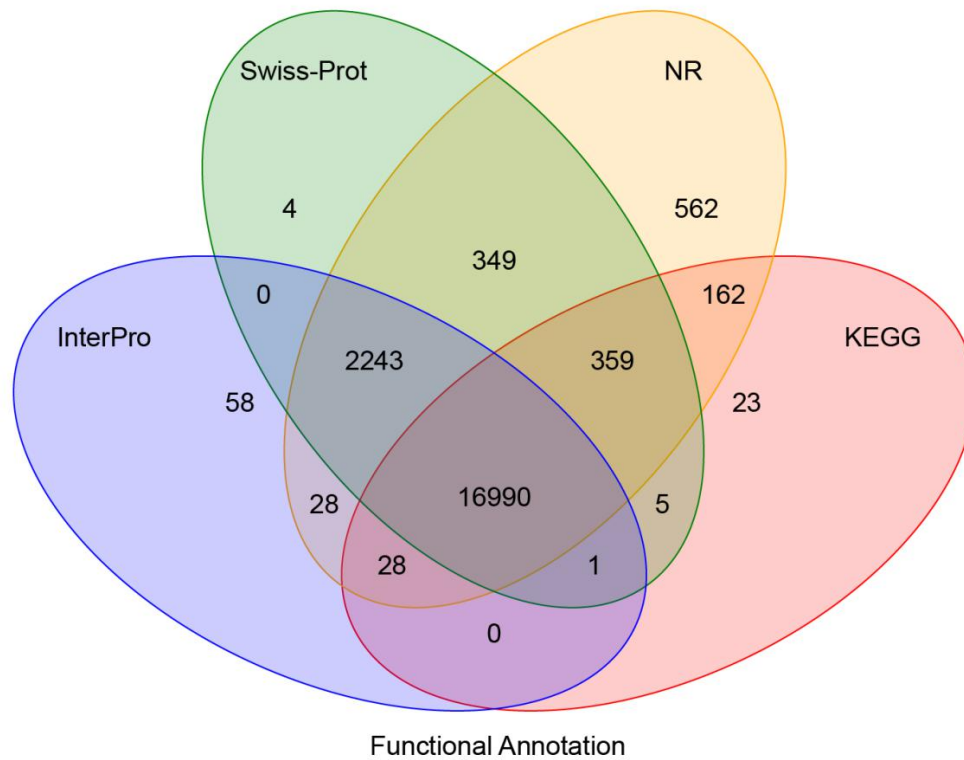


Figure S7. Number of functionally annotated genes of the BM genome using various methods, Related to Figure 3.

The alignment of gene set from annotation to protein databases including SwissProt (<http://www.uniprot.org/>) , NR (<https://ftp.ncbi.nlm.nih.gov/blast/db/>), KEGG (<http://www.genome.jp/kegg/>) and InterPro (<https://www.ebi.ac.uk/interpro/>).

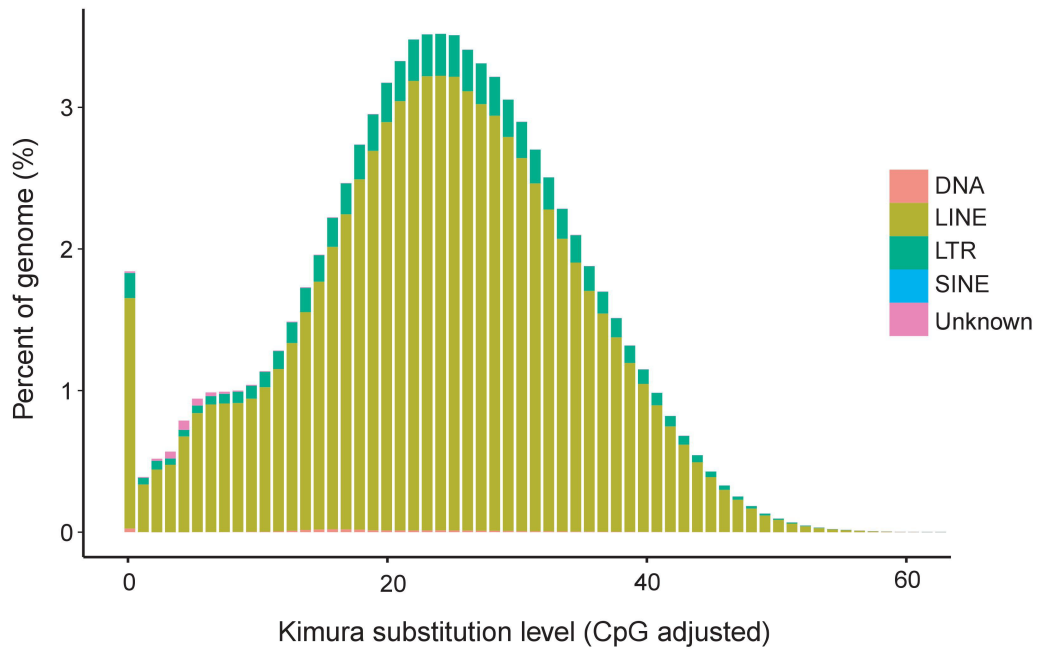


Figure S8. Divergence distribution of classified families of transposable elements of BM, Related to Figure 3. Transposable elements (TEs), including long terminal repeat (LTR), long interspersed element (LINE) and short interspersed elements (SINE), were identified by using *de novo*-based and Rebased-based methods.

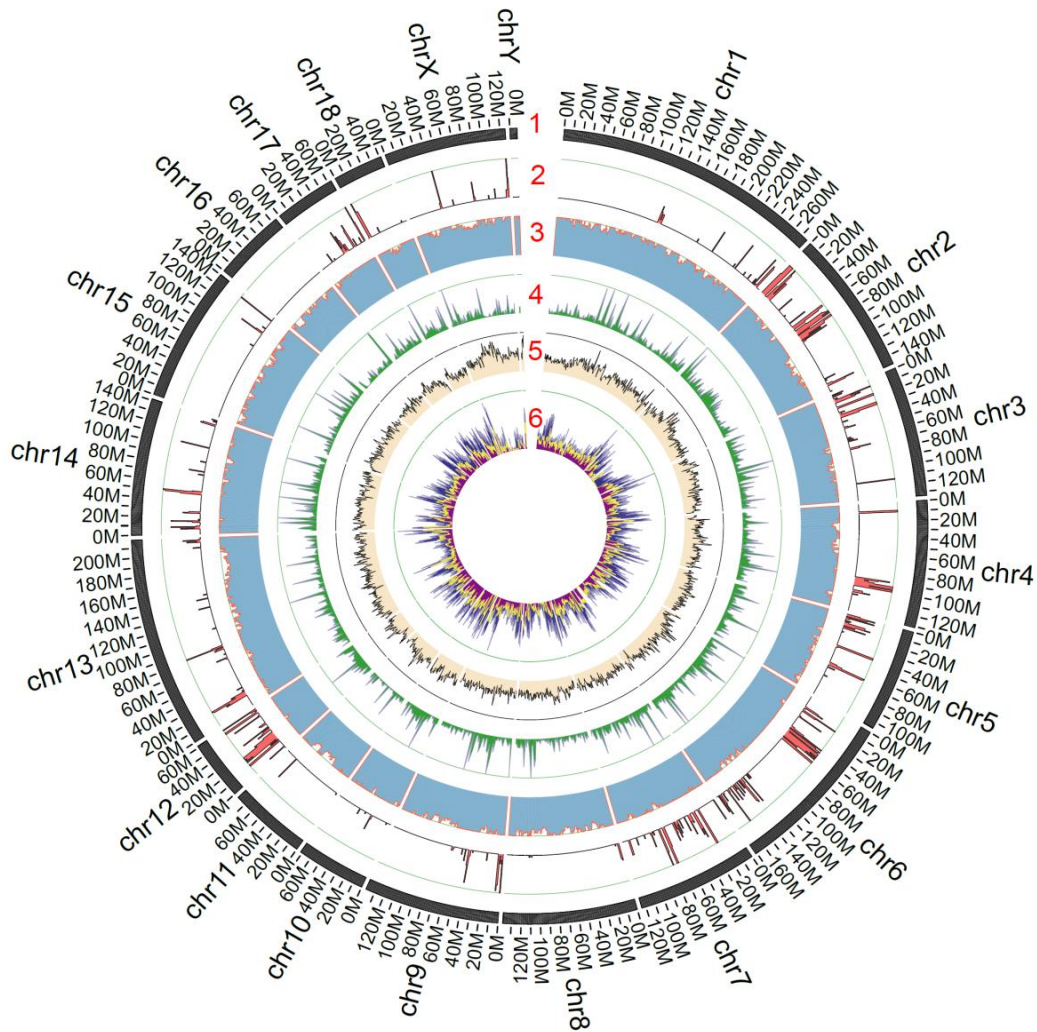


Figure S9. Genomic landscape of BM and single nucleotide divergence between BMs and BXs, Related to Figure 3.

From outer to inner circles: (1), marker distribution along the 18 autosomes and X, Y chromosomes at the megabyte (Mb) scale. (2) and (3), Gene density and GC content across the genome drawn in 1-Mb nonoverlapping windows. (4), The transcription level of each gene estimated by averaging the reads per kilobase (Kb) of exon model per million mapped reads from different tissues in nonoverlapping 2-Mb windows. (5), Repeat density across the genome, bin = 1Mb. (6), Distribution of heterozygous SNPs in the BM and BX genomes in 1-Mb nonoverlapping windows; SNPs in BM are shown in purple and SNPs in BX are in blue.

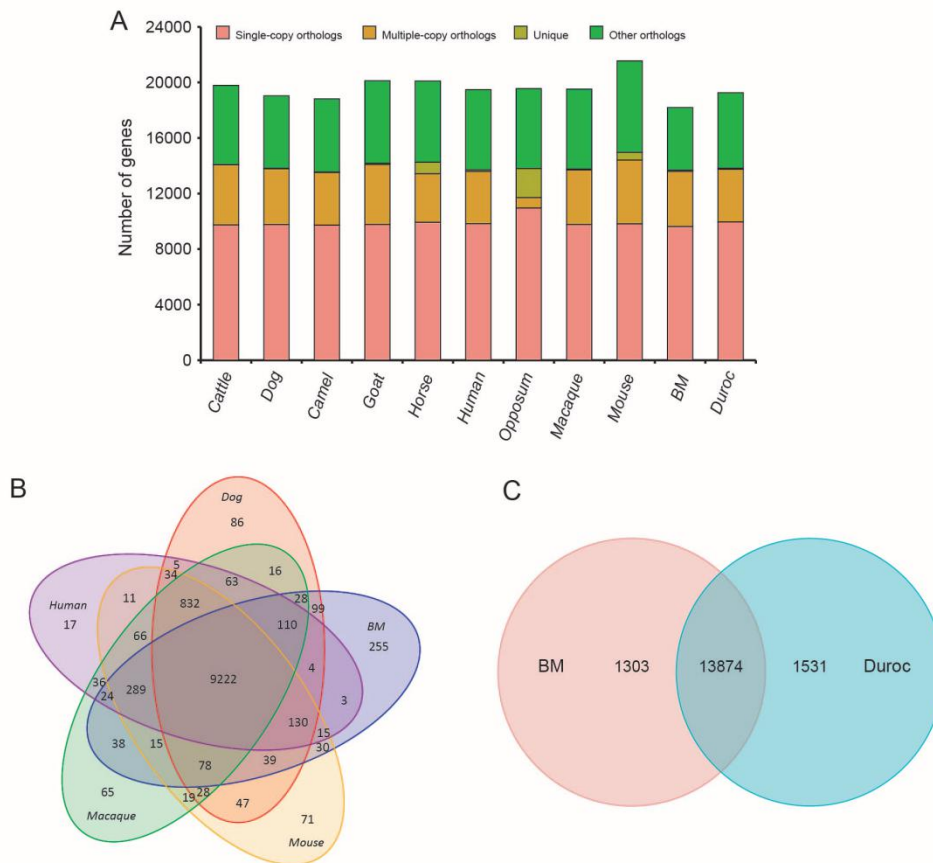


Figure S10. Analysis of lineage-specific genes, Related to Figure 3.

(A) Orthology delineation among the protein-coding gene family repertoires of the BM and other mammal. (B) Shared gene families among the BM, dog, mouse, macaque and human genomes. The BM has the most lineage-specific families compared with the five other mammals. (C) Shared gene families between BM and Duroc genomes. The number specific gene families of Duroc is higher than that of BM.

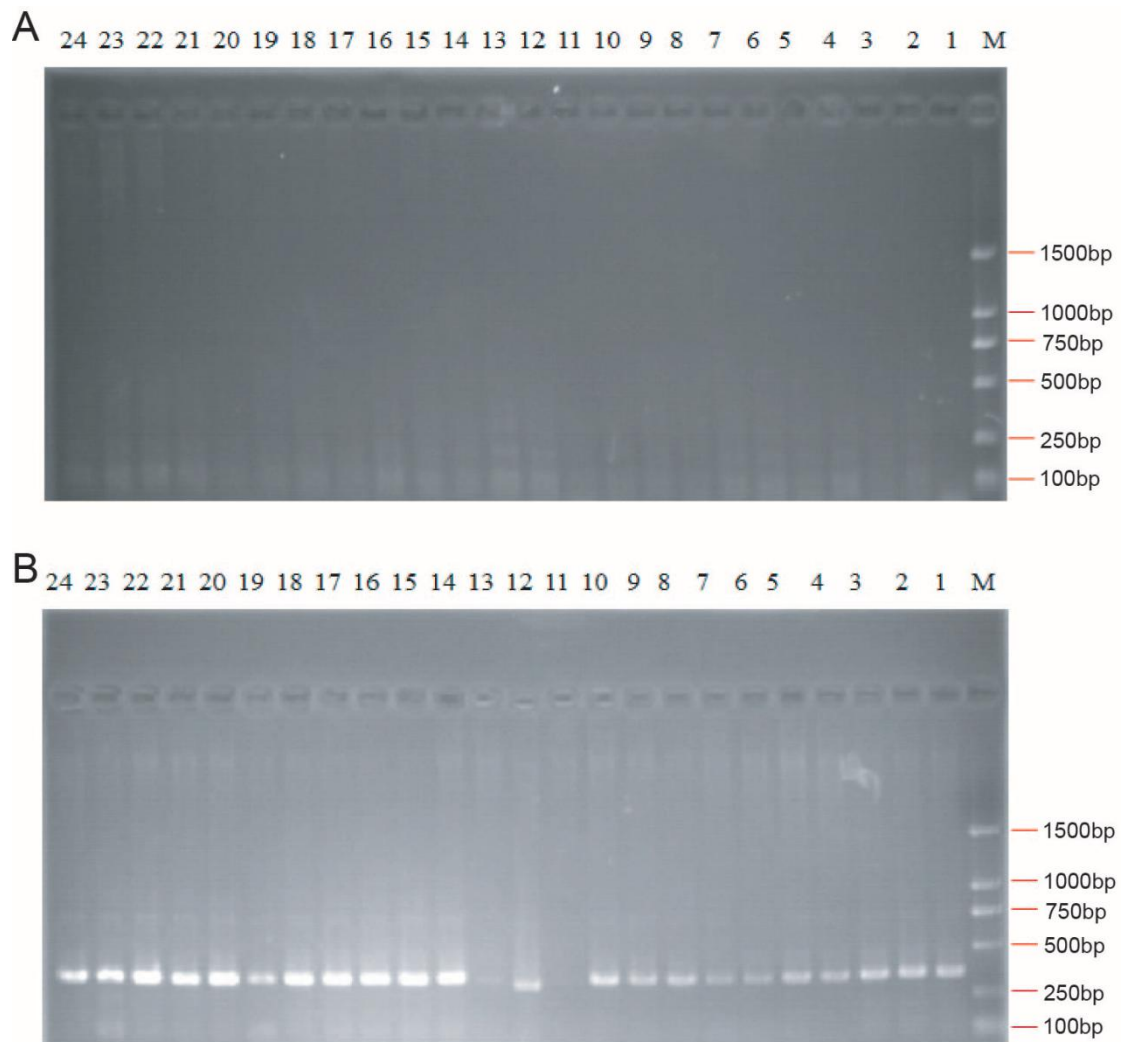


Figure S11. Electrophoresis for BM and Duroc PERV-C-*env* genotyping, Related to Figure 3.

(A) BM PERV-C-*env* detection in BMs. (B) Duroc PERV-C-*env* detection in Durocs. M represents DL2000 bp Marker and 1-24 represent PCR products from different samples.

