

<https://doi.org/10.1038/s42003-026-09557-3>

# The Neijiang pig T2T genome reveals domestication history and germplasm traits of Southwest Chinese local breeds

Check for updates

Dong Chen <sup>1,2,5</sup>, Shengdi Cui <sup>1,2,5</sup>, Zhenjian Zhao <sup>1,2</sup>, Pingxian Wu <sup>3</sup>, Wenxuan Zhou <sup>1,2</sup>, Patrick Kofi Makafui Tecku <sup>1,2</sup>, Kai Wang <sup>4</sup>, Junge Wang <sup>1,2</sup>, Ziyang Chen <sup>1,2</sup>, Shixin Yu <sup>1,2</sup>, Jiamiao Chen <sup>1,2</sup>, Yaoxi Zhou <sup>1,2</sup>, Runjie Huang <sup>1,2</sup>, Xuewei Li <sup>1,2</sup> & Guoqing Tang <sup>1,2</sup> ✉

A complete telomere-to-telomere (T2T) pig genome serves as a high-precision reference for functional genomics and structural variation studies due to its high level of completeness and minimal error rate. Here we present a comprehensive framework for genomic research aimed at the effective utilization of Neijiang pig genetic resources. The highly repetitive centromeric regions of the Neijiang pig are identified, and their characteristic centromeric landscapes are constructed using chromosomal landmark identification combined with centromeric repeat monomer localization strategies. Genome-wide association studies (GWAS) based on the T2T reference genome identify key genetic loci associated with reproductive traits, offering potential molecular targets for genetic improvement. Gene family analysis and genetic investigation into head morphology in Neijiang pigs reveal selection signals within olfactory receptor (OR) genes that are associated with head shape differentiation, highlighting the potential role of environmental adaptation in the phenotypic domestication of Chinese indigenous pig breeds. The Neijiang pig T2T genome (NJP-T2T) assembly provides a critical reference resource and foundational dataset for genetic improvement and functional genomic studies in indigenous pig breeds.

Over millennia, natural selection and artificial breeding have led to significant phenotypic diversification, with over 700 domestic pig breeds globally, mainly in Asia and Europe<sup>1,2</sup>. Regional pig breeds with distinct traits emerged through long-term socio-cultural evolution<sup>3</sup>. Head morphology, a key trait linked to domestication and genetics, influences behaviors like foraging, feeding, and aggression. Chinese indigenous pigs are vital genetic resources due to their environmental adaptability and distinct phenotypes. According to China's National Germplasm Center (<https://cdad-is.org.cn/>), over twenty native breeds exist, such as Neijiang, Hang, and Wannan Black pigs, all showing significant head morphology differences. Yet, scientific studies are scarce.

The Neijiang pig, one of the typical representatives of Chinese local pig breeds, originates from the Neijiang region in Sichuan Province and has a domestication history of ~1800 years<sup>4</sup>. The region's subtropical climate, agricultural practices, and traditional breeding culture together have shaped its unique evolutionary path. The Neijiang pig is known for its high fertility,

strong stress resistance, remarkable adaptability, and distinct morphology, making it highly valuable for conservation and sustainable utilization<sup>5</sup>. However, modern commercial breeding has drastically reduced its population, threatening genetic diversity. Although swine genomics has made significant progress, the absence of a breed-specific, complete reference genome continues to hinder the effective use of these local genetic resources.

The current mainstream pig reference genome (Sscrofa11.1)<sup>6</sup> is renowned for its high contiguity, but its genetic foundation is derived from the European Duroc breed, which exhibits approximately 7% genomic divergence from Asian pig breeds<sup>7</sup>. This discrepancy significantly limits its application in analyzing genetic variations in indigenous Chinese pigs, especially those related to specific adaptive variations. The integration of third-generation sequencing technologies (such as PacBio High-Fidelity (HiFi) and Oxford Nanopore ultra-long reads) with chromatin conformation capture (Hi-C) has enabled the achievement of T2T genome assemblies<sup>8–13</sup>. HiFi reads combine long read lengths (>20 kb) with high

<sup>1</sup>State Key Laboratory of Swine and Poultry Breeding Industry, College of Animal Science and Technology, Sichuan Agricultural University, Chengdu, Sichuan, China. <sup>2</sup>Farm Animal Genetic Resources Exploration and Innovation Key Laboratory of Sichuan Province, College of Animal Science and Technology, Sichuan Agricultural University, Chengdu, Sichuan, China. <sup>3</sup>National Center of Technology Innovation for Pigs, Chongqing, China. <sup>4</sup>Sichuan New Hope Liuhe Pig Breeding Technology Co. Ltd, Chengdu, Sichuan, China. <sup>5</sup>These authors contributed equally: Dong Chen, Shengdi Cui. ✉e-mail: [tyq003@163.com](mailto:tyq003@163.com)

accuracy (>99.9%), allowing precise resolution of complex repeat units, while Nanopore ultra-long reads (>100 kb) span large structural variation regions. Their synergistic use substantially reduces assembly fragmentation<sup>14</sup>. These technologies are expected to achieve complete genome assemblies for indigenous pig breeds, fully resolving centromeres, telomeres, and repetitive gene clusters, thereby establishing a paradigm for precise genetic resource characterization and functional genomic studies.

In this study, we constructed the nearly gap-free NJP-T2T by integrating PacBio HiFi reads, Nanopore ultra-long reads, and Hi-C sequencing data. This assembly enabled systematic characterization of centromeric and telomeric architectures, while comparative genomic analyses revealed unique features distinguishing the Neijiang pig genome. Population genetic analyses delineated the genetic structure and inbreeding patterns within conservation populations. Through GWAS and selection signature analyses, we identified candidate genes associated with reproductive performance and head morphology. The NJP-T2T genome serves as both a high-precision molecular toolkit for swine genetic resource utilization and a critical foundation for deciphering adaptive evolution mechanisms in East Asian pig breeds and advancing molecular breeding strategies.

## Results

### NJP-T2T genome assembly and annotation

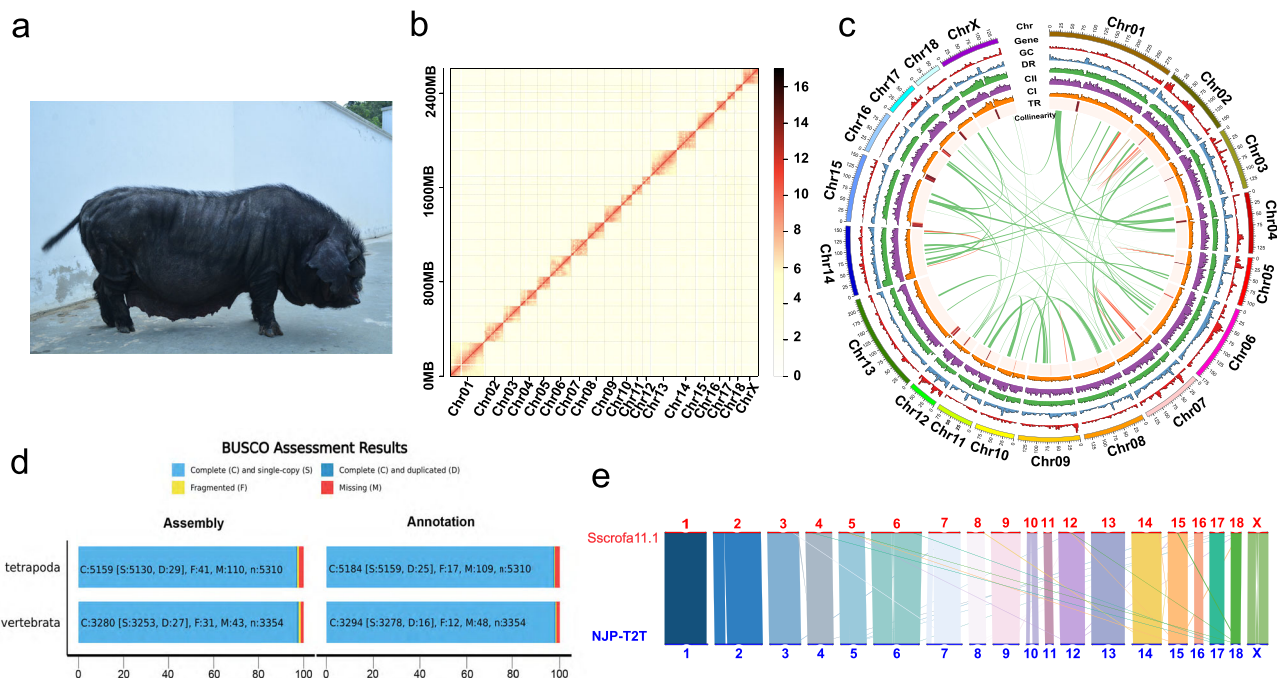
A Neijiang pig, one of China's most iconic indigenous breeds (Fig. 1a), was selected from the National Neijiang Pig Genetic Resources Conservation Farm for genome assembly. PacBio HiFi sequencing generated 2,984.26 Gb of raw data, from which 176.13 Gb of high-quality reads (68× coverage, N50 = 18 kb) were retained after filtering (Supplementary Data 1). Nanopore sequencing yielded 462.65 Gb of ultra-long reads (179× coverage), including 33.95 Gb >100 kb (Supplementary Fig. 1 and Supplementary Data 2). Hi-C data (396.57 Gb, ~152× coverage) were additionally acquired (Supplementary Data 3).

The initial contig-level assembly was constructed using HiFi sequencing data, resulting in 161 contigs (total length 2696.18 Mb, N50 = 90.71 Mb). Meanwhile, an Oxford Nanopore Technologies (ONT)

assembly was generated using Nanopore ultra-long reads, producing 844 contigs (total length 2523.41 Mb, N50 = 59.65 Mb; Supplementary Data 4). Integration of Hi-C chromatin conformation capture data (Supplementary Data 5) anchored 96.69% of the HiFi sequences to 19 pseudochromosomes, including 18 autosomes and one X chromosome (Supplementary Data 6). In the heatmap, intrachromosomal interaction signals along the diagonal regions were markedly stronger than interchromosomal signals (Fig. 1b), confirming the accuracy of the assembly. Gap closure using the ONT assembly achieved a near complete telomere-to-telomere assembly, with only one 500 bp gap remaining on chromosome 16 (Fig. 1c). The final genome size was approximately 2.55 GB (including only sequences anchored to autosomes and the X chromosome), with contig and scaffold N50 values both reaching 142.12 Mb (Supplementary Data 7).

