BMC Genomics

RESEARCH

Open Access

Integrative genomic analysis reveals shared loci for reproduction and production traits in Yorkshire pigs

Ran Wei[1], Zhenyang Zhang[1], He Han[1], Jian Miao[1], Pengfei Yu[1], Hong Cheng[1], Wei Zhao[1,3], Xiaoliang Hou[3], Jianlan Wang[3], Yongqi He[3], Yan Fu[3], Zhen Wang[1], Qishan Wang[1,2], Zhe Zhang[1*] and Yuchun Pan[1,2*]

Abstract

Background Improving reproductive performance in Yorkshire pigs, a key maternal line in three-way crossbreeding systems, remains challenging due to low heritability and historical selection pressures favoring production traits. Identifying pleiotropic genetic variants that influence both reproduction and production traits is crucial for understanding their genetic interplay and enhancing molecular breeding strategies.

Results Genome-wide association studies (GWAS) using 2,764 individuals identified 264,660 significant loci associated with reproduction traits and 12,460 loci for production traits, with 73 independent signals, including genes such as *SCLT1* and *CAPN9*. A total of 465,047 independent loci were identified, resulting in a genome-wide significance threshold of 2.15×10^{-6}. Genetic correlations analysis between reproduction and production traits across parities revealed varying trends, including a strengthening negative correlation between mean litter weight (MLW) and backfat thickness (BFT) with increasing parity (P1:r_g=-0.0376; P2:r_g=-0.1371; P3:r_g=-0.1475). Given 1062 shared significant loci between MLW and BFT, local genetic correlation was calculated within the corresponding genomic regions, resulting in a weak correlation of 0.014. Transcriptome-wide association studies (TWAS) leveraging data from the PigGTEx project, which includes 9,530 RNA-sequencing samples across 34 tissues, revealed 2,143 significant genes, with 31 linked to total number of piglets born (TNB) and 133 to number of piglets born alive (NBA). These results highlight the importance of these genes in reproductive performance, with *SCLT1* being notably significant in reproductive tissues. For MLW, integrating results from multiple analyses revealed *CENPE* as a strong candidate gene, exhibiting significant association and colocalization. Validation in an independent population (n = 300) showed that incorporating the top 0.2% of significant single nucleotide polymorphisms (SNPs) in the GFBLUP model improved predictive accuracy, increasing from 0.0168 to 0.0242 for MLW.

Conclusion This study provides new insights into the pleiotropic genetic architecture underlying reproduction and production traits in Yorkshire pigs. Genetic correlations, shared loci, and candidate genes inform breeding program design. The increased accuracy of genomic selection using these significant loci highlights their practical utility in improving breeding efficiency. These findings suggest opportunities for refining selection strategies, although further research is warranted to fully realize their potential for enhancing breeding programs.

*Correspondence:
Zhe Zhang
zhe_zhang@zju.edu.cn
Yuchun Pan
panyuchun1963@aliyun.com
Full list of author information is available at the end of the article

© The Author(s) 2025. **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/.

Highlights

- Seventy-three independent signals identified for reproduction and production traits in Yorkshire pigs.
- One thousand sixty-two shared loci indicate genetic correlation between reproduction and production traits.
- Explored the relationship between reproduction and production traits across three parities.
- Multi-omics integration (GWAS, TWAS, COLOC, and SMR) revealed functional roles of key genes, such as SCLT1, CENPE, COL9A1.
- Incorporating significant loci as features in GFBLUP significantly enhanced the accuracy of genomic selection.

Keywords Genetic pleiotropy, Reproduction traits, Multi-omics integration, Cross-parity genetic correlation, Genomic selection enhancement, Yorkshire pigs, Shared genetic loci

Introduction

Sow reproductive performance is crucial for achieving higher economic benefits in modern commercial breeding programs. Traits like total number born (TNB) and number born alive (NBA) are prioritized in breeding selection indices worldwide due to their significant economic impact [1]. However, these reproduction traits have low heritability, resulting in slow genetic improvement using conventional breeding methods [2]. Genomic selection has emerged as a powerful tool to accelerate genetic progress. It offers higher accuracy and enables early selection, shortening the generational interval [3]. Despite these advances, accurately estimating breeding values for reproduction traits remains challenging. Many studies show that a comprehensive understanding of the underlying genetic mechanisms is crucial for improving the accuracy of genomic selection [4–6].

Advancements in sequencing technologies have propelled genome-wide association studies (GWAS) to the forefront of identifying candidate genes associated with economically important traits in livestock [7–10]. GWAS has significantly contributed to understanding the genetic basis of these traits. For example, the Animal QTL database lists over 55,166 QTLs reported for pigs [11]. SNP chips are the most common genotyping method in livestock genomic breeding. However, their limited number of markers hinders comprehensive identification of associated loci. Previous studies [12–14] on reproduction and production traits in Yorkshire pigs used limited SNPs, potentially overlooking important genomic regions. Whole-genome sequencing (WGS) offers a more comprehensive approach [15], but its high cost limits widespread use. The increasing availability of haplotype reference panels, such as PHARP [16], SWIM [17], AGIDB [18], and PGRP [19], facilitates imputation from sparse SNP datasets to high-density, genome-wide SNPs. This improves the identification of trait-associated loci.

However, GWAS often identify significant signals in intergenic or non-coding regions. This complicates the functional interpretation of these genetic variants [20]. Resources like the GTEx project provide expression quantitative trait loci (eQTLs) across various crucial tissues [19], facilitating interpretation of these signals. This information benefits GWAS, genomic selection programs, and genome editing strategies, advancing both practical and theoretical livestock genetics.

Production traits, with their higher heritability and direct economic impact, have been under strong selection pressure in recent decades. This often came at the expense of reproduction traits. Consequently, neglecting the complex interplay between production and reproduction has reduced reproductive performance in many livestock species. Studies on dairy cattle show that selection for increased milk yield negatively impacts reproduction [21, 22]. In pig breeding, Yorkshire pigs are crucial as a maternal line. While the relationship between reproduction and production traits has been investigated across different pig breeds, including Yorkshire pigs, most studies have focused on genetic correlation estimates and heritability analyses [23, 24]. There remains a gap in identifying specific genes and their functional roles in shaping this relationship.