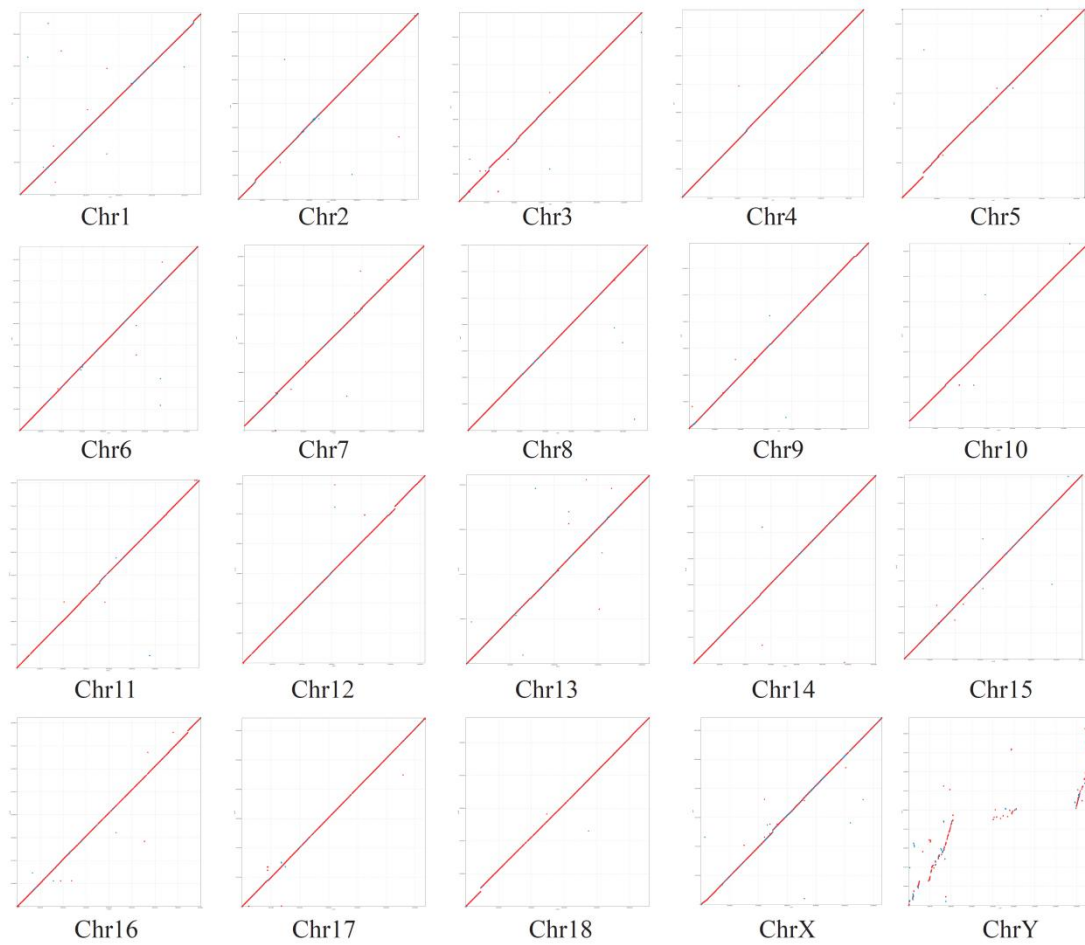


Figure S12. Homologous alignment of 20 chromosomes (18 autosomes + XY) between the BM and Duroc genomes, Related to Figure 4.

The horizontal and vertical coordinates represent the chromosome sequences of the Duroc and BM genomes, respectively.

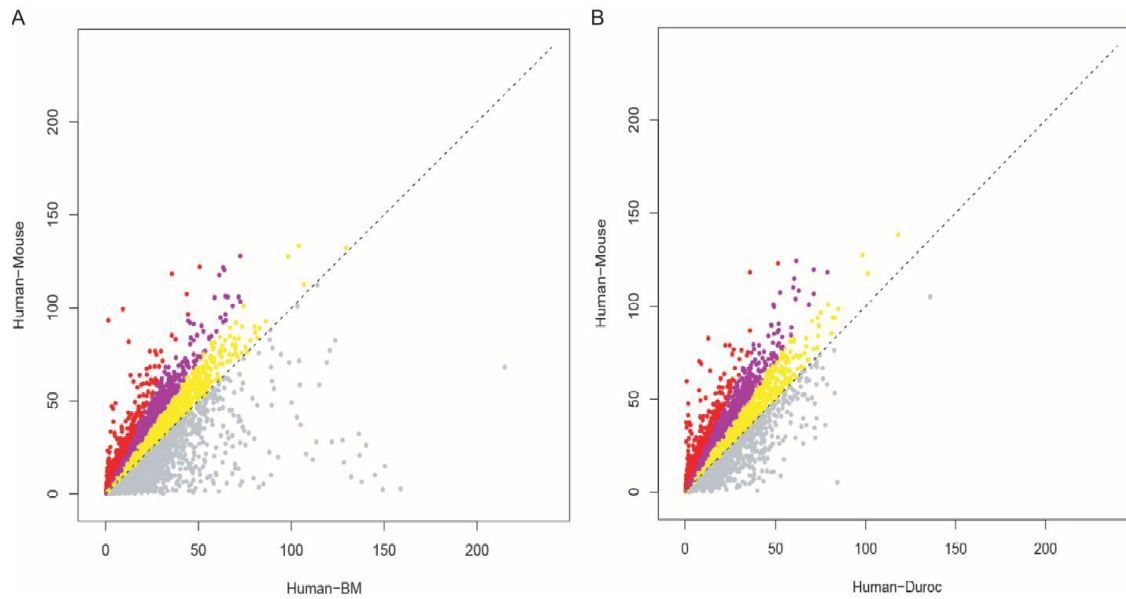


Figure S13. Orthologous protein sequence comparisons show that BM/Duroc protein sequence are similar to that of human, Related to Figure 5.

(A) Protein sequence comparisons among human, BM, and mouse. (B) Protein sequence comparisons among human, Duroc, and mouse. Scatter plot shows divergence between human and mouse protein sequences in terms of the Point Accepted Mutation (PAM) metric (y-axis) against the corresponding human vs. BM/Duroc protein sequence divergence (x-axis). Each point represents a ratio of two protein divergence. BM/Duroc proteins appearing above the 45° diagonal (gray dashes) represent those closer to the human sequence than the corresponding mouse sequence. The angle of the line to each protein from the origin is directly related to the ratio of mouse/pig divergence from the human sequence. A greater angle from the origin indicates greater divergence. The yellow part: $45^\circ \cong X < 55^\circ$. The purple part: $55^\circ \cong X < 65^\circ$. The red part: $65^\circ < X$.

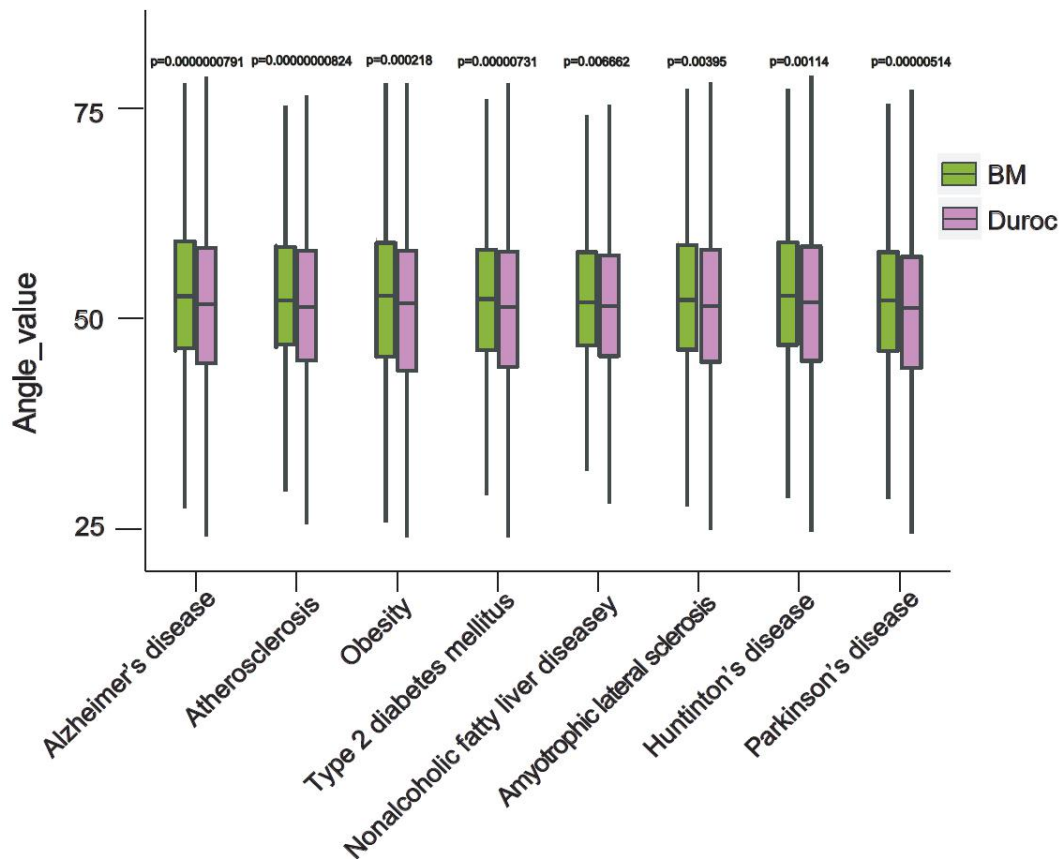


Figure S14. Boxplots show that the angles for BM proteins involved in eight human diseases are higher than those of Duroc, Related to Figure 5.

Boxplots of the angles (y-axis) represented in Figure S13 for proteins in eight common human diseases from GeneCards. A greater angle from the origin indicates greater similarity in human vs. BM than in human vs. Duroc. Boxes represent the interquartile range between the first and third quartiles and median (internal line). Whiskers denote the lowest and highest values within 1.5 times the range of the first and third quartiles, respectively. P-value was determined using Student *t* test.

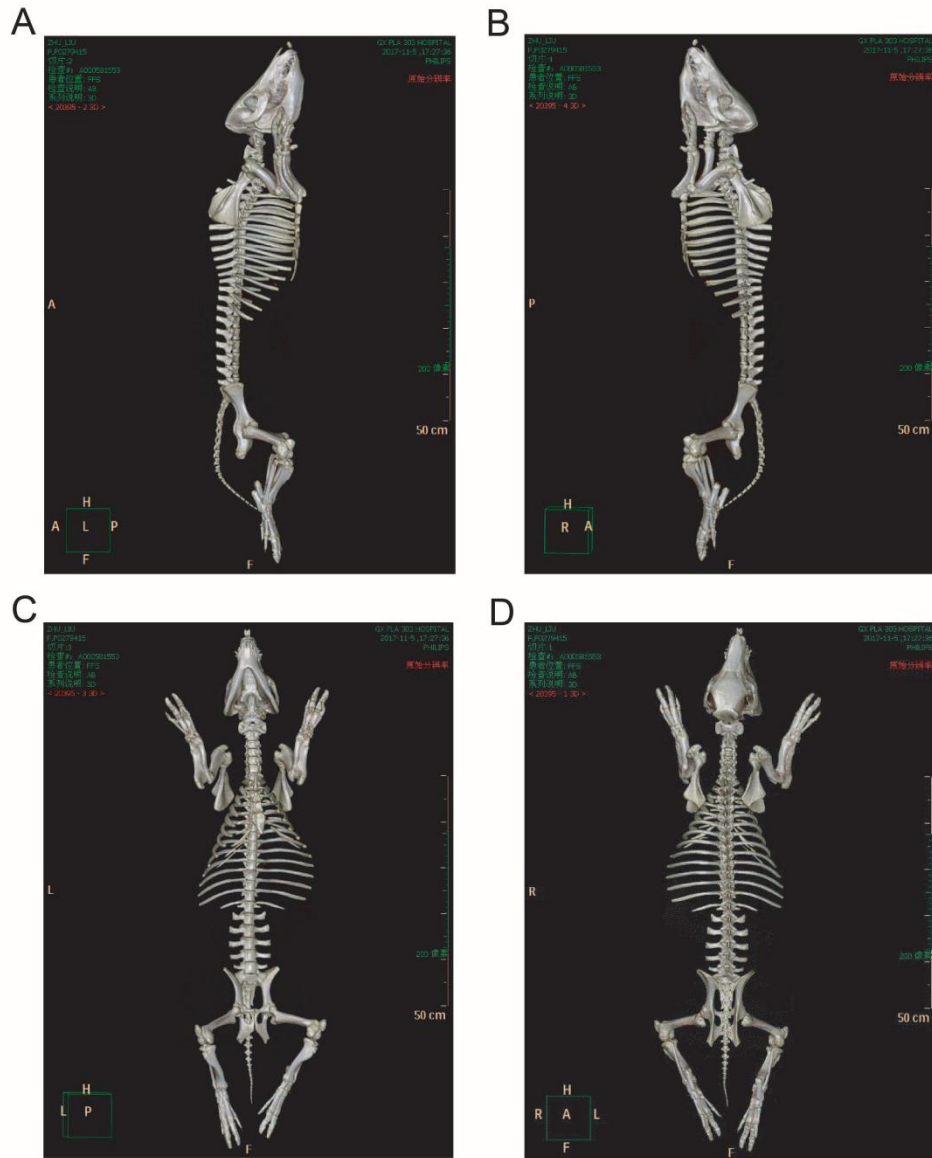


Figure S15. Skeleton of BM, Related to Figure 5.

(A) and (B) Lateral view of the skeleton. (C) and (D) Ventral and dorsal views of the skeleton. The images were obtained from a male BM (12 month old) by using Brilliance iCT (Philips, Netherlands).

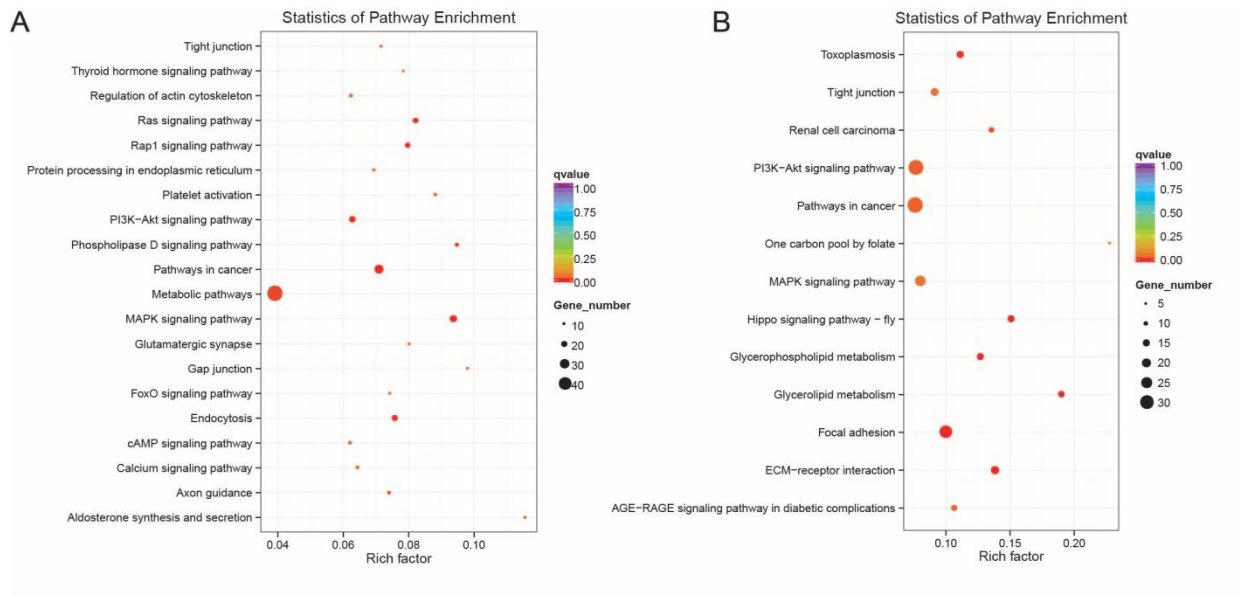


Figure S16. KEGG enrichment of the positively selected genes (PSGs) of BM and Duroc, Related to Figure 5.

(A) KEGG enrichment of the PSGs of BM. (B) KEGG enrichment of the PSGs of Duroc.

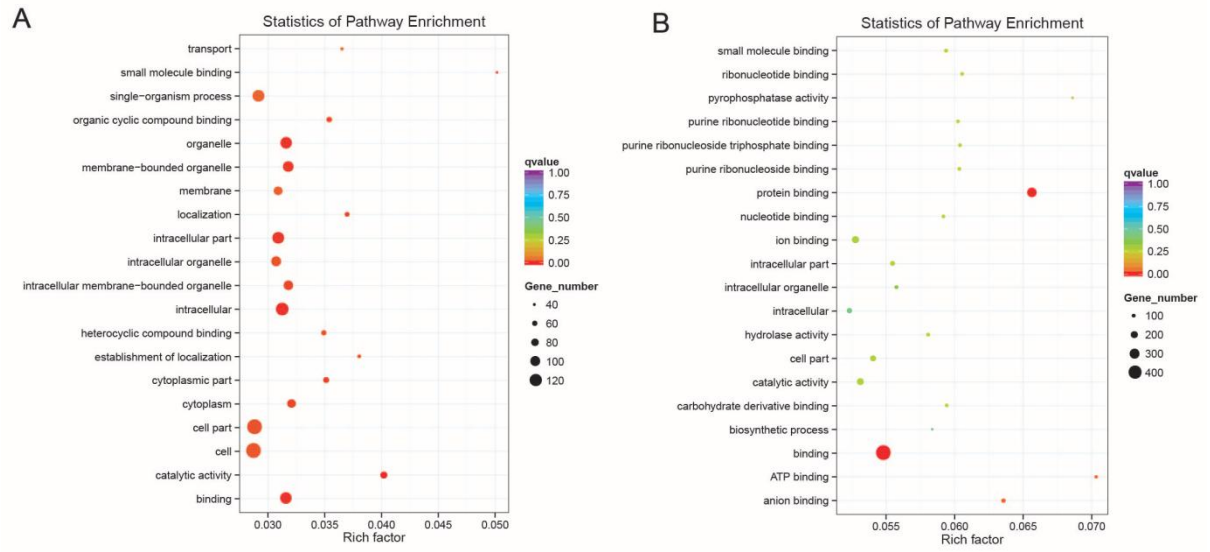


Figure S17. GO enrichment of the PSGs of BM and Duroc, Related to Figure 5.
 (A) GO enrichment of BM PSGs. (B) GO enrichment of Duroc PSGs.

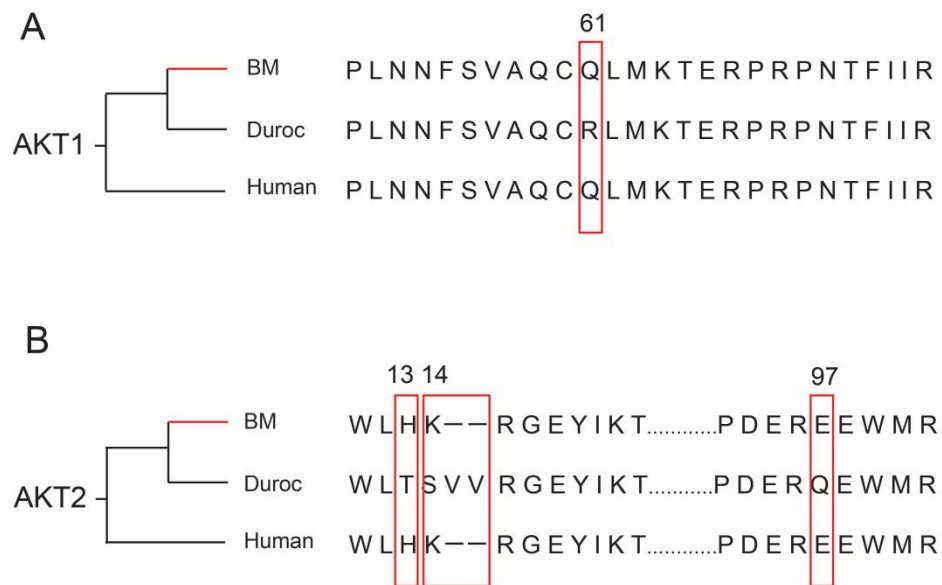


Figure S18. The alignment of the AKT1 and AKT2 proteins from BM, Duroc and human, Related to Figure 5.

