BMC Biology

COMMENT

Open Access

Pangenome graph mitigates heterozygosity overestimation from mapping bias: a case study in Chinese indigenous pigs

Jian Miao[1], Qingyu Wang[1], Zhe Zhang[1], Qishan Wang[1,2], Yuchun Pan[1,2*] and Zhen Wang[1*]

Abstract

Background Breeds genetically distant from the reference genome often show considerable differences in DNA fragments, making it difficult to achieve accurate mappings. The genetic differences between pig reference genome (Sscrofa11.1) and Chinese indigenous pigs may lead to mapping bias and affect subsequent analyses.

Results Our analysis revealed that pangenome exhibited superior mapping accuracy to the Sscrofa11.1, reducing false-positive mappings by 1.4% and erroneous mappings by 0.8%. Furthermore, the pangenome yielded more accurate genotypes of SNP (F1: 0.9660 vs. 0.9607) and INDEL (F1: 0.9226 vs. 0.9222) compared to Sscrofa11.1. In real sequencing data, the inconsistent SNPs called from the pangenome exhibited lower genome heterozygosity compared to those identified by the Sscrofa11.1, including observed heterozygosity and nucleotide diversity. The same reduction of heterozygosity overestimation was also found in the chicken pangenome.

Conclusions This study quantifies the mapping bias of Sscrofa11.1 in Chinese indigenous pigs, demonstrating that mapping bias can lead to an overestimation of heterozygosity in Chinese indigenous pig breeds. The adoption of a pig pangenome mitigates this bias and provides a more accurate representation of genetic diversity in these populations.

Keywords Mapping bias, Pig, Pangenome, Genome graph, Variant calling

Background

The reference genome is essential for medical, comparative, and population genomic analyses. It provides a standardized coordinate system that facilitates the comparison of various experimental results and establishes a consistent framework for genome mapping, annotation, and interpretation [1]. Nearly all sequencing-related studies start by mapping sequence reads to the reference genome. However, relying solely on a single reference genome is insufficient for capturing the complete genomic diversity within a species [2]. This limitation gives rise to a phenomenon known as reference bias, where aligning reads containing non-reference alleles to a single reference genome often results in missing or incorrect mappings [3, 4]. Such biases can significantly impact downstream analyses like calling variants [5–8], quantifying gene expression [9], and accurately determining epigenomic peaks [6, 10].

China hosts a rich diversity of indigenous pig breeds, accounting for approximately one-third of the world's pig breeds [11]. These breeds are distributed across diverse geographical regions in China and display a wide range

*Correspondence:
Yuchun Pan
panyc@zju.edu.cn
Zhen Wang
wangzhen20@zju.edu.cn
[1] College of Animal Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China
[2] Hainan Institute of Zhejiang University, Yazhou Bay Science and Technology City, Building 11, Yongyou Industrial Park, Yazhou District, Sanya, Hainan 572025, China

of phenotypic traits, such as variations in body size, coat color, reproductive performance, and disease resistance. The extensive genetic diversity of Chinese indigenous pigs makes them invaluable resources for understanding the genetic basis of economically important traits and improving commercial pig breeds through crossbreeding and genetic introgression. The current pig reference genome (Sscrofa11.1) was assembled from a female European Duroc pig [12]. However, European pigs exhibit substantial genetic divergence from Chinese domestic pigs, as a consequence of approximately 10,000 years of separate domestication [13]. This divergence means that the Sscrofa11.1 may not accurately represent the genetic diversity of Chinese pig breeds. To date, almost all genomic studies of Chinese indigenous pigs have used Sscrofa11.1 for read mapping [14–17], but the impact of reference bias in mapping reads from these pigs to Sscrofa11.1 remains unclear. Given the complex genetic characteristics of Chinese pigs, there is a need for more representative genomic references for accurate genetic and genomic studies [18, 19].

The recently developed graph genome structure intuitively incorporates additional genomic variations into the linear reference genome, offering a promising approach to mitigate reference bias [20]. Although graph genomes have been constructed for several farm animals [21–25], few studies have assessed their contribution to reducing mapping bias. Previous research comparing the bovine graph genome with the linear reference genome demonstrated that incorporating additional variants from whole-genome sequencing (WGS) into a graph genome significantly improved mapping and genotyping accuracy [26]. However, the standard variant calling pipeline for WGS still requires mapping reads to a linear reference genome, which introduces its own biases. Graph genomes generated directly from multiple assemblies may avoid such biases by avoiding mapping short sequencing reads to the reference genome [27–30]. Therefore, there is a need to compare the performance of augmented graph genomes and assembly-derived graph genomes in reducing reference bias.

In this study, we initially constructed a pig pangenome using ten assemblies from different pig breeds. Specifically, for the Chinese indigenous Meishan (MS) pig, we generated several customized graph genomes by incorporating various types of genetic variations from WGS of MS pigs into the Sscrofa11.1. Using simulated reads, we first quantified the mapping bias of Sscrofa11.1 against the MS pig. We then evaluated the impact of different graph genomes on the accuracy of read mapping and variant genotyping. Finally, we compared the performance of the pangenome and Sscrofa11.1 in analyzing WGS and epigenetic sequencing data. Our findings highlight the potential for graph genomes to transform genomic research, particularly in fields requiring precise genetic analysis.

Results

Construction of the graph genomes

The average sequence coverage of the 28 MS WGS was 23×, ranging from 15×to 71×. We identified a total of 1,265,789 SNPs, 232,502 short INDELs, and 4391 SVs on Sscrofa11.1-chr5 to generate the four customized MS graphs. By considering a wide range of pig breeds, the single-chromosome pangenome graph contains more genetic variation than customized MS graphs. Using the Minigraph-Cactus pipeline, we identified 1,346,364 SNPs, 305,380 INDELs, 1,336,706 multiple nucleotide polymorphisms, and 28,585 other complex variants (including complex substitutions and SVs). Phylogenetic analysis of the assemblies used in pangenome construction highlights the clear genetic difference between Sscrofa11.1 and Chinese local pig breeds (Additional file 2: Fig. S1). For the complete pangenome, we identified 30,086,854 SNPs, 7,894,354 INDELs, and 721,192 complex variants. In total, 42,061,158 non-reference sequences (NRSs) with a cumulative length of 160 Mb were added to the Sscrofa11.1.

Mapping accuracy

The mapping performance of the six graph genomes (including four customized genomes from MS pigs, the Sscrofa11.1-graph, and the pangenome) and the linear reference genome (Sscrofa11.1-linear) were evaluated using 10×reads simulated from MS-chr5. The Sscrofa11.1-linear yielded obviously higher number of mapped reads than Sscrofa11.1-graph (Fig. 1A). This is likely due to the default parameters of BWA-MEM having a higher tolerance for alignment errors compared to VG Giraffe. For the mapping rates of graph genomes, the four customized MS genomes (ranging from 97.72 to 97.82%) were slightly higher than the Sscrofa11.1-graph (97.67%), while the pangenome (98.14%) was notably higher than all other graph genomes (Additional file 1: Table S1). We noticed that some reads lost their mapped positions during the process of surjecting from graph structure to linear space. The reduction in mapped reads is negligible for SNP-graph and SHORT-graph. However, a substantial number of reads lost their mapping coordinates for SV-graph, ALL-graph, and pangenome (Fig. 1A, Additional file 1: Table S1). The pangenome, which includes the most SVs, lost about 1.8% mapped reads after surjection, resulting in notably lower mapping rates compared to other graph genomes. This indicates that adding more SVs to the reference genome may result in a greater loss of mapping information for reads during the surjection.