The complex relationship between reproduction and production traits in Yorkshire pigs significantly influences breeding efforts aimed at improving productivity and economic viability. Pleiotropic genetic effects, where a single genetic variant influences multiple traits, play a crucial role in shaping these relationships. This study investigates the genetic correlations between these traits using advanced genomic tools and large-scale data analysis to address current knowledge gaps. Specifically, we aim to identify pleiotropic genetic variants that contribute to both reproductive success and productive efficiency, shedding light on their shared genetic architecture. By integrating genome-wide association study findings with functional genomic insights, we aim to inform more effective selection strategies that improve reproductive and productive performance.

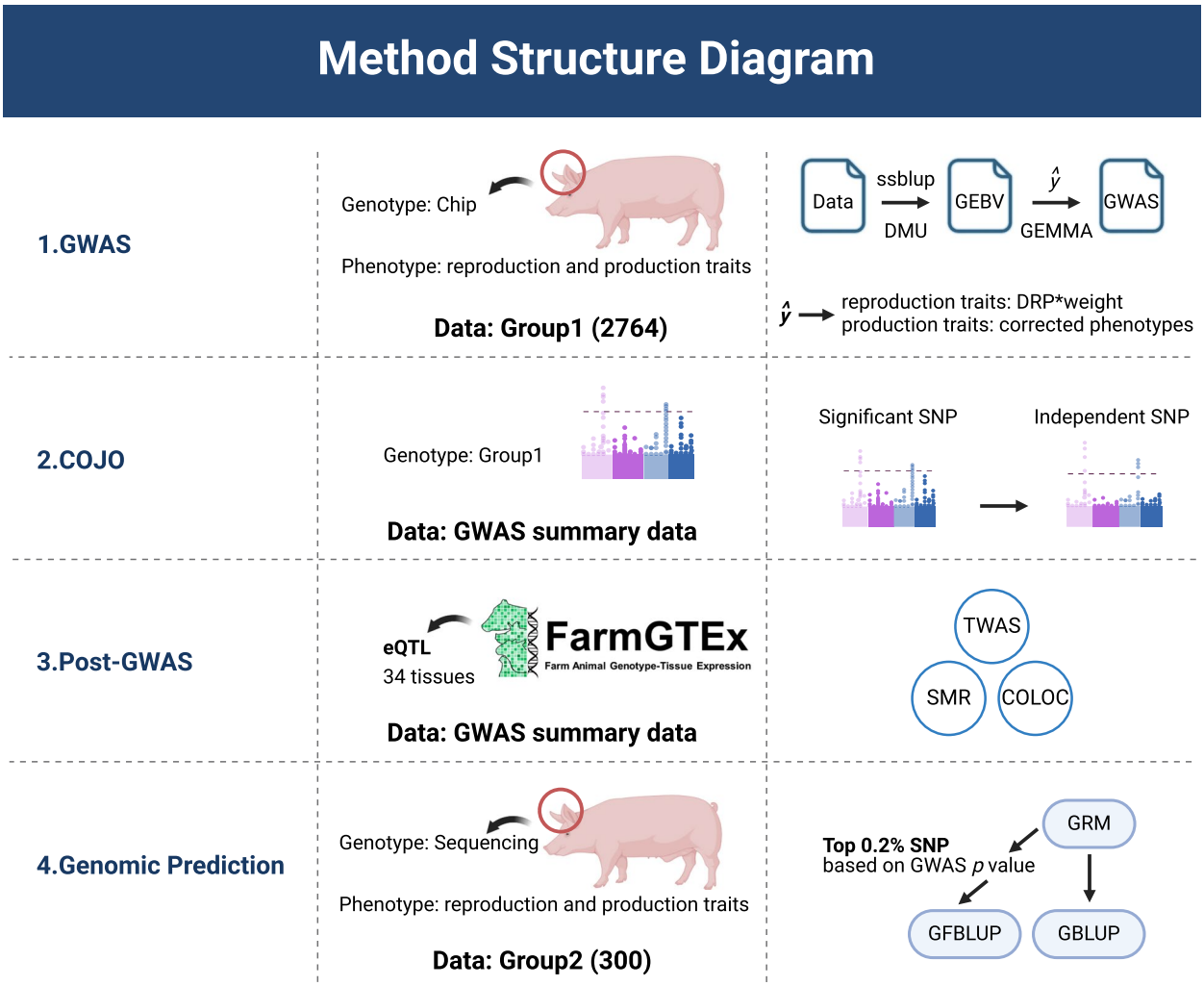

Fig. 1 Structural diagram of the methodological framework. Created in BioRender. ›O, q6. (2025) https://BioRender.com/a41l822

Materials and methods

For clarity, our materials and methods are summarized (Fig. 1).

Animals and data

The dataset used in this study consisted of phenotype records of four reproduction and six production traits measured in a population of 3,064 Yorkshire pigs from a nucleus pig farm in Chizhou, Anhui Province, China.

Genotypic data

Total DNA of all Yorkshire pigs ($n = 3{,}064$) was extracted from ear tissue samples, which were collected using the following procedure: First, the ear was cleaned sequentially with warm water, physiological saline, and 70% alcohol to ensure it was free from contaminants. A small piece of ear tissue was then carefully excised using sterile scissors. The tissue sample was immediately placed into a centrifuge tube containing 75% ethanol for preservation. After collection, the samples were stored at −20°C for long-term preservation until further processing.

Of the 3,064 pigs, 2,764 individuals (referred to as **Group1**) were genotyped using the GGP 50K Porcine v1 Genotyping BeadChip (Neogen), acquiring 50,697 SNP markers distributed across the genome. We firstly removed the SNPs with missing rate greater greater than 0.05 or minor allele frequencies (MAF) lower than 0.05, and only autosomal SNPs were considered in this study. Then, conform-gt program (http://faculty.washington.edu/browning/conform-gt.html) was used to address potential strand issues for the remaining SNPs based on the reference SNPs from haplotype reference panel of

PGRP [16]. At last, the remaining SNPs were imputed to the whole genome level using Beagle (v5.1) with the help of PGRP [25]. After imputation, SNPs with Dosage R-Squared (DR2) lower than 0.8 were discarded. Using PLINK v1.90, SNPs with MAF less than 0.05 or not in Hardy–Weinberg equilibrium (HWE, $p < 1 \times 10^{-6}$) were further filtered out [26]. Ultimately, 11,794,966 autosomal SNPs were retained for analysis.