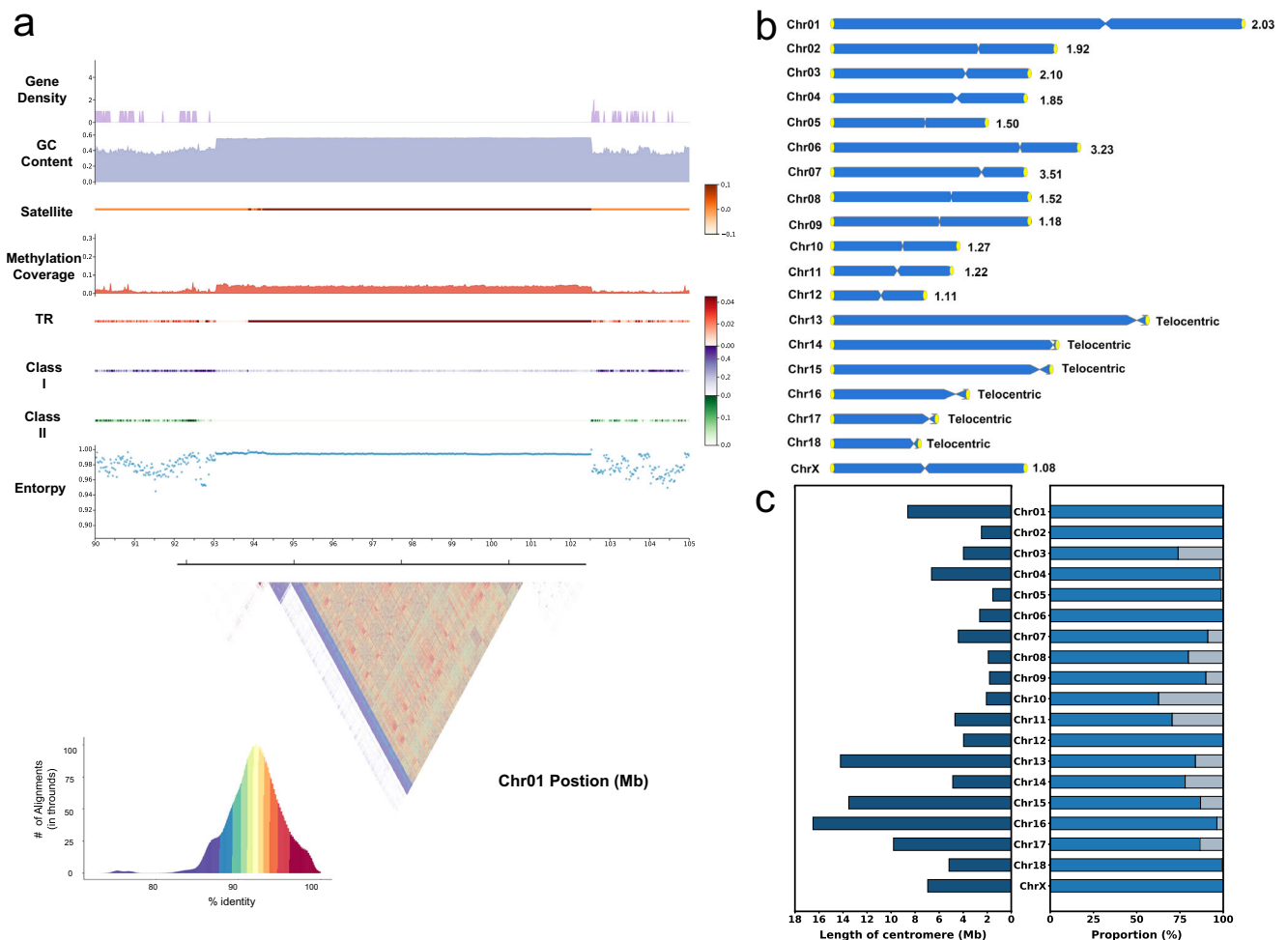
Repetitive elements were annotated across 42.26% of the genome, comprising 19.33% long interspersed nuclear elements (LINEs), 1.53% short interspersed nuclear elements (SINES), 4.41% long terminal repeats (LTRs), and 8.60% satellite repeats (Fig. 1c, Supplementary Data 8). Protein-coding genes were predicted through integrative annotation (Supplementary Fig. 2), identifying 23,375 high-confidence genes. Functional annotation was achieved for 95.79% of genes (Supplementary Fig. 3 and Supplementary Data 9), with enrichment observed in intracellular components (cellular anatomical entities), molecular binding/catalytic activities, and carbohydrate utilization pathways (Supplementary Fig. 4). Among all protein-coding genes, the previously unresolved regions (PURs) regions contained 1199 genes, including 1040 with known functional annotations and 159 without annotation (Supplementary Data 10). These genes are distributed across all chromosomes, with relatively high densities observed in the posterior regions of the p-arms of chromosomes 2 and 7 (Supplementary Fig. 5). Non-coding RNA annotation revealed 4421 tRNAs, 632 rRNAs, and 954 miRNAs (Supplementary Data 11), alongside 173 pseudogenes averaging 5700 bp.

The quality of genome assembly and annotation was evaluated through sequencing read coverage, mapping rates, consensus quality value (QV), Benchmark of Universal Single-Copy Orthologs (BUSCO) scores, and



**Fig. 1 | NJP-T2T genome assembly.** **a** Neijiang pig. **b** Heat map of interactive intensity between chromosome sequences of Neijiang pig by Hi-C. **c** The genome circos plot of NJP-T2T, displaying the genomic features of NJP-T2T. From innermost to outermost, it represents the chromosome length of NJP-T2T, genes, GC

content, dispersed repeats, tandem repeats, Class 1 retrotransposons, Class 2 DNA transposons, tandem repeats, and synteny. **d** BUSCO scores of genome assembly and annotation. **e**: Genome synteny between NJP-T2T and Scrofa11.1.



**Fig. 2 | Centromere Region Identification and Features.** **a** Genomic features at the centromere of Neijiang pig. From top to bottom, these include gene density, GC content, centromere satellite sequence alignment results, 5mC methylation, tandem repeats, Class 1 retrotransposons, Class 2 DNA transposons, sequence entropy

distribution, and centromere sequence identity heatmap. **b** The position and arm ratio of Neijiang pig centromeres. **c** Composition of satellite sequences in the centromere region length of NJP-T2T.

genomic synteny. HiFi and ultra-long ONT reads were mapped back to NJP-T2T to validate assembly accuracy, which found that all reads were uniformly distributed across nearly all chromosomes (Supplementary Fig. 6), with mapping rates reaching 100% and 99.99%. Consensus quality value (QV) was determined using a Kmer-based method to assess assembly accuracy, showing QV values of 54.25–78.07 across all chromosomes (mean = 62.38; Supplementary Data 12). The BUCSO score indicated that 97.90% of the benchmark universal single-copy orthologs were annotated (Fig. 1d). Additionally, high synteny was observed between NJP-T2T and the NCBI pig reference genome Sscrofa11.1, these results demonstrate the high quality and reliability of the NJP-T2T genome (Fig. 1e).

### Characterization of centromeres and telomeres

Centromeres play essential roles in eukaryotic cell division<sup>15</sup>, yet their complete resolution in porcine genomes has been hindered by highly repetitive sequences. In this study, a two-step strategy was implemented: preliminary localization of centromeres through regional genomic features, followed by precise demarcation using centromeric repeat monomers. The centromeric regions were characterized by the absence of gene clusters, elevated GC content, methylation enrichment, and dense tandem repeat coverage. The sequences demonstrated remarkably uniform complexity in terms of entropy. Notably, previously documented satellite DNA should be observed in the centromeric regions (Supplementary Data 13). Through these defining characteristics, the centromeric positions in the NJP-T2T genome were preliminarily mapped, and a comprehensive centromeric

feature landscape was established (Fig. 2a). The analysis revealed dense accumulation of tandem repeats in centromeric regions, while Class I retrotransposons and Class II DNA transposons exhibited minimal representation. Subsequently, nine major centromeric repeat units (CRUs) were identified within the preliminary centromere regions (Supplementary Data 14). CRU1 (335 bp), CRU2 (1679 bp), CRU3 (336 bp), CRU4 (672 bp), and CRU5 (336 bp) corresponded to Mc1 satellite sequences predominantly enriched in metacentric chromosomes (Chr1–12). In contrast, CRU6 (1342 bp), CRU7 (907 bp), CRU8 (2242 bp), and CRU9 (1344 bp) were identified as Ac2 satellite sequences specifically localized to acrocentric chromosomes (Chr13–18). Centromere positions were definitively mapped based on CRU distributions (Supplementary Data 15), with Chr1–12 classified as metacentric (arm ratio: 1.11–3.51; Fig. 2b). The longest metacentric centromere was observed in Chr1 (8.62 Mb), while the shortest was detected in Chr5 (1.54 Mb). Chr13–18 were categorized as acrocentric, with the longest centromere spanning 16.5 Mb in Chr16. ChrX is a standard metacentric chromosome, with an arm ratio of a uniform 1.08. Tandem repeat sequences are enriched along the centromeric position on the chromosome, with different distribution patterns across various chromosomes (Fig. 2c). In addition, our assembly detected telomere structures at the ends of all chromosomes (Supplementary Data 16).

### Assembly improvement of NJP-T2T

Figure 3a illustrates the structural distinctions between the NJP-T2T and Sscrofa11.1 genomes. By aligning the NJP-T2T assembly with the

Sscrofa11.1 reference genome, we identified a set of PURs that correspond to sequences either absent or misassembled in the earlier reference. Compared with Sscrofa11.1, the NJP-T2T assembly was extended by 181.35 Mb in autosomes and the X chromosome, with 532 gaps closed and the telomeric regions of the Neijiang pig resolved for the first time (Supplementary Fig. 7). In addition, structural errors present in Sscrofa11.1 were corrected in NJP-T2T (Supplementary Data 17). Further sequence characterization for PURs revealed that these additional regions are enriched for centromeric satellite sequences (43.21%), Segmental duplications (SDs) (14.21%), and dispersed repeats (13.26%), with 3.51% overlapping both centromeric satellites and segmental duplications, and 25.81% remaining unclassified (Fig. 3b). Taken together, the NJP-T2T assembly represents one of the most contiguous and robust swine genomes currently available. The remarkably low number of gaps and the presence of exceptionally long contiguous sequences provide a highly valuable, high-resolution genomic resource for the Neijiang pig (Supplementary Data 18).

To evaluate the practical utility of NJP-T2T, we downloaded whole-genome sequencing (WGS) data for 400 individuals from public databases, including 200 indigenous pigs and 200 exotic pigs across four breeds, with sequencing depths ranging from 8.5× to 38.27× (Supplementary Data 19). When aligned to NJP-T2T and Sscrofa11.1, significantly higher mapping rates were observed for NJP-T2T in both indigenous (99.21% for NJP-T2T and 98.64% for Sscrofa11.1) and exotic pigs (99.18% for NJP-T2T and 98.84% for Sscrofa11.1) (Fig. 3c). The properly mapped read rates further demonstrated NJP-T2T's superiority, achieving 97.44% in indigenous and 96.39% in exotic pigs, whereas Sscrofa11.1 showed lower rates of 95.86% and 95.91%, respectively (Fig. 3d). This indicates enhanced accuracy in identifying functional genomic regions and reduced misalignment events.

Base error rate analysis revealed that NJP-T2T exhibited a significantly lower error rate in indigenous populations (0.98% for NJP-T2T and 1.27% for Sscrofa11.1) (Fig. 3e), confirming its higher assembly accuracy for local breeds. However, a marginally elevated error rate was observed in European pigs (1.11% for NJP-T2T and 0.90% for Sscrofa11.1), likely attributable to greater genetic divergence between breeds.

RNA-seq data (Supplementary Data 20) alignment showed improved mapping rates for NJP-T2T (90.26% for indigenous pigs using NJP-T2T and 89.47% using Sscrofa11.1; 90.67% for exotic pigs using NJP-T2T and 89.85% using Sscrofa11.1) (Fig. 3f), underscoring its utility in transcriptomic studies. Structural variant (SV) detection using short-read data from three individuals per group demonstrated NJP-T2T's increased sensitivity, with higher counts of deletions/insertions (Del/Ins) and inversions/translocations (Inv/Tra) identified compared to Sscrofa11.1 (Supplementary Fig. 8).

In summary, the NJP-T2T genome demonstrates superior accuracy and applicability in both DNA and RNA sequencing analyses, particularly for genetic studies of indigenous pig breeds. These experimental validations establish NJP-T2T as a comprehensive and reliable genomic resource for advancing swine genetics research and breeding optimization.