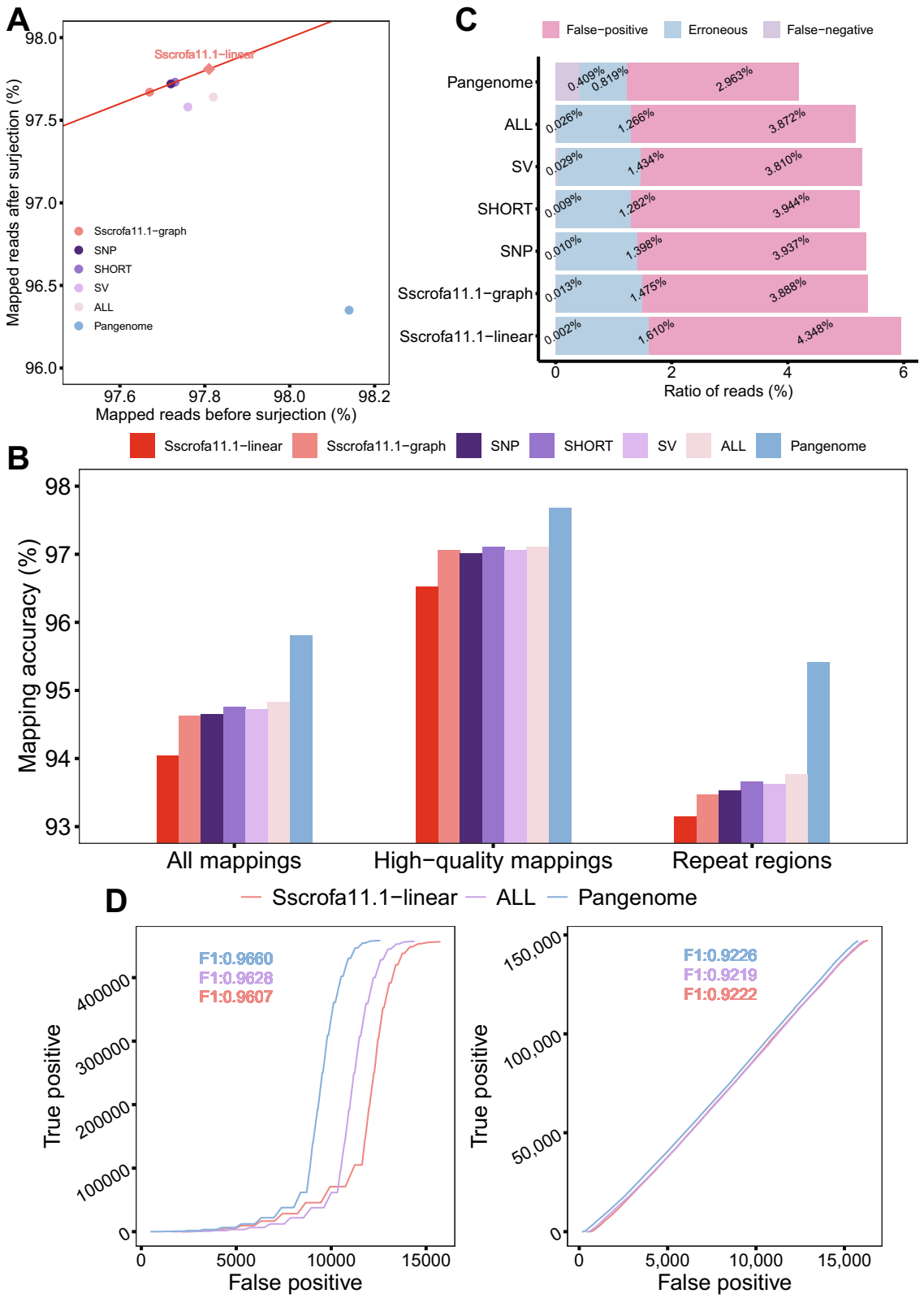

Fig. 1 The mapping and genotyping performance of different genomes. **A** The ratio of mapped reads in the graph space (x-axis) and in the linear space (y-axis). **B** The accuracy of all mappings, high-quality mappings (mapping quality > 30), and mappings in repeat regions. **C** The ration of different mapping bias. The mapping bias was classified into three types: false-positive mappings, false-negative mappings, and erroneous mappings. **D** Genotyping accuracy of SNPs (left panel) and short INDELs (right panel)

We first evaluated the mapping performance of BWA-MEM and VG by comparing the mapping accuracy of Sscrofa11.1-linear and Sscrofa11.1-graph (Additional file 1: Table S2). Despite that VG exhibited lower mapping rates than BWA-MEM, it achieved higher mapping accuracy (94.62% vs. 94.04%, Fig. 1B). We also compared the mapping accuracy of different graph genomes using VG. All customized genomes achieved slightly higher mapping accuracy than Sscrofa11.1-graph, with improvements ranging from 0.03 to 0.22%. The pangenome achieved a notable improvement of mapping accuracy compared to Sscrofa11.1-graph (95.81% vs. 94.62%). For reads with high-mapping quality (quality score > 30) and those mapped to repeat regions, the pangenome maintained a superior mapping accuracy compared to all other genomes (Fig. 1B). Compared to Sscrofa11.1-linear, using the pangenome for mapping improved accuracy in repetitive regions by 2.27%, which is significantly higher than improvements in high-quality mappings (1.16%) and all mappings (1.77%).

To mitigate potential biases that may arise from simulating only the chr5, we selected the longest chromosome (chr1) and the shortest chromosome (chr18) in the pig reference genome for the same simulation analysis. Overall, compared to the linear reference, the pangenome graph still significantly improved the alignment accuracy of reads (Additional file 1: Table S3). Specifically, on chr1 and chr18, the alignment accuracy increased by approximately 1.3% and 1%, respectively.

The sources of mapping bias

We classified the sources of mapping bias into three types: false-positive mappings, false-negative mappings, and erroneous mappings (Fig. 1C, Additional file 1: Table S4). In Sscrofa11.1-linear mappings, about 4.35% of reads were false-positively mapped, and 1.6% of reads were mapped to incorrect positions. The false-negative mapping rate for Sscrofa11.1-linear was almost negligible. Compared to Sscrofa11.1-linear, the false-positive mapping rate for Sscrofa11.1-graph decreased to approximately 0.46% (Fig. 1C). This higher rate of false-positive mappings in Sscrofa11.1-linear may be attributed to the relaxed default mapping parameters of BWA-MEM. Therefore, we tuned the penalty of mismatch penalty (parameter B) and read clipping (parameter L) in BWA to adjust its tolerance for mapping errors. We found that while strict mapping criteria effectively decreased the false-positive mappings, it also reduced the overall mapping rate and the number of correctly mapped reads (Additional file 1: Table S5). All graph genomes reduced erroneous and false-positive mapping to some extent. The pangenome was the most effective genome, decreasing erroneous mappings of Sscrofa11.1-linear by 49%. However, the false-negative mappings of pangenome were higher than those of Sscrofa11.1-linear (0.4% vs. 0.002%). We found that about 99.4% of false-negative mappings in pangenome were mapped to NRSs, losing their coordinates during surjection. Therefore, the unusually high false-negative mapping rate in the pangenome is due to the interference from NRSs.