The other 300 Yorkshire individuals (referred as **Group2**) were sequenced using DNBSEQ-T7 platform at low coverage (~1X). Subsequently, The raw sequencing reads were processed with fastp v0.20.0 to remove the low-quality reads using default filtering criteria [27]. Clean reads were then aligned to the Sus scrofa 11.1 reference genome using GTX v2.1.5 [28, 29]. The resulting BAM files were utilized for imputation to the PGRP level using GLIMPSE2 v2.0.0 [30]. After imputation, SNPs with an information quality score (INFO_SCORE) greater than 0.7 were retained. PLINK v1.90 was used to remove SNPs with minor allele frequencies (MAF) less than 0.05 or those extremely deviated from Hardy–Weinberg equilibrium (HWE, $p < 1 \times 10^{-10}$). After these quality control steps, 8,554,664 autosomal SNPs were retained for Group2.

Reproductive data

The reproduction traits included total number of piglets born (TNB), number of piglets born alive (NBA), coefficient of variation of piglets' weight at birth (WeightCV), and mean litter weight (MLW). TNB is the total number of piglets born per litter, while NBA is the total number of piglets born within 24 h, excluding stillborn piglets. Farrowing records were excluded from analysis if either TNB or NBA fell below 6 or exceeded 30. No data filtering step was applied to WeightCV and MLW (Table 1).

Productive dataset

The production traits included two growth traits, off-test body weight (BW) and average daily gain (ADG), two body composition traits, backfat thickness (BFT) and loin muscle depth (LMD), and two feed traits, average daily feed intake (ADFI) and feed conversion ratio (FCR). BW was measured at off-test age (168.6 ± 7.5 days) and ADG was the average weight gain across the period from on-test age (121.6 ± 5.3 days) to the off-test age. BFT and LMD were obtained at off-test stage using the BioSoft Toolbox (v2.6.0.1) from ultrasound images, which were captured between the 3rd and 4th last rib of pigs.

Table 1 The descriptive statistics of 4 reproduction and 6 production traits

Trait	Group	N_{record}[a]	$N_{individual}$[b]	Mean[c]	SD[c]	SE[c]	Min[d]	Max[d]
TNB	Group1	13068	744	14.53	3.21	0.118	6.00	29.00
	Group2	862	300	14.27	3.17	0.183	6.00	22.00
NBA	Group1	12897	746	13.25	2.93	0.107	6.00	26.00
	Group2	853	300	13.41	3.02	0.175	6.00	21.00
WeightCV	Group1	11764	744	20.73	9.65	0.352	1.76	224.63
	Group2	869	300	17.32	5.86	0.339	0.00	52.10
MLW	Group1	11874	742	1.38	0.24	0.009	0.54	4.03
	Group2	876	300	1.29	0.24	0.014	0.00	2.35
BW (kg)	Group1	2614	2614	104.44	12.27	0.240	80.00	146.96
	Group2	300	300	119.29	9.36	0.541	94.50	146.50
BFT (mm)	Group1	2614	2614	10.88	2.22	0.043	4.20	24.90
	Group2	300	300	12.74	2.72	0.157	6.20	23.80
LMD (mm)	Group1	2614	2614	60.61	6.38	0.125	39.10	81.00
	Group2	300	300	62.86	5.09	0.294	42.40	77.50
ADFI (kg/days)	Group1	823	823	0.66	0.10	0.0035	0.34	1.14
	Group2	281	281	0.75	0.11	0.0066	0.46	1.19
ADG (kg/days)	Group1	2016	2016	0.89	0.16	0.0035	0.38	1.84
	Group2	281	281	0.95	0.10	0.0059	0.72	1.24
FCR	Group1	821	821	2.47	0.36	0.0126	1.25	4.09
	Group2	281	281	2.71	0.33	0.0197	1.82	3.88

[a] N_{record} is the number of phenotype records

[b] $N_{individual}$ is the number of individuals with phenotype records

[c] Mean, SD and SE represent the mean, standard deviation and standard error, respectively

[d] Min and Max represent the minimum and maximum values observed

ADFI was measured by the total feed intake divided by the length of the period from the on-test age to the off-test age. FCR was calculated by ADFI divided by ADG (Table 1).

Pedigree data

A complete five-generation pedigree was constructed for Group1, encompassing all individuals with available genotypic or phenotypic data, totaling 251,543 individuals.

Genetic parameters estimation

Genetic parameters for reproduction traits were estimated using data from Group1 based on single step best linear unbiased prediction (ssBLUP) model as follows:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Z}_1\boldsymbol{a} + \boldsymbol{Z}_2\boldsymbol{s} + \boldsymbol{Z}_3\mathbf{pe} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{y}$ is the vector of phenotypic values for all individuals; $\boldsymbol{b}$ is the vector of the effects of fixed effects, including year season of delivery, herd, parity; $\boldsymbol{a}$ is the vector of additive genetic effects, which is assumed to follow normal distribution $N(0, \boldsymbol{H}\sigma_a^2)$, where $\boldsymbol{H}$ is the combined relationship matrix built from the pedigree and genotypic data, and σ_a^2 is the additive genetic variance. The inverse of $\boldsymbol{H}$ is:

$$\boldsymbol{H}^{-1} = \boldsymbol{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{G}^{-1} - \boldsymbol{A}_{22}^{-1} \end{bmatrix} \tag{2}$$

where $\boldsymbol{A}$ is the numerator relationship matrix constructed based on pedigree and $\boldsymbol{A}_{22}$ is the pedigree-based relationship matrix of the genotyped animals. $\boldsymbol{G}$ is the genomic relationship matrix, built using GCTA v1.92.4 software [31] based on pruned SNPs (465,047 loci obtained by the command "–indep-pairwise 50 5 0.3" in PLINK v1.90). $\boldsymbol{G}$ was calculated using the following formula:

$$\boldsymbol{G}_{jk} = \frac{1}{N}\sum_i \boldsymbol{G}_{ijk} = \begin{cases} \frac{1}{N}\sum_i \frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-p_i)}, j \neq k \\ 1 + \frac{1}{N}\sum_i \frac{x_{ij}^2-(1+2p_i)x_{ik}+2p_i^2}{2p_i(1-p_i)}, j = k \end{cases} \tag{3}$$

where $\boldsymbol{G}_{ijk}$ is the genetic relationship between j th and k th individual at locus i, and N is the number of SNPs. $\boldsymbol{x_i}$ is the genotypic vector for locus i, which is composed elements coded as 0, 1 or 2 for A_1A_1, A_1A_2 and A_2A_2. p_i is the allele frequency of A_2.