### Gene family analysis

Gene family clustering analysis involving Large White, Landrace, Duroc, Berkshire, African Warthog, Jinhua, Bamei, Tibetan, and Rongchang pigs identified 14,376 core gene families shared across these breeds (Fig. 4a). These shared gene families form the genetic foundation for essential biological functions and physiological processes. Their universal presence reflects a high degree of evolutionary conservation during critical stages of development, establishing a solid framework for comparative genomic investigations among diverse pig lineages. Strikingly, 75 gene families were found to be unique to the Neijiang pig genome, the highest number of breed-specific families identified in this analysis (Fig. 4a). This notable enrichment is likely driven by both the exceptional completeness and accuracy of the NJP-T2T genome assembly, as well as the distinctive genetic background and adaptive traits of the Neijiang breed. Together, these unique gene families provide important insights into the genetic distinctiveness and potential functional specialization of the Neijiang pig.

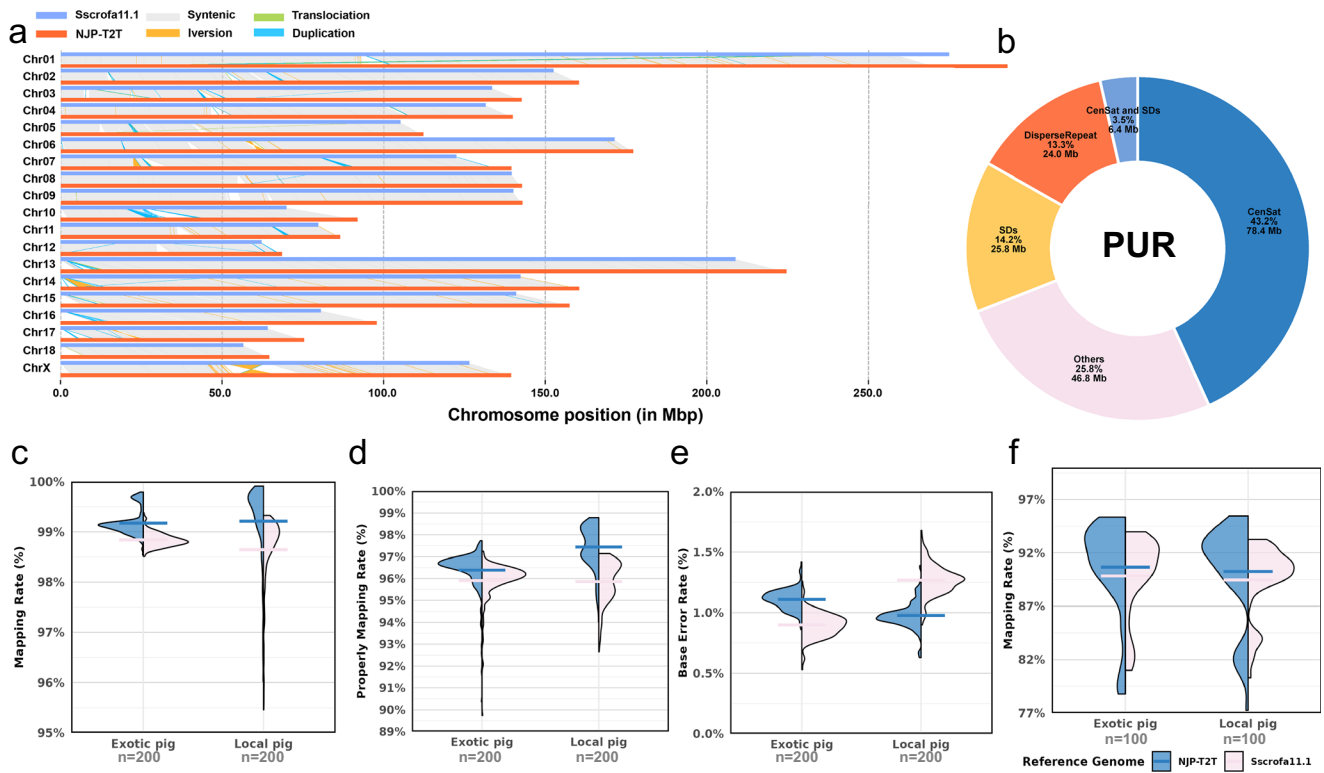
Comparative genomic analysis revealed a substantial number of multicopy orthologous genes in the NJP-T2T genome (Fig. 4b). Notably, 318 genes were found to be unique to the Neijiang pig, which were absent in all other breeds (Supplementary Data 21), underscoring the genomic distinctiveness of the Neijiang breed. These genes were distributed across all chromosomes of the NJP-T2T genome, with dense clustering observed on the short arms (p-arms) of chromosomes 7 and 10, as well as on chromosome 13 (Fig. 4c), suggesting potential functional or adaptive specialization in these regions. Functional enrichment analysis demonstrated that these genes are involved in a wide range of processes, from basic cellular components and metabolic pathways (Fig. 4d) to complex physiological systems such as sensory perception, immunity, and aging regulation (Fig. 4e). The enrichment profiles indicate that Neijiang pigs possess genetic advantages in environmental adaptation, health maintenance, and developmental regulation. In particular, the enrichment of sensory-related genes is closely associated with their enhanced environmental perception, such as olfactory sensitivity. Direct comparison with other indigenous breeds (Jinhua, Bamei, Tibetan, Rongchang) identified 803 Neijiang-specific gene families encompassing 1330 genes, exclusively present in the Neijiang genome (Fig. 4f). Functional analysis of these families revealed profiles comparable with those in Fig. 4, further evidencing Neijiang pigs' unique adaptive specializations (Supplementary Fig. 9).

### Population genetic structure and inbreeding analysis of Neijiang pigs

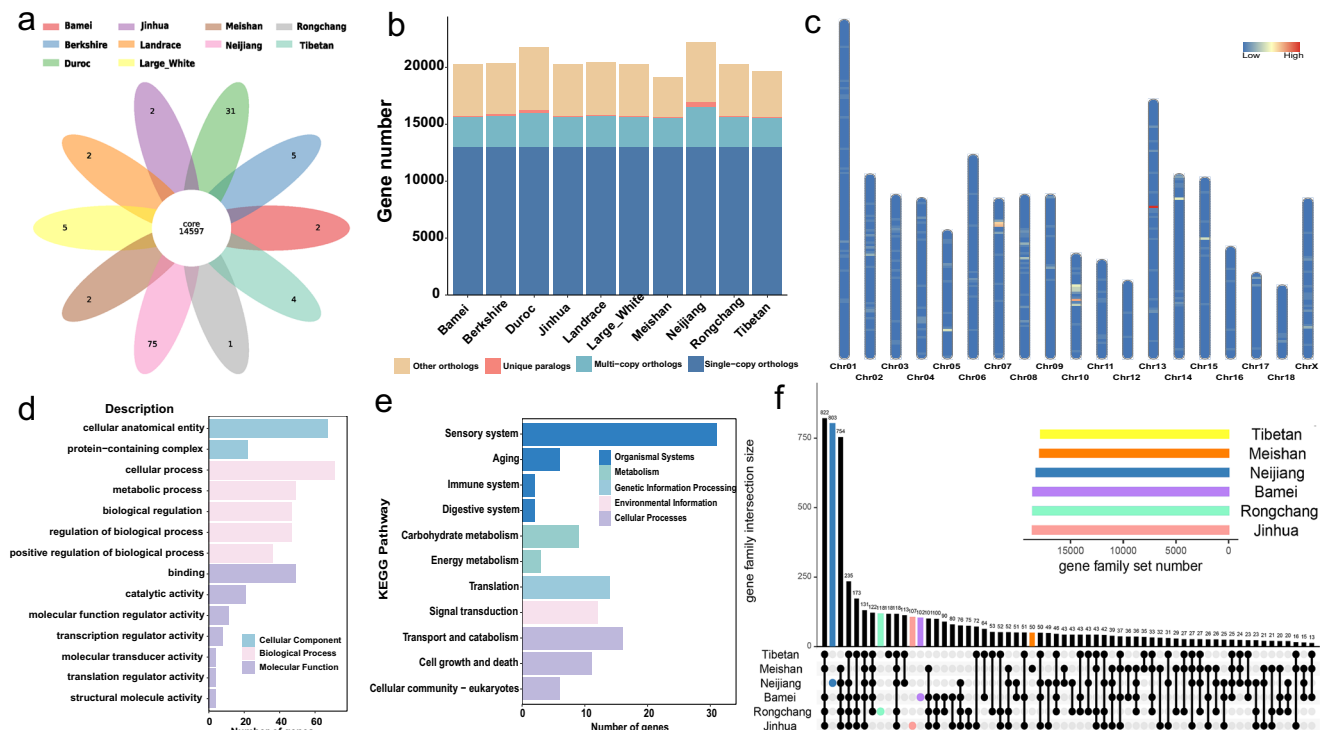
Based on the genomic data of the Chinese National Neijiang Pig Genetic Resources Conservation Farm, the genetic diversity, population structure, and inbreeding level of the population were analyzed. The results of genetic diversity analysis (Supplementary Data 22) show that the average frequency of minor alleles in this population is 0.126, which is moderately low, indicating that the population retains some genetic variation, but the major allele dominates. The effective number of alleles is 1.286, suggesting that alleles are unevenly distributed across many loci, yet the population still possesses some level of diversity. The observed heterozygosity and expected heterozygosity are 0.205 and 0.206, respectively (Supplementary Data 23), reflecting moderate genetic variation both within individuals and the overall population. Taken together, the Neijiang pig population exhibits a moderate level of genetic diversity, providing a solid foundation for conservation and sustainable use, highlighting its genetic value as a local breed.

Principal component analysis (PCA) of 115 individuals from the Neijiang pig population reveals that the first three principal components (PC1, PC2, and PC3) explain 37.93% of the genetic variation. The PCA scatter plot (Fig. 5a) shows that the population forms a relatively uniform cluster, reflecting that there is no obvious subpopulation differentiation within the Neijiang pig population. While the population possesses some genetic diversity, the level of diversity is not particularly rich. The phylogenetic tree, constructed based on genetic distance (Supplementary Fig. 10), divides the Neijiang pig population into six distinct lineages. Furthermore, the kinship matrix (Fig. 5b) constructed from the same dataset further reveals the genetic structure of the population: most individuals have distant kinship, indicating diversity in the overall genetic background; however, there are also closer relationships between some individuals, suggesting the possible presence of local inbreeding.