(A) The alignment of the AKT1 protein from BM, Duroc and human. Compared with the 61 amino acid Q in BM and human AKT1 protein, 61 amino acid of Duroc AKT1 protein changed to R. (B) The alignment of the AKT2 protein from BM, Duroc and human. The changes 13-16 amino acid in AKT2 protein of Duroc are corresponded to the part of the structure enclosed in the dashed box in Figure 4d. Compared with the 97 amino acid E in BM and human AKT2 protein, 97 amino acid of Duroc AKT2 protein changed to Q.

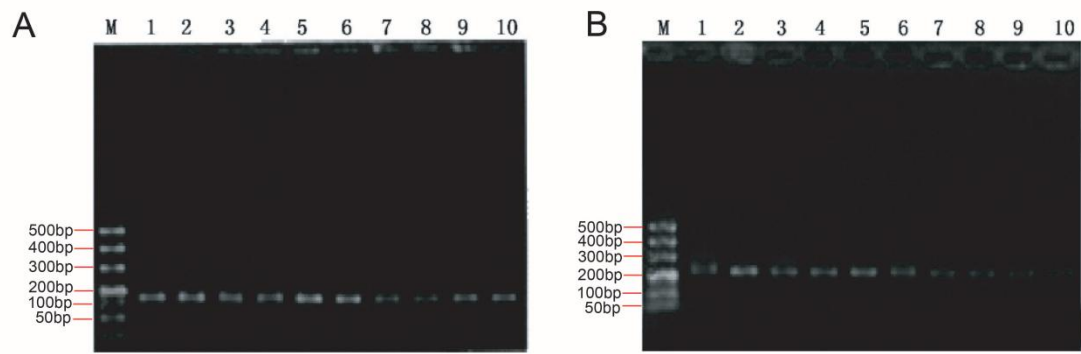


Figure S19. Electrophoresis for detection BM microsatellite loci, Related to Figure 6.

(A) PCR product representing locus S0155 (length: 155–166 bp). (B) PCR product representing locus SW0005 (length: 205–248 bp). M represents 500 bp DNA Marker and 1-10 represent PCR products from different samples.

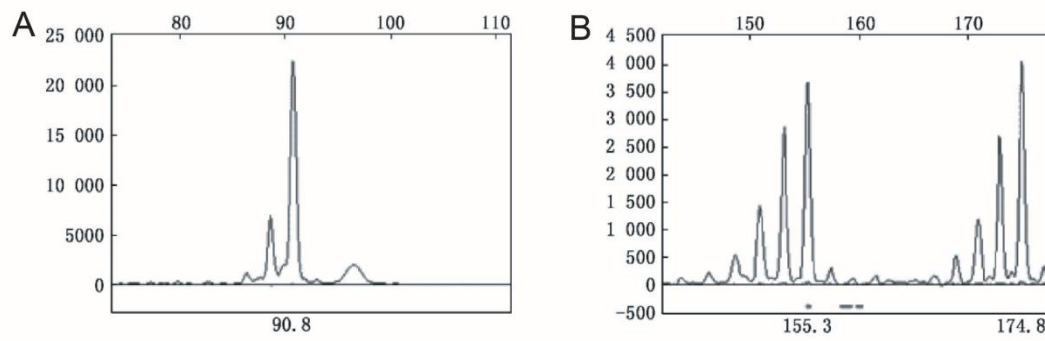


Figure S20. The part of microsatellite DNA typing using the ABI-3730XL DNA analyzer, Related to Figure 6.

(A) Detection of individual (F11, NO.1110) in microsatellite locus SW240; this locus in the BM genome is 91 bp and homozygous. (B) Detection of individual (F13, NO.1310) in microsatellite locus SW1119; this locus in the BM genome is heterozygous, including the genotypes associated with 155 bp and 175 bp sequences. Y-axis represents fluorescence intensity. X-axis represents length of microsatellite DNA (bp).

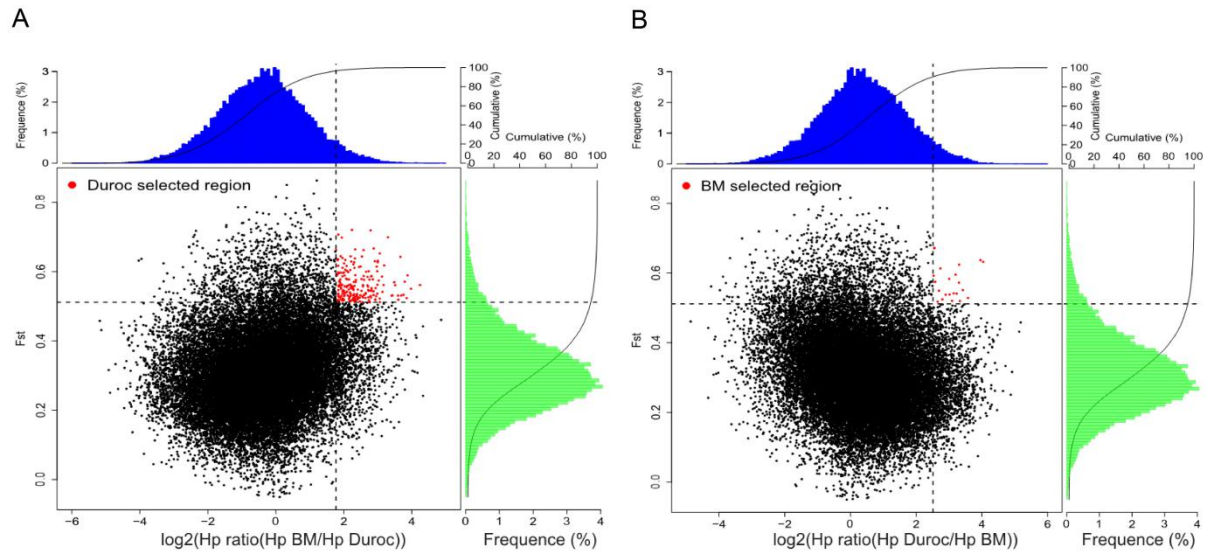
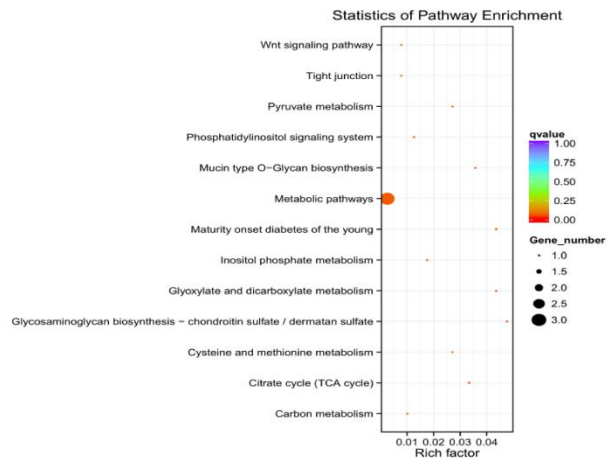


Figure S21. Distribution of Hp ratios and Fst values, Related to Figure 6.

(A) The distribution of Duroc selected sweep regions. (B) The distribution of BM selected sweep regions. Data points located to the right of the right vertical dashed line and above the horizontal dashed line were identified as selected sweep regions for Durocs and BMs, respectively.

A



B

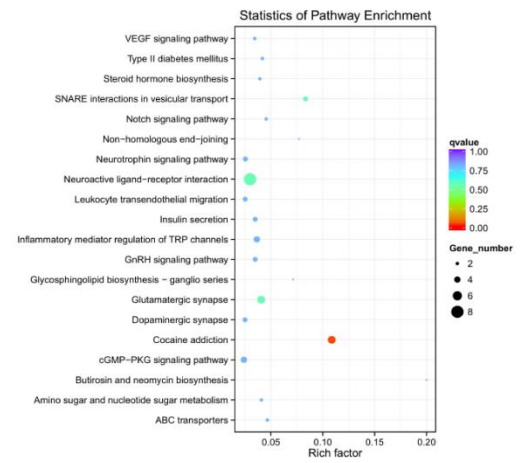
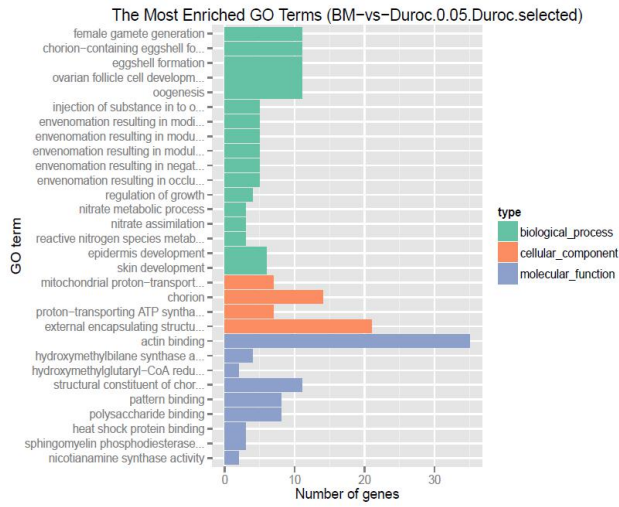


Figure S22. KEGG enrichment of the BM and Duroc genes located in the selective sweep regions, Related to Figure 6.

(A) KEGG enrichment of BM genes. (B) KEGG enrichment of Duroc genes.

A



B

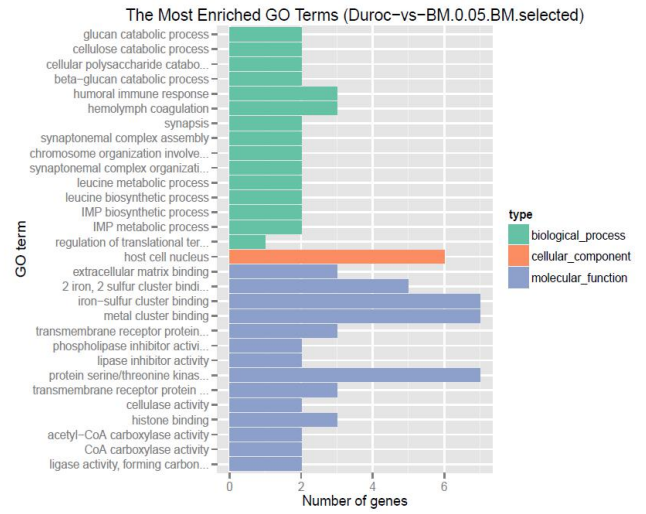


Figure S23. GO enrichment of the BM and Duroc genes located in the selective sweep regions, Related to Figure 6.

(A) GO enrichment of Duroc genes. (B) GO enrichment of BM genes.

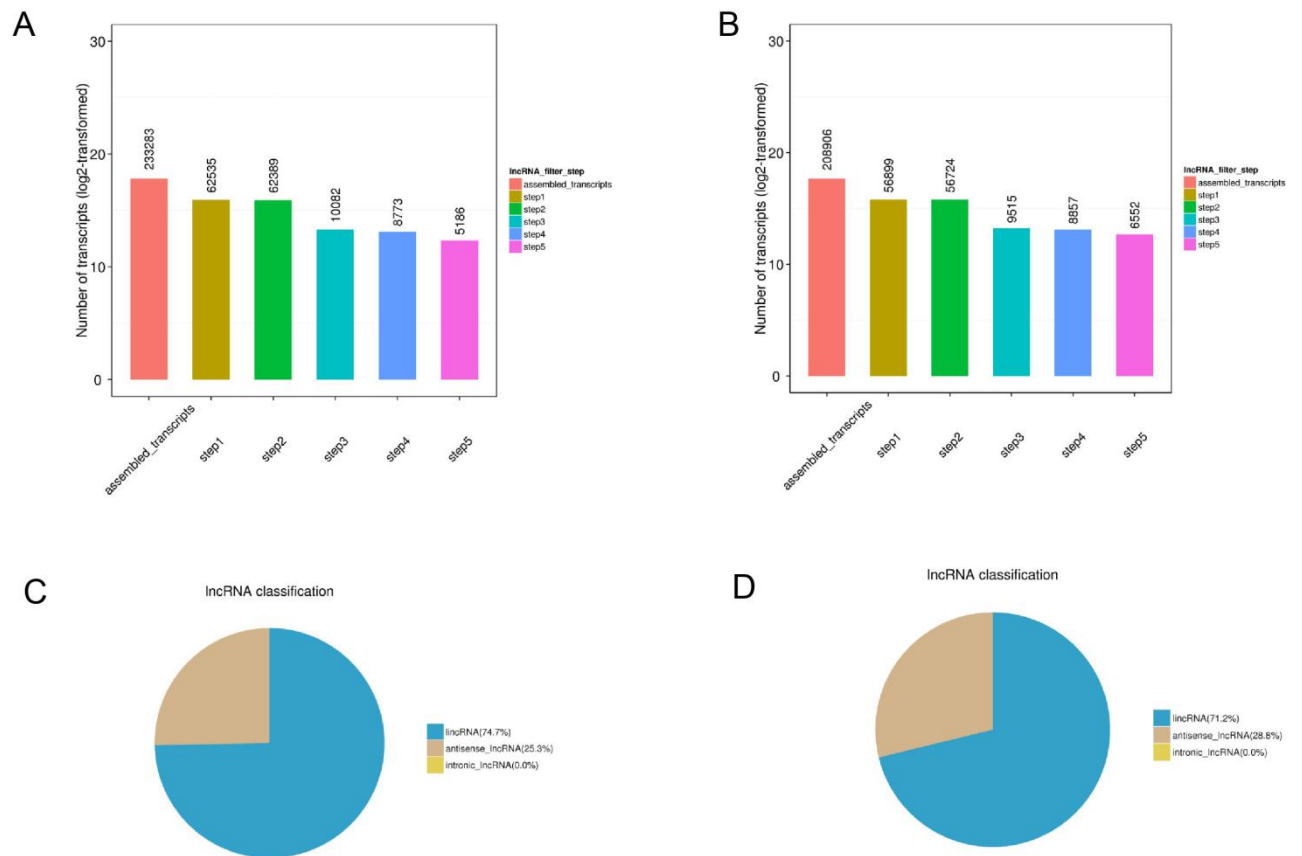


Figure S24. The results of lncRNAs identification in liver and skeletal muscle of individuals in BM-induced and Duroc-induced groups, Related to Figure 2.

BMs and Durocs were fed with a high fat and carbohydrate diet and limited activities for 12 months. RNA was extracted from the liver and skeletal muscle at the end of the experiment and was subjected to RNA-seq. (A) lncRNA identifying results of the BMs and Durocs liver after several filtering steps (Transparent Methods). (B) lncRNA identifying results of the BMs and Duroc skeletal muscle after several filtering steps (Transparent Methods). (C) Scattergram of different type of lncRNAs in the BMs and Durocs liver. (D) Scattergram of different type of lncRNAs in the BMs and Durocs skeletal muscle. The numbers above the column in (A) and (B) represent the amounts of lncRNAs after different filtering steps. Classification of the lncRNAs detected in liver (C) and skeletal muscle (D), mainly including lincRNA, anti-sense-lincRNA, intronic-lincRNA.

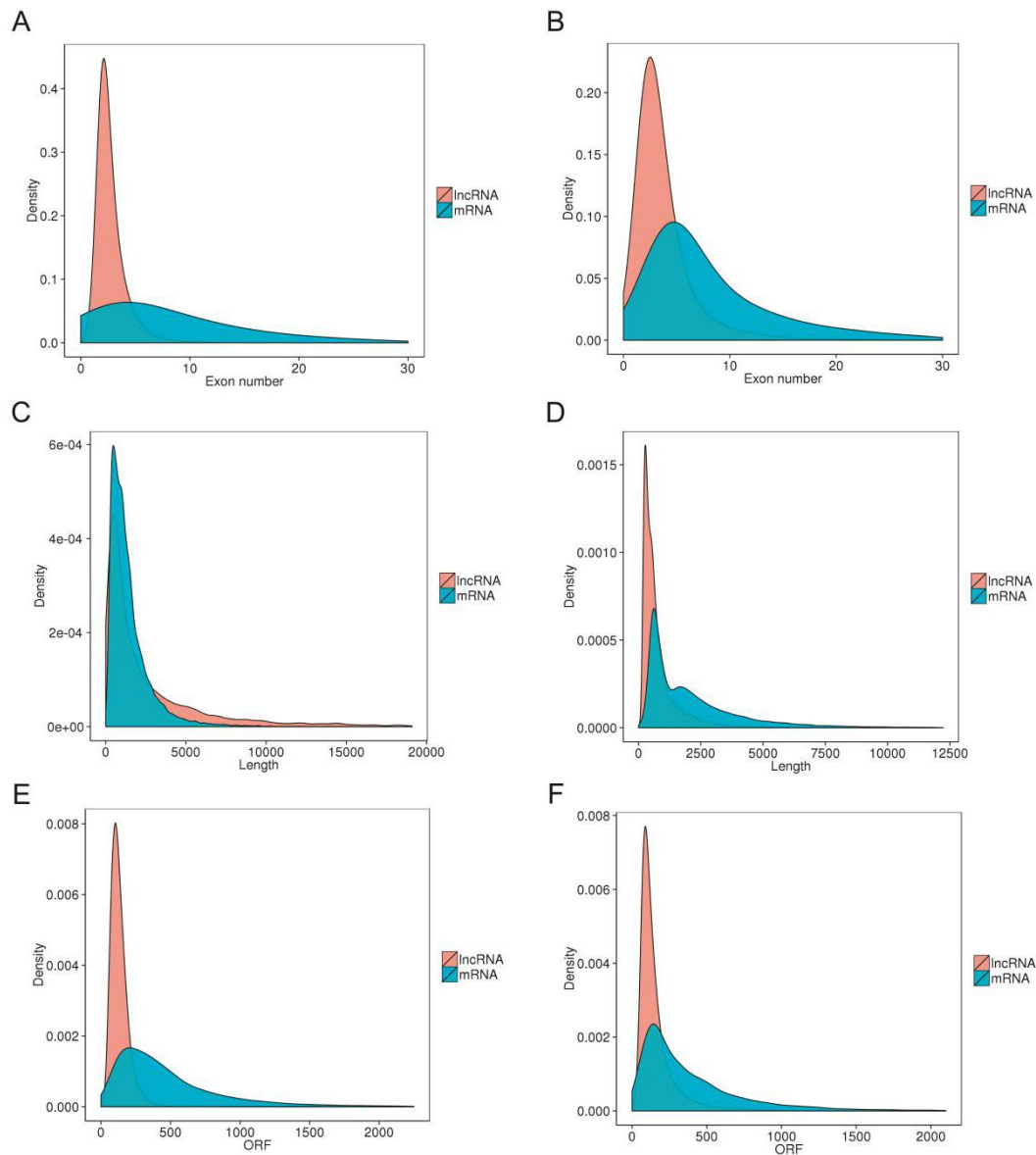


Figure S25. The characteristics of lncRNAs and mRNAs identified in liver and skeletal muscle of individuals in BM-induced and Duroc-induced groups, Related to Figure 2.