In model (1), $\boldsymbol{s}$ is the vector of mated sire effects, which is assumed to follow $N(0, \boldsymbol{I}\sigma_s^2)$ with $\boldsymbol{I}$ denoting identity matrix and σ_s^2 is the sire effect variance. **pe** is the vector of permanent environmental effects, which is assumed to follow normal distribution $N(0, \mathbf{I}\sigma_{\text{pe}}^2)$, where σ_{pe}^2 permanent environmental variance. **e** is the vector of random residual effects, and is assumed to follow $N(0, \mathbf{I}\sigma_e^2)$, where σ_e^2 denoting the residual variance. $\boldsymbol{X}$, $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$, and $\boldsymbol{Z}_3$ are the incidence matrices assigning observations to corresponding effects. The variance components were estimated using average information restricted maximum likelihood estimation (AI-REML) implemented in DMU software [32]. Based on the estimates, we can further get the estimated heritability for each reproduction trait using the following formula:

$$h^2 = \frac{\widehat{\sigma}_a^2}{\widehat{\sigma}_a^2 + \widehat{\sigma}_s^2 + \widehat{\sigma}_{\text{pe}}^2 + \widehat{\sigma}_e^2} \tag{4}$$

For production traits, the single-trait model was as follows:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{e} \tag{5}$$

where $\boldsymbol{y}$ is the vector of phenotypic values for production traits, including BW, ADG, BFT, LMD, ADFI and FCR. $\boldsymbol{b}$ is the vector of fixed effects. For BW, the fixed effects include growth days, year season of birth, herd, and birth parity. For BFT, ADG, FCR, LMD and ADFI, the fixed effects include off-set weight, year season of birth, herd, and birth parity. $\mathbf{a} \sim N(0, \boldsymbol{H}\sigma_a^2)$ is the vector of additive genetic effects, with $\boldsymbol{H}$ and σ_a^2 denoting the combined genetic relationship matrix calculated using (3) and additive genetic variance, respectively; $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the incidence matrices assigning observations to corresponding effects; $\mathbf{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$ is the vector of random residual effects, with $\boldsymbol{I}$ and σ_e^2 denoting the identity matrix and the residual variance. The variance components in (5) were also estimated using AI-REML implemented in DMU and the estimated heritability for each production trait using the following formula:

$$h^2 = \frac{\widehat{\sigma}_a^2}{\widehat{\sigma}_a^2 + \widehat{\sigma}_e^2} \tag{6}$$

The genetic correlations (r_g) between individual pairs of reproduction and production traits were independently calculated for the initial three reproductive parities using two-trait model, which is in the form of Eq. (2), where $\boldsymbol{y}$ includes each pair of reproduction and production traits. $\boldsymbol{a}$ is now assumed to follow normal distributions N, $\left(0, \boldsymbol{H} \otimes \begin{pmatrix} \sigma_{\alpha_1}^2 & \sigma_{\alpha_1\alpha_2} \\ \sigma_{\alpha_1\alpha_2} & \sigma_{\alpha_2}^2 \end{pmatrix}\right)$, where $\sigma_{a_i}^2$ is the genetic variance for trait i, $\sigma_{a_1a_2}$ is the genetic covariance between the two traits and $\otimes$ is the Kronecker product. $\boldsymbol{e}$ is now assumed to follow normal distributions N, $\left(0, \boldsymbol{I} \otimes \begin{pmatrix} \sigma_{e_1}^2 & \sigma_{e_1e_2} \\ \sigma_{e_1e_2} & \sigma_{e_2}^2 \end{pmatrix}\right)$, where $\sigma_{e_i}^2$ is the residual variance for trait i and $\sigma_{e_1e_2}$ is the residual covariance between the two traits. The equations for calculating r_g is as follows [33]:

$$\widehat{r}_g = \frac{\widehat{\sigma}_{a_1 a_2}}{\sqrt{\widehat{\sigma}^2_{a_1}\widehat{\sigma}^2_{a_2}}} \tag{7}$$

where $\widehat{\sigma}_{a_1 a_2}$ is the estimated additive genetic covariance between each pair of traits; $\widehat{\sigma}^2_{a_i}$ is the estimated additive genetic variance for trait i. The estimates were all calculated by the AI-REML algorithm using DMU software [32].

Genome-wide association study

GWAS was undertaken to identify genetic variants associated with reproduction and production traits in Group1, and the whole procedure was divided into two steps.

First step For reproduction traits, we calculated the deregressed proofs (DRP) and weights for each pig based on the GEBV estimated from (1) using the methods described by Garrick et al. [34]. The formula is as follows:

$$\begin{bmatrix} \boldsymbol{Z}'_{PA}\boldsymbol{Z}_{PA} + 4k & -2k \\ -2k & \boldsymbol{Z}'_i\boldsymbol{Z}_i \end{bmatrix} \begin{bmatrix} \mathrm{PA} \\ \mathrm{GEBV}_i \end{bmatrix} = \begin{bmatrix} y^*_{PA} \\ y^*_i \end{bmatrix} \tag{8}$$

where y^*_i is information equivalent to a right-hand-side element pertaining to the individual. PA is the parental average of genomic estimated breeding values (GEBV). $\mathrm{GEBV_i}$ is the GEBV for animal i. $\boldsymbol{Z}'_{PA}\boldsymbol{Z}_{PA}$ and $\boldsymbol{Z}'_i\boldsymbol{Z}_i$ reflect the unknown information content of the parental average and individual (plus information from any of its offspring and/or subsequent generations). They were calculated using the following formulas:

$$\begin{aligned} \boldsymbol{Z}'_{PA}\boldsymbol{Z}_{PA} &= k(0.5\alpha - 4) + 0.5k\sqrt{\alpha^2 + 16/\delta}, \\ \boldsymbol{Z}'_i\boldsymbol{Z}_i &= \delta\boldsymbol{Z}'_{PA}\boldsymbol{Z}_{PA} + 2k(2\delta - 1) \end{aligned} \tag{9}$$