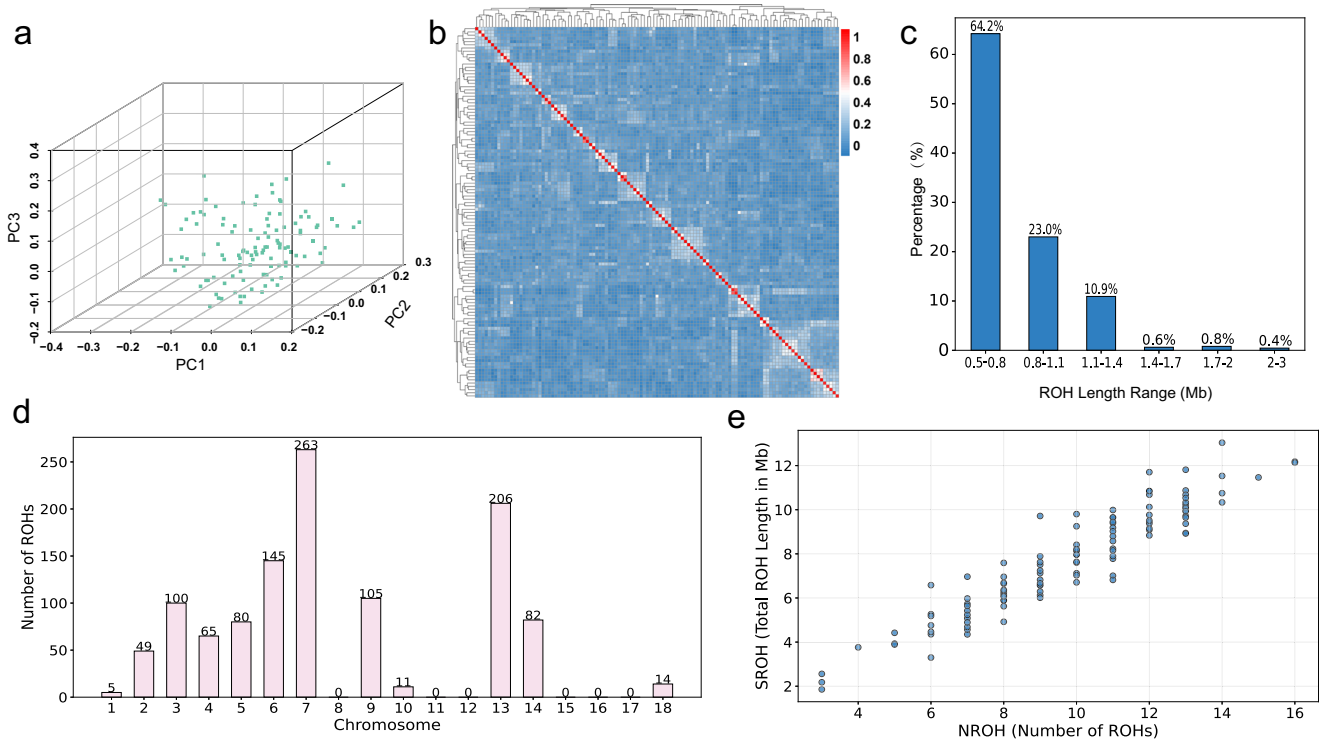
Inbreeding coefficients for the Neijiang pig population were calculated using the Plink (v1.90) software based on genomic data, with an average value of 0.119, indicating a moderately low inbreeding risk. A total of 1126 runs of homozygosity (ROHs) were detected in the analysis (Supplementary Data 24), with an average length of 790,724 bp, reflecting the commonality of short ROHs in the population. 64.2% of the ROHs are distributed within the shorter range of 0.5 to 0.8 Mb (Fig. 5c), likely resulting from distant historical inbreeding or random homozygosity in large populations, indicating that the population faces some homozygosity pressure while still retaining moderate genetic diversity. Figure 5d shows that ROHs are distributed across different chromosomal regions but are unevenly spread, with chromosomes 7 (263 ROHs) and 13 (206 ROHs) exhibiting significantly



**Fig. 3 | Improvement of the NJP-T2T Genome.** **a** Differences in details between the NJP-T2T and Sscrofa11.1 genomes. **b** Sequence characteristics of previously unresolved regions. **c** Comparison of DNA resequencing data mapping rates. **d** Comparison of correctly mapped resequencing data. **e** Comparison of base error rates in resequencing data. **f** Comparison of RNA resequencing data mapping rates.



**Fig. 4 | Gene family analysis results.** **a** Gene family clustering analysis of Large White, Landrace, Duroc, Berkshire, Neijiang pig, Mei shan pig, Jinhua pig, Baimei pig, Tibetan pig, and Rongchang pig. **b** Summary of homologous genes. **c** Distribution of Neijiang pig-specific genes on chromosomes. **d** GO functional enrichment analysis of Neijiang pig-specific genes. **e** KEGG pathway enrichment analysis of Neijiang pig-specific genes. **f** Unique gene families of Neijiang pig, Jinhua pig, Baimei pig, Tibetan pig, and Rongchang pig.



**Fig. 5 | Population structure and inbreeding analysis of Neijiang pig. a** PCA analysis of Neijiang pig. **b** Phylogenetic relationship of Neijiang pig population. **c** ROH length distribution. **d** Distribution of ROH on different chromosomes. **e** SROH-NROH scatter plot.

higher numbers compared to other chromosomes (e.g., chromosomes 11 and 18 show almost no ROHs), reflects the effects of local selection or genetic drift. The sum of ROH lengths and number of ROHs (SROH-NROH) scatter plot (Fig. 5e) reflects the individual variations in genetic diversity and inbreeding levels, with each point representing an individual. Individuals in the high N (ROH count) and high S (total ROH length) region exhibit higher inbreeding levels, while those in the low N and low S region show better genetic diversity, indicating that the population overall exhibits a certain degree of inbreeding variation. Overall, the Neijiang pig population has a relatively low level of inbreeding, but the high ROH count and length in some individuals suggest a certain inbreeding risk. The population maintains a moderate level of genetic diversity, although short ROHs dominate, and overall homozygosity still exerts some pressure on diversity.

**GWAS of Reproductive Traits in Neijiang Pigs**

This study conducted a GWAS using 115 individuals from the Chinese National Neijiang Pig Genetic Resources Conservation Farm (Sichuan, China) with well-documented phenotypic records of reproductive traits, aiming to identify key genetic loci influencing these traits in Neijiang pigs. Descriptive statistics of four parity records (total number born, TNB; number born alive, NBA) are provided in Supplementary Data S25. Histograms of reproductive traits exhibited approximately normal distributions with high inter-parity variability (Fig. 6a). Imputation of low-coverage sequencing data was performed using the STITCH-Beagle pipeline, yielding 19,785,303 SNPs with an imputation accuracy of 0.85 (Fig. 6b). Assessment of the genomic inflation factor ( $0.996 < \lambda_{GC} < 1.034$ ) and QQ plots (Fig. 6c-d) indicated no substantial inflation, suggesting that the GWAS results were unlikely to be driven by systematic bias.

The GWAS revealed significant genetic heterogeneity across parities, with the first parity analysis detecting only one significant locus (Supplementary Data 26) lacking direct gene associations (Fig. 6c), while no significant loci were identified in the fourth parity, potentially due to limited sample size, data resolution, or environmental confounders. In contrast,

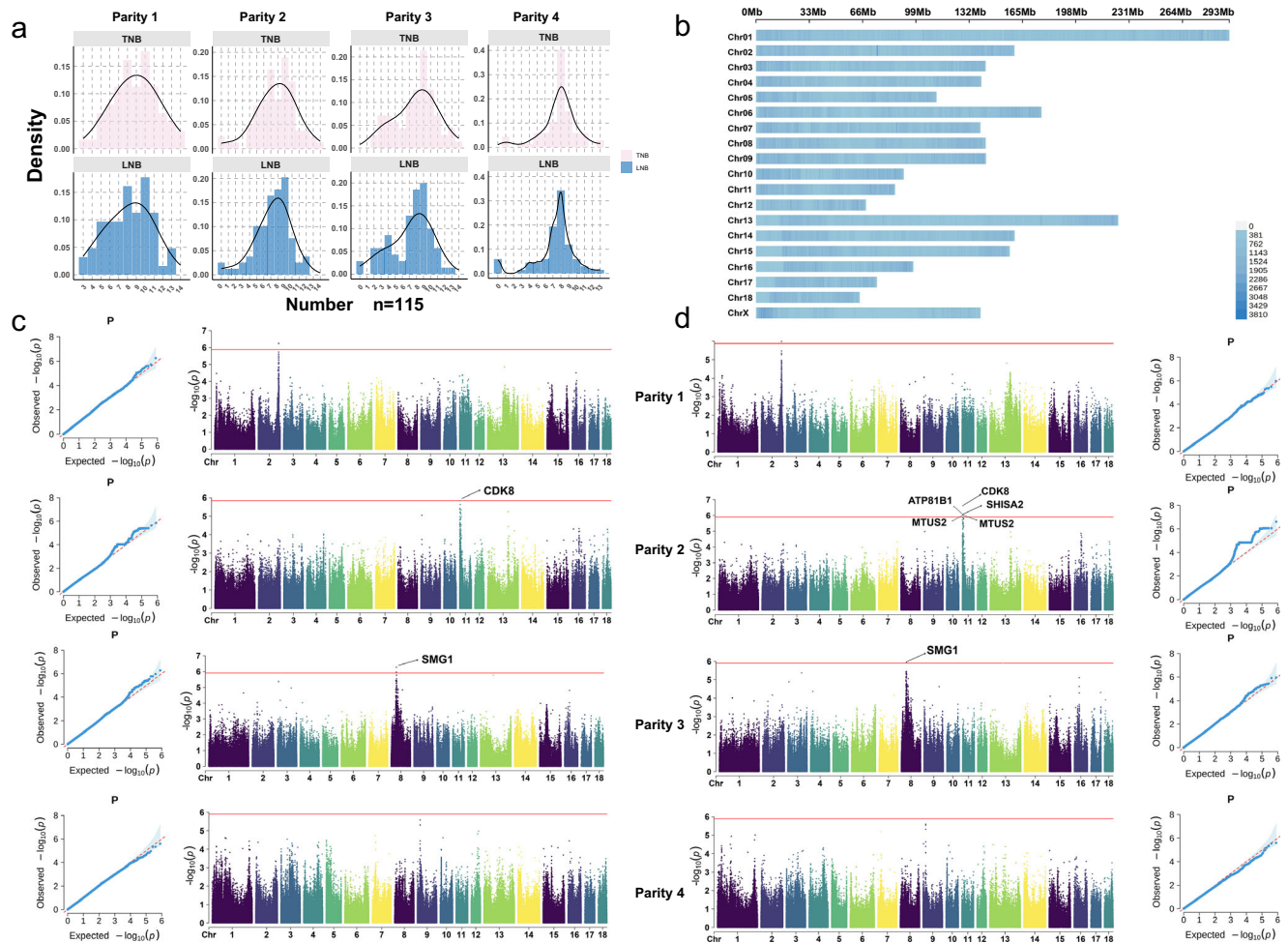
analyses of the second and third parities detected multiple significant loci on chromosomes 11 and 8, respectively (Fig. 6).

For NBA in the second parity, 11 significant loci were mapped to chromosome 11, including *SHISA2* (rs511232), *ATP8B1* (rs511340), *CDK8* (rs511480), *GPR12* (rs511692), and *MTUS2* (rs513144). Notably, the *CDK8* locus (rs511480) showed significance in both NBA and TNB analyses. Previous studies have reported that *CDK8* is required for embryo implantation in mice<sup>16</sup>, and functional validation following GWAS has also demonstrated its role in granulosa cell proliferation in pigs<sup>17</sup>. Furthermore, *SMG1* is essential for embryogenesis and early embryonic development<sup>18</sup>. Taken together, these results provide evidence for deciphering the genetic characteristics underlying the reproductive performance of Neijiang pigs and identify potential loci as well as their associated genes. These molecular markers may offer support for future breeding strategies.

**Selective characteristics of head morphology in Neijiang pigs**

Neijiang pigs exhibit the characteristic head shapes typical of Chinese indigenous pig breeds, particularly the Shizitou-type (SZT) and Erfangtoutype (EFT) (Fig. 7). SZT Neijiang pigs have very short snouts, prominent facial wrinkles, and generally short head shapes. In contrast, EFT pigs possess longer snouts, fewer facial wrinkles, and overall longer head shapes<sup>4</sup>. These distinctive head types are not only key phenotypic markers of the breed but also reflect adaptive evolutionary processes. To elucidate the genetic basis of head differentiation in Neijiang pigs, we performed selection signature analysis using high-depth (>100x) resequencing data from individuals with divergent extreme head morphotypes. Investigating the genetic underpinnings of these morphotypes is critical for understanding their environmental adaptation and phenotypic evolution.