(A) Number of exons in transcripts in the BMs and Durocs liver. (B) Number of exons in transcripts in the BMs and Durocs skeletal muscle. (C) Distribution of transcript length in the BMs and Durocs liver. (D) Distribution of transcript length in the BMs and Durocs skeletal muscle. (E) Distribution of open read frame (ORF) length in the BMs and Durocs liver. (F) Distribution of open read frame (ORF) length in the BMs and Durocs in skeletal muscle.

In the liver and skeletal muscle, the average numbers of exon were all lower in lncRNAs than in mRNAs; the average transcript lengths of transcripts of lncRNAs were similar to those of mRNAs; the average lncRNA length was shorter than the average mRNA length.

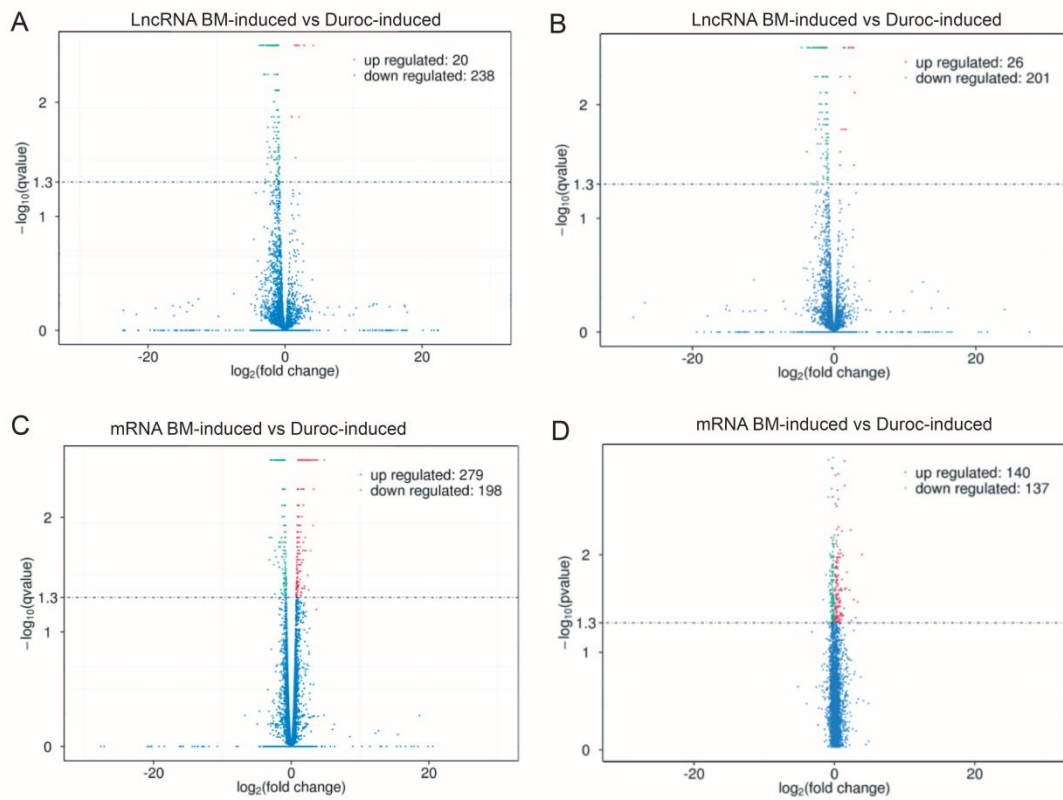


Figure S26. Volcano plot of differentially expressed of lncRNAs and mRNAs in individuals in BM-induced group vs. individuals in Duroc-induced group, Related to Figure 2.

(A) Volcano plot of differential lncRNA expression in the liver. (B) Volcano plot of differential mRNA expression in the liver. (C) Volcano plot of differential lncRNA expression in skeletal muscle. (D) Volcano plot of differential mRNA expression in the skeletal muscle.

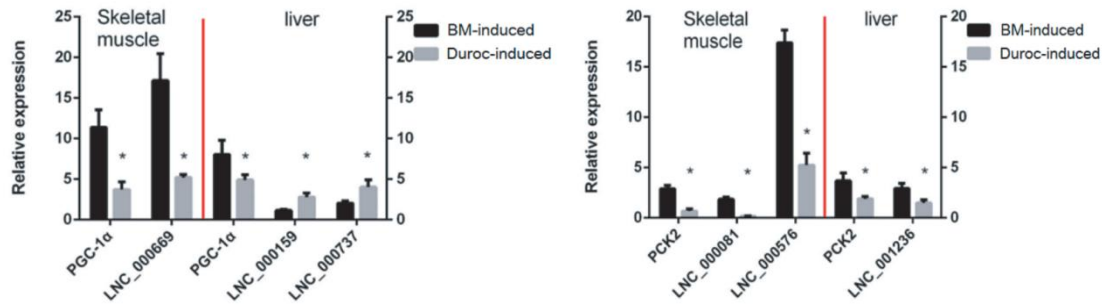





Figure S27. Validation of genes expression levels using qRT-PCR, Related to Figure 2.

Transcript expressions in liver and skeletal muscle were quantified relative to the expression level of *GAPDH* using the comparative cycle threshold (ΔCT) method. Error bars indicate Standard Error of Mean (SEM) of three pigs per group. Asterisk indicate significant differences ($P < 0.05$) using Student *t* test.

Supplementary Tables

Table S1. General characteristics of Bama miniature pig compared to Bama xiang pig (from which they derive) and the commercial breed Duro pig commonly used in pork production, Related to Figure 1.

			
	Bama miniature pig (BM)	Bama xiang pig (BX)	Duroc
Breeding history	Inbred since 1987 to establish an ideal experimental animal strain with a clear genetic background, high homozygosity and inbreeding tolerance.	A primitive breed, indigenous to the south of China, living at an average North latitude of 22°. BX had undergone a low degree of inbreeding (frequently female parent × male offspring mating) for the last several centuries, which shaped the relative higher genetic stability and adaptation to inbreeding than those of other pig breeds.	Found mainly in North America but originated in Europe at an average northern latitude of 50°. This breed has been intensively artificially selected for fast growth, and efficient accumulation of lean meat (muscle). In general, the inbreeding coefficient of Duroc is less more than 12%.
Coat color	Uniform two-end-black fur without other black spot. White fur occupies over 82% of the surface area (inbred line F10-F19).	Two-end-black fur with black spots of different sizes on shoulders, back and waist.	Reddish-brown.

Body size	The average body weights of newborn, 12-month-old, and 20-month-old pigs are approximately 0.276 ± 0.053 , 36.443 ± 0.49 , and 43.158 ± 0.42 kg (inbred line F10-F19), respectively. The average adult (20 months old) body length and height are 73.67 ± 5.59 cm and $= 42.75\pm1.21$ cm (inbred line F10-F19), respectively, which are significantly smaller than these of BXs and commercial Durocs.	The average adult body weight (51.92 ± 1.92 kg), length (90.14 ± 4.89 cm), and height (50.29 ± 4.18 cm) are greater than those of BM pigs but less than those of commercial pigs.	The average adult body weight, length, and height are > 300 kg, 167.5 ± 5.36 cm, and 66.83 ± 3.26 cm, respectively.
Skeleton	There are only 19 or 20 thoracic and lumbar vertebrae, which is fewer than those in other large commercial pigs (21–23 thoracic and lumbar vertebrae in Duroc and other pigs). The average adult length of the fibula (12.24 ± 0.92 cm) is shorter than that (17.27 ± 0.95 cm) of the Duroc pig.	The numbers of thoracic and lumbar vertebrae are the same as those in the BM pig. The length of the fibula (14.24 ± 0.76 cm) is longer than that of the BM pig but shorter than that of the Duroc pig.	There are 21–23 thoracic and lumbar vertebrae, and the total length of the cervical, thoracic, lumbar, and sacral vertebrae of Duroc swine is 162.3 ± 4.36 cm. The average adult length of the fibula is 17.27 ± 0.95 cm.
Sexual maturity	Early maturity: at 30 and 60 days old, the weight of the testes (5.43 and 13.80 g, respectively) and seminiferous tubule diameter (100.47 and 188.5 μ m, respectively) are higher than those of Durocs at the same ages (2.94 and 11.55 g; 80.00 and 146.43 μ m, respectively). Spermiogenesis of spermatids in the testes (30 days old) and mature sperm in the epididymis (48 days old) have been documented.	The reproductive system is the same as that of BMs.	Late maturity: The reproductive system of commercial pigs such as the Duroc is clearly not as developed as that of BMs/BXs. Sexual maturity is reached between 6 and 8 months, meaning that mature sperm do not appear in the epididymis until after 180 days of age.
Feeding	Pigs (> 6 kg) are fed a standard diet composed of 12.95 MJ/kg with	BXs is under extensive feeding.	Pigs are fed according to the

standard	<p>the full value of feed digestion and 14.5% crude protein. All of the pigs are fed twice a day with drinking water provided.</p> <p>The nutrient requirements of miniature pigs according to size¹ are as follows:</p> <p>0.4–6 kg: Total digestive energy of one miniature pig/day = 2.38 MJ, and energy concentration of pig milk = 5.31 MJ/kg. We chose to supply sufficient milk (> 0.45 kg) to piglets.</p> <p>6–20 kg: Total digestive energy of one miniature pig/day = 4.67 MJ. In this phase, food intake of one pig/day = 2% of body weight (~average 0.26 kg), which provides 72% (~3.37 MJ) of the daily nutrient requirement.</p> <p>20–40 kg: Total digestive energy of one miniature pig/day = 6.91 MJ. In this phase, food intake of one pig/day = 1.2% of body weight (~average 0.36 kg), which provides 67% (~4.47 MJ) of the daily ad libitum intake.</p>		“Nutrient requirements of swine (NRC 2012)”.
Temperament	After selection, behavior is dramatically tamer than that of BXs.	BXs have aggressive behaviors, including jumping over a 1-m-high fence and attacking human beings.	Temperament has become very docile after hundreds of years of domestication.
Distribution	Currently, most pigs, especially closed colonies and inbred strains, are distributed in Guangxi Province, south China.	Indigenous to Guangxi Province, south China.	Internationally used breed (93 countries).

Note: Values represent the mean±SD.

Table S2. FBG levels in BMs, and Durocs fed a high-fat/high-carbohydrate diet and limited activities for 0–12 months, Related to Figure 2.

	Generation	Time	0M	1M	2M	3M	4M	5M	6M	7M	8M	9M	10M	11M	12M
BM-induced group	Inbred F17	BM-1	4.6	5.3	5.6	5.6	6.1	5.4	5.8	5.9	6.4	6.7	6.8	6.9	6.7
	<i>Inbred F17</i>	<i>BM-2</i>	<i>4.5</i>	<i>5</i>	<i>4.4</i>	<i>5.5</i>	<i>6</i>	<i>6</i>	<i>6</i>	<i>5.5</i>	<i>5.6</i>	<i>6.2</i>	<i>7.1</i>	<i>7.4</i>	<i>7.5</i>
	<i>Inbred F17</i>	<i>BM-3</i>	<i>4.7</i>	<i>4.8</i>	<i>5.8</i>	<i>5.6</i>	<i>5.7</i>	<i>5.6</i>	<i>5.2</i>	<i>5.4</i>	<i>5.7</i>	<i>5.6</i>	<i>5.6</i>	<i>6.3</i>	<i>7</i>
	<i>Inbred F17</i>	<i>BM-4</i>	<i>4.8</i>	<i>4.6</i>	<i>4.9</i>	<i>5.9</i>	<i>5.6</i>	<i>5.9</i>	<i>5.6</i>	<i>6.1</i>	<i>6</i>	<i>5.8</i>	<i>6.3</i>	<i>7.1</i>	<i>7</i>
	<i>Inbred F17</i>	<i>BM-5</i>	<i>4.7</i>	<i>5.4</i>	<i>5.3</i>	<i>4.9</i>	<i>5</i>	<i>5.3</i>	<i>5.2</i>	<i>6.2</i>	<i>6.7</i>	<i>6</i>	<i>7.3</i>	<i>7.8</i>	<i>7.9</i>
	Inbred F18	BM-6	4.1	4.5	5.2	5.1	5.3	5	5.9	6.1	6.5	6.6	6.5	6.9	6.6
	<i>Inbred F18</i>	<i>BM-7</i>	<i>4.7</i>	<i>5.2</i>	<i>5.6</i>	<i>5.6</i>	<i>6.1</i>	<i>6.3</i>	<i>6.3</i>	<i>6.7</i>	<i>6.9</i>	<i>6.8</i>	<i>7.4</i>	<i>7.5</i>	<i>8.1</i>
	<i>Inbred F18</i>	<i>BM-8</i>	<i>3.9</i>	<i>5</i>	<i>5.4</i>	<i>5.6</i>	<i>5.3</i>	<i>5.1</i>	<i>5.2</i>	<i>5.4</i>	<i>5.6</i>	<i>5.8</i>	<i>6.2</i>	<i>6.3</i>	<i>7.2</i>
	<i>Inbred F18</i>	<i>BM-9</i>	<i>4.4</i>	<i>6.1</i>	<i>5.6</i>	<i>5.5</i>	<i>6.4</i>	<i>5.1</i>	<i>5.4</i>	<i>5.9</i>	<i>6.5</i>	<i>6.8</i>	<i>6.8</i>	<i>6.6</i>	<i>7.6</i>
	<i>Inbred F18</i>	<i>BM-10</i>	<i>3.9</i>	<i>5.5</i>	<i>4.7</i>	<i>4.9</i>	<i>5.8</i>	<i>7.2</i>	<i>5</i>	<i>5.9</i>	<i>6.8</i>	<i>6.4</i>	<i>6.9</i>	<i>7.8</i>	<i>7.8</i>
	<i>Inbred F19</i>	<i>BM-11</i>	<i>3.5</i>	<i>5.7</i>	<i>5.3</i>	<i>5.6</i>	<i>5.4</i>	<i>5.6</i>	<i>5.9</i>	<i>5.7</i>	<i>6.2</i>	<i>6.7</i>	<i>7.8</i>	<i>7.9</i>	<i>8</i>
	<i>Inbred F19</i>	<i>BM-12</i>	<i>3.8</i>	<i>4.6</i>	<i>5.3</i>	<i>5.1</i>	<i>5.5</i>	<i>5.8</i>	<i>6</i>	<i>6.7</i>	<i>6.4</i>	<i>6.3</i>	<i>7</i>	<i>7.9</i>	<i>8.4</i>
	Inbred F19	BM-13	4	3.8	4.9	5	4.7	5.1	5.4	5.9	5.3	5.7	5.8	5.8	5.4
	Inbred F19	BM-14	3.8	4.6	5.9	5.1	4.8	5.8	6	5.7	5.4	5.3	5	5.9	5.7
	Inbred F19	BM-15	4.3	4.5	4.6	4.7	5.4	5.5	5.6	5.5	5.4	5.5	5.2	5.4	5.6
Duroc-induced group	-	Duroc-1	3.9	4.2	4.2	4.3	4.5	4.8	4.4	4.7	4.7	4.9	4.3	4.8	4.4
	-	Duroc-2	4.3	4.4	4.4	4.7	4.6	4.3	4.4	4.7	4.5	4.8	5	4.7	4.7
	-	Duroc-3	4.9	4.6	4.5	4.2	3.9	4.3	4.2	4.7	5	4.7	4.7	4.9	4.5
	-	<i>Duroc-4</i>	<i>3.7</i>	<i>4.7</i>	<i>4.3</i>	<i>4.6</i>	<i>4.2</i>	<i>4.8</i>	<i>5</i>	<i>5.4</i>	<i>5.7</i>	<i>5.9</i>	<i>6.3</i>	<i>6.8</i>	<i>7</i>
	-	Duroc-5	4	4.2	3.8	4.1	4.2	4.4	4.3	5	5.1	4.8	4.6	5.2	4.8
	-	Duroc-6	4.3	4.2	4.3	4.5	4.7	4.3	4.9	4.3	4.7	4.3	5.1	4.8	5
	-	Duroc-7	4	4.2	4.6	4.7	4.3	4.3	4.1	4.3	4.4	4.7	4.1	4.7	4.1

-	Duroc-8	3.6	4.4	5.4	5.1	5.3	4.6	4.9	4.9	4.3	4.6	4.9	4.8	4.9
-	Duroc-9	3.8	4	4.1	4.4	4.3	4.7	4.2	4.7	4.4	4.9	4.3	4.2	4.1
-	Duroc-10	4.1	4.8	5.2	5.3	5.1	4.8	4.7	4.6	4.8	5	5.1	5.1	5
-	<i>Duroc-11</i>	<i>3.9</i>	<i>4.8</i>	<i>4.8</i>	<i>5.4</i>	<i>5.1</i>	<i>5.5</i>	<i>5.3</i>	<i>5.3</i>	<i>5.8</i>	<i>6.4</i>	<i>6.8</i>	<i>7</i>	<i>7.1</i>
-	Duroc-12	3.6	3.8	4.6	4	4.8	4.4	4.7	4.3	4.5	5.1	5.3	4.4	4.8
-	Duroc-13	3.9	4.9	4.5	3.9	4.6	4.6	4.9	4.5	5.5	5.3	5.4	4.3	5.4
-	Duroc-14	4.2	4.9	4.6	4.3	4.4	4.1	4.3	4.7	4.9	5	5.2	5.3	5.6
-	Duroc-15	4.3	4.1	4.4	4.9	4.6	5	4.8	4.8	5	5.3	5.4	5.5	5.2
BM-control group														
	BM-control-1	4.1	4.6	4.5	4.2	4.7	4.4	5.0	4.8	5.0	4.5	4.9	4.7	4.6
	BM-control-2	4.4	4.4	4.3	4.3	4.5	4.6	4.6	4.9	4.6	4.8	4.2	4.8	4.7
	BM-control-3	4.0	4.4	4.4	4.6	4.6	4.9	4.2	4.3	4.4	5.0	5.2	4.7	4.5
Duroc-control group														
	Duroc-control-1	3.9	4.2	4.1	4.3	4.1	4.2	4.5	4.5	4.7	4.9	4.6	4.8	4.4
	Duroc-control-2	4.3	4.4	4.3	4.3	4.4	4.3	4.4	4.1	4.2	4.6	4.3	4.4	4.3
	Duroc-control-3	4.1	4.2	4.6	4.5	4.4	4.7	4.3	4.6	4.3	4.4	4.5	4.7	4.6

Note: Bold face and italic indicate those pigs whose FBG level reached 7 mmol/L (human diabetic standard) at 12 months.