where k, α and δ can be calculated using the following formulas:

$$\begin{aligned} k &= \left(1 - h^2\right)/h^2, \alpha = 1/(0.5 - \mathrm{REL_{PA}}), \\ \delta &= (0.5 - \mathrm{REL_{PA}})/(1 - \mathrm{REL}_i) \end{aligned} \tag{10}$$

where $\mathrm{REL_{PA}}$ is the reliability of the parental average, and REL_i is the reliability of the GEBV for animal i. Then, the DRP can be obtained as follows:

$$\begin{aligned} \mathrm{DRP} = \mathrm{y}^*_\mathrm{i} &= [-2k\mathrm{PA} + \left(\mathbf{Z}'_\mathrm{i}\mathbf{Z}_\mathrm{i} + 2k\right)\mathrm{GEBV_i}]/\mathbf{Z}'_\mathrm{i}\mathbf{Z}_\mathrm{i}, \\ \mathrm{REL_{DRP}} &= 1 - \mathrm{k}/(\mathbf{Z}'_\mathrm{i}\mathbf{Z}_\mathrm{i} + \mathrm{k}) \end{aligned} \tag{11}$$

where $\mathrm{REL_{DRP}}$ the reliability of the DRP for animal i.

The weights of DRP can be derived from $w_i = \frac{1-h^2}{[\mathrm{c}+(1-\mathrm{REL_{DRP}})/\mathrm{REL_{DRP}}]h^2}$, where c is assumed to be known as the proportion of genetic variation for which genotypes cannot account. In this study we set c to be 0.2.

For production traits, we calculated corrected phenotypes through adding GEBV and residual estimated from Model (2).

Second step We conducted the individual GWAS using GEMMA (v 0.98.5) [35]:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Pf} + \boldsymbol{e} \tag{12}$$

where $\boldsymbol{y}$ is the vector of DRP $*$ weight for reproduction traits and $\boldsymbol{y}$ is the vector of corrected phenotypes for production traits; $\boldsymbol{X}$ is the vector of marker genotypes, $\boldsymbol{\beta}$ is the effect size of the marker effects; $\boldsymbol{P}$ is a matrix containing the top five principal components (PCs) of genomic relationships, which were calculated using PLINK v1.9 [26], $\boldsymbol{f}$ is the vector of corresponding regression coefficients; $\boldsymbol{e}$ is a vector of random residual errors with $e \sim N(0, \boldsymbol{R}\sigma^2_e)$, where σ^2_e is residual error variance.

Model (12) was run for each SNP, with the significance threshold calculated as 1/N, where N denotes the number of relatively independent SNPs used to construct the genomic relationship matrix using Eq. (4) by GCTA v1.92.4 software [31], employing a leave-one-chromosome-out approach. Specifically, it utilized pruned genomic SNPs that were not located on the chromosome being tested. A total of 465,047 relatively independent loci were identified using the command "–indep-pairwise 50 5 0.3" in PLINK v1.90. Consequently, the genome-wide significance threshold was determined to be 2.15×10^{-6} (1/465,047).

The proportion of variance in phenotype explained (PVE) of significant SNPs was estimated as:

$$\begin{aligned} PVE &= \frac{\beta^2\mathrm{Var}(X)}{\mathrm{Var}(Y)} = \frac{\beta^2\mathrm{Var}(X)}{\beta^2\mathrm{Var}(X)+\sigma^2} \\ &= \frac{2\widehat{\beta}^2 p(1-p)}{2\widehat{\beta}^2 p(1-p)+(se(\widehat{\beta}))^2 2Np(1-p)} \end{aligned} \tag{13}$$

where $\widehat{\beta}$ is the effect size of for genetic matrix X, p is minor allele frequency, $\mathrm{se}(\widehat{\beta})$ is standard error of $\widehat{\beta}$; and N is the sample size.

Post-GWAS analysis

Conditional and joint analysis

To identify independent significant signals accurately, we employed the conditional and joint association analysis (COJO) method implemented in the GCTA v1.92.4 software to select lead SNPs ($r^2 < 0.1$) from those that achieved the genome-wide significance threshold on each chromosome. The closest gene to each lead SNP was identified as candidate gene through mapping analysis with the reference genome. In order to understand the candidate genes of the lead SNPs, we manually queried PubMed (https://pubmed.ncbi.nlm.nih.gov/), PigGTEx (https://piggtex.farmgtex.org/), and GeneCards

(https://www.genecards.org/) to obtain information on the associations between candidate genes for all lead SNPs and the traits studied.

Transcriptome-wide association study

TWAS using data from the PigGTEx project, which includes 9,530 public RNA-sequencing samples across 34 tissues, including Adipose, Blastocyst, Blood, Cartilage, Duodenum, Fetal thymus, Heart, Ileum, Kidney, Liver, Lymph node, Milk, Muscle, Ovary, Placenta, Spleen, Testis, Artery, Blastomere, Brain, Colon, Embryo, Frontal cortex, Hypothalamus, Jejunum, Large intestine, Lung, Macrophage, Morula, Oocyte, Pituitary, Small intestine, Synovial membrane, and Uterus. We conducted single-tissue TWAS using the FUSION method based on GWAS summary statistics [36]. We utilized the FarmGTEx TWAS-Server (v1, https://twas.farmgtex.org/, accessed on 1 December 2023) [37] to identify associations between genetically regulated gene expression and phenotypic traits. The server imputed gene expression levels (transcripts per million, TPM) for 26,908 genes across 34 pig tissues, sourced from the FarmGTEx project [19]. To control for multiple testing, we applied Bonferroni correction, setting the significance threshold at a corrected $\boldsymbol{P} < 0.05$.

Colocalization and summary mendelian randomization

We determined SNPs used by GWAS colocalized with eQTLs using the COLOC package (v5.1.0, https://cran.r-project.org/web/packages/coloc/) in R [38]. For COLOC, predictions were made on the basis of the reported posterior probability of colocalization (PP4), and PP4 > 0.9 was considered as significance. To further investigate the putative causal relationships between the identified SNPs and gene expression levels (eQTLs), we applied the SMR method [39] based on the linkage disequilibrium (LD) information of the Yorkshire pigs, which was calculated using PLINK v1.90 with the parameters, " –bfile, –make-bld, –r, –ld-wind 4000 and –out ". In order to identify potential association signals, a significant threshold of 0.1 was applied to identify SNPs with potential causal effects on gene expression.