Genotyping performance

The BAM files from BWA-MEM or those surjected from graph space were supplied to GATK4 HaplotypeCaller for variant calling. Owing to the similar mapping accuracy across customized graph genomes, we selected only the "ALL" customized genome for genotyping. The known variants included 477,930 SNPs and 159,026 INDELs, with 97.9% of the INDELs being less than 30 bp (Additional file 2: Fig. S2). We independently evaluated the genotyping accuracy of SNPs and INDELs shorter than 30 bp. Generally, the graph genome showed higher precision and recall than the linear genome. The ROC plot indicated that both the pangenome and "All" genome improved SNP genotyping accuracy by reducing false-positive SNPs (Fig. 1D). The precision and recall for "ALL" genome were slightly higher than for the Sscrofa11.1-linear genome (precision: 96.95% vs. 96.66%; recall: 95.61% vs. 95.48%), while the pangenome obviously outperformed the Sscrofa11.1-linear genome (precision: 97.33%; recall: 95.89%). The genotyping accuracy of INDELs showed minor differences among genomes. The pangenome exhibited a slightly higher F1 value than the Sscrofa11.1-linear genome, which was also slightly higher than the "ALL" genome (Fig. 1D).

We removed erroneous mappings from the BAM files of both the pangenome and the Sscrofa11.1-linear genome, and subsequently performed SNP genotyping using GATK4 HaplotypeCaller. We observed that after filtering out erroneous mappings from the BAM file of the linear reference genome, its F1 score increased from 0.9607 to 0.9655. In contrast, the pangenome exhibited a negligible improvement, with its F1 score rising from 0.9660 to 0.9663. This minimal increase may be attributed to the already high mapping accuracy of the pangenome.

Genotyping difference between linear reference genome and pangenome

Similar to the simulation analysis, the mapping ratios of Sscrofa11.1-linear were higher than those of the pangenome (99.46% vs. 97.16% on average, Fig. 2A). We assessed mapping errors using the proportion of mate reads that mapped to different chromosomes. The mapping errors of the pangenome were negligible, while Sscrofa11.1-linear showed significantly higher mapping errors (1.42% vs. 0.08% on average, Fig. 2A). After

removing non-autosome and low-quality SNPs, we found that the pangenome identified 322,555 more SNPs than the Sscrofa11.1-linear. Most SNPs (>95%) identified by the pangenome and Sscrofa11.1-linear were common, with fewer being specific to each genome (Additional file 2: Fig. S3). Among the 19,765,180 common SNPs, the overall genotyping concordance between pangenome and Sscrofa11.1-linear was 99.28%. There were 95,873 SNPs with more than 30% inconsistent genotypes, and 632 SNPs showing completely different genotypes (Fig. 2B). To explore if different genotypes enriched in certain genomic regions, we used sliding windows to compare genotyping concordance. We identified 570 high-inconsistent widows with genotyping difference greater than 0.3 (Additional file 1: Table S6). We suggest that association results within these genomic regions should be interpreted with caution. We identified some genomic regions with obviously different genotypes, such as chr2: 50–72 Mb, chr7: 77–80 Mb, chr9: 44–49 Mb, and chr13: 160–162 Mb (Fig. 2C). These high-inconsistent regions may be more susceptible to the reference mapping bias. A total of 24 protein-coding genes were overlap with high-inconsistent widows (Additional file 1: Table S7). We detected a total of 1472 distinct QTL loci overlapping with high-inconsistent widows, among which 962 QTLs (accounting for 65%) are associated with meat quality and carcass traits (Additional file 2: Fig. S4 and Additional file 1: Table S8). In the QTL database we utilized, there are 31,013 QTLs in total, with 11,568 of them related to meat quality and carcass traits. Consequently, through hypergeometric distribution testing, these high-inconsistent widows are significantly enriched for meat quality and carcass traits ($P < 1 \times 10^{-20}$).

Mapping bias leads to overestimation of genome heterozygosity

We compared the differences in observed heterozygosity and nucleotide diversity of the common SNPs genotyped from the pangenome and Sscrofa11.1-linear. A total of 6.17% and 6.3% of SNPs showed different values of observed heterozygosity and nucleotide diversity, respectively. The top 10,000 SNPs with the largest differences in either observed heterozygosity or nucleotide diversity showed higher values in Sscrofa11.1-linear (Fig. 2D–E). We found that the SNPs called by Sscrofa11.1 have higher observed heterozygosity and nucleotide diversity, leading to a noticeably smaller number of runs of homozygosity (ROH). The number of identified ROHs from pangenome was greater than that from Sscrofa11.1-linear (52 vs. 16, Fig. 2F). By mapping genomic sequencing reads of Lueyang black-bone chickens to the chicken linear reference genome and the chicken pangenome, we also found the similar phenomenon of heterozygosity overestimation in chickens (Fig. 3A–D). A total of 1.08% and 1.11% of SNPs showed different values of observed heterozygosity and nucleotide diversity, respectively. The top 10,000 SNPs with the largest differences in either observed heterozygosity or nucleotide diversity showed significantly higher values in the linear reference genome.

Epigenomic data analysis

We compared the differences in peak calling for ATAC-seq and ChIP-seq (H3K27ac and H3K4me3) using the pangenome and the Sscrofa11.1-linear. The results from ATAC-seq and ChIP-seq analyses showed similar patterns in terms of proportion of significantly different peaks. Our analysis revealed that a substantial majority (ranging from 97.81 to 99.54%) of the identified peaks were not significantly different when comparing peaks called against the pangenome and the Sscrofa11.1-linear (Fig. 4A–C). This high proportion of non-significant peaks suggests that the overall landscape of accessible chromatin and protein-DNA binding sites is largely conserved between the two genomic frameworks. The number of Sscrofa11.1-linear-specific peaks was higher than that of pangenome-specific peaks (Fig. 4A–C, Additional file 1: Tables S9–S11), which may be due to the high mapping rates of BWA-MEM. The significantly different peaks identified were predominantly concentrated in unplaced contigs for both ATAC-seq and ChIP-seq (Fig. 4D–F). These unplaced contigs, which often contain complex, repetitive, or highly variable regions, may suffer from inaccurately mapping and therefore pose challenges for peak identification.

(See figure on next page.)