Table S3. Statistics of glucose disappearance rate in intravenous glucose tolerance test at 12th month of BMs and Durocs with FBG > 7, Related to Figure 2.

Time (min)	10	30	60	90	120
BM-2	32.1%	53.8%	68.9%	72.2%	75.3%
BM-3	19.3%	40.0%	49.8%	66.9%	73.8%
BM-4	17.4%	48.4%	60.9%	70.8%	74.4%
BM-5	26.7%	47.4%	62.8%	69.1%	72.6%
BM-7	18.3%	43.1%	60.7%	69.2%	71.5%
BM-8	17.1%	45.1%	55.9%	71.0%	73.8%
BM-9	19.7%	39.8%	51.3%	66.9%	73.2%
BM-10	20.5%	47.3%	62.2%	70.7%	74.2%
BM-11	25.4%	47.8%	60.5%	68.7%	72.2%
BM-12	30.3%	44.9%	59.2%	70.1%	73.1%
Duroc-4	47.0%	60.6%	65.3%	70.5%	72.5%
Duroc-11	47.3%	63.6%	68.2%	69.8%	72.9%
BM-control1	50.4%	68.1%	73.6%	79.5%	81.1%
BM-control2	48.2%	68.4%	74.5%	76.5%	80.6%
BM-control3	47.0%	68.1%	74.9%	76.5%	80.5%
Duroc-control1	52.8%	70.2%	79.8%	80.6%	81.5%
Duroc-control2	52.2%	65.9%	77.2%	79.3%	80.6%
Duroc-control3	52.1%	68.1%	80.5%	80.9%	82.9%

Note: Glucose disappearance rates (%) were calculated based on the FBGs (mmol/L) of different time compared with the FBGs at 3 minutes.

Table S4. Estimation of the BM genome size using K-mer analysis, Related to Figure 3 and Table 1.

K-mer	K-mer number	K-mer depth	Revised genome Size (Mbp)
17	107,425,255,608	41	2583.64

Note: The estimated size of BM genome is ~2.58Gb. “Revised genome size” is the accurate estimation without error K-mers. “Repeat” is the proportion of the same K-mer fragments in all K-mers.

Table S5. Statistics of the genome sequencing data of BM, Related to Figure 3 and Table 1.

Pair-end libraries	Insert size (bp)	Total data (G)	Read length (bp)	Sequence coverage (X)
	250	98.90		38.33
	500	70.56		27.35
Illumina reads	2000	96.05	150	37.23
	5000	63.28		24.53
	10000	65.62		25.43
10 × Genomics	--	184.83	--	71.64
Pacbio reads	--	53.00	--	20.54
Nanopore	--	26.10	--	10.12
Hi-C reads	350	273.7	150	106.09
Total	--	932.04	--	361.25

Note: In total, 932.04 Gb sequencing data were used for *de novo* assembly. Sequencing depth was calculated based on a genome size of 2.58 Gb according to survey analysis.

Table S6. Two assembled versions of the BM genome, Related to Figure 3 and Table 1.

Version	Title	Total length(bp)	N50 length(bp)	N50 Number	N90 length(bp)	N90 Number
Primary genome	Contig	2,434,455,766	258,261	2,713	62,333	9,916
	Scaffold	2,462,711,326	21,133,597	35	5,711,122	120
Chromosome genome	Contig	2,475,389,976	1,010,657	736	211,363	2,697
	Scaffold	2,491,207,002	140,438,739	7	75,503,400	16

Note: The primary genome was assembled using Illumina reads and 10 × Genomics data. Advanced genome assembly was based on the primary genome assembled using PacBio, Nanopore and Hi-C data.

Table S7. BM chromosome-level genome assembly, Related to Figure 3 and Table 1.

Type	Length		Number	
	Contig(bp)	Scaffold(bp)	Contig	Scaffold
Total	2,475,389,976	2,491,207,002	13,676	6,638
Max	6,189,594	283,123,735	-	-
Number \geq 2000	-	-	9,378	3,569
N50	1,010,657	140,438,739	736	7
N60	785,283	135,930,596	1,012	9
N70	593,748	126,657,800	1,371	11
N80	409,075	107,564,314	1,866	13
N90	211,363	75,503,400	2,697	16

Note: Only scaffolds \geq 100 bp are included in the genome assembly.

Table S8. Lengths of all chromosomes of BM, Related to Figure 3 and Table 1.

Total base (bp)	Total base anchored to chromosome (bp)	Anchored rate	Chromosome	Length	Chromosome	Length
2,499,755,479	2,436,957,224	97.49%	1	283,123,735	11	80,778,500
			2	153,537,847	12	62,080,114
			3	135,930,596	13	220,391,114
			4	132,416,539	14	143,097,157
			5	107,564,314	15	141,856,286
			6	173,039,949	16	75,503,400
			7	126,657,800	17	63,961,684
			8	139,843,531	18	51,975,951
			9	140,438,739	X	123,307,586
			10	76,766,902	Y	4,685,480

Table S9. GC contents of the BM chromosome-level genome, Related to Figure 3 and Table 1.

Type	Number (bp)	% of genome
A	718,891,702	28.86
T	719,386,200	28.88
C	518,571,214	20.82
G	518,540,860	20.81
N	15,817,026	0.63
Total (bp)	2,491,207,002	100
GC	1,037,112,074	41.90

Note: GC content of the genome after excluding N nucleotides. The proportions of the four bases A, T, G and C conform with the normal proportions. GC contents account for 41.90% of the genome, and the Ns constitute only 0.63% of the genome.

Table S10. Integrity evaluation of the chromosome-level BM genome assembly according to the read remapping ratio and coverage, Related to Figure 3 and Table 1.

Type1	Type2	Percentage (%)
Short Reads	Mapping rate (%)	97.69
	Average sequencing depth	65.85
Genome	Coverage (%)	98.32
	At least 4×Coverage (%)	98.20
	At least 10×Coverage (%)	98.10
	At least 20×Coverage (%)	97.75

Note: A total of 97.69% of the total high-quality short-insert reads were realigned with the chromosome-anchored genome with 98.32% coverage, suggesting a high level of consistency between the sequencing reads and assembly of the chromosome-level BM genome.

Table S11. The accuracy of the chromosome-scale genome at the single-nucleotide level, Related to Figure 3 and Table 1.

Type	Number	Rate (%)
All SNPs	1,835,478	0.0758
Heterozygous SNPs	1,826,136	0.0755
Homozygous SNPs	9,342	0.0004

Note: To evaluate the accuracy of the BM genome at the single-nucleotide level, all short-insert reads were mapped back to the chromosome-level genome. A total of 7,959 homozygous SNPs (0.0004%) were identified, suggesting high accuracy in the assembly of the chromosome-scale BM genome.

Table S12. Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment of the BM genome, Related to Figure 3 and Table 1.

Species	BUSCO notation assessment result
BM	C: 93.9% [S:93.3%, D:0.6%], F:3.2%, M:2.9%, n:4104

Note: C: complete, [S: single copy], [D: duplicated], F: fragmented, M: missing, n: total BUSCO groups searched. 3,829 (93.9%) of 4,104 complete benchmarking universal single-copy orthologs were assembled. Orthologs dataset of vertebrata_odb9 (Benchmarking Universal Single-Copy Orthologs, BUSCO v3, <https://busco.ezlab.org/>) was used to evaluate integrity degree of assembled genome.

Table S13. Summary of predicted protein-coding genes in the BM genome compared with other representative mammalian genomes, Related to Figure 3.

Species	Number	Average gene length (bp)	Average CDS length (bp)	Average exon number per gene	Average exon length (bp)	Average intron length (bp)
Duroc	20667	44706.36	1702.51	9.44	180.42	5097.27
Human	20320	52095.71	1620.79	9.51	170.44	5931.70
Mouse	22612	38318.08	1550.91	8.77	176.74	4728.93
Dog	19851	38426.08	1643.99	9.53	172.44	4310.13
Cattle	19994	35359.53	1609.58	9.64	167.03	3907.72
Sheep	20921	35305.75	1559.03	9.62	162.13	3916.74
BM	21334	37279.78	1489.06	8.57	173.79	4729.18

Table S14. Number of BM genes with functional classification, Related to Figure 3.

Database	Number annotated	Percentage annotated (%)	
NR	20381	95.5	
Swiss-Prot	19917	93.4	
KEGG	17486	82.0	
InterPro	All	19272	90.3
	Pfam	17271	81.0
	GO	14068	65.9
Annotated	20409	95.7	
Total	21334	-	

Table S15. Repeat sequences in the BM genome determined using various software programs, Related to Figure 3.

Type	Repeat size(bp)	% of genome
TRF	30,766,805	1.248890
RepeatMasker	896,054,192	36.372733
RepeatProteinMask	238,460,691	9.679623
Total	919,268,654	37.315057

Note: "Type" refers to the different software programs used to predict repeat sequences.

Table S16. Transposable elements (TEs) in the BM genome, Related to Figure 3.

Type	Denovo+Rebase		TE Proteins		Combined TEs	
	Length (bp)	% of genome	Length (bp)	% of genome	Length (bp)	% of genome
DNA	7,269,477	0.295083	4590113	0.186322	10,165,012	0.412619
LINE	811,781,584	32.951929	225,800,882	9.165735	830,452,895	33.709837
SINE	38,778	0.001574	0	0.000000	38778	0.001574
LTR	148,223,141	6.016690	8,108,891	0.329157	153,178,127	6.217824
Unknown	7,737,034	0.314063	0	0.000000	7,737,034	0.314063
Total	896,054,192	36.372733	238,460,691	9.679623	908,867,268	36.892842

Note: The genomic TEs were predicted by RepeatModeler and RepeatMasker based on *ab initio* determination and database. The TE libraries predicted by RepeatModeler and Rebase were used to detect the TEs of Denovo+Rebase. TE proteins in the genome were predicted by RepeatProteinMask based on Rebase. Finally, 'Combined TEs' is represent the nonredundant consensus TEs between the above two prediction methods.

Table S17. The distribution and annotation of noncoding RNAs in the BM pig genome, Related to Figure 3.

Type		Copy number	Average length (bp)	Total length (bp)	% of genome
miRNA		8001	95.72	765841	0.031087
tRNA		4369	75.68	330646	0.013422
rRNA	rRNA	248	124.70	30926	0.001255
	18S	18	347.44	6254	0.000254
	28S	72	177.54	12783	0.000519
	5.8S	3	107	321	0.000013
	5S	155	74.63	11568	0.000470
snRNA	snRNA	2076	101.43	210570	0.008547
	CD-box	277	85.28	23622	0.000959
	HACA-box	288	135.16	38927	0.001580
	splicing	1481	96.67	143166	0.005811

Note: The microRNA (miRNA), small nuclear RNA (snRNA) and tRNA sequences located in repeat or gap regions were filtered. rRNAs (< 50 bp) with an identity < 85% were also filtered. The average and total lengths were calculated using the integrated data.

Table S18. Copy number of PERV-derived genes in the Duroc and BM genomes, Related to Figure 3.

Subtype	Duroc	BM
PERV-A- <i>gag</i>	25	6
PERV-A- <i>pol</i>	18	6
PERV-A- <i>env</i>	12	0
PERV-B- <i>gag</i>	26	6
PERV-B- <i>pol</i>	27	9
PERV-B- <i>env</i>	10	6
PERV-C- <i>gag</i>	26	6
PERV-C- <i>pol</i>	18	6
PERV-C- <i>env</i>	9	0

Table S19. Annotation of the SVs between BM and Duroc genomes, Related to Figure 4.

Number of SVs	Percentage	Locations of SVs
257	0.004328567	downstream
622	0.010476142	exonic
37080	0.6245263	intergenic
21092	0.35524565	intronic
4	0.0000673707	splicing
314	0.005288599	upstream
4	0.0000673707	upstream; downstream

Note: (1) downstream: 1 kb downstream of gene; (2) upstream: 1 kb upstream of gene; (3) upstream; downstream: both in 1 kb downstream and 1 kb upstream of gene.

Table S20. Annotation of the SVs located in exonic regions between BM and Duroc genomes, Related to Figure 4.

Number of SVs	Type	Percentage (%)	Number of related genes
284	Deletion	45.66%	914
208	Duplication	33.44%	3252
28	Insersion	4.50%	28
101	Inversion	16.24%	1629
1	Inverted DUP	0.16%	1

Table S21. Genes loated in overlapping exonic regions of SVs between BM and Duroc genomes, Related to Figure 4.

Gene symbol	SV category	SV length	Reference			Supporting read number
			Chromosome	Starting position	Ending position	
<i>ADAM7</i>	Deletion	29561bp	NC_010456.5	8506604	8536165	5
<i>OVOL1</i>	Deletion	62476bp	NC_010444.4	6491564	6554040	5
<i>AQN-1,PSP1,PSP2</i>	Inversion	65033bp	NC_010456.5	132257806	132322839	6
<i>HSP70-2</i>	duplication	10778bp	NC_010449.5	23916445	23927223	5
<i>UBE2B</i>	duplication	63241bp	NC_010444.4	136591883	136655124	25
<i>AHNAK</i>	Deletion	705bp	NC_010444.4	9211343	9212048	6
<i>ADGRF5/GPR116</i>	Deletion	52092bp	NC_010449.5	41671321	41723413	9
<i>ATP10D</i>	Deletion	4821bp	NC_010450.4	37524612	37529433	8

Table S24. Average weights of different visceral organs in adult BMs and Durocs, Related to Figure 5.

Organs	Average organ weight (g)		
	BM (n =15)	Duroc (n =15)	Human (50 kg) (Young et al., 2009)
Liver	1081.5±43.0	1613.1±41.7	1228.5
Heart	251.3±11.2	480±31.2	258.5
Kidney	239.6±18.2	390±37.8	247.5
Spleen	135.7±10.2	211±16.7	135.5
Lung	817±43.5	1095±38.2	805.5

Note: Values represent mean±SD.

Table S25. Analysis of microsatellite genetic diversities of individuals in F11-F19 inbred line of BM family, Related to Figure 6.

Family	Average effective allele number	Average heterozygosity	Average polymorphism information content	Average inbreeding coefficient
F11 (n=7)	1.2709	0.1737	0.1307	0.8263
F13 (n=5)	1.1653	0.1290	0.0985	0.8710
F15 (n=7)	1.1158	0.0882	0.0702	0.9118
F17 (n=11)	1.0739	0.0596	0.0541	0.9378
F19 (n=9)	1.0185	0.0166	0.0157	0.9825

Note: There were a total of 29 alleles at the 19 microsatellite loci, and the number of allele on F11, F13, F15, F17 and F19 groups were 29, 27, 25, 25 and 22, respectively; the average inbreeding coefficients of animals from the 11th, 13th, 15th, 17th 19th generations were 0.8263, 0.8710, 0.9118, 0.9378 and 0.9825, respectively. Seven, five, seven, eleven and nine individuals from each generation were used for microsatellite genetic analysis, respectively.

Table S27. GO enrichment of mRNAs targets of the differentially expressed lncRNAs in the liver, Related to Figure 2.

GO accession	Description	Term type	Over represented p Value
GO:0022402	cell cycle process	Biological process	0.0005942
GO:0016049	cell growth	Biological process	0.00071391
GO:0071840	cellular component organization or biogenesis	Biological process	0.0011208
GO:0072511	divalent inorganic cation transport	Biological process	0.0012766
GO:0005520	insulin-like growth factor binding	Molecular function	0.0015053
GO:0005488	binding	Molecular function	0.0015327

Table S28. GO enrichment of mRNA targets of the differentially expressed lncRNAs in skeletal muscle, Related to Figure 2.

GO accession	Description	Term type	Over represented p Value
GO:0006072	glycerol-3-phosphate metabolic process	Biological process	2.3423e-05
GO:0052646	alditol phosphate metabolic process	Biological process	0.00040261
GO:0006333	chromatin assembly or disassembly	Biological process	0.00046626
GO:0009331	glycerol-3-phosphate dehydrogenase complex	Cellular component	0.00056468
GO:0008219	cell death	Biological process	0.00086296
GO:0016265	death	Biological process	0.00086296

Table S29. KEGG enrichment of mRNA targets of the differentially expressed lncRNAs in the liver, Related to Figure 2.