We first validated our grouping of eight pigs through PCA. The PCA clearly segregated the samples into two clusters (Fig. 7a), concordant with phenotypic classifications. To further resolve phylogenetic relationships, we constructed a phylogenetic tree using a Duroc pig as an outgroup. The tree revealed distinct branches for SZT and EFT (Fig. 7b), confirming the robustness of our grouping strategy.



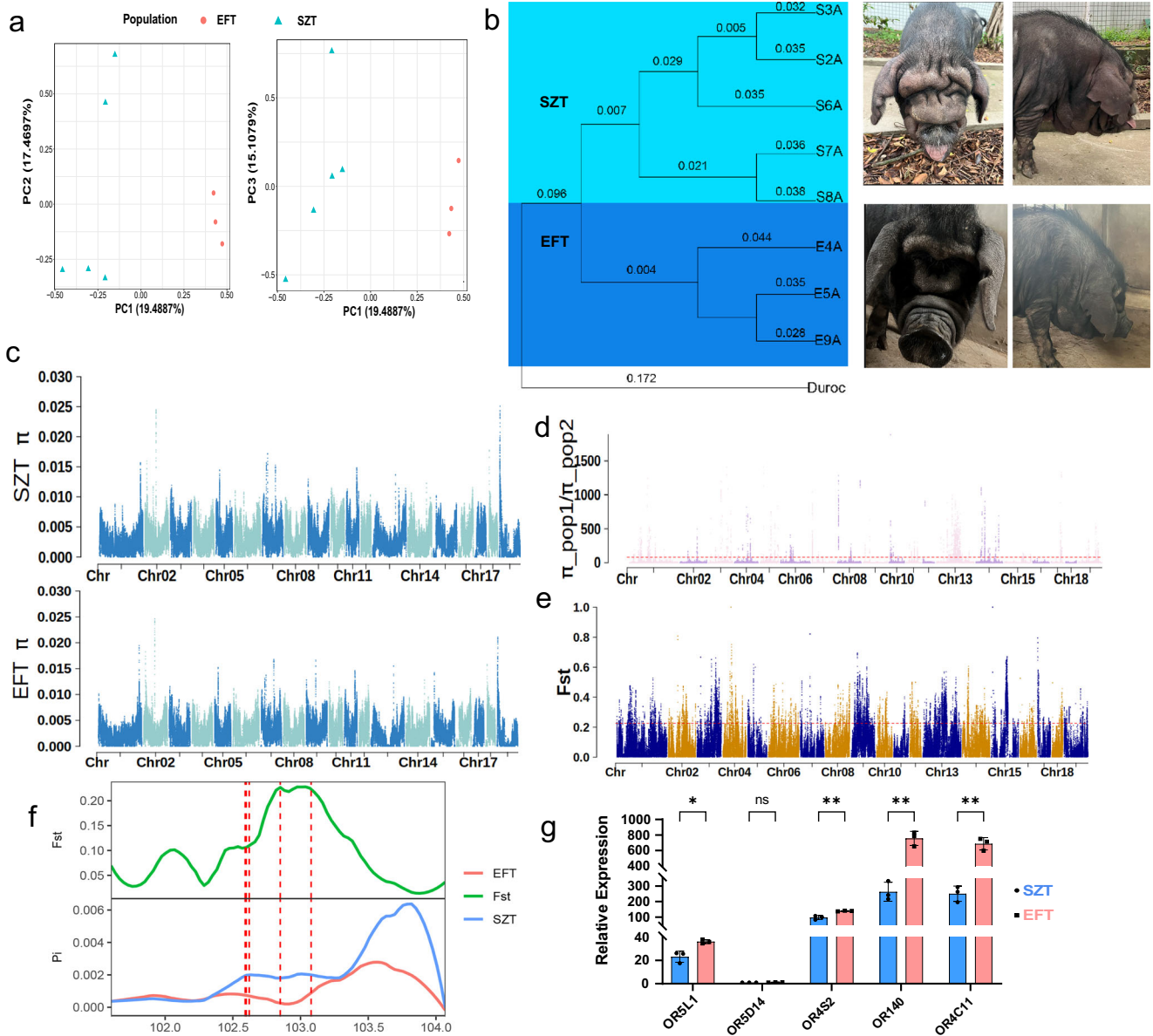
**Fig. 6 | Genome-wide association analysis of reproductive traits. a** Distribution of reproductive traits in Neijiang pig. **b** Filled SNP density. **c** Manhattan plot and QQ plot of total litter size GWAS. **d** Manhattan plot and QQ plot of live birth GWAS.

Genetic divergence between the two morphotypes was assessed using sliding-window nucleotide diversity ( $\pi$ ) analysis. While  $\pi$  distributions were largely overlapping, significant divergence was observed on chromosomes 1, 6, 9, and 13 (Fig. 7c), suggesting selective pressures in these regions. ROD (ratio of observed to expected differences) analysis identified additional selection signals on chromosomes 1, 4, 8, and 13 (Fig. 7d). Fixed differentiation index (Fst) analysis further highlighted multiple regions under selection (Fig. 7e). Integrating Fst and ROD thresholds (top 5%), we identified 148 high-confidence selection windows containing 301 candidate genes (Supplementary Data 27). Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment revealed significant associations with sensory perception, immune response, and environmental adaptation (Supplementary Fig. 11), implicating these processes in head differentiation. Notably, a selection window on chromosome 1 contains a gene cluster consisting of five olfactory receptor (OR) genes (*OR5L1*, *OR5D14*, *OR4S2*, *OR140*, *OR4C11*), which are not included in the current NCBI pig reference genome assembly (Fig. 7f). To verify the involvement of these olfactory receptor (OR) genes in head morphology differentiation of Neijiang pigs, we analyzed their expression patterns in the SZT and EFT groups using RT-qPCR. The results showed that *OR5L1* exhibited significantly higher relative expression in the EFT group compared to the SZT group (\*,  $P < 0.05$ ); *OR5D14* showed no significant difference in expression between SZT and EFT (ns); whereas *OR4S2*, *OR140*, and *OR4C11* displayed markedly higher expression in EFT than in SZT (\*\*,  $P < 0.01$ ) (Fig. 7g, Supplementary Data 28). These findings confirm that these previously unassembled OR genes exhibit transcriptional activity in Neijiang pigs and further support their functional role in shaping the head features of Neijiang pigs.

## Discussion

T2T genome assembly has proven to be a powerful tool for elucidating the genetic basis of various species<sup>19</sup>. For the Neijiang pig genome, the T2T assembly generated a near-complete sequence, with only a 500 bp gap remaining on chromosome 16. We compiled the chromosome-anchored sequence lengths of four existing pig T2T genomes: NJP-T2T (2.61 Gbp), T2T\_Mpig1.0 (2.61 Gbp), T2T\_RCpig1.0 (2.63 Gbp)<sup>20</sup>, and JH-T2T (2.57 Gbp)<sup>21</sup>. The genome size of NJP-T2T is similar to those of T2T\_Mpig1.0, T2T\_RCpig1.0, and JH-T2T. Differences in genome size may be attributed to breed variation, differences in telomere and centromere lengths of the assembled individuals, as well as assembly methodologies. Compared to other representative indigenous pig genomes, it represents one of the most contiguous and reliable assemblies currently available<sup>20–23</sup>. Moreover, compared to the NCBI pig reference genome (Sscrofa11.1), the NJP-T2T assembly demonstrates higher alignment accuracy and a lower error rate in variant detection. This genome exhibits broad applicability in genetic analyses of Chinese indigenous pigs, particularly the Neijiang pig.

Centromeric regions are essential for chromosome segregation and stability, but their highly repetitive sequences pose significant challenges for analysis<sup>24</sup>. Currently, commonly used approaches for centromere detection include software-based predictions and ChIP-seq<sup>25–27</sup>. Software-based prediction typically localizes centromeres by enriching for tandem repeat sequences. Although this approach is convenient, its precision is often limited. ChIP-seq, combined with CENH3 or CENP antibodies, offers more accurate centromere mapping, but species-specific antibodies are unavailable for many organisms. For example, pigs lack a CENP antibody. In our study, we selected a rabbit-derived anti-CENPA antibody (ab217622,



**Fig. 7 | Selection signal analysis of Neijiang pig's special head shape.** **a** PCA validation of Neijiang pig special head shape grouping. **b** Phylogenetic tree and feature display of individuals with special head shapes in Neijiang pig. **c**  $\pi$  value comparison between SZT and EFT populations. **d** ROD comparison between SZT and EFT populations. **e** Fixation index of differentiation between SZT and EFT

populations. **f** Selected features of OR gene family. **g** RT-qPCR of selected OR genes ( $n = 3$  biologically independent animal experiments). Error bars represent the mean  $\pm$  SD from three independent experiments. \*\* indicates  $p < 0.001$ ; \* indicates  $p < 0.05$ ; ns indicates no statistically significant difference.

Abcam, Cambridge, UK), which shares 96% homology with Neijiang pig CENPA, for ChIP-seq analysis. However, Western blotting failed to effectively detect CENPA in Neijiang pig. Additionally, the centromere localization results from tools such as quaraTET were inconsistent with prior studies in pigs. Ultimately, we preliminarily located centromeric regions based on their characteristic genomic features and confirmed their accuracy by aligning with previously validated pig centromeric satellite sequences. Subsequently, we performed clustering analysis of repeat monomers<sup>28</sup>, successfully identifying and mapping the centromeric regions of the Neijiang pig genome for the first time.

As one of the representative indigenous pig breeds in China, Neijiang pigs are known for their exceptional meat quality, reproductive performance (relative to other local breeds), and environmental adaptability<sup>29</sup>. However, with the advancement of modern farming practices, their genetic diversity and resources are at risk of loss and endangerment. Genetic diversity analysis reveals that the Neijiang pig population exhibits moderate genetic

variation, with a predominant presence of fixed alleles, though some diversity is still retained. However, certain individuals show a higher number of ROH, suggesting potential local inbreeding risks. These results indicate that the breed is susceptible to genetic drift, underscoring the need to control inbreeding within the limited gene pool to ensure the sustainable development of the population.

In the GWAS analysis, we detected several significant loci associated with reproductive traits in the second and third parities, primarily on chromosomes 11 and 8. These loci involved candidate genes such as *CDK8* and *SMG1*, which are known to participate in critical biological processes, including embryonic development and cell fate regulation<sup>16-18</sup>, suggesting that they may contribute to the genetic regulation of reproductive traits in Neijiang pigs. In the analyses of both the number born alive and the total number born, *CDK8* and *SMG1* showed potential associations. Further queries in the PigGTEx database showed that these genes also exhibited associations with reproductive indicators such as teat number and uterine

capacity, providing external evidence in support of our GWAS findings. It should be noted, however, that the statistical power of this study is limited by the relatively small sample size, and these results will require validation in larger populations and through integration with transcriptomic or other multi-omics data. Nevertheless, our findings provide foundational insights into the genetic basis of reproductive performance in Neijiang pigs and highlight candidate genes that may serve as valuable references for future molecular breeding research.

In the historical and evolutionary context of Neijiang pigs, significant head differentiation reflects the development of breed characteristics, mainly including two groups: “SZT” and “EFT.” The SZT type shows a lion-like head with a full face and short snout, while the EFT type has a longer snout, similar to common local pigs. Due to environmental changes and neglect, the SZT phenotype has gradually disappeared. Field investigations of several conservation farms and breeding bases revealed that only five individuals currently show clear SZT features, making them rare materials for studying the genetic mechanism of head morphology in this breed.

The Neijiang pig population has been influenced by environmental changes and feeding practices throughout its history. The breed can be traced back to the Han Dynasty. During the Ming Dynasty, the expansion of sugarcane cultivation increased feed availability, reduced the pigs’ reliance on olfaction, and helped retain cranial traits such as the SZT type. However, with social changes and shifts in breeding practices, traditional free-range feeding was restored, increasing reliance on olfaction and leading to a rise in the proportion of the EFT group.