Fig. 2 The comparison of real SNPs called from pangenome and Sscrofa11.1. **A** The ratio of mapped reads (left panel) and mate reads that mapped to different chromosomes (right panel). **B** The number of SNPs under different genotyping differences. **C** The Manhattan plot showing the distribution of sliding windows under different genotyping differences. The colors of the points are used to distinguish different chromosomes. **D** Distribution of observed heterozygosity for the 10,000 SNPs with the largest differences in observed heterozygosity between the pangenome and Sscrofa11.1. **E** Distribution of nucleotide diversity for the 10,000 SNPs with the largest differences in nucleotide diversity between the pangenome and Sscrofa11.1. **F** The distribution of ROHs identified by pangenome and Sscrofa11.1. The blue squares represent the pangenome, while red circles represent Sscrofa11.1

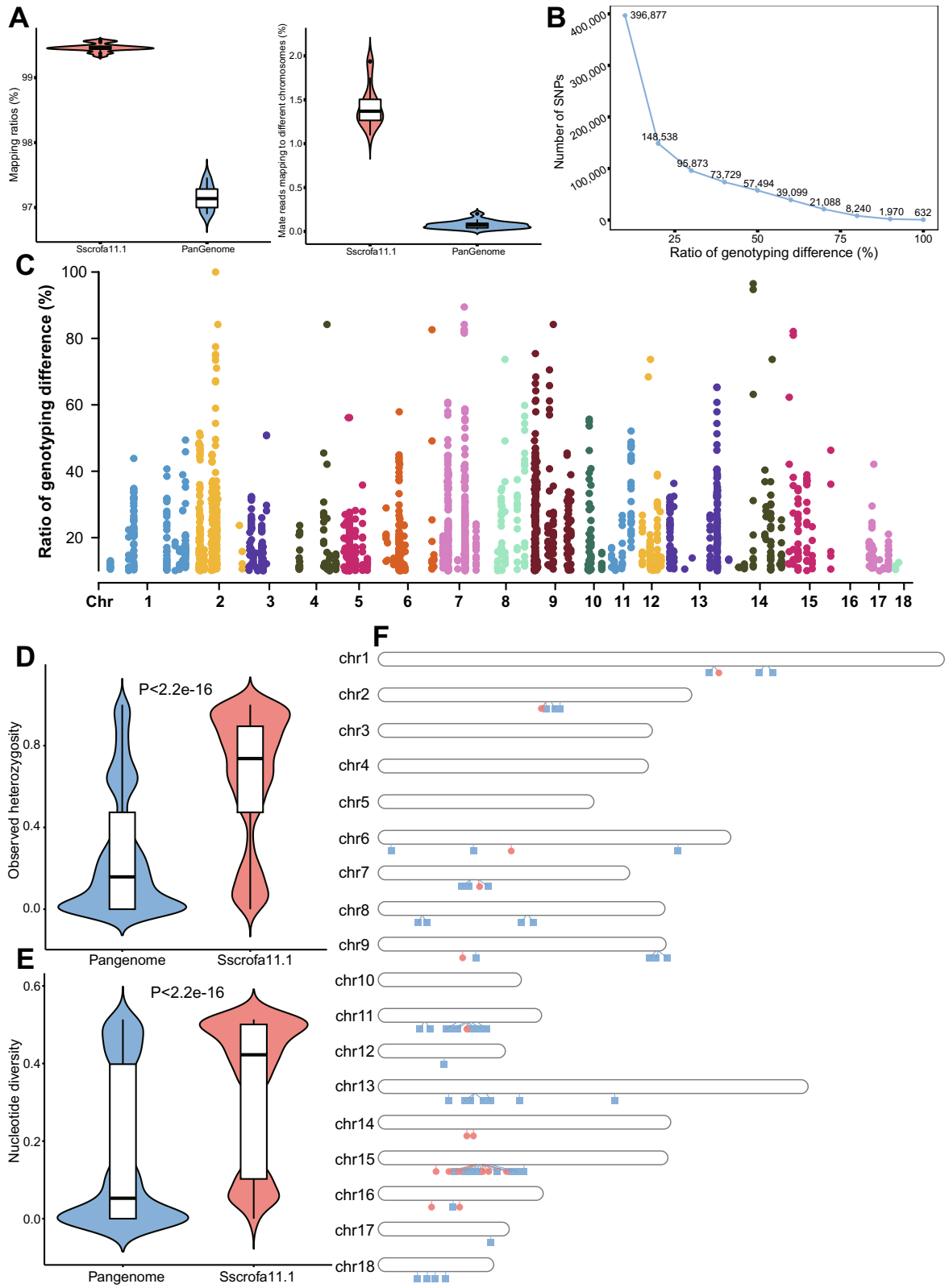

Fig. 2 (See legend on previous page.)

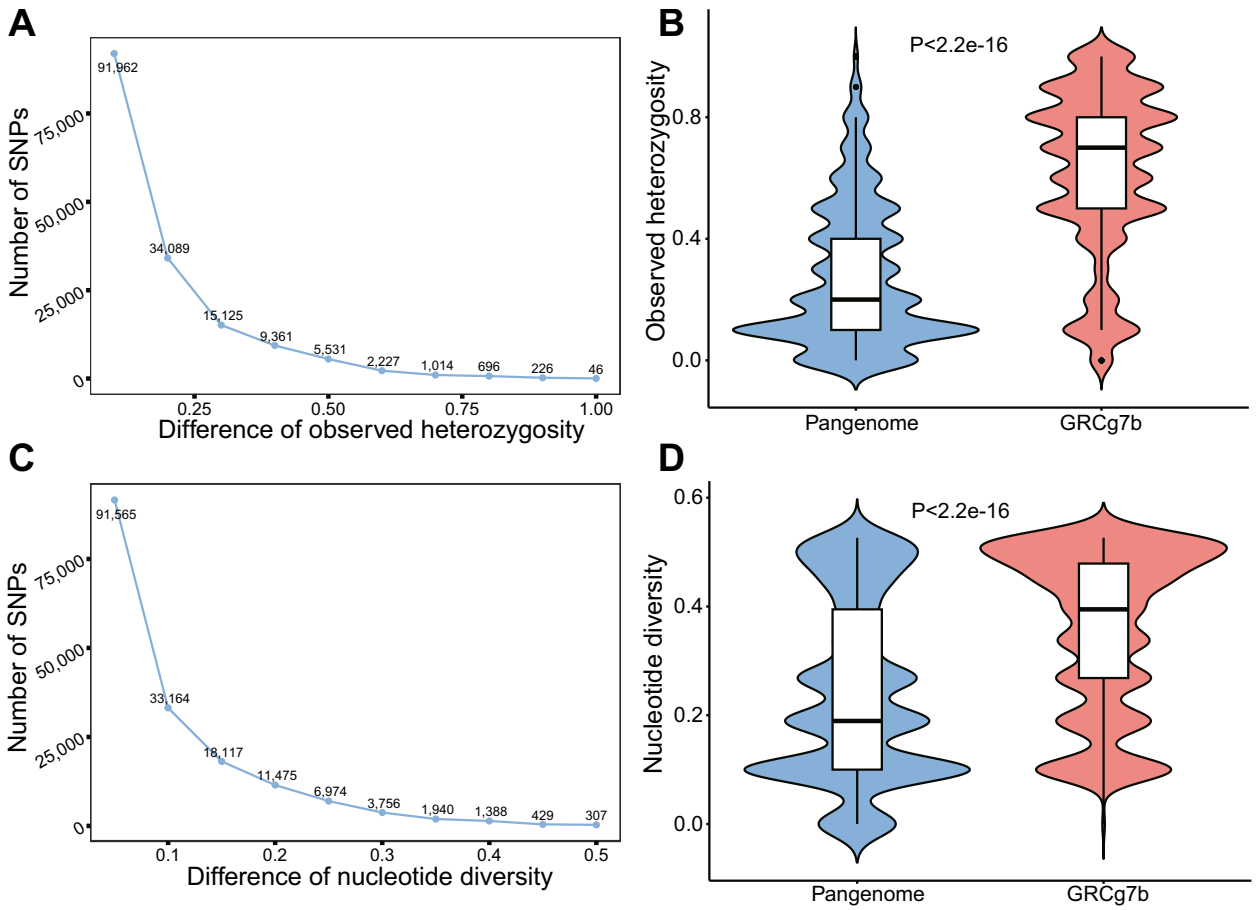

Fig. 3 The comparison of real SNPs called from chicken pangenome and GRCg7b. **A** The number of SNPs with different observed heterozygosity between the pangenome and GRCg7b under different thresholds. **B** Distribution of observed heterozygosity for the 10,000 SNPs with the largest differences in observed heterozygosity between the pangenome and GRCg7b. **C** The number of SNPs with different nucleotide diversity between the pangenome and GRCg7b under different thresholds. **D** Distribution of nucleotide diversity for the 10,000 SNPs with the largest differences in observed heterozygosity between the pangenome and GRCg7b