Term	Database	KO ID	P-Value
PI3K-Akt signaling pathway	KEGG PATHWAY	ssc04151	0.0165623900777
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	KEGG PATHWAY	ssc05412	0.0188879984899
Oxytocin signaling pathway	KEGG PATHWAY	ssc04921	0.0198506827977
Proteasome	KEGG PATHWAY	ssc03050	0.0245177106915
Dilated cardiomyopathy	KEGG PATHWAY	ssc05414	0.0257824691069
Transcriptional misregulation in cancer	KEGG PATHWAY	ssc05202	0.0293816654376

Table S30. KEGG enrichment of mRNA targets of the differentially expressed lncRNAs in skeletal muscle, Related to Figure 2.

Term	Database	ID	P-Value
ErbB signaling pathway	KEGG PATHWAY	ssc04012	0.0159508247002
Oxytocin signaling pathway	KEGG PATHWAY	ssc04921	0.020511339311
AMPK signaling pathway	KEGG PATHWAY	ssc04152	0.0229409555906
Tight junction	KEGG PATHWAY	ssc04530	0.0293978723578
Pertussis	KEGG PATHWAY	ssc05133	0.0353475193857
Epstein-Barr virus infection	KEGG PATHWAY	ssc05169	0.0415830623246

Table S31. GO enrichment of the differentially expressed mRNAs in the liver, Related to Figure 2.

GO accession	Description	Term type	Over represented p Value
GO:0043565	sequence-specific DNA binding	Molecular function	7.3414e-06
GO:0031012	extracellular matrix	Cellular component	5.2445e-05
GO:0008716	D-alanine-D-alanine ligase activity	Molecular function	0.00014597
GO:0043169	cation binding	Molecular function	0.00023738
GO:0046872	metal ion binding	Molecular function	0.00032501
GO:0020037	heme binding	Molecular function	0.00046217

Table S32. GO enrichment of the differentially expressed mRNA of in skeletal muscle, Related to Figure 2.

GO accession	Description	Term type	Over represented pValue
GO:0043565	sequence-specific DNA binding	Molecular function	6.6919e-05
GO:0005634	nucleus	Cellular component	9.3464e-05
GO:0001071	nucleic acid binding transcription factor activity	Molecular function	0.00010936
GO:0003700	sequence-specific DNA binding transcription factor activity	Molecular function	0.00010936
GO:0005576	extracellular region	Cellular component	0.000498
GO:0005520	insulin-like growth factor binding	Molecular function	0.000763

Table S33. KEGG enrichment of the differentially expressed mRNAs in the liver, Related to Figure 2.

Term	Database	KO ID	P-Value
PI3K-Akt signaling pathway	KEGG PATHWAY	ssc04151	1.47555511446e-05
Metabolic pathways	KEGG PATHWAY	ssc01100	0.000404997150282
Chemical carcinogenesis	KEGG PATHWAY	ssc05204	0.00074161616888
Steroid hormone biosynthesis	KEGG PATHWAY	ssc00140	0.00111680207926
Protein digestion and absorption	KEGG PATHWAY	ssc04974	0.00136626411415
Retinol metabolism	KEGG PATHWAY	ssc00830	0.00155083160804

Table S34. KEGG enrichment of the differentially expressed mRNAs in skeletal muscle, Related to Figure 2.

Term	Database	KO ID	P-Value
MAPK signaling pathway	KEGG PATHWAY	ssc04010	0.00209329237638
Estrogen signaling pathway	KEGG PATHWAY	ssc04915	0.00222407814906
p53 signaling pathway	KEGG PATHWAY	ssc04115	0.00248411459897
Complement and coagulation cascades	KEGG PATHWAY	ssc04610	0.0113899481044
Circadian rhythm	KEGG PATHWAY	ssc04710	0.0134645687392
AMPK signaling pathway	KEGG PATHWAY	ssc04152	0.0209932852213

Table S35. Amplification primers for transcriptome qRT-PCR, PERV-C-*env* and 19 microsatellite DNA PCR, Related to Figures 2 and 6.

	Gene symbol	Primers
Transcriptome primer	<i>PGC1-α</i>	F:5'-AGGGACTTGTCTCCGTTG -3' R:5'-AGGGACACTTGTCTCCGTTG-3'
	<i>LNC_000669</i>	F:5'- ATGCCAATGTAGTTTAGGT -3' R:5'- CTGTGGATTTTCTACTGGTCA -3'
	<i>LNC_000159</i>	F:5'- GCTTTCACCCGGTACGCTG -3' R:5'- TGCCAAGTTGTATCCGTGCT -3'
	<i>LNC_000737</i>	F:5'- ATAATAAGGACCACGAGGAC -3' R:5'- CATTTTGCTAACGAAACCAGA -3'
	<i>PEPCK</i>	F:5'-TGTAACCTCTTCTCAACGGGACAC-3' R:5'- TTTCCCCAGTCGGGTCATAATCC -3'
	<i>LNC_000081</i>	F:5'- ACTATGGGTAATAATCCTG -3' R:5'- TATTGCGATACAGTCAAC -3'
	<i>LNC_000576</i>	F:5'- TTGTCTGGTTGTCTCGGGTC -3' R:5'- TCGCACCCCTCGTGAAAAT -3'
	<i>LNC_001236</i>	F:5'- ATATAACAGCCTACTGATGAG -3' R:5'- GCTTTCACCCGGTACGCTG -3'
	<i>GAPDH</i>	F:5'- GCCATCACCATCTTCCAGG -3' R:5'- TCACGCCCATCACAAACAT -3'
	PERV-C- <i>env</i> primer	PERV-C- <i>env</i>
19 microsatellite DNA primer	S0155	F:5'- TGTTCTCTGTTTCTCCTCTGTTTG -3' R:5'- AAAGTGGAAGAGTCAATGGCTAT -3'
	SW240	F:5'- AGAAATTAGTGCCTCAAATTGG -3' R:5'- AAACCATTAAGTCCCTAGCAAA -3'
	S0007	F:5'- TTACTTCTTGGATCATGTC -3' R:5'- GTCCCTCCTCATAATTTCTG -3'
	SW1057	F:5'- TCCCCTGTTGTACAGATTGATG -3' R:5'- TCCAATTCCAAGTCCACTAGC -3'
	S0225	F:5'- GCTAATGCCAGAGAAATGCAGA -3' R:5'- CAGGTGGAAAGAATGGAATGAA -3'
	S0227	F:5'- GATCCATTTATAATTTTAGCACAAAGT -3' R:5'- TGCATGGTGTGATGCTATGTCAAGC -3'
	S0355	F:5'- TCTGGCTCCTACACTCCTTCTTGATG -3' R:5'- TTGGGTGGGTGCTGAAAAATAGGA -3'
	S0090	F:5'- CCAAGACTGCCTTGTAGGTGAATA -3' R:5'- GCTATCAAGTATTGTACCATTAGG -3'
	S0218	F:5'- GTGTAGGCTGGCGGTTGT -3' R:5'- CCCTGAAACCTAAAGCAAAG -3'
	S0226	F:5'- GCACTTTTAACTTTTCATGATACTCC -3' R:5'- GGTAAACTTTTNCCTCAATACA -3'

SW911	F:5'- CTCAGTTCTTTGGGACTGAACC -3' R:5'- CATCTGTGGAAAAAAAAAAGCC -3'
S0005	F:5'- TCCTTCCCTCCTGGTAACTA -3' R:5'- GCACTTCCCTGATTCTGGGTA -3'
SW951	F:5'- TTTCACA ACTCTGGCACCAG -3' R:5'- GATCGTGCCCAAATGGAC -3'
SW857	F:5'- TGAGAGGTCAGTTACAGAAGACC -3' R:5'- GATCCTCCTCCAAATCCCAT -3'
S0002	F:5'- GAAGCCCAAAGAGACA ACTGC -3' R:5'- GTTCTTTACCCACTGAGCCA -3'
S0228	F:5'- GGCATAGGCTGGCAGCAACA -3' R:5'- AGCCACCTCATCTTATCTACAC -3'
S0026	F:5'- AACCTTCCCTTCCCAATCAC -3' R:5'- CACAGACTGCTTTTTACTCC -3'
SW61	F:5'- GAGAGGGATGAGCACTCTGG -3' R:5'- AGAGCATTCCAGGCTTCTCA -3'
SW1119	F:5'- CAACCTCAAAAATGGAGAAAGG -3' R:5'- GTTCTTGCGGTGTTTGGC -3'

Supplemental URLs

Genecards, <http://www.genecards.org>

DisGeNET, <http://www.disgenet.org/web/DisGeNET/>

RepeatMasker, RepeatProteinMask and RepeatModeler, <http://www.RepeatMasker.org>

Solar, <http://treesoft.svn.sourceforge.net/viewvc/treesoft/>

Picard, <http://sourceforge.net/projects/picard/>

Transparent Methods

Ethics approval

Animal breeding and care and all experiments were performed in accordance with “Regulations for the administration of affairs concerning experimental animals of China (CNAS-CL60)”. All experimental procedures used in this study were carried out in accordance with the Experimental Animal Management Regulations (amendment on March 1st, 2017, China).

Establishment of BM strain

Two male and fourteen female primitive BXs were initially used to establish a laboratory inbred line (BM) (SCXK (GUI) 2018-0003) beginning in 1987. In phase 1 (1987-1997), the method of random mating of individuals in the group and occasional mating between half siblings was used to establish a closed colony, under the conditions of limited feeding (Table S1), elimination of descendants with deformity, and directional selection for smaller body size, tamer behaviors (excluding those that jumped over a 1-m-high fence), and uniform black coat color on the head and tail. In phase 2 (1997-), the inbred line (BM) has been generated using consecutive mated brother × sister (known as full siblings) mating interspersed with a low degree of parent × offspring mating sometimes, which has been conducted so far for 19 generations (inbred F19, or bred F29). Breeding was initiated using four male and four female F10 closed colony individuals (as inbred generation 0 (F0)), with a smaller size, tamer behavior, uniform two-end-black coat color but without malformations. All extant individuals in inbred line can be traced to a single ancestral breeding pair.

Detection of resistance to diabetogenic environment

Fifteen male BMs (inbred line F17-F19, 6-month-old), as BM-induced group, and fifteen male Durocs (6-month-old), as Duroc-induced group, were fed with high fat and carbohydrate diet (standard material (Table S1) with 37% sucrose and 10% fat) for 12 months. Sucrose and fat were gradually added to reduce dietary stress and the diet reached the above criteria one month later. Three BMs and three Durocs were selected to be fed with standard material as control groups. All of pigs were fed twice a day and the drinking water was provided for free (Table S1). Single animal was raised in a limited area—a cage of 6m³ for a Duroc and 3.75m³ for a BM, respectively, to limit their activities.

Blood samples were collected every month from the pigs after 16 h of fasting and then centrifuged at 3000 g for 10 min at 4°C to obtain the serum. The fasting blood glucose (FBG) level in serum was detected by an automatic biochemical analyzer using the oxidase method. ELISA was performed to detect the fasting insulin level. To measure the insulin sensitivity of pigs in each group, a simple IVGTT was carried out with the following modifications: after 16 h fasting, 50% glucose (1.2 mL/kg) was injected via the ear vein for 3 min, and blood glucose levels were measured at 0, 3, 10, 30, 60, 90, and 120 min (Zhang et al., 2002). Pancreases, kidney, liver and skeletal muscle tissues were removed immediately, fixed in 10% neutral formalin solution for at least 24 h, embedded in paraffin wax, and then sectioned (4 μm thickness) for histopathological evaluation. Pancreases, kidney, liver and skeletal muscle sections were stained with hematoxylin and eosin using the Hematoxylin and Eosin Staining kit (C0105; Beyotime Institute of

Biotechnology, Haimen, China) and analyzed by light microscopy (CKX41 Optical Microscope; Olympus, Japan). The animals, who have FBG level of > 126 mg/dL (7 mmol/L) (according to human diabetic standards), and abnormal pathology (section of pancreas, kidney, liver and skeletal muscle), increased FINS level (compared to control) and decreased glucose disappearance rate, were considered to be non-resistant to diabetogenic environment.

DNA and RNA isolation for *de novo* sequencing

(1) Preparation of genomic DNA for Illumina paired-end read and mate-pair sequencing. Skeletal muscle was collected from a 1-year-old male BM (19th inbred generation) for genome sequencing. Genomic DNA was extracted using the BioSprint 96 Kit on the BioSprint 96 Workstation (Qiagen, Crawley, UK) according to the manufacturer's protocol. Extracted DNA was quantified using the Qubit 3.0 Fluorometer (Thermo Fisher Scientific, USA) according to the manufacturer's protocol.

(2) Preparation of high molecular weight (HMW) genomic DNA for 10 × Genomics linked reads and PacBio long-read sequencing. HMW DNA was prepared from the abovementioned BM skeletal muscle samples by Amplicon Express (Pullman, WA, USA) using their proprietary protocol for HMW grade (megabase scale) DNA preparation. This protocol involves isolation of cell nuclei and yields pure HMW DNA. This DNA was used to obtain the 10 × Genomics linked reads. Genomic DNA suitable for PacBio long-read sequencing was prepared from BM skeletal muscle samples by Amplicon Express using their proprietary next generation sequencing-grade DNA isolation protocol.

(3) Preparation of genomic DNA for Nanopore sequencing. DNA was extracted from BM skeletal muscle samples using the QIAamp DNA mini kit (Qiagen) for sequencing using the MinION sequencer (Oxford Nanopore Technologies Ltd, Oxford, UK). DNA quality was assessed by running 1 µl on a genomic ScreenTape on the TapeStation 2200 (Agilent Technologies, Santa Clara, CA, USA) to ensure a DNA integrity value > 7 (value for NA12878 was 9.3). The DNA concentration was assessed using the dsDNA HS assay on the Qubit 3.0 Fluorometer (Thermo Fisher Scientific, UK).

(4) Preparation of DNA for Hi-C sequencing. Skeletal muscle tissues were fixed with formaldehyde and then lysed. Crosslinked chromatin was recovered by centrifugation at 13 KRPM in the AccuSpin Micro17 centrifuge (Fisher) and rinsed with 1×TBS buffer. Chromatin was digested overnight with 100 U of either HindIII or NcoI restriction endonuclease (NEB) at 37°C. To enrich for long-range interactions, digested chromatin was centrifuged for 10 min at 13 KRPM, rinsed in 1× NEBuffer 2 (NEB), centrifuged again, and resuspended in 1× NEBuffer 2. Restriction fragment overhangs were filled in using biotinylated dCTP (Invitrogen) and Klenow (NEB) (Van Berkum et al., 2010). The DNA concentration of the chromatin suspension was quantitated using the QuBit fluorometer (Thermo Fisher Scientific, UK), and an 8-mL ligation reaction consisting of a final DNA concentration of 0.5 ng/µL was carried out using T4 DNA Ligase (NEB). Ligation reactions were incubated at room temperature for 4 h and then overnight at 70°C to reverse crosslinks. DNA was purified by a standard phenol/chloroform purification procedure followed by ethanol precipitation and resuspended in water with 1× NEBuffer 2 and 1× BSA (NEB). To remove biotin from unligated DNA ends, T4 Polymerase (NEB) was added to the DNA sample and incubated at 25°C for 10 min followed by 12°C for 1 h. DNA was purified using the DNA Clean and Concentrator-5 Kit (Zymo Research) and then physically sheared to a length of 300–500 bp for construction of the Hi-C library (Van Berkum et al., 2010; Bickhart et al.,

2017).

(5) RNA isolation for transcriptome sequencing. BM tissues (brain, liver, heart, spleen, lung, kidney, pancreas, stomach, skeletal muscle and adipose) were harvested in liquid nitrogen and stored at -80°C until extraction. Total RNA was extracted from the tissues using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. RNA degradation and contamination were monitored by 1 % agarose gel electrophoresis. RNA purity was assessed using the NanoPhotometer spectrophotometer (Implen, Los Angeles, CA, USA). RNA was quantified using a Nanodrop® 2000 spectrophotometer (ThermoFisher Scientific, UK). RNA integrity was assessed using an RNA Nano 6000 Assay Kit with the Bioanalyzer 2100 system (Agilent Technologies, Santa Clara, CA, USA).

Genome sequencing, assembly and annotation of the BM reference genome

A one-year-old male BM (19th inbred generation) from BM breeding center, Guangxi University, was used to isolate DNA and RNA for genome sequencing and annotation. To obtain a high-quality BM genome assembly, we adopted a combination of sequencing methods including Illumina paired-end and mate-pair sequencing, $10 \times$ Genomics linked reads, single molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio), MinION reads (Oxford Nanopore sequencing Technology) and Hi-C data. As detailed in Table S5, a total of ~ 932 Gb sequencing data (equivalent to $361 \times$ genomic coverage, based on an estimated genome size of 2.58 Gb) were generated.