Local genetic correlation estimation by SUPERGNOVA

Given the computational constraints, we focused solely on the common significant loci intervals between production traits (BFT) and reproduction traits (MLW) for estimating local genetic correlations. We first harmonized all GWAS summary data using the munge_sumstats.py function of the linkage disequilibrium score regression (LDSC v1.0.1) with parameters: "–sumstats, –N and –out" [40]. Next, we used the SUPERGNOVA (v1.0.1) software to calculate the local r_g between BFT and MLW [41].

Genomic prediction

To validate the application of the significant loci identified in GWAS for genomic prediction, we utilized the genotypic and phenotypic data from Group2 to conduct a fivefold cross-validation. This approach allowed us to compare the predictive performance of the genomic feature best linear unbiased prediction (GFBLUP) model, which incorporates the significant loci as features, against GBLUP model.

For GBLUP in reproduction trait, the statistical model was in the form of Eq. (1), where $\boldsymbol{y}$, $\boldsymbol{b}$, $\boldsymbol{s}$, $\mathbf{pe}$, $\boldsymbol{e}$, $\boldsymbol{X}$, $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$ and $\boldsymbol{Z}_3$ are same as model (1), $\boldsymbol{a} \sim N(0, \boldsymbol{A}\sigma_a^2)$ is the vector of additive genetic effects, with $\boldsymbol{A}$ denoting additive genetic relationship matrix calculated using GCTA v1.92.4 software based on Eq. (4) [31].

For GFBLUP in reproduction trait, the statistical model was:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}_{11}\boldsymbol{f} + \boldsymbol{Z}_{12}\boldsymbol{r} + \boldsymbol{Z}_2\boldsymbol{s} + \boldsymbol{Z}_3\mathbf{pe} + \boldsymbol{e} \tag{15}$$

where $\boldsymbol{y}$, $\boldsymbol{b}$, $\boldsymbol{s}$, $\mathbf{pe}$, $\boldsymbol{e}$, $\boldsymbol{X}$, $\boldsymbol{Z}_2$ and $\boldsymbol{Z}_3$ are same as GBLUP, $\boldsymbol{f}$ is the vector of the genomic values captured by the genetic markers linked to the genomic feature of interest, following a normal distribution of $\boldsymbol{f} \sim N(0, \boldsymbol{G}_f\sigma_f^2)$; and $\boldsymbol{r}$ is a vector of genomic values captured by the remaining set of genetic markers, following a normal distribution of $\boldsymbol{r} \sim N(0, \boldsymbol{G}_r\sigma_r^2)$. $\boldsymbol{Z}_{11}$ and $\boldsymbol{Z}_{12}$ are the incidence matrices. $\boldsymbol{G_f}$ was constructed according to the preselected markers which included the significant GWAS SNPs identified from Group1, while $\boldsymbol{G_r}$ was constructed according to the remaining markers.

For GBLUP in production trait, the statistical model was as model (2), where $\boldsymbol{y}$, $\boldsymbol{b}$, $\boldsymbol{e}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the same, but $\boldsymbol{a}$ is now following the same settings as GFBLUP in reproduction trait.

For GFBLUP production trait, the statistical model was:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}_1\boldsymbol{f} + \boldsymbol{Z}_2\boldsymbol{r} + \boldsymbol{e} \tag{16}$$

where $\boldsymbol{y}$, $\boldsymbol{b}$, $\boldsymbol{e}$, and $\boldsymbol{X}$ are same as GBLUP, $\boldsymbol{f}$ and $\boldsymbol{r}$ are the same as GFBLUP in reproduction trait, $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are the incidence matrices relating the additive genetic values ($\boldsymbol{g}$ and $\boldsymbol{f}$) to the phenotypic records. $\boldsymbol{G_f}$ and $\boldsymbol{G_r}$ were constructed according to using the preselected and remaining markers.

The predictive accuracy was assessed by calculating the Pearson correlation coefficient between the

Table 2 Estimates of additive genetic variance (σ^2_α), residual variance (σ^2_e), sire effect variance (σ^2_s), permanent environmental effect variance (σ^2_{pe}), heritability (h^2) and standard error (h^2(SE))

	σ^2_α	σ^2_e	σ^2_s	σ^2_{pe}	h^2	h^2(SE)
TNB	1.152	7.442	0.556	0.898	0.115	0.000169
NBA	0.923	6.471	0.360	0.622	0.110	0.000163
WeightCV	3.037	76.156	1.345	0.951	0.0373	0.0000592
MLW	0.0165	0.0317	0.00428	0.00501	0.287	0.000278
BW	25.861	78.483	-	-	0.248	0.0316
BFT	1.379	2.447	-	-	0.360	0.0321
LMD	7.748	16.157	-	-	0.324	0.0307
ADFI	30.892	106.091	-	-	0.226	0.0652
ADG	0.00204	0.0119	-	-	0.146	0.0319
FCR	0.0152	0.0895	-	-	0.145	0.0550

estimated breeding values (EBVs) and the corrected phenotypes. Both the GFBLUP and GBLUP models were calculated using the HIBLUP software [42]. The whole procedure was repeated ten times.

Result

Genetic parameters

Except for MLW ($h^2 = 0.287$), the heritability of the other three reproduction traits were quite low (Table 2). The production traits generally exhibited higher heritability estimates, with 0.248, 0.360, 0.324, 0.226, 0.146, and 0.145 for BW, BFT, LMD, ADFI, ADG and FCR, respectively.

Genetic correlations between reproduction and production traits were calculated separately for parities 1–3 (P1-P3) due to variation in reproduction traits across parities (Fig. 2a, Table S1). The correlations were generally low and predominantly negative. However, a notable positive correlation between ADG and WeightCV was observed in P1 (r_g= 0.2925), which decreased in subsequent parities (P2:r_g= 0.0724; P3:r_g= −0.022), highlighting the crucial role of parity in selection strategies. Correlations between TNB, NBA, and production traits were weak, although all were positive in P2. Conversely, the negative genetic correlation between MLW and BFT increased in magnitude across parities (P1:r_g= −0.0376; P2:r_g = −0.1371; P3:r_g = −0.1475), suggesting a biological constraint or trade-off that intensifies with increasing parity.