Discussion

Advancements in genome sequencing and assembly technologies have shown that a single reference genome cannot encompass the full genetic diversity of a species. Mapping short reads to a single reference genome introduces mapping bias, also known as reference bias. Although this bias is acknowledged, its impact in pigs remains unclear. In this study, we quantified the mapping bias of the Sscrofa11.1 reference genome on Chinese indigenous pigs using simulated data. We found that the permissive default parameters of BWA-MEM lead to a high rate of false positives, where sequences mapping to NRSs are incorrectly mapped to their similar regions in the reference genome. Adjusting BWA-MEM parameters can reduce these false-positive mappings [31]. Our simulation analysis indicated that if unliftable regions were considered, the reference bias would be even greater than current estimation. Previous research on the 1000 Genomes Project showed that mapping bias led to overestimation of allele frequencies in *HLA* genes [32]. Similarly, our study found that mapping bias in Sscrofa11.1 leads to overestimated SNP heterozygosity in Chinese indigenous pigs. However, the pangenome graph can help mitigate the overestimation of allele frequencies.

Pig genome researches have predominantly relied on Sscrofa11.1, despite the availability of some high-quality Chinese pig assemblies [22, 33–35]. Using a local assembly often suffers from inconsistent genomic coordinates compared to other studies based on Sscrofa11.1, as well as from insufficient annotations. Therefore, adopting the new genomic coordinates solely for marginal SNP accuracy improvements is generally unacceptable. However, lifting over variants to another genomic coordinates not only results in a loss of variant numbers but also leads to incorrect genotyping due to potential inconsistencies in the reference alleles [36].

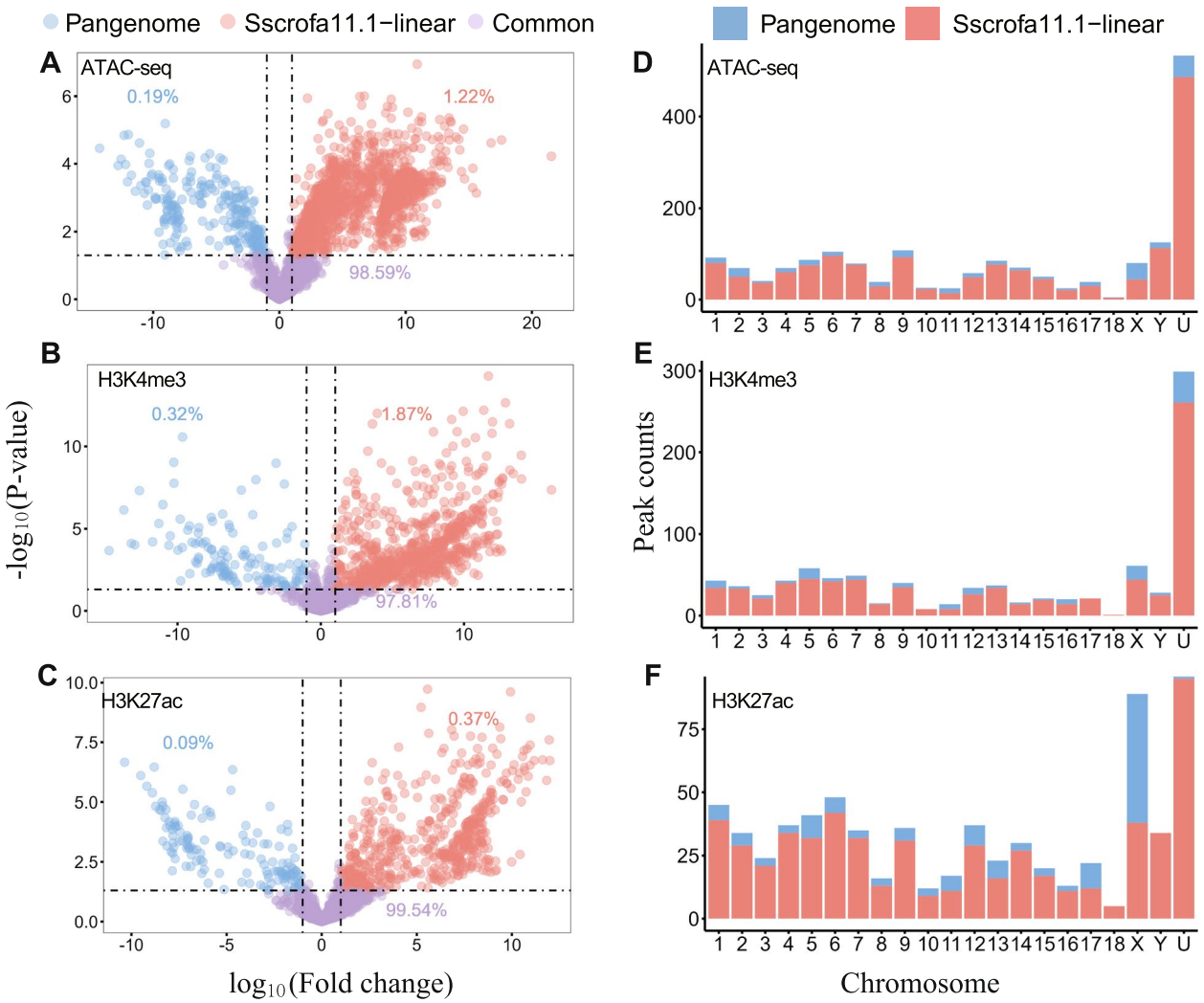

Fig. 4 The comparison of peaks called from Sscrofa11.1 and pangenome. **A**–**C** Volcano plots showing the significantly different peaks identified by pangenome and Sscrofa11.1. **D**–**F** The number significantly different peaks in each chromosome. The character "U" represents unplaced contigs

The pangenome graph effectively reduces mapping bias while preserving the coordinate system of the reference genome. Our findings demonstrate that the pangenome graph significantly outperforms breed-specific graphs, likely due to its larger number and higher quality of variants. Previous studies have shown that breed-specific graphs constructed using variants derived from WGS can outperform pangenome graphs when using the same variants [25, 35]. Moreover, as long-read sequencing costs continue to decrease, obtaining multiple high-quality assemblies for a single breed will become increasingly feasible. Thus, we infer that customized graphs created from multiple assemblies of a specific breed would perform better than a general pangenome graph for mapping reads from that breed.