DNA fragments longer than 50 kb were used to construct one GeMcode library using the Chromium instrument ($10 \times$ Genomics, Pleasanton, CA, USA). This library was sequenced on the Illumina HiSeqX platform to produce 2×150 bp reads, producing a total of ~ 185 Gb of $10 \times$ Chromium library sequencing data. Furthermore, short-insert (250 bp and 500 bp, paired-End) and long-insert (2, 5 and 10 kb, mate-Pair) DNA libraries were constructed using standard Illumina library prep protocols and sequenced on the Illumina HiSeqX platform as 2×150 bp reads, producing a total of ~ 394 Gb Illumina library sequencing data. In addition, we constructed the 20 kb PacBio libraries using the BluePippin™ Size-Selection System recommended by Pacific Biosciences. In total, 20 μg DNA were sheared to ~ 20 kb fragments by ultrasonication (Covaris, Woburn, MA, USA) to construct the libraries. The quality of the sheared DNA was examined by FEMTO Pulse pulse field capillary electrophoresis (Advanced Analytical Technologies, Inc.). The sheared DNA was filtered using AMPure PB paramagnetic beads (Beckman Coulter Inc., Beverly, MA, USA) with a recovery rate of 80%. The constructed libraries were sequenced using the Sequel system, and a total of six SMRT cells were used to yield ~ 53 Gb sequencing data, including 5.5 million clean subreads with an average length of 9.61 kb and an N50 of 15.826 kb. Moreover, sequencing libraries were prepared using the Oxford Nanopore Technologies sequencing kit SQK-LSK108 and sequenced by the MinION sequencer to produce a total of ~ 26 Gb clean data by using MinION flow cells, with a read N50 of 10.6 kb. Finally, Hi-C libraries were constructed from the purified DNA (van Berkum et al., 2010) using reagents from the Illumina Mate Pair Sample Preparation Kit. Paired-end sequencing was performed using the Illumina HiSeq X platform to yield a total of 273.7 Gb. All libraries were constructed and sequenced at Novogene (Tianjing, China). To remove the negative effects of sequencing, short-insert read (250 and 500 bp), long-insert read (2, 5, and 10 kb), $10 \times$ Genomics, and Hi-C data were filtered according to the following criteria: reads with (1) low-quality bases ($> 50\%$

bases with Q-value ≤ 8), (2) a rate of N bases $> 10\%$, and (3) adaptor contamination were removed.

We estimated the genome size of BM using K-mer frequency analysis with a K-mer size of 17 (Marçais and Kingsford, 2011). The estimated size of the BM genome is 2,583.64 Mb (~ 2.58 Gb) (Table S4 and Figure S4). Given the challenges with this large animal genome, we adopted a hybrid assembly strategy for the project. Contiguous scaffolding is essential for capturing the whole gene structure to allow downstream analyses, such as genome annotation. Genome scaffolding relies on long DNA fragments, and in recent years, the application of barcoded linked reads from the 10X Genomics platform has begun to replace earlier scaffolding methods that rely on mate-pair data. The third-generation sequencing platforms PacBio and Nanopore provide long reads spanning repeat-rich genomic regions and ensure longer sequence continuity. Hi-C is an adaptation of the chromosome conformation capture (3C) methodology (Dekker et al., 2002) that identifies long-range chromosome interactions in an unbiased fashion without *a priori* target site selection. The frequency of long-range consensus interactions decreases rapidly as the linear distance along a chromosome increases, allowing Hi-C data to scaffold assembled contigs at the scale of full chromosomes (Burton et al., 2013). HMW DNAs offer long fragments up to 1 Mb in length (Murchison et al., 2012), which help produce a number of high-quality and contiguous assemblies (Seo et al., 2016; Hulsekemp et al., 2018; Mostovoy et al., 2016; Avni et al., 2017; Lu et al., 2015; Weisenfeld et al., 2017). Here, we present a *de novo* assembly using all of the technologies described above. Filtered $10 \times$ genomic data were used to assemble the initial genome sketch using Supernova-v1.1.3 (Weisenfeld et al., 2017), and long-insert reads (2 kb, 5 kb and 10 kb) were used to generate longer genomic fragments using SSPACE-v3.0 (Li et al., 2010). In brief, the assembly was performed according to the following steps:

- (1) We used a de Bruijn graph approach (Pevzner et al., 2001), adopting the method of DISCOVAR (Weisenfeld et al., 2014). k-mers ($k = 48$) were prefiltered to remove those present in only one barcode. The remaining k-mers were used to construct an initial directed graph, in which edges represent unbranched DNA sequences and abutting edges overlap by $k-1$ bases.
- (2) We then used the read pairs to effectively increase k to approximately 200, so that the new graph represents an approximation of what would be obtained by collapsing the true sample genome sequence along identical 200-base sequences, thus achieving considerably greater resolution.
- (3) We decomposed the graph into units called lines that are extended linear regions, punctuated only by “bubbles.” We used these lines to scaffold the assembly graph, determining the relative order and orientation of two lines, then breaking the connections at their ends, and inserting a special “gap” edge between the lines. The end result is a new line, which has a special “bubble” consisting only of a gap edge.
- (4) Scaffolding was carried out using read pairs initially. If the right end of one line is unambiguously connected by read pairs to the left end of another line, they can be connected. Read pairs can span short gaps. To scaffold across larger gaps, we used barcodes. Briefly, if two lines are actually near each other in the genome, then with high probability, multiple molecules (in the partitions) bridge the gap between the two lines. Therefore, for any line, we may find candidate lines nearby by looking for other lines sharing many of the same barcodes. By scoring the alternative orders and orientations of these lines, we can scaffold the lines by

choosing their most likely configuration, excluding short lines whose positions are uncertain.

(5) The next step in the assembly process is to phase the lines. First, for each line, we found all of its simple bubbles, i.e., bubbles having just two branches. Then, we defined a set of molecules. These are defined by a series of reads from the same barcode, incident upon the line and without very large gaps (> 100 kb) between successive reads. A given molecule then “votes” at certain bubbles, and the totality of this voting (across all molecules on each line) is then used to identify phaseable sections of the line, which are then separated into “megabubble” arms.

(6) Finally, we used Illumina long-insert reads (2, 5, and 10 kb) to elongate contigs and generate the draft genome.

The $10\times$ assembly consists of 2.46 Gb and is highly contiguous with scaffold N50 at 21.1 Mb and with contig N50 at 258 kb. After the primary scaffold assembly using 10X genomics data and Illumina long-insert reads, the next assembly was subjected to PBJelly from PBSuite v15.8.24 (English et al., 2012) using the original PacBio and Nanopore long reads to resolve remaining scaffold gaps, and short-insert (250 and 500 bp) reads were used to correct the inaccurate bases in the genome introduced by PacBio and Nanopore data using pilon-v1.22 (Walker et al., 2014; Bickhart et al., 2017). Finally, we generated reasonably accurate chromosome-scale *de novo* assemblies for the BM pig genome by combining the above improved scaffolds and the Hi-C data using LACHESIS (Burton et al., 2013). LACHESIS functions in three steps. In the first step, LACHESIS uses hierarchical agglomerative clustering to group scaffolds that are likely derived from the same chromosome, exploiting the fact that intrachromosomal contacts are on average more probable than interchromosomal contacts in Hi-C datasets (Lieberman-Aiden et al., 2009). An average-linkage metric (Eisen et al., 1998) is used for this clustering, with linkage defined as the normalized density of Hi-C read-pairs linking any given pair of scaffolds. The final number of groups is prespecified, ideally set to the expected number of chromosomes. In the second step, LACHESIS orders scaffolds linearly within each chromosome group by taking advantage of the higher Hi-C link densities expected between closely located scaffolds. For each chromosome group, a graph is built with vertices representing scaffolds and edge weights corresponding to the inverse of the normalized Hi-C linkage density between pairs of scaffolds. A minimum spanning tree is found in this graph, and the longest path in this tree is extracted as the ‘trunk’, an incomplete but high-confidence ordering of scaffolds within each chromosome group. To generate full ordering, scaffolds excluded from the trunk are reinserted into it at sites that maximize the amount of linkage between adjacent scaffolds. In the third step, the ordered scaffolds are oriented with respect to one another by taking into account precisely where the Hi-C reads map on each scaffold. For each chromosome group, a weighted directed acyclic graph is built representing all possible ways to orient the scaffolds, given the predicted order. The weights are calculated as the log-likelihood of the observed Hi-C links between adjacent scaffolds in a given combined orientation, assuming that the probability of a link connecting two reads at a genomic distance of x decays as $1/x$ for $x \geq \sim 100$ kb (Lieberman-Aiden et al., 2009). The maximum likelihood path in this graph yields a predicted orientation for each scaffold.

Based on the primary assembly, we incorporated the PacBio, Nanopore, and Hi-C data to generate a chromosome-level genome with 2.49 Gb and contig N50 = 1.01 Mb and scaffold N50 = 140.4 Mb. The proportion of N bases in the genome is only 0.6%, and the total bases of ~ 2.25 Gb (97.8%) are anchored to the 20 chromosomes (18 autosomes + XY). We calculated the GC content

and sequencing depth of the assembled BM pig genome in a 10-kb nonoverlapping sliding window. Only one island was observed, indicating no obvious GC separation (Figure S5). To evaluate the accuracy of the BM pig genome at the single-nucleotide level, we realigned the ~169.5 Gb high-quality reads from the short-insert libraries (250 and 500 bp) with the chromosome-level genome assembly using the package BWA-v0.7.8 (Li and Durbin, 2009). Then, we performed SNP calling using the package Samtools-v0.1.19 (Li et al., 2009) and finally obtained ~1.8 M heterozygous and 24,838 homozygous SNPs for the BM pig genome with high confidence (i.e., a coverage depth ≥ 4 and ≤ 150 , a genotype quality ≥ 20 , copy number ≤ 2 , and distance between adjacent SNPs ≥ 5) (Table S11), which represents a homozygous SNP rate of 0.001% in the BM pig genome. The completeness of the BM genome assembly was further assessed using the BUSCO (database: OrthoDBv9 Vertebrata) methods (Hu et al., 2017; Table S12).

After genome assembly, we performed repeat annotation for the BM pig genome (Figure S8). Transposable elements (TEs) were identified by using *de novo*-based and Rebased-based methods and the total process was split into three steps. Firstly, three *de novo*-based softwares of LTR_FINDER, RepeatScout and RepeatModeler were used to identify TEs in BM genome. Secondly, RepeatMasker-v3.3.0 took above three *de novo* library results and Rebase-rb20170127 as dataset to predict TE set. Thirdly, RepeatProteinMask- (Supplemental URLs) were performed WU-BLASTX against the TE protein database. Finally, TE set predicted RepeatMasker and TE set predicted RepeatProteinMask were combined as final TE set (Table S16). Tandem Repeat Finder-v4.07b (Benson, 1999) was used to predicted tandem repeats. Above three results of Tandem Repeat Finder, RepeatMasker and RepeatProteinMask were combined to calculate repeat ratio of BM genome, respectively. Finally, nonoverlapped repeat dataset was obtained by filtering overlapped results in combined repeat dataset (Table S15).

To optimize the genome annotation, the transcriptomes of 10 tissues (heart, stomach, fat, kidney, brain, lung, liver, pancreas, muscle and spleen) were performed on Illumina platform. The genes' structure in the BM genome were predicted by using three ways: homology-based, *de novo*-based and transcriptome-based methods. (a) Homology-based prediction. Protein sequences from six related species (Duroc, human, mouse, dog, cattle and sheep) downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/>) were regarded as homologous database to align to BM genome by TBLASTn-v2.2.26 with an E-value cutoff of $1e-5$. With above protein alignments, GeneWise-v2.4.1 was used to predict accurate spliced position of coding regions. (b) *De novo* prediction. We used three *ab initio* predicted softwares—Augustus-v3.2.3 (Stanke and Waack, 2003), GlimmerHMM-v3.0.1 (Majoros et al., 2004) and SNAP-2013-11-29 (Korf, 2004) to perform *de novo* predication. (c) Transcriptome-based prediction. In our annotation protocol, transcriptome data were used by two different ways. In one hand, RNA data of ten tissues were performed to *de novo* assembly by TRINITY-v2.1.1 with parameters set to “--seqType fq --normalize_reads --full_cleanup --min_glue 2 --min_kmer_cov 2 --KMER_SIZE 25”. Then unigenes from assembly results were used to predict gene structure with PASA-r20140417. In the other hand, RNA data of ten tissues were mapped into BM genome to obtain some evidences as gene models using TopHat-v2.0.13 (Kim et al., 2013) and Cufflinks-v2.1.1 (Trapnell et al., 2010). The final nonredundant reference gene set was generated by merging genes predicted by above three ways in (a), (b) and (c) using EvidenceModeler-v1.1.1 (Haas et al., 2008), genes encoding ≤ 50 amino acids, or with only *de novo* predictive support were removed. The final reference gene

set of the BM comprised of 21,334 genes which is comparable with the gene repertoire of the Duroc genome (20,667 genes) (Table S13).

Gene functions were assigned according to the best alignments in the SwissProt and NR databases (Bairoch and Apweiler, 1999) using BLASTP (Kent, 2002). We annotated motifs and domains using InterPro (Mulder and Apweiler, 2007) by searching against publicly available databases, including Pfam (Punta et al., 2012), using InterProScan (Mulder and Apweiler, 2007). Gene ontology (GO) terms (Ashburner et al., 2000) for each gene were retrieved from the corresponding InterPro descriptions (Table S14). Furthermore, we mapped these BM genes to the KEGG pathway (Kanehisa and Goto, 2000) to identify the best-match category for each gene. tRNA genes were predicted by tRNAscan-SE (Lowe and Eddy, 1997) using eukaryote-specific parameters. rRNA, microRNA (miRNA), and small nuclear (snRNA) genes were identified using the Infernal software (Nawrocki et al., 2009) by searching against the Rfam database (Griffiths-Jones et al., 2005) using the default parameters (Table S17). In addition, we filtered the miRNAs, snRNAs, and tRNAs that were located in the repeat or gap regions, as well as the rRNAs of short length (≤ 50 bp) and low identity ($\leq 85\%$).

Comparative genomic analysis

All CDS and protein data from the BM, Duroc, sheep, cattle, camle, horse, macaque, dog, mouse, human and opossum were downloaded from NCBI database. For genes with alternative splicing variants, we chose the longest transcripts (≥ 30 amino acids) to represent the genes. We used the TreeFam methodology (Li et al., 2006) to define a gene family as a group of genes that descended from a single gene of the last common ancestor of the considered species. An all-against-all BLASTP (Kent, 2002) was applied to determine the similarities among the genes of eleven mammalian genomes with an e-value of $1e-7$ and conjoined fragmental alignments for each gene pair by Solar (Figures S12 and S14; Supplemental URLs).

Genome analysis of SVs

To compare the differences between Duroc and BM genomes, $\sim 20 \times$ single-molecule sequencing data from BMs were mapped to the Duroc genome (Sscrofa11.1) to detect SVs in BMs. NGMLR software (Sedlazeck et al., 2018) was used to locate SVs, and ANNOVAR software (Wang et al., 2010) was used to annotate the SVs. Then, SVs located within genes were extracted to determine sequence similarities using Miropeats software (Parsons, 1995).

Cross-species comparisons of gene sets

We extracted orthologous genes among human, mouse, and BM/Duroc pig genomes to examine the divergence among BM/Duroc pig genes and their human and mouse counterparts. MUSCLE software was used to align the protein sequences from humans, mice and BM/Duroc pigs in triplicate, and gaps within aligned sequences were removed using Gblocks (Talavera and Castresana, 2007). Point accepted mutation (PAM) unit estimated by pairwise maximum likelihood distance estimation was implemented using TKF package (Thorne et al., 1991) in R. To further determine the conservation between BM and Duroc pig genes involved in eight common human diseases (Obesity, type 2 diabetes mellitus, Nonalcoholic fatty liver disease, Atherosclerosis, Parkinson's disease, Huntington's disease, Alzheimer's disease, Amyotrophic lateral sclerosis), we obtained gene lists associated with human diseases from GeneCards and

extracted BM and Duroc protein sequences from PAM results according to shared gene names for subsequent comparison (Figure S14).

Phylogenetic tree construction and identification of gene family expansion and contraction

We constructed a phylogenetic tree based on the sequences from 11 species (BM, Duroc, sheep, cattle, camel, horse, macaque, dog, mouse, human and opossum) using maximum likelihood analysis of concatenated alignments of 7,928 single-copy orthologous protein-coding gene sequences shared with their genomes. MUSCLE software was used to generate multiple sequence alignments, and RAxML (Luo et al., 2015) were used to optimize and reconstruct the phylogenetic tree. We determined the expansion and contraction of the gene orthologue clusters by comparing the cluster size differences between the ancestor species and BM, Duroc, sheep, cattle, camel, horse, macaque, dog, mouse, human and opossum, respectively using the Café program (De Bie et al., 2006). P-value ≤ 0.05 was selected as cutoff value and the number of randomizations was set to be 10000, and parameter λ value was searched globally.

Detection of PSGs

To detect PSGs, we extracted 11,042 single-copy orthologous proteins shared among BM, Duroc, and human for multiple sequence alignments using MUSCLE software and branch-site models and likelihood ratio tests (LRTs) as implemented in PAML software (Yang, 2007). To control false discovery rate, we employed strict filtering standards for the results detected by PAML as follows: (1) we extracted positively selected sites and the two adjacent amino acid sites previous and next them and removed the candidate sites if there is any gap in the three amino acid sites; (2) the candidate sites were removed if the amino acids at the previous one or next one site of positively selected sites, differs in human, BM and Duroc. Finally, 789 BM genes of and 990 Duroc genes as forward branch were obtained. GO and KEGG pathway enrichment analysis of positive selected genes were performed using EnrichPipeline (Chen et al., 2010).

Whole-genome re-sequencing and SNP calling

We sampled a total of 100 pigs, including 50 BMs and 50 BXs. Genomic DNA was extracted from blood samples from the 50 individuals of each breed using the DNeasy Blood & Tissue Kit (Qiagen). Pool-seq was performed on the Illumina HiSeq platform, generating a total of 325.01 Gb paired-end DNA sequencing data. Similarly, the blood of 10 BMs and 10 Durocs from the individuals in Table S2 were selected to extract genomic DNA that was used to generate about 545.13 Gb Illumina paired-end sequencing data.

The criteria for quality checking and filtering of the sequences were also applied. Consequently, 324.24 Gb (from BMs and BXs) and 544.98 Gb (from BMs and Durocs) high-quality paired-end reads were mapped to the Duroc genome assembly (Sscrofa11.1) using BWA software (Li and Durbin, 2009). The command 'aln -o 1 -e 10 -t 4 -l 32 -i 15 -q 10' was used to find the suffix array coordinates of good matches for each read, and the best alignments were generated in the SAM format given paired-end reads with command 'sampe'. Next, we improved the alignment results using the following three steps: (a) the alignment reads were filtered according to mismatches ≤ 5 and mapping quality = 0; (b) the alignment results were corrected using the package Picard (Supplemental URLs) with two core commands; and (c) potential PCR duplications were removed. If multiple read pairs had identical external coordinates, only the pair with the highest mapping

quality was retained.