To investigate the genetic mechanisms underlying head variation in Neijiang pigs, we performed PCA and phylogenetic analysis. The results revealed significant genetic differences between the SZT and EFT groups, providing genetic evidence for head-type differentiation. Selection signal analysis showed that head-type variation is closely related to gene pathways involved in sensory function and environmental adaptation. Notably, an unassembled cluster of OR genes, absent from Sscrofa1.1.1, was identified on chromosome 1, with stronger selective pressure observed in the EFT population for this region. Real-time quantitative PCR analysis further confirmed significant transcriptional activity of these previously unassembled OR genes in the olfactory epithelium of Neijiang pigs. Moreover, the expression levels of *OR5L1*, *OR4S2*, *OR140*, and *OR4C11* were significantly higher in the EFT group compared to the SZT group. This transcriptional evidence reinforces the functional role of OR genes in shaping head traits in Neijiang pigs and suggests that differences in olfactory capacity may be an important factor driving changes in head morphology. In addition, gene family analysis revealed that Neijiang pig-specific gene families are also enriched in sensory-related functions, further supporting the uniqueness of this breed in sensory traits.

In summary, this study not only generated the first high-quality T2T reference genome of the Neijiang pig, but also provided the first systematic characterization of its centromeric regions, offering a technical framework for studying complex genome structures in pigs. Based on this genomic platform, we achieved high-resolution detection of genetic variation and structural evaluation in conservation populations. Combined with GWAS analysis, we precisely identified key genomic regions associated with reproductive traits, filling a gap in genetic research on Neijiang pigs in this area. These findings provide a theoretical foundation for the conservation and utilization of Neijiang pig germplasm resources and establish a methodological paradigm for exploring the genetic mechanisms of Chinese indigenous pig breeds and developing precision breeding strategies. Future research will focus on the co-evolutionary mechanisms between head morphology and olfactory perception, and will integrate multi-omics approaches to further explore the structural and functional relationships of centromeric regions, paving reliable paths for the innovative use of local pig genetic resources.

## Materials and methods

### Ethics approval

All experimental procedures in this study were approved by the Institutional Review Board of Sichuan Agricultural University (IRB14044) and

conducted in accordance with the ethical guidelines and regulations established by the Institutional Animal Care and Use Committee. The study was performed under permit number DKY-B20220358. The ethical review and approval covered all samples and treatments reported in this paper. We have complied with all relevant ethical regulations for animal use.

### Sample collection and sequencing

A 12-month-old female Neijiang pig was selected from the standardized breeding population at the Chinese National Neijiang Pig Genetic Resources Conservation Farm (Neijiang, Sichuan, China). The longissimus dorsi muscle was collected within 30 min post-slaughter for DNA sequencing to support T2T genome assembly and subsequent analyses. In addition, to facilitate gene annotation, tissue samples from seven organs of one 12-month-old female Neijiang pig were collected for RNA-seq, and tissue samples from six organs of another 12-month-old female Neijiang pig were collected for Iso-seq (Isoform Sequencing). All pigs in this study were humanely slaughtered according to the Chinese National Standard GB/T 22569-2008, the Technical Specification for the Humane Slaughter of Pigs. The procedure involved immediately rendering the pigs unconscious via electrical stunning, followed by exsanguination through sticking within 15 seconds after stunning.

To analyze the population structure and inbreeding level of Neijiang pigs and further investigate their reproductive performance, a total of 115 individuals with complete breeding records were selected from the National Neijiang Pig Genetic Resources Conservation Farm (Neijiang, China). Ear tissue samples were collected and subjected to whole-genome DNA sequencing. The resulting data were used for population structure analysis, inbreeding coefficient estimation, and GWAS.

The Shizitou-type (SZT) Neijiang pig represents a nearly extinct sub-population within the breed. To define the extreme morphological differences between SZT and Erfangtou-type (EFT) individuals, classification criteria were established based on typical head features: individuals with snouts nearly level with the forehead and with prominently raised forehead wrinkles were defined as extreme SZT; individuals with snouts clearly protruding and misaligned with the forehead, and with few or no forehead wrinkles, were defined as extreme EFT. Based on these criteria, eight Neijiang pigs with markedly different head morphologies were selected from the conservation population and subjected to high-depth whole-genome sequencing (100×) to investigate the genetic basis underlying head phenotype variation.

### Library preparation and sequencing

High-molecular-weight DNA was extracted from muscle tissue using a long-fragment column-based method for constructing PacBio HiFi, Hi-C, and ONT libraries (Table 1).

For PacBio HiFi sequencing, DNA was sheared to 15–20 kb fragments using a Megaruptor system, followed by damage repair and enzymatic processing to generate SMRTbell libraries. Sequencing was performed on the PacBio Sequel II platform with five SMRT Cells, yielding high-quality circular consensus sequencing (CCS) data after filtration.

For Hi-C sequencing employed in chromosome-level scaffolding, muscle tissue was crosslinked with 1% formaldehyde (Sigma) and resuspended in lysis buffer. Chromatin was fragmented using the MboI restriction enzyme, followed by biotin labeling with T4 DNA ligase (Enzymatics). The biotinylated DNA fragments were captured using streptavidin-coated magnetic beads (Thermo Fisher Scientific) and sequenced on one lane of the DNBseq platform.

For ONT sequencing aimed at gap filling, high-molecular-weight DNA was processed as above. DNA fragments longer than 200 kb were selected using the BluePippin automated size selection system (Sage Science) and processed with the Ligation Sequencing 1D kit. Sequencing was conducted on the Nanopore platform at BGI. For RNA-seq that supported genome annotation, total RNA was extracted from seven tissues (heart, liver, spleen,

**Table 1 | Sequencing data statistics**

Library type	Platform	Application	Tissue	Total bases (Gb)	Coverage (X)	Number of animals
HiFi	PacBio Sequell	Genome assembly	muscle	176.13	68	1
Hi-C	DNBseq	Chromosome construction	muscle	396.57	152	1
Ultralong ONT	Nanopore	Genome filling	muscle	466.04	179	1
RNA-seq	IlluminaNovaSeq 6000	Genome Annotation	Heart, Liver, Spleen, Lung, Kidney, Intestines, Muscle	59.28	\	1
Iso-seq	PacBio Sequell	Genome Annotation	Heart, Liver, Spleen, Lung, Kidney, Muscle (Mix)	107.14	\	1
Resequencing	Illumina NovaSeq 6000	Population genetic analysis	muscle	443	1	115
	DNBSEQ		muscle	2150.4	100	8

lung, kidney, intestines, and muscle) using TRIzol reagent (Invitrogen). mRNA libraries were constructed using the Optimal Dual-mode mRNA Library Prep Kit (BGI-Shenzhen, China) and sequenced on the Illumina NovaSeq 6000 platform

For full-length transcriptome (Iso-seq) analysis also used in annotation, pooled total RNA from six tissues (heart, liver, spleen, lung, kidney, and muscle) was used to prepare SMRTbell libraries with two size-selected fractions (0–5 kb and 4.5–10 kb). Libraries passing quality control were sequenced on the PacBio Sequel II platform to capture full-length isoform diversity.

For resequencing applied in population genetic analyses, genomic DNA was extracted from 123 muscle samples via the cetyltrimethylammonium bromide (CTAB) method. Libraries for 115 individuals with breeding trait records were prepared using the TruSeq Nano DNA HT Kit (Illumina) and sequenced at approximately 1× coverage on the Illumina NovaSeq 6000 platform for population genomic analysis and GWAS. Separately, libraries for the remaining eight individuals with distinctive head morphology were processed with the BGI Optimal DNA Library Prep Kit and subjected to >100× deep resequencing on the DNBSEQ platform for selection sweep analysis.

**Genome assembly and evaluation**

Primary contig-level genome assembly was performed using Hifiasm (v0.16.1-r375)<sup>30</sup> with PacBio HiFi reads preprocessed by CCS v6.2 (Pacific Biosciences, CA, USA) under parameters: --minPasses 3 --min-rg 0.99 --min-length 500. Hi-C reads were processed to remove adapters and low-quality sequences, then aligned to the contig-level assembly using Juicer (v1.5.6)<sup>31</sup> to generate valid chromatin interaction pairs. Chromosome-level scaffolding was achieved with 3D-DNA (v180922)<sup>32</sup>, leveraging Hi-C-derived genomic proximity signals. To resolve residual gaps, Nanopore ultra-long reads were filtered by Nanopore\_IOC.pl (BGI, Shenzhen, China; parameters: -minlen 5000 -mingv 7), error-corrected with NECAT (v2020)<sup>33</sup>, and integrated into the assembly using LR\_gapcloser (v3) for gap closure (Supplementary Fig. 12).

The final gap-free genome assembly was rigorously evaluated through multiple metrics. Read mapping rates and coverage depths were calculated by aligning PacBio HiFi and ONT reads to the assembly using Minimap2 (v2.24)<sup>34</sup>. Continuity was assessed via contig and scaffold N50 statistics. Base-level accuracy was quantified using Merqury (v1.3)<sup>35</sup> with a k-mer size of 21 to generate quality value (QV) scores. Completeness was validated through BUSCO (v5.4.7)<sup>36</sup> analysis against the Tetrapoda and Vertebrata databases.

**Repetitive sequence annotation**

The annotation process began with de novo prediction using RepeatModeler2 (v2.0.1)<sup>37</sup>, which integrates outputs from RECON (v1.0.8)<sup>38</sup> and RepeatScout (v1.0.6)<sup>39</sup>. Predicted repeats were classified via RepeatClassifier with reference to the Dfam (v3.5) database. Long terminal repeat (LTR) retrotransposons were specifically annotated using LTR\_retriever (v2.9.0)<sup>40</sup>, which combines predictions from LTRharvest (v1.5.10)<sup>41</sup> and LTR\_FINDER (v1.07)<sup>42</sup>.

A custom repeat library was constructed by merging de novo predictions with known repeat entries from Dfam, followed by deduplication. This library was then used as input for RepeatMasker (v4.1.2)<sup>43</sup> to systematically annotate transposable elements (TEs) across the genome (Supplementary Fig. 13 and 14a).

Tandem repeats were identified through complementary approaches: microsatellites were predicted using MISA (v2.1)<sup>44</sup>, while larger tandem repeats were annotated with Tandem Repeat Finder (TRF, v4.09)<sup>45</sup> under parameters: 2 7 7 80 10 50 500 -d -h.

SDs were identified using BISER (v1.4)<sup>46</sup>. To ensure reliability, candidate SDs were retained only if they satisfied the following criteria: (1) sequence identity greater than 90%; (2) alignment gaps covering no more than 50% of the sequence; (3) aligned length of at least 1 kb; and (4) satellite

repeats, as annotated by RepeatMasker, comprising no more than 70% of the sequence.

### Protein-coding gene annotation and evaluation

Protein-coding genes were annotated through an integrated approach combining de novo prediction, homology-based inference, and transcriptome evidence (Supplementary Fig. 14). De novo gene structures were predicted using Augustus (v3.1.0)<sup>47</sup> and SNAP (2006-07-28 release)<sup>48</sup>, leveraging intrinsic genomic features. Homology-based predictions were generated with GeMoMa (v1.7)<sup>49</sup> through cross-species alignment of orthologous sequences. Transcriptome-guided annotation utilized two complementary strategies: (1) Illumina RNA-seq reads were aligned to the genome via HISAT2 (v2.1.0)<sup>50</sup>, assembled into transcripts with StringTie (v2.1.4)<sup>51</sup>, and processed by GeneMarkS-T (v5.1)<sup>52</sup> for gene prediction; (2) full-length Iso-seq transcripts assembled through RNA-Bloom (v2.0.0)<sup>53</sup> were integrated with PASA (v2.4.1)<sup>54</sup> to model isoform diversity. Additionally, long-read transcripts were aligned to the genome using GMAP (2020-06-30 release)<sup>55</sup>, followed by splice junction refinement with PASA.