The shift from single reference genomes to more customized references, like breed-specific or personalized references, may become a future trend. For humans, the concept of a "Personalized pangenome Reference" involves extracting subgraphs from a high-quality pangenome using *k*-mer similarity to create personalized graph for each WGS sample [8]. However, the high-quality pangenome for farming animals are not yet widely recognized, highlighting the need for continued collaboration among pangenome consortia, such as the Bovine Pangenome Consortium [37].

To date, both customized genomes and pangenome have their drawbacks. Customized genomes are costly to generate for each breed, and accurately identifying variants without reference mapping is challenging. Conversely, pangenome include multiple paths, which results in longer alignment times compared to customized genomes. For example, when aligning MS pig sequencing data, reads with significant differences from the

pangenome took an unusually long time to process. This issue likely stems from overly complex graph structures in certain regions. More efforts are needed to optimize the pig pangenome by pruning some of the complex bubbles in the graph.

Conclusions

In this study, we quantified the mapping bias of the pig reference genome on a well-known Chinese indigenous pig using simulated data. Our findings show that a pangenome reduces this bias, enhancing variant detection accuracy. Variants called from the linear genome show higher heterozygosity, which can distort genome-wide analyses and genetic parameter estimates, such as runs of homozygosity.

Methods

Read simulation

To accurately assess the impact of reference mapping bias, we simulated sequencing reads from a Chinese local pig assembly (MS pig) [34] and subsequently mapped these simulated reads to the Sscrofa11.1. To simplify the simulation process, only chromosome 5 from both MS assembly (MS-chr5) and Sscrofa11.1 (Sscrofa11.1-chr5) were selected for simulation analysis. We employed nf-LO (v1.6.0) [38] with GSAlign [39] as aligner to build a chain file from MS-chr5 to Sscrofa11.1-chr5. The "unliftable" regions in MS-chr5 were hard-masked by BEDTools (v2.30.0) [40]. We simulated 7.6 million (~10× coverage) of 150-bp pair-end reads on the masked MS-chr5 using mason2 (v2.0.9) [41]. To establish the ground truth for Sscrofa11.1-chr5 coordinates of the simulated reads, we adopted CrossMap (v0.7.0) [42] to liftover the coordinates from MS-chr5 to Sscrofa11.1-chr5, using the chain file we previously constructed.

To assess the impact of reference bias on genotyping accuracy, we also simulated 7.6 million reads based on a set of known variants and Sscrofa11.1-chr5 using mason2. The set of known variants were derived by mapping the masked MS-chr5 to Sscrofa11.1-chr5 using minimap2 (v2.22-r1101) [43], followed by variants calling with Paftools.js.

Customized graph genomes for MS pigs

We collected WGS data from 28 MS pigs, each with a sequence coverage greater than 10× (Additional file 1: Table S12). To identify SNPs and INDELs, we employed the GTX.CAT germline variants calling pipeline (http://www.gtxlab.com/en/product/cat), using the Sscrofa11.1 reference genome. The identified variants were filtered using GATK (v4.0.5.1) [44], with SNPs filtered based on the criteria "QD < 2.0 || MQ < 40.0 || FS > 60.0 || SOR > 3.0," and INDELs filtered using "QD < 2.0 || FS > 200.0 || SOR > 10.0." Variants with genotyping rates < 0.8 and alternate allele number > 1 were filtered by PLINK (v1.90) [45]. The remaining variants were then phased using Beagle (v5.4) [46]. We additionally downloaded long-read sequencing data (~50× coverage) produced by Oxford Nanopore Technologies (ONT) from a MS pig [23]. The ONT long reads were aligned to the Sscrofa11.1 using NGMLR (v0.2.7) [47], and SVs were called using cuteSV (v1.0.8) [48] to generate a SV set for MS pigs. Insertions and deletions between 50 bp and 100 kb were retained in the SV set. The variants identified from both whole-genome sequencing and ONT data were integrated and categorized into four sets: SNPs, short variants (SNPs and INDELs), SVs, and all variants (including SNPs, INDELs, and SVs). To construct MS pig-specific graph genomes, we augmented each variant set onto the Sscrofa11.1 reference genome using VG (v1.5.6) [20], resulting in four customized graph genomes: SNP-graph, SHORT-graph, SV-graph, and ALL-graph.

Building a pig pangenome

We collected 10 high-quality pig assemblies, including seven Chinese indigenous pigs, one Korea pig and, two European pigs from NCBI and CNGD (Additional file 1: Table S13). We employed the Mash (v2.3) [49] to calculate genetic distances between the assemblies and Sscrofa11.1. A phylogenetic tree was built using the genetic distance obtained from Mash as input and visualized by ggtree (v2.4.2) [50]. To generate the pig pangenome graph, the 10 assemblies were additionally integrated to the Sscrofa11.1 reference using the Minigraph-Cactus pipeline (v2.4.2) [27]. Briefly, the pipeline first constructed a graph only containing SVs by progressively mapping other assemblies to the reference genome using Minigraph [28]. These assemblies were then remapped to the structural variation graph, and the mapping results were used as input for Cactus [51] to construct a comprehensive graph that includes all types of variants. A pangenome graph containing only chr5 were also built for simulation analysis. Since the simulated reads were derived from the MS assembly, the MS assembly was excluded from the construction of this single-chromosome pangenome.

Evaluation of mapping performance

We mapped the simulated reads to pangenome and MS customized graph genomes using VG Giraffe (v1.5.6) [52]. The mappings from graph genome (GAM files) were then surjected to linear coordinate space of Sscrofa11.1 to generate BAM files. For comparison, we also mapped the simulated reads to the original linear reference genome (Sscrofa11.1-linear) using BWA-MEM (v0.7.17-r1188) [53] and to a flat graph genome (Sscrofa11.1-graph, with

no additional variants) using VG Giraffe. This allowed us to compare the mapping performance across different genome representations under the unified coordinate system of Sscrofa11.1. Mapping accuracy was assessed by comparing the mapped start position of each read to its simulated position. A read was considered correctly mapped if the difference between the mapped start position and the simulated position was within 10 bp. Additionally, reads without simulated positions (those that unable to be lifted over to Sscrofa11.1) were considered correctly mapped if they failed to map to any position. Incorrectly mapped reads were categorized into three scenarios: (1) erroneous mappings, where reads mapped to positions inconsistent with their simulated positions; (2) false-positive mappings, where reads without simulated positions mapped to specific locations; and (3) false-negative mappings, where reads with simulated positions failed to map to any position.

Evaluation of genotyping performance

We employed GATK4 HaplotypeCaller to call variants from BAM files produced by VG Giraffe and BWA-MEM. We used the submodule vcfeval in RTG Tools (v3.9.1) [54] to evaluate the genotyping accuracy by comparing the called variants and the known variants provided for simulation.