After alignment, we performed SNP calling at the population-scale for the two groups (50 BMs and 50 BXs; 10 BMs and 10 Durocs) using a Bayesian approach as implemented in the package SAMtools (Li et al., 2009). The genotype likelihoods from reads for each individual at each genomic location were calculated, and the allele frequencies were also estimated. The ‘mpileup’ command was used to identify SNPs using the parameters ‘-q 1 -C 50 -S -D -m 2 -F 0.002 -u’. Then, only the high-quality SNPs (coverage depth 4–1,000, RMS mapping quality ≥ 20 , distance between adjacent SNPs ≥ 5 bp, and rate of missing samples within each group $< 50\%$) were retained for the subsequent analysis.

Microsatellite DNA analysis

To investigate the genetic structure of the inbred BM genome from generations F11 to F19, 39 pigs and 19 microsatellite markers (Table S35), recommended by food and agriculture organization of the united nations/international society for animal genetics (FAO/ISAG), were used to obtain short tandem repeat/microsatellite DNA typing data using the ABI 3730XL DNA analyzer (ABI, USA). Then, we used Popgene32 (v1.31) to interpret the short tandem repeat typing data to analyze indicators of genetic diversity, including effective allele number, heterozygosity, polymorphic information content, and inbreeding coefficient.

Selective sweep for BMs and BXs

Allele counts and frequencies for all SNP positions were used to search signatures in the BM pig genome that have undergone selection in the closed colony and inbreeding processes compared with BXs, which live outdoors. First, pooled heterozygosity (H_p) values were calculated in 100-kb sliding windows with a 50-kb step size for pooled resequencing samples from BXs and BMs, respectively, following a previously reported method (Rubin et al., 2010). The H_p value of every window was estimated using the following formula:

$$H_p = 2 \frac{\sum n_{MAJ} \sum n_{MIN}}{(\sum n_{MAJ} + \sum n_{MIN})^2}$$

where n_{MAJ} is the read count from the major allele for each SNP position, $\sum n_{MAJ}$ is the sum of the total reads from the major alleles for all SNPs in every window, n_{MIN} is the read count from the minor allele for each SNP position, and $\sum n_{MIN}$ is the sum of the total reads from the minor alleles for all SNPs in every window. In-house Perl scripts were used to calculate the H_p value according to the above formula.

Second, the value of the fixation index (F_{st}) was also calculated for two pooled populations using in-house Perl scripts and the following formula (Weir and Hill):

$$F_{st} = \frac{\sum_{k=1}^K N_k}{\sum_{k=1}^K D_k}$$

where

$$N_K = \left(\frac{a_1}{n_1} - \frac{a_2}{n_2} \right)^2 - \frac{h_1}{n_1} - \frac{h_2}{n_2}$$

$$D_K = N_K + h_1 + h_2$$

$$h_i = \frac{a_i(n_i - a_i)}{n_i(n_i - 1)} \quad (i = 1, 2)$$

where k is the individual SNP; i is population 1 or population 2; a_1 and a_2 are the number of reads for allele 1 in population 1 and population 2, respectively; and n_1 and n_2 are the total number of reads for each SNP position in population 1 and population 2, respectively.

Afterwards, selective sweep tests were performed to identify those genes that had experienced selection using a combination of F_{st} values and H_p results. We first calculated the ratio between the H_p value of BXs and that of BMs using the formula: $H_p \text{ ratio} = H_p_BX_{pig}/H_p_BM_{pig}$. Then, the \log_2 H_p ratio and F_{st} were calculated to identify the candidate genes. Some maximum (≥ 0.81) and minimum (≤ 0.07) \log_2 H_p ratio values and the top maximum F_{st} values (≥ 0.49) were used to identify those selected genes belonging to BMs and BXs, respectively.

Selective sweep for BMs and Durocs

Each allele counts (SNP-REF and SNP-ATL) of total SNP positions were used to calculate each allele frequency, then allele frequencies were used to search signatures of BMs compared with Durocs. Heterozygosity (H_p) value and fixation index (F_{st}) were calculated as the same the above method. Subsequently, selective sweep regions were also identified using a combination of F_{st} values and H_p results.

RNA-seq analysis

The liver and skeletal muscle tissues from each individual in BM-induced group ($n=15$, at month 12) and Duroc-induced group ($n=15$, at month 12) (Table S2) were collected for RNA-seq. The total RNA (excluding ribosomal RNA) of liver tissues/skeletal muscles were extracted and then pooled as 3 portions in equimolar quantities, one portions for 5 RNAs.

A total of 3 μg RNA per sample was used as input material for the RNA sample preparations. First, ribosomal RNA was removed using the Epicentre Ribo-zero™ rRNA Removal Kit (Epicentre, USA), and the rRNA-free RNA was cleaned up by ethanol precipitation. Subsequently, sequencing libraries were generated using rRNA-depleted RNA with the NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina® (NEB) following the manufacturer's recommendations. Briefly, fragmentation was carried out using divalent cations under an elevated temperature in NEBNext First Strand Synthesis Reaction Buffer (5X). First-strand cDNA was synthesized using random hexamer primers and M-MLV Reverse Transcriptase (RNase H-). Second-strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H. In the reaction buffer, dTTP was replaced with dUTP among the dNTPs. The remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. After adenylation

of the 3' ends of the DNA fragments, an NEBNext Adaptor with a hairpin loop structure was ligated in preparation for hybridization. To select cDNA fragments of 150–200 bp in length, the library fragments were purified using the AMPure XP system (Beckman Coulter). Then, 3 μ L USER Enzyme (NEB) were used with size-selected, adaptor-ligated cDNA at 37°C for 15 min followed by 5 min at 95°C before PCR. PCR was then performed with Phusion High-Fidelity DNA polymerase, Universal PCR primers, and Index (X) Primer. Finally, the products were purified (AMPure XP system), and library quality was assessed on the Agilent Bioanalyzer 2100 system. The clustering of the index-coded samples was performed on the cBot Cluster Generation System using the TruSeq PE Cluster Kit v3-cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the libraries were sequenced on the Illumina HiSeq platform, and 150 bp paired-end reads were generated.

Raw data were first processed using in-house Perl scripts. In this step, clean data were obtained by removing reads containing adapters, reads containing > 10% N bases, and low-quality reads (> 50% of bases whose Phred scores were < 5%) from the raw data. The Phred scores (Q20, Q30) and GC content of the clean data were calculated. All subsequent analyses were based on high-quality data.

The BM pig genome and gene model annotation files were used as direct references. The index of the reference genome was built using Bowtie v2.0.6 (Langmead et al., 2009), and paired-end clean reads were aligned to the reference genome using TopHat v2.0.9 (Trapnell et al., 2012). The mapped reads of each sample were assembled using both Scripture (beta2) (Guttman et al., 2010) and Cufflinks (v2.1.1) (Trapnell et al., 2012) via a reference-based approach. Scripture was run using the default parameters. Cufflinks was run using 'min-frags-per-transfrag=0' and '-library-type fr-firststrand', and other parameters were set as the defaults. In the present study, single-exon lncRNAs were filtered out from the BM pig genome due to limitations of the algorithm. This operation, in which at least two exons are preferred, is a purely technical one. To avoid false-positive results as much as possible, the transcripts containing a single exon were usually considered as background transcripts and were discarded, whereas multi-exon lncRNAs were retained (Prensner et al., 2011).

Cuffdiff (v2.1.1) was used to calculate the fragments per kb per million reads (FPKM) of both lncRNAs and coding genes in each sample. Differential expression analysis in the two groups (three biological replicates per condition) was performed using the DESeq R package. For biological replicates, transcripts or genes with an adjusted P value of < 0.05 were defined as differentially expressed between the BM-induced and Duroc-induced groups.

To achieve high-quality data, we used four analytic tools, including CNCI (v2) (Sun et al., 2013), CPC (0.9-r2) (Kong et al., 2007), Pfam-scan (v1.3) (Punta et al., 2011), and PhyloCSF (v20121028) (Lin et al., 2011) to identify candidate lncRNAs. Transcripts predicted to have coding potential by any of these four tools were removed, and those without coding potential were retained. Then, we selected those commonly identified by the four tools as the final candidate lncRNAs and included them in further analyses. Quantification of gene expression levels was estimated by calculating the FPKMs of the transcripts. To investigate the sequence conservation of the transcripts, we used the phyloFit program in the Phast (v1.3) package (Siepel et al., 2005) to compute phylogenetic models for conserved and nonconserved regions among species. Then, we used phastCons to compute a set of conservation scores for the lncRNAs and coding genes.

To explore the function of lncRNAs, we first predicted the target genes of lncRNAs regulated in

cis and trans. Cis regulation involves lncRNAs acting on neighboring target genes. We searched for coding genes 10 kb/100 kb upstream and downstream of lncRNAs and then analyzed their functions. Trans-regulated target genes of lncRNAs were identified by gene expression analyses. While there were no more than 25 samples, we calculated the expression correlations between lncRNAs and coding genes using custom scripts. Otherwise, we clustered the genes from different samples using WGCNA (Langfelder and Horvath, 2008) to search for common expression modules. Then, we performed functional enrichment analysis of the lncRNA target genes using the EnrichPipeline (Chen et al., 2010). Significance was assessed by P values, which were calculated using the EASE score (P value < 0.05 was considered indicative of significance).

Quantitative real-time PCR analysis

Total RNAs from liver and skeletal muscle in BM-induced and Duroc-induced groups were used for quantitative real-time PCR analysis. Briefly, the first cDNA strains were obtained using a One Step cDNA Synthesis Kit (Bio-Rad, USA), and were then subjected to quantification of mRNAs or lncRNAs with *GAPDH* as an endogenous control using a standard SYBR Green PCR kit (Bio-Rad) on the Bio-Rad CFX96 Touch™ Real-Time PCR Detection System. The quantitative PCR was performed using the following conditions: 95 °C for 30 s, 40 cycles of 95 °C for 5 s, and the optimized annealing temperature for 30 s. The primers and annealing temperatures for 9 genes are listed in Table S35. All reactions were performed in triplicate for each sample. Gene expressions were quantified relative to *GAPDH* expression using the comparative cycle threshold (Δ CT) method. Differences in gene expressions between the two groups were detected by using Student *t* test.

Data Availability

This genome project has been registered in NCBI under the BioProject (<http://www.ncbi.nlm.nih.gov/bioproject>) accession PRJNA478804. The sequencing data of BM have been deposited in NCBI Sequence Read Archive (SRA; <http://www.ncbi.nlm.nih.gov/sra>) under the accession numbers of SRR7760074 to SRR7760081, SRR7760083 to SRR7760090, SRR7760103 to SRR7760123 and SRR7759992-SRR7759997, for genomic data; and SRR7760082, SRR7760091, SRR7760092, SRR7760095-SRR7760099, SRR7760101, SRR7760102 and SRR8490196-SRR8490207 for transcriptomic data; and SRR8449986-SRR8449987 and SRR8523736-SRR8523756 for genome resequencing data. The assembled whole-genome sequences have been submitted to NCBI GenBank (<http://www.ncbi.nlm.nih.gov/genbank>) under accession numbers SIDA00000000.

Supplemental References

- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., and Eppig, J.T. (2000). Gene ontology: tool for the unification of biology. *Nature genetics* 25, 25.
- Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S.O., Gundlach, H., Hale, I., Mascher, M., Spannagl, M., and Wiebe, K. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* 357, 93-97.
- Bairoch, A., and Apweiler, R. (1999). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Research* 28, 49--54.
- Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* 27, 573-580.
- Bickhart, D.M., Rosen, B.D., Koren, S., Sayre, B.L., Hastie, A.R., Chan, S., Lee, J., Lam, E.T., Liachko, I., and Sullivan, S.T. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics* 49, 643-650.
- Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology* 31, 1119.
- Chen, S., Yang, P., Jiang, F., Wei, Y., Ma, Z., and Kang, L. (2010). De novo analysis of transcriptome dynamics in the migratory locust during the development of phase traits. *PloS one* 5, e15633.
- De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269-1271.
- Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science* 295, 1306-1311.
- Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* 95, 14863-14868.
- English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., and Worley, K.C. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PloS one* 7, e47768.
- Griffithsjones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research* 33, 121-124.
- Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., and Nusbaum, C. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology* 28, 503-510.
- Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome biology* 9, R7.
- Hu, Y., Wu, Q., Ma, S., Ma, T., Shan, L., Wang, X., Nie, Y., Ning, Z., Yan, L., and Xiu, Y. (2017). Comparative genomics reveals convergent evolution between the bamboo-eating giant and red

- pandas. *Proceedings of the National Academy of Sciences* *114*, 1081-1086.
- Hulsekemp, A.M., Maheshwari, S., Stoffel, K., Hill, T.A., Jaffe, D., Williams, S.R., Weisenfeld, N., Ramakrishnan, S., Kumar, V., and Shah, P. (2018). Reference quality assembly of the 3.5-Gb genome of *Capsicum annuum* from a single linked-read library. *Hortic Res* *5*, 4.
- Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* *28*, 27-30.
- Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. *Genome Research* *12*, 656-664.
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* *14*, R36.
- Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic acids research* *35*, W345-W349.
- Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* *9*, 559.
- Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* *10*, R25.
- Li, H., Coghlan, A., Ruan, J., Coin, L.J., Hériché, J.K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., and Bolund, L. (2006). TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Research* *34*, D572.
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map (SAM) Format and SAMtools. *Bioinformatics* *25*, 1653-1654.
- Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., and Kristiansen, K. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* *20*, 265-272.
- Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragozy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., and Dorschner, M.O. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science* *326*, 289-293.
- Lin, M.F., Jungreis, I., and Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* *27*, i275-i282.
- Lu, F., Romay, M.C., Glaubitz, J.C., Bradbury, P.J., Elshire, R.J., Wang, T., Li, Y., Li, Y., Semagn, K., and Zhang, X. (2015). High-resolution genetic mapping of maize pan-genome sequence anchors. *Nature Communications* *6*, 6914.
- Luo, Y.J., Takeuchi, T., Koyanagi, R., Yamada, L., Kanda, M., Khalturina, M., Fujie, M., Yamasaki, S., Endo, K., and Satoh, N. (2015). The *Lingula* genome provides insights into brachiopod evolution and the origin of phosphate biomineralization. *Nature Communications* *6*.
- Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* *20*, 2878-2879.
- Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* *27*, 764-770.
- Mostovoy, Y., Levysakin, M., Lam, J., Lam, E.T., Hastie, A.R., Marks, P., Lee, J., Chu, C., Lin, C., and

- Džakula, Ž. (2016). A hybrid approach for de novo human genome sequence assembly and phasing. *Nature Methods* *13*, 587-590.
- Mulder, N., and Apweiler, R. (2007). InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods in Molecular Biology* *396*, 59.
- Murchison, E.P., Schulzrieglaff, O.B., Ning, Z., Alexandrov, L.B., Bauer, M.J., Fu, B., Hims, M., Ding, Z., Ivakhno, S., and Stewart, C. (2012). Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. *Cell* *148*, 780-791.
- Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics* *25*, 1335.
- Parsons, J.D. (1995). Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci* *11*, 615-619.
- Pevzner, P.A., Tang, H., and Waterman, M.S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the national academy of sciences* *98*, 9748-9753.
- Prensner, J.R., Iyer, M.K., Balbin, O.A., Dhanasekaran, S.M., Cao, Q., Brenner, J.C., Laxman, B., Asangani, I.A., Grasso, C.S., and Kominsky, H.D. (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nature Biotechnology* *29*, 742-749.
- Punta, M., Coghill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., and Clements, J. (2012). The Pfam protein families database. *Nucleic acids research* *40*, D290-D301.
- Rubin, C.-J., Zody, M.C., Eriksson, J., Meadows, J.R., Sherwood, E., Webster, M.T., Jiang, L., Ingman, M., Sharpe, T., and Ka, S. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* *464*, 587.
- Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., and Schatz, M.C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* *15*, 461-468.
- Seo, J.S., Rhie, A., Kim, J., Lee, S., Sohn, M.H., Kim, C.U., Hastie, A., Cao, H., Yun, J.Y., and Kim, J. (2016). De novo assembly and phasing of a Korean human genome. *Nature* *538*, 243-247.
- Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., and Richards, S. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* *15*, 1034-1050.
- Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* *19*, 215--225.
- Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., Liu, Y., Chen, R., and Zhao, Y. (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Research* *41*, e166.
- Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* *56*, 564-577.
- Thorne, J.L., Kishino, H., and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *J Mol Evol* *33*, 114-124.
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* *7*, 562-578.
- Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Baren, M.J.V., Salzberg, S.L., Wold,

- B.J., and Pachter, L. (2010). Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nature Biotechnology* 28, 511-515.
- Van Berkum, N.L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L.A., Dekker, J., and Lander, E.S. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. *JoVE (Journal of Visualized Experiments)*, e1869.
- Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., and Young, S.K. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* 9, e112963.
- Wang, K., Li, M.Y., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38.
- Weir, B.S., and Hill, W.G. ESTIMATING F-STATISTICS - Annual Review of Genetics, 36(1):721. Population Structure Forensic Profiles Inbreeding Relatedness.
- Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M., and Jaffe, D.B. (2017). Direct determination of diploid genome sequences. *Genome research* 27, 757-767.
- Weisenfeld, N.I., Yin, S., Sharpe, T., Lau, B., Hegarty, R., Holmes, L., Sogoloff, B., Tabbaa, D., Williams, L., and Russ, C. (2014). Comprehensive variation discovery in single human genomes. *Nature genetics* 46, 1350.
- Yang, Z.H. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24, 1586-1591.
- Young, J.F., Luecke, R.H., Pearce, B.A., Lee, T., Ahn, H., Baek, S., Moon, H., Dye, D.W., Davis, T.M. and Taylor, S.J. (2009). Human organ/tissue growth algorithms that include obese individuals and black/white population organ weight similarities from autopsy data. *Journal of Toxicology and Environmental Health* 72, 527-540.
- Zhang, F., Ye, C., Li, G., Ding, W., Zhou, W., Zhu, H., Chen, G., Luo, T., Guang, M., and Liu, Y. (2002). The rat model of type 2 diabetic mellitus and its glycometabolism characters. *Acta Laboratorium Animalis Scientia Sinica* 52, 401-407.