GWAS identified significant loci and candidate genes for reproduction and production traits

After quality control, 11,794,966 variants from Group1 were utilized in GWAS, which included 4 reproduction traits and 6 production traits. Ultimately, 264,660 significant loci were identified for reproduction traits, while 12,460 significant loci were found for production traits (Figure S1, Table S3-S7). Additionally, COJO analysis identified 73 independent signals (Fig. 2b), including 62 for reproduction traits (6 in TNB and NBA, 50 in MLW) and 11 for production traits (7 in BFT, 3 in LMD, 1 in BW) (Table 3). Annotation revealed 3,697 candidate genes and 6,014 QTL intervals, with several genes like *EGR2*, *BMPR1B*, and *FSHR* previously linked to reproductive performance. Furthermore, 1,062 shared significant loci were identified for MLW and BFT, predominantly clustered on chromosome 1 (1,059 loci) within the 50.0–54.0 Mb region (Fig. 2c), with three additional loci on chromosome 13.

Local genetic correlation in co-significant loci regions

The distinct genetic correlation and presence of shared significant loci between MLW and BFT prompted further investigation of their genetic association within these regions. Using SUPERGNOVA [41], we calculated genetic correlations in the shared loci. No significant association was detected in the co-significant region on chromosome 13. In contrast, the co-significant region on chromosome 1 exhibited weak correlations ranging

(See figure on next page.)

Fig. 2 Estimates of genetic correlations and GWAS result of 10 traits. **A** Estimates of genetic correlations. Blue, positive genetic correlation; red, negative genetic correlation. **B** A 4Mb interval in Sscrofa11.1 Chr1 enriched significant associations for BFT and MLW. **C** Circular plot representing the positions and quantities of lead SNPs obtained from GWAS analysis processed through COJO. The points on the ring indicate the locations of lead SNPs on the chromosomes associated with each trait, with gene names labeled accordingly. The traits are displayed in the circular plot from the inner to the outer ring in the order: MLW, NBA, TNB, BFT, Weight, and LMD

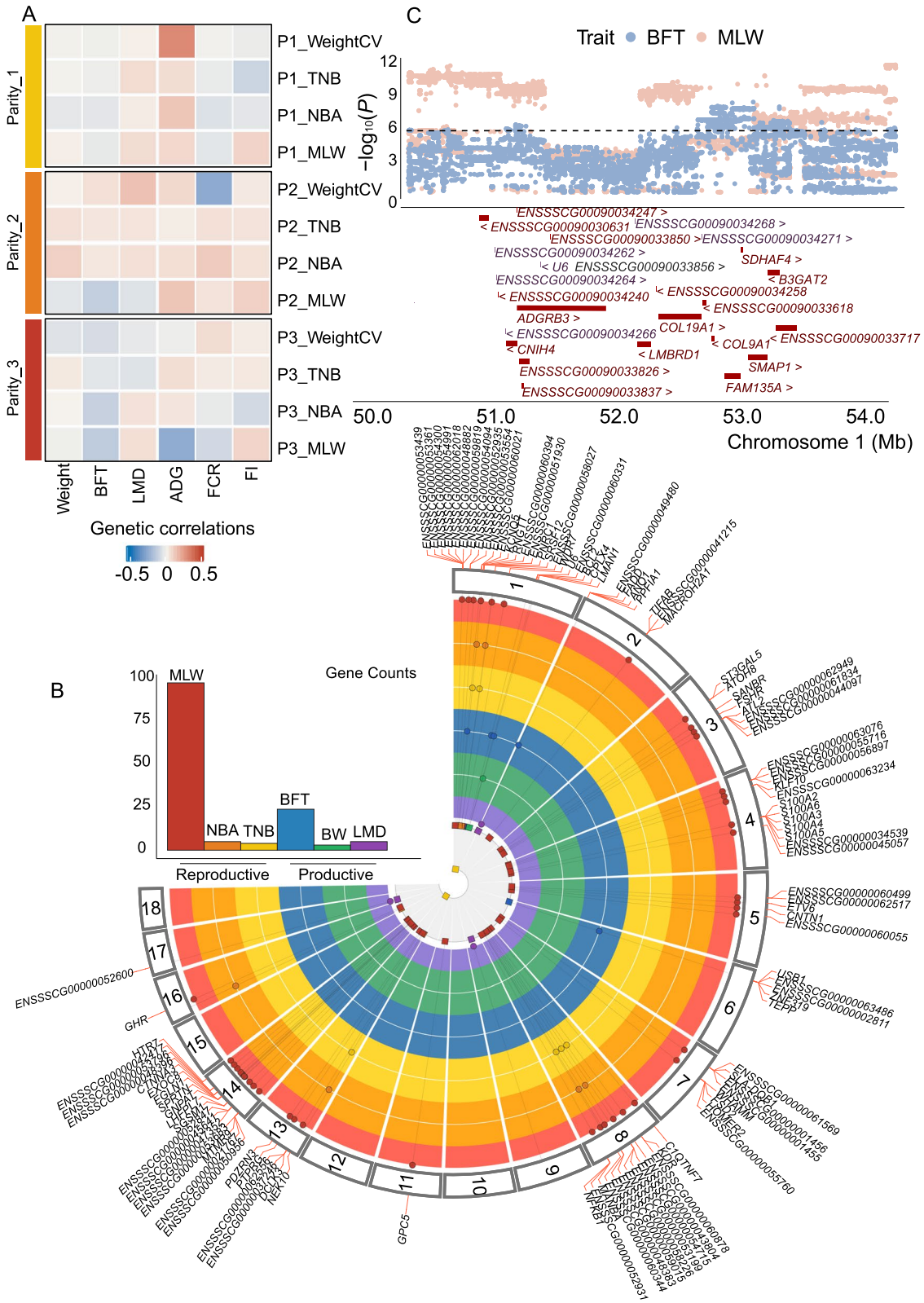

Fig. 2 (See legend on previous page.)