Real WGS data for pigs

A total of 19 MS pigs with sequence coverage > 6 × (Additional file 1: Table S14) [55] were used to compare the difference of BWA-GATK (Sscrofa11.1) and Giraffe-GATK (pangenome) pipeline. The raw sequencing data were obtained from our PHARP database [56]. The adapters and low-quality reads were removed by fastp (v0.23.0) [57], and the filtered reads were mapped to Sscrofa11.1-linear and pangenome using BWA-MEM and VG Giraffe, respectively. We employed the Sambamba (v1.0.1) [58] to mark the duplicates, and then genotyped the variants in Genomic Variant Call Format (GVCF) mode using GATK haplotype Caller. Multi-sample VCF files were generated by merging individual GVCF files for joint variant calling. We used Beagle to phase the multi-sample VCF file and then extracted SNPs that fulfilled the criteria "QD < 2.0 || MQ < 40.0 || FS > 60.0 || SOR > 3.0." PLINK (v1.90) was employed to perform the analysis of ROH with parameters "–homozyg-snp 100 –homozyg-kb 500 –homozyg-density 50 –homozyg-gap 1000 –homozyg-het 1." The shared SNPs from the two pipelines located on autosomes were extracted to compare genotyping consistency. The VCFtools (v0.1.17) [59] and PLINK (v1.90) were used to calculate the observed heterozygosity and nucleotide diversity for each common SNP, respectively. To identify genomic regions with high genotype differences, we used a 50-kb sliding window with 5-kb step to measure the genotype differences. To explore the impact of the genomic regions with high genotype differences (> 30%), we overlapped these genomic regions with known pig QTL and genes using GALLO (v1.3) [60]. The known pig QTLs and gene annotation were collected from animal QTL database [61] and Ensembl (http://ftp.ensembl.org/pub/release111/gtf/sus_scrofa/Sus_scrofa.Sscrofa11.1.111.gtf.gz).

Real WGS data for chicken

To compare the genotype difference of linear reference genome and pangenome graph in chicken, we downloaded WGS data of 10 Lueyang black-feathered blackbone chickens from [62] and the chicken pangenome from [21]. The downloaded WGS data were processed using the same pipeline as that applied to the WGS data of pigs. The chicken reference genome (bGalGal1.mat.broiler.GRCg7b) was used as the linear reference genome. We calculated the observed heterozygosity and nucleotide diversity for the shared SNPs identified by linear reference genome and pangenome graph.

Epigenomic data analysis

We downloaded ATAC-seq data from three tissues and ChIP-seq data from five tissues (sequenced with H3K4me3 and H3K27ac antibodies) of an MS pig from [63] (Additional file 1: Table S15). The quality control and mapping of the epigenomic sequencing data were performed the same as WGS data. Peaks for each sample were called using MACS2 (v2.2.9.1) [64]. To compare peak differences, we employed DiffBind (v3.12.0) [65] to normalize the peaks and then identified statistically significant differences between Ssrofa11.1 and pangenome using edgeR (v4.0.16). Peaks with P value < 0.05 and log10(fold-change) > 1 were considered as significantly different (either Sscrofa11.1-specific or pangenome-specific), while others were treated as common.

Abbreviations

MS	Meishan
MS-chr5	Chromosome 5 of Meishan assembly
NRS	Non-reference sequences
ONT	Oxford Nanopore Technologies
QTL	Quantitative trait locus
ROH	Runs of homozygosity
Sscrofa11.1-chr5	Chromosome 5 of Sscrofa11.1 assembly
WGS	Whole-genome sequences

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-025-02194-y.

Additional file 1. Tables S1–S15. Table S1 The number of mapped reads in the graph space (before surjection) and in the linear space (after surjection). Table S2 The mapping accuracy of linear reference and graph

genomes. Table S3 The mapping accuracy of linear reference and pangenome on chr1 and chr18. Table S4 The sources of mapping errors. Table S5 The mapping accuracy of BWA-MEM with different mapping parameters. Table S6 The position of 570 high-inconsistent widows (genotype difference > 0.3). Table S7 Overlapped coding proteins with the high-inconsistent widows. Table S8 Overlapped QTLs with the high-inconsistent widows. Table S9 Significantly different ATAC-seq peaks identified by pangenome and Sscrofa11.1-linear. Table S10 Significantly different H3K4me3 peaks identified by pangenome and Sscrofa11.1-linear. Table S11 Significantly different H3K27ac peaks identified by pangenome and Sscrofa11.1-linear. Table S12 Meishan pigs used for construction of customized MS graphs. Table S13 Assemblies used in pangenome construction. Table S14 Meishan pigs used for assessment of variants calling. Table S15 Meishan pigs for epigenomic sequencing.

Additional file 2. Figures S1–S4. Fig. S1 Phylogenetic tree of assemblies used in pangenome construction. Fig. S2 The length distribution of variants. Fig. S3 Venn plot of common SNPs identified by pangenome and Sscrofa11.1-linear. Fig. S4 The proportion of different types of QTLs overlapped with high-inconsistent widows.

Acknowledgements
We thank all the researchers worldwide who made their assembly and sequencing data publicly available.

Authors' contributions
Z.W. and Y.P. conceived and supervised the study. J.M. carried out most of the analyses and wrote the manuscript. Q.W. performed the variants calling of WGS data. Z.Z. and Q.S.W. participated in the discussion of the results. All authors read and approved the final manuscript. All authors reviewed the manuscript.

Funding
This work was supported by the National Key Research and Development Program of China (2021YFD1200802, 2022YFF1000500, and 2023YFF1001100), Natural Science Foundation of China (Grant No. 32172691 and 32372831), Sanya Science and Technology Innovation Project (2022KJCX49), and the Hainan Province Science and Technology Special Fund (ZDYF2024XDNY186).

Data availability
The data that support the results of this research are available within the article and its Supplementary Information files. The newly generated pig pangenome graph is available at Figshare (https://figshare.com/s/afc83176af09b8e2e4c8). Sequencing datasets analyzed and their accessions are available in Additional File 1: Tables S12-S15. Code associated to this work can be found at https://github.com/JanMiao/Pig-reference-bias.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 15 November 2024 Accepted: 18 March 2025
Published online: 26 March 2025