Predictions from all three approaches (de novo, homology-based, and transcriptome-guided) were consolidated using EvidenceModeler (EVM, v1.1.1)<sup>56</sup> and further polished with PASA. The final gene set was functionally annotated against multiple databases, including NR, eggNOG (v5.0)<sup>57</sup>, Gene Ontology (GO), KEGG<sup>58</sup>, TrEMBL, KOG, SWISS-PROT<sup>59</sup>, and Pfam (v35.0)<sup>60</sup>.

### Non-coding RNA and pseudogene annotation

Non-coding RNAs were systematically annotated using a combination of tool-based predictions and database alignments. Transfer RNAs (tRNAs) were identified using tRNAscan-SE (v1.3.1)<sup>61</sup>, while ribosomal RNAs (rRNAs) were predicted via barnap (v0.9). Small non-coding RNAs, including microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), and small nuclear RNAs (snRNAs), were annotated by aligning sequences to the Rfam database (v14.9)<sup>62</sup> with Infernal (v1.1)<sup>63</sup>.

Pseudogenes, defined as genomic loci homologous to functional genes but rendered non-functional by frameshift mutations or premature stop codons, were identified through a two-step pipeline: (1) GenBlastA (v1.0.4)<sup>64</sup> was used to detect homologous sequences in the genome after masking functional gene loci; (2) putative pseudogenes were validated by screening for disruptive mutations (e.g., premature termination codons, indels) using GeneWise (v2.4.1)<sup>65</sup>. Protein domain motifs were annotated with InterProScan (v5.34-73.0)<sup>66</sup>.

### Centromere and telomere identification

Centromeric regions were identified based on their distinct genomic signatures, including minimal gene density, reduced interspersed repeat content, stable sequence entropy, elevated tandem repeat density, GC-rich sequences, and cytosine methylation patterns. Initial localization of centromeres was performed by aligning NCBI-curated porcine centromeric repeats to the NJP-T2T genome using BLAST (v2.15.0)<sup>67</sup> with stringent thresholds (identity  $\geq 90\%$ , e-value  $< 1e-10$ ), followed by calculating coverage within 20 kb sliding windows. Sequence complexity was assessed in 20 kb windows using NeSSie<sup>68</sup>, while genomic features such as gene density, tandem/inverted repeats, and GC content were quantified using prior annotation data. Cytosine methylation (5mC) profiles were generated from ONT fast5 data through Nanopolish (v0.14.0)<sup>69</sup> and Minimap2<sup>34</sup> alignments, with methylation coverage evaluated in 20 kb windows via BEDTools (v2.31.1)<sup>70</sup>. Putative centromeric regions were further analyzed by constructing k-mer libraries with KMC<sup>71</sup> (parameters: -fm -k151 -ci10 -cs100000), followed by identification of centromeric repeat monomers using SRF<sup>72</sup> based on k-mer frequency distributions. Final centromere boundaries were determined by integrating chromosomal architectural features and repeat unit patterns. Telomeres were annotated using quarTeT (v1.2.4)<sup>73</sup> with species-specific parameters (-c animal -m 10).

### Comparative analysis with NCBI reference genome

Given that the NJP-T2T genome was derived from a female individual and therefore lacks the Y chromosome, comparative analyses were restricted to autosomes and the X chromosome. Structural comparisons between NJP-T2T and Sscrofa11.1 were performed by first aligning the genomes using Minimap2<sup>34</sup>, followed by structural annotation and visualization through syri (v1.6.3)<sup>74</sup>. PURs were detected by aligning the NJP-T2T assembly to the Sscrofa11.1 reference using Winnowmap (v2.03)<sup>75</sup>, converting the alignments into PAF format with paftools.js, and extracting unresolved intervals with BEDtools (v2.31.1)<sup>70</sup>.

Resequencing data from 200 individuals (encompassing four indigenous and four exotic pig breeds) were obtained from public repositories. RNA-seq datasets from five additional indigenous and five exotic breeds (10 individuals per breed) were included, combining publicly available data with newly generated sequences from this study. Resequencing reads were aligned to the reference genome using BWA (0.7.17-r1188)<sup>76</sup>, with mapping statistics (mapping rate, base error rate, proper pair rate) calculated via SAMtools (v1.9)<sup>77</sup>. SV detection was implemented through a pipeline integrating discordant read pairs and split reads extracted from BWA alignments after duplicate removal with Picard. SVs were identified using Lumpy-sv (v0.2.13)<sup>78</sup> and genotyped with svtyper (v0.7.1)<sup>79</sup>. RNA-seq reads were aligned to the genome using HISAT2 (v2.2.1)<sup>80</sup>, with alignment metrics systematically quantified.

### Gene family analysis

Genome assemblies and GFF3 annotation files for Bamei pig (GCA\_001700235.1), Jinhua pig (GCA\_001700295.1), Meishan pig (GCA\_001700195.1), Tibetan pig (GCA\_000472085.2), Rongchang pig (GCA\_001700155.1), Berkshire (GCA\_001700575.1), Duroc (GCA\_000003025.6), Landrace (GCA\_001700215.1), and Large White pig (GCA\_001700135.1) were obtained from public repositories (Ensembl). Gene family identification was performed using OrthoFinder v2.5.4<sup>81</sup> by integrating the NJP-T2T genome assembly with datasets from the 10 aforementioned species. Unannotated genes in NJP-T2T were functionally characterized via eMapper (v2.1.9) against the EggNOG database (parameters: -m mmseqs -evalue 1e-5), which provides comprehensive functional annotations including Gene Ontology (GO) terms, KEGG pathways, and COG categories. The chromosomal distribution of unassigned genes was visualized using the RIdeogram<sup>82</sup> package with a sliding window size of 1000 kb.

### Variant detection

Raw sequencing data were first quality-assessed using FastQC (v0.11.9)<sup>83</sup>, followed by adapter trimming and low-quality sequence removal with Trimmomatic (v0.39)<sup>84</sup> to generate high-quality clean reads. Clean reads were aligned to NJP-T2T using the BWA-MEM algorithm in BWA (0.7.17-r1188)<sup>76</sup>. Resulting BAM files were sorted with SAMtools (v1.9)<sup>77</sup> and deduplicated using Picard's MarkDuplicates module (parameter: REMOVE\_DUPLICATES=true). The deduplicated BAM files were then used for variant detection with GATK (v4.1.9.0)<sup>85</sup>. Single nucleotide polymorphism (SNP) loci were extracted using SelectVariants (-select-type SNP) and filtered via VariantFiltration with thresholds: QD  $< 2.0$  || MQ  $< 40.0$  || FS  $> 60.0$  || SOR  $> 3.0$  || MQRankSum  $< -12.5$  || ReadPosRankSum  $< -8.0$ . Sites with  $>10\%$  missing genotypes were removed using VCFtools (v0.1.17)<sup>86</sup>, yielding a variant set for population genetic analysis. Further filtering excluded loci with minor allele frequency (MAF)  $< 0.05$ , retaining high-quality biallelic SNPs for genome-wide association and selective sweep analyses.

### Population genetic analysis

Genetic diversity, inbreeding levels, and runs of homozygosity (ROH) within the Neijiang pig conservation population were analyzed using PLINK (v1.90)<sup>87</sup>. SNP loci were filtered to exclude sites with missing rates  $>10\%$  or significant deviations from Hardy-Weinberg equilibrium (HWE;  $p < 1e-6$ ). Allele frequency spectra were generated via PLINK's --freq

parameter, with effective allele numbers ( $N_e$ ) calculated as  $1/(p^2 + q^2)$ , where  $p$  and  $q$  represent allele frequencies. Observed heterozygosity ( $H_o$ ), expected heterozygosity ( $H_e$ ), and inbreeding coefficients ( $F$ ) were derived using --het.

Based on genome-wide SNP data, a multiple sequence alignment file in PHYLIP format was generated. Genetic distances were calculated using the dnadist program in the PHYLIP package (v3.697) with the default substitution model. A neighbor-joining (NJ) phylogenetic tree was then constructed using the neighbor program. The resulting tree was exported in Newick format and visualized using MEGA11<sup>88</sup>. Genetic relatedness matrices were generated via GCTA (v1.94.1)<sup>89</sup> to assess kinship patterns. ROH segments were identified with PLINK parameters: --homozyg --homozyg-density 50 --homozyg-gap 100 --homozyg-kb 500 --homozyg-snp 50 --homozyg-window-het 1 --homozyg-window-snp 50 --homozyg-window-threshold 0.05.

### Genome-wide association study

To enhance genotype accuracy from low-coverage sequencing data, the STITCH-Beagle imputation pipeline was implemented. The STITCH algorithm (v1.7.1)<sup>90</sup>, optimized for genotype imputation using low-depth data, was first applied with parameters  $K = 10$  --nGen=320 --method = "diploid" --nIterations=40 to perform self-imputation on variant data from 115 phenotypically well-characterized Neijiang pigs, achieving an imputation accuracy of  $DR^2 = 0.85$ . Subsequently, Beagle (v5.4)<sup>91</sup> performed reference-based imputation using a dedicated reference panel of eight deeply sequenced Neijiang pig genomes (>100× coverage).

Imputed data were filtered through PLINK<sup>87</sup> using thresholds --geno 0.1 --maf 0.05 --indep-pairwise 50 10 0.3 to retain high-confidence SNPs for analysis. In the GWAS, phenotype data was first corrected by fitting a linear model that controlled for fixed effects, including Farrowing year and Farrowing season. The residuals from this model were used directly as the corrected phenotype data for subsequent GWAS analysis. GWAS analysis was performed using the EMMAX (v.20120210)<sup>92</sup>, employing a mixed linear model (MLM) for association analysis. The model is expressed as:

$$y = X\beta + G_\gamma + Zu + \epsilon$$

where  $y$  is the phenotype vector,  $X$  is the fixed effect design matrix,  $\beta$  is the fixed effect parameters,  $G$  is the genotype matrix,  $\gamma$  is the genotype effect parameters,  $Z$  is the random effect matrix, representing the kinship matrix,  $u$  is the random effect, and  $\epsilon$  is the error term. The kinship matrix  $Z$  was used to account for population structure, and genome-wide significance thresholds were defined as  $1/n$ , where  $n$  represents the effective number of independent SNPs.

### Selective signatures of head morphology in Neijiang pigs

Genome-wide scans for selection signatures were performed using VCFtools (v0.1.17)<sup>86</sup>. Nucleotide diversity ( $\pi$ ) was calculated in 100 kb sliding windows with 10 kb step size (--window-pi 100000 --window-pi-step 10000). The ratio of nucleotide diversity (ROD) between two phenotypic groups was derived from  $\pi$  values. Population differentiation was assessed through fixation index (FST) analysis using VCFtools parameters --fst-window-size 100000 --fst-window-step 10000.

Genomic regions under selection were identified by intersecting the top 5% of both FST and ROD values. Adjacent candidate windows were merged using BEDtools (v2.31.1)<sup>70</sup> merge command. Genes overlapping these regions by at least 10% (assessed via BEDtools intersect -F 0.1) were retained as candidate selected genes. Gene structure visualization was implemented using GSDS (v2.0)<sup>93</sup>.

### RT-qPCR assay

Three individuals each from the extreme head shape EFT group and the SZT group were selected for the collection of olfactory epithelium tissue. Total RNA was extracted from the cells using Trizol reagent (Invitrogen Corporation, Grand Island, NY, USA) following the manufacturer's

instructions. The extracted RNA was reverse-transcribed into cDNA using the PrimeScript RT reagent Kit (RR037Q) from Takara Bio, according to the recommended protocol provided with the kit.

Real-time fluorescence quantification of the synthesized cDNA was conducted using the Hieff UNICON® Universal Blue qPCR SYBR Green Master Mix kit (NextGen). The reaction mixtures were prepared, mixed, and centrifuged following the kit instructions. Fluorescence signal detection and quantification were performed on a BIO-RAD real-time PCR instrument. The primer sequences used for each gene in the qPCR analysis are listed in Supplementary Data 29.

### Statistics and reproducibility

Statistical analyses and reproducibility information for this study are summarized as follows. Population genetic parameters and selective sweeps were analyzed using PLINK and VCFtools. Genome-wide association studies (GWAS) were performed using EMMAX under a mixed linear model, with the genome-wide significance threshold set at  $p < 1/n$  (where  $n$  is the number of effective independent SNPs). Gene expression validation by RT-qPCR (three biological replicates per group) was assessed using a two-tailed Student's t-test ( $p < 0.05$ ). All raw sequencing data generated in this study have been deposited in public repositories. Specific sample sizes are provided in the respective methodological sections: de novo genome assembly ( $n = 1$ ), population/GWAS cohort ( $n = 115$ ), and selection sweep cohort ( $n = 8$ ).

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

All data generated in this study have been deposited in public repositories. The HiFi, Hi-C, Ultralong ONT, Iso-seq, and RNA-seq data and the haplotype-resolved genome assembly are available under BioProject PRJNA1126453 at the National Center for Biotechnology Information (NCBI). The whole-genome resequencing data for population genetic analysis are archived under BioProject PRJNA1373490. GWAS summary statistics, and the numerical source data for the figures are available via the Figshare repository at <https://doi.org/10.6084/m9.figshare.28645085>. For comparative analyses, publicly available genomic data were retrieved from BioProjects PRJNA947797, PRJNA531381, PRJNA842867, and PRJNA860033, and transcriptomic data from BioProjects PRJNA322309, PRJNA998604, and PRJNA309102.

### Code availability

The codes used in this study are available on GitHub at <https://github.com/ccchendong/NJP-T2T>.

Received: 19 June 2025; Accepted: 9 January 2026;

Published online: 16 January 2026

### References

- Giuffra, E. et al. The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics* **154**, 1785–1791 (2000).
- Chen, K., Baxter, T., Muir, W. M., Groenen, M. A. & Schook, L. B. Genetic resources, genome mapping and evolutionary genomics of the pig (*Sus scrofa*). *Int. J. Biol. Sci.* **3**, 153–165 (2007).
- Lukić, B. et al. Conservation genomic analysis of the Croatian indigenous black Slavonian and Turpolje pig breeds. *Front. Genet.* **11**, 261 (2020).
- Li, Q., Zhu, L. & He, Z. P. *Sichuan Animal and Poultry Genetic Resources (in Chinese)*. (China Agriculture Press, 2024).
- Dan, H. et al. Integrated transcriptomic and metabolomic analyses reveal heterosis for meat quality of Neijiang pigs. *Front. Vet. Sci.* **11**, 1493284 (2024).

6. Warr, A. et al. An improved pig reference genome sequence to enable pig genetics and genomics research. *Gigascience* **9**, <https://doi.org/10.1093/gigascience/giaa051> (2020).
7. Li, M. et al. Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple de novo assemblies. *Genome Res.* **27**, 865–874 (2017).
8. Zhang, S. et al. Integrated analysis of the complete sequence of a macaque genome. *Nature*, <https://doi.org/10.1038/s41586-025-08596-w> (2025).
9. Yang, C. et al. The complete and fully-phased diploid genome of a male Han Chinese. *Cell Res.* **33**, 745–761 (2023).
10. Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
11. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
12. Chen, J. et al. A complete telomere-to-telomere assembly of the maize genome. *Nat. Genet.* **55**, 1221–1231 (2023).
13. Huang, Z. et al. Evolutionary analysis of a complete chicken genome. *Proc. Natl. Acad. Sci. USA* **120**, e2216641120 (2023).
14. Wang, Y. et al. Telomere-to-telomere and haplotype-resolved genome of the kiwifruit *Actinidia chinensis*. *Mol. Hortic.* **3**, 4 (2023).
15. Putnam, N. H. et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
16. Westerling, T., Kuuluvainen, E. & Mäkelä, T. P. Cdk8 is essential for preimplantation mouse development. *Mol. Cell. Biol.* **27**, 6177–6182 (2007).
17. Wu, P. et al. Whole-genome sequence association study identifies cyclin dependent kinase 8 as a key gene for the number of mummified piglets. *Anim. Biosci.* **36**, 29–42 (2023).
18. McIlwain, D. R. et al. Smg1 is required for embryogenesis and regulates diverse genes via alternative splicing coupled to nonsense-mediated mRNA decay. *Proc. Natl. Acad. Sci. USA* **107**, 12186–12191 (2010).
19. Luo, L. et al. Telomere-to-telomere sheep genome assembly identifies variants associated with wool fineness. *Nat. Genet.* **57**, 218–230 (2025).
20. Zong, W. et al. Two telomere-to-telomere pig genome assemblies and pan-genome analyses provide insights into genomic structural landscape and genetic adaptations. *iMeta* **4**, e70013 (2025).
21. Cao, C. et al. A near telomere-to-telomere genome assembly of the Jinhua pig: enabling more accurate genetic research. *Gigascience* **14**, <https://doi.org/10.1093/gigascience/giaf048> (2025).
22. Wang, Y. et al. A chromosome-level genome of Chenghua pig provides new insights into the domestication and local adaptation of pigs. *Int J. Biol. Macromol.* **270**, 131796 (2024).
23. Zhou, R. et al. The Meishan pig genome reveals structural variation-mediated gene expression and phenotypic divergence underlying Asian pig domestication. *Mol. Ecol. Resour.* **21**, 2077–2092 (2021).
24. Altemose, N. et al. Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
25. Li, B. et al. A gap-free reference genome reveals structural variations associated with flowering time in rapeseed (*Brassica napus*). *Hortic. Res.* **10**, <https://doi.org/10.1093/hr/uhad171> (2023).
26. Sun, M. et al. Telomere-to-telomere pear (*Pyrus pyrifolia*) reference genome reveals segmental and whole genome duplication driving genome evolution. *Hortic. Res.* **10**, uhad201 (2023).
27. Feng, Y. et al. The haplotype-resolved T2T genome assembly of the wild potato species *Solanum commersonii* provides molecular insights into its freezing tolerance. *Plant Commun.* **5**, 100980 (2024).
28. Wu, H. et al. Telomere-to-telomere genome assembly of a male goat reveals variants associated with cashmere traits. *Nat. Commun.* **15**, 10041 (2024).
29. Li, Y. Comparing of backfat microRNAomes of Landrace and Neijiang pig by high-throughput sequencing. *Genes Genomics* **43**, 543–551 (2021).
30. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
31. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
32. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. **356**, 92–95, (2017).
33. Chen, Y. et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat. Commun.* **12**, 60 (2021).
34. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
35. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
36. Seppely, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019).
37. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
38. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
39. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
40. Ou, S. & Jiang, N. LTR\_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
41. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.* **9**, 18 (2008).
42. Xu, Z. & Wang, H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* **35**, W265–W268 (2007).
43. TarailoGraovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **4**, 4.10.11–14.10.14 (2009).
44. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585 (2017).
45. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
46. Išerić, H., Alkan, C., Hach, F. & Numanagić, I. Fast characterization of segmental duplication structure in multiple genome assemblies. *Algorithms Mol. Biol. AMB* **17**, 4 (2022).
47. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
48. Korf, I. Gene finding in novel genomes. *BMC Bioinform.* **5**, 59 (2004).
49. Keilwagen, J. et al. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**, e89 (2016).
50. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
51. Perteira, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
52. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **43**, e78 (2015).
53. Nip, K. M. et al. RNA-Bloom enables reference-free and reference-guided sequence assembly for single-cell transcriptomes. *Genome Res.* **30**, 1191–1200 (2020).

54. Haas, B. J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
55. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
56. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
57. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
58. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
59. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
60. Finn, R. D. et al. Pfam: clans, web tools and services. *Nucleic Acids Res.* **34**, D247–D251 (2006).
61. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).
62. Griffiths-Jones, S. et al. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2005).
63. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
64. She, R., Chu, J. S., Wang, K., Pei, J. & Chen, N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res.* **19**, 143–149 (2009).
65. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
66. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
67. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
68. Berselli, M., Lavezzo, E. & Toppo, S. NeSSie: a tool for the identification of approximate DNA sequence symmetries. *Bioinformatics* **34**, 2503–2505 (2018).
69. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).
70. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
71. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
72. Zhang, Y., Chu, J., Cheng, H. & Li, H. De novo reconstruction of satellite repeat units from sequence data. *ArXiv* (2023).
73. Lin, Y. et al. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic. Res.* **10**, uhad127 (2023).
74. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
75. Jain, C., Rhie, A., Hansen, N. F., Koren, S. & Phillippy, A. M. Long-read mapping to repetitive reference sequences using Winnommap2. *Nat. Methods* **19**, 705–710 (2022).
76. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
77. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
78. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
79. Chiang, C. et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**, 966–968 (2015).
80. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
81. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
82. Hao, Z. et al. Rldeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput. Sci.* **6**, e251 (2020).
83. de Sena Brandine, G. & Smith, A. D. Falco: high-speed FastQC emulation for quality control of sequencing data. *F1000Res* **8**, 1874 (2019).
84. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
85. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
86. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
87. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
88. Koichiro, T., Glen, S., Sudhir, K. J. M. B. & Evolution. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. 7 (2021).
89. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
90. Davies, R. W. et al. Rapid genotype imputation from sequence with reference panels. *Nat. Genet.* **53**, 1104–1111 (2021).
91. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).
92. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
93. Hu, B. et al. GSDS 2.0: an upgraded gene feature visualization server. *Bioinformatics* **31**, 1296–1297 (2014).

## Acknowledgements

The study was supported by grants from the Sichuan Science and Technology Program (No.2021ZDZX0008, 2020YFN0024, 2021YFYZ0030) and the National Center of Technology Innovation for Pigs (NCTIP-XD/B01); and was also supported by Chinese National Neijiang Pig Genetic Resources Conservation Farm, Sichuan Market Swine Genetic Big Data Evaluation Laboratory of the State Pig E-market and Sichuan Hengtong Neijiang Pig Breeding and Conservation Co., Ltd.

## Author contributions

Dong Chen: Writing—Original Draft, Methodology, Conceptualization. Shengdi Cui: Methodology, Validation. Zhenjian Zhao: Formal analysis. Pingxian WU: Resources, Supervision. Wenxuan Zhou: Software. Patrick Kofi Makafui Tecku: Visualization. Kai Wang: Project administration, Resources. Junge Wang: Validation. Ziyang Chen: Data Curation. Shixin Yu: Investigation. Jiamiao Chen: Validation. Yaoxi Zhou: Formal analysis. Runjie Huang: Investigation. Xuewei Li: Funding acquisition. Guoqing Tang: Conceptualization, Project administration, Resources.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s42003-026-09557-3>.

**Correspondence** and requests for materials should be addressed to Guoqing Tang.

**Peer review information** *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Matteo Dell'Acqua & Rosie Bunton-Stasyshyn. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026