References

1. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376:44–53.
2. Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, et al. The Human Pangenome Project: a global resource to map genomic diversity. Nature. 2022;604:437–46.
3. Chen N-C, Solomon B, Mun T, Iyer S, Langmead B. Reference flow: reducing reference bias using multiple population genomes. Genome Biol. 2021;22:8.
4. Lin M-J, Iyer S, Chen N-C, Langmead B. Measuring, visualizing, and diagnosing reference bias with biastools. Genome Biol. 2024;25:101.
5. Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. Genome Res. 2017;27:665–76.
6. Thorburn D-MJ, Sagonas K, Binzer-Panchal M, Chain FJJ, Feulner PGD, Bornberg-Bauer E, et al. Origin matters: using a local reference genome improves measures in population genomics. Mol Ecol Resour. 2023;23:1706–23.
7. Vaddadi K, Mun T, Langmead B. Minimizing reference bias with an impute-first approach. bioRxiv. 2024;2023.11.30.568362.
8. Sirén J, Eskandar P, Ungaro MT, Hickey G, Eizenga JM, Novak AM, et al. Personalized pangenome references. Nat Methods. 2024;21:2017–23.
9. Coombes B, Lux T, Akhunov E, Hall A. Introgressions lead to reference bias in wheat RNA-seq analysis. BMC Biol. 2024;22:56.
10. Groza C, Kwan T, Soranzo N, Pastinen T, Bourque G. Personalized and graph genomes reveal missing signal in epigenomic data. Genome Biol. 2020;21:124.
11. Fang X, Mu Y, Huang Z, Li Y, Han L, Zhang Y, et al. The sequence and analysis of a Chinese pig genome. Gigascience. 2012;1:16.
12. Warr A, Affara N, Aken B, Beiki H, Bickhart DM, Billis K, et al. An improved pig reference genome sequence to enable pig genetics and genomics research. Gigascience. 2020;9:giaa051.
13. Ai H, Fang X, Yang B, Huang Z, Chen H, Mao L, et al. Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. Nat Genet. 2015;47:217–25.
14. Teng J, Gao Y, Yin H, Bai Z, Liu S, Zeng H, et al. A compendium of genetic regulatory effects across pig tissues. Nat Genet. 2024;56:112–23.
15. Zhong Z, Wang Z, Xie X, Tian S, Wang F, Wang Q, et al. Evaluation of the genetic diversity, population structure and selection signatures of three native Chinese pig populations. Animals. 2023;13:2010.
16. Jang S, Ros-Freixedes R, Hickey JM, Chen C-Y, Holl J, Herring WO, et al. Using pre-selected variants from large-scale whole-genome sequence data for single-step genomic predictions in pigs. Genet Sel Evol. 2023;55:55.
17. Miao J, Chen Z, Zhang Z, Wang Z, Wang Q, Zhang Z, et al. A web tool for the global identification of pig breeds. Genet Sel Evol. 2023;55:18.
18. Zhang L, Zhang S, Zhan F, Song M, Shang P, Zhu F, et al. Population genetic analysis of six Chinese indigenous pig meta-populations based on geographically isolated regions. Animals. 2023;13:1396.
19. Gong Y, Li Y, Liu X, Ma Y, Jiang L. A review of the pangenome: how it affects our understanding of genomic variation, selection and breeding in domestic animals? J Anim Sci Biotechnol. 2023;14:73.
20. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018;36:875–9.
21. Rice ES, Alberdi A, Alfieri J, Athrey G, Balacco JR, Bardou P, et al. A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex structural variants. BMC Biol. 2023;21:267.
22. Miao J, Wei X, Cao C, Sun J, Xu Y, Zhang Z, et al. Pig pangenome graph reveals functional features of non-reference sequences. J Animal Sci Biotechnol. 2024;15:32.
23. Jiang Y-F, Wang S, Wang C-L, Xu R-H, Wang W-W, Jiang Y, et al. Pangenome obtained by long-read sequencing of 11 genomes reveal hidden functional structural variants in pigs. iScience. 2023;26:106119.
24. Leonard AS, Mapel XM, Pausch H. Pangenome genotyped structural variation improves molecular phenotype mapping in cattle. Genome Res. 2024;gr.278267.123.
25. Milia S, Leonard A, Mapel XM, Bernal Ulloa SM, Drögemüller C, Pausch H. Taurine pangenome uncovers a segmental duplication upstream of KIT associated with depigmentation in white-headed cattle. Genome Res. 2024;gr.279064.124.
26. Crysnanto D, Pausch H. Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery. Genome Biol. 2020;21:184.

27. Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. Nat Biotechnol. 2024;42:663–73.
28. Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with minigraph. Genome Biol. 2020;21:265.
29. Garrison E, Guarracino A. Unbiased pangenome graphs. Bioinformatics. 2023;39:btac743.
30. Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, et al. Building pangenome graphs. Nat Methods. 2024;21:2008–12.
31. Yao Z, You FM, N'Diaye A, Knox RE, McCartney C, Hiebert CW, et al. Evaluation of variant calling tools for large plant genome re-sequencing. BMC Bioinformatics. 2020;21:360.
32. Brandt DYC, Aguiar VRC, Bitarello BD, Nunes K, Goudet J, Meyer D. Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes Project phase I data. G3 (Bethesda). 2015;5:931–41.
33. Ma H, Jiang J, He J, Liu H, Han L, Gong Y, et al. Long-read assembly of the Chinese indigenous Ningxiang pig genome and identification of genetic variations in fat metabolism among different breeds. Mol Ecol Resour. 2022;22:1508–20.
34. Zhou R, Li S-T, Yao W-Y, Xie C-D, Chen Z, Zeng Z-J, et al. The Meishan pig genome reveals structural variation-mediated gene expression and phenotypic divergence underlying Asian pig domestication. Mol Ecol Resour. 2021;21:2077–92.
35. Wang Y, Gou Y, Yuan R, Zou Q, Zhang X, Zheng T, et al. A chromosome-level genome of Chenghua pig provides new insights into the domestication and local adaptation of pigs. Int J Biol Macromol. 2024;270:131796.
36. Chen N-C, Paulin LF, Sedlazeck FJ, Koren S, Phillippy AM, Langmead B. Improved sequence mapping using a complete reference genome and lift-over. Nat Methods. 2024;21:41–9.
37. Smith TPL, Bickhart DM, Boichard D, Chamberlain AJ, Djikeng A, Jiang Y, et al. The Bovine Pangenome Consortium: democratizing production and accessibility of genome assemblies for global cattle breeds and other bovine species. Genome Biol. 2023;24:139.
38. Talenti A, Prendergast J. nf-LO: a scalable, containerized workflow for genome-to-genome lift over. Genome Biol Evol. 2021;13:evab183.
39. Lin H-N, Hsu W-L. GSAlign: an efficient sequence alignment tool for intra-species genomes. BMC Genomics. 2020;21:182.
40. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
41. Holtgrewe M. Mason–a read simulator for second generation sequencing data. Technical Report FU Berlin. 2010.
42. Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics. 2014;30:1006–7.
43. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
44. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
45. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
46. Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. Am J Hum Genet. 2021;108:1880–90.
47. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15:461–8.
48. Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, et al. Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. 2020;21:189.
49. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17:132.
50. Xu S, Li L, Luo X, Chen M, Tang W, Zhan L, et al. Ggtree: a serialized data object for visualization of a phylogenetic tree and annotation data. Imeta. 2022;1: e56.
51. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 2020;587:246–51.
52. Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science. 2021;374:abg8871.
53. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
54. Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, et al. Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. BioRxiv. 2015;023754.
55. Chen M, Su G, Fu J, Wang A, Liu J-F, Lund MS, et al. Introgression of Chinese haplotypes contributed to the improvement of Danish Duroc pigs. Evol Appl. 2019;12:292–300.
56. Wang Z, Zhang Z, Chen Z, Sun J, Cao C, Wu F, et al. PHARP: a pig haplotype reference panel for genotype imputation. Sci Rep. 2022;12:12645.
57. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:884–90.
58. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31:2032–4.
59. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
60. Fonseca PAS, Suárez-Vega A, Marras G, Cánovas Á. GALLO: an R package for genomic annotation and integration of multiple data sources in livestock for positional candidate loci. Gigascience. 2020;9:giaa149.
61. Hu Z-L, Park CA, Reecy JM. Bringing the Animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. Nucleic Acids Res. 2022;50:D956–61.
62. Xue Z, Wang L, Tian Y, Yang Y, Li P, Yang G, et al. A genome-wide scan to identify signatures of selection in Lueyang black-bone chicken. Poult Sci. 2023;102: 102721.
63. Zhao Y, Hou Y, Xu Y, Luan Y, Zhou H, Qi X, et al. A compendium and comparative epigenomics analysis of cis-regulatory elements in the pig genome. Nat Commun. 2021;12:2217.
64. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.
65. Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature. 2012;481:389–93.

Publisher's Note