Table 3 Seventy-three independent signals

Chr	SNP	bp	refA	freq	b	*p*	LD_r	trait
1	1_17650409	17650409	A	0.478	−0.0786043	1.89E-11	0	MLW
1	1_31489383	31489383	T	0.484	0.0724781	8.51E-11	0	MLW
1	1_41616008	41616008	A	0.358	0.0868325	4.30E-14	0	MLW
1	1_52903638	52903638	G	0.27	0.382757	4.53E-09	0	BFT
1	1_57076537	57076537	T	0.325	−0.0993716	6.51E-18	0	MLW
1	1_57097746	57097746	A	0.323	−2.65513	1.57E-08	0	TNB
1	1_57137635	57137635	G	0.448	−2.48957	3.02E-09	0	NBA
1	1_78566750	78566750	C	0.47	−2.88334	2.21E-09	0	TNB
1	1_78566750	78566750	C	0.468	−2.60303	2.23E-08	0	NBA
1	1_80060057	80060057	C	0.388	−0.0976229	1.49E-14	0	MLW
1	1_106088048	106088048	T	0.057	−0.130498	3.34E-08	0	MLW
1	1_148348537	148348537	C	0.327	−0.360743	1.02E-08	0	BFT
1	1_158552874	158552874	A	0.341	−0.385012	7.82E-10	0	BFT
1	1_161595589	161595589	C	0.381	−1.90171	3.37E-09	0	WG
2	2_3198096	3198096	T	0.116	−1.32569	5.47E-09	0	LMD
2	2_3366709	3366709	C	0.053	0.810718	2.55E-10	0	BFT
2	2_137672690	137672690	C	0.12	−0.0953401	1.33E-08	0	MLW
3	3_58937886	58937886	T	0.477	−0.0794469	3.78E-10	0	MLW
3	3_80716636	80716636	A	0.45	−0.0747885	8.68E-09	0	MLW
3	3_91642164	91642164	A	0.395	0.0800743	1.52E-10	0	MLW
3	3_101916599	101916599	A	0.276	0.0723421	3.93E-08	0	MLW
4	4_23806331	23806331	G	0.24	0.0793875	8.59E-10	0	MLW
4	4_34414347	34414347	C	0.338	0.0803028	1.87E-12	0	MLW
4	4_46540547	46540547	T	0.15	0.101251	6.51E-10	0	MLW
4	4_96126396	96126396	G	0.182	−0.112455	1.82E-15	0	MLW
4	4_112953751	112953751	G	0.293	−0.0798908	5.91E-10	0	MLW
5	5_50376752	50376752	A	0.488	−0.0657739	6.64E-09	0	MLW
5	5_60783991	60783991	G	0.146	0.116623	1.26E-12	0	MLW
5	5_72398125	72398125	A	0.483	−0.0623669	4.44E-09	0	MLW
5	5_85715468	85715468	A	0.128	0.0879312	3.39E-08	0	MLW
6	6_19840130	19840130	T	0.187	−0.422517	6.32E-09	0	BFT
7	7_24931790	24931790	G	0.369	−0.068572	2.13E-08	0	MLW
7	7_52127214	52127214	C	0.244	0.077862	1.39E-08	0	MLW
8	8_10707246	10707246	G	0.42	0.0615775	1.32E-08	0	MLW
8	8_41865526	41865526	C	0.438	0.0672975	2.35E-08	0	MLW
8	8_57742391	57742391	T	0.28	−2.97796	2.20E-09	0	TNB
8	8_66522154	66522154	A	0.206	−0.087649	1.55E-09	0	MLW
8	8_77092738	77092738	A	0.511	2.68296	2.42E-09	0	NBA
8	8_77106783	77106783	G	0.334	−2.77227	2.41E-08	0	TNB
8	8_88457598	88457598	A	0.42	−0.0788588	8.06E-11	0	MLW
8	8_95554276	95554276	G	0.493	2.99159	8.64E-11	0	TNB
8	8_95747367	95747367	C	0.46	−2.6576	1.50E-09	0	NBA
8	8_98775172	98775172	C	0.429	−0.105222	4.29E-17	0	MLW
8	8_118381837	118381837	G	0.146	0.0940498	8.51E-10	0	MLW
9	9_83499731	83499731	A	0.18	−1.03361	2.83E-08	0	LMD
11	11_61034879	61034879	T	0.462	0.0674645	3.90E-09	0	MLW
13	13_13856063	13856063	A	0.343	2.80159	5.07E-09	0	TNB
13	13_13856063	13856063	A	0.342	2.6391	1.38E-08	0	NBA
13	13_22140758	22140758	C	0.252	0.0762411	4.88E-09	0	MLW

Table 3 (continued)

Chr	SNP	bp	refA	freq	b	p	LD_r	trait
13	13_43720049	43720049	A	0.244	0.0828626	8.83E-10	0	MLW
13	13_54913160	54913160	A	0.37	−0.0836799	5.00E-13	0	MLW
13	13_134114070	134114070	T	0.446	−0.355036	1.16E-09	0	BFT
13	13_139889196	139889196	G	0.162	0.0908292	1.73E-08	0	MLW
13	13_160434261	160434261	A	0.392	−0.0740589	1.42E-09	0	MLW
13	13_181391118	181391118	C	0.419	−0.318899	1.09E-08	0	BFT
13	13_183265993	183265993	T	0.347	−0.0697992	2.73E-08	0	MLW
14	14_14535867	14535867	A	0.371	0.076048	2.08E-10	0	MLW
14	14_31085113	31085113	T	0.128	−0.103933	1.71E-09	0	MLW
14	14_42954489	42954489	C	0.184	−0.111539	1.96E-14	0	MLW
14	14_57830778	57830778	T	0.112	0.0622371	0.000659189	−0.170976	MLW
14	14_59154730	59154730	C	0.201	−0.144115	4.90E-25	0.496021	MLW
14	14_63097471	63097471	C	0.323	−0.0228217	0.0557781	0.107681	MLW
14	14_63169107	63169107	T	0.279	0.0156105	0.231277	−0.0833084	MLW
14	14_64575780	64575780	G	0.262	−0.0461018	0.000420251	0.542911	MLW
14	14_66090588	66090588	A	0.407	−0.0343522	0.00318331	0.439825	MLW
14	14_70468838	70468838	A	0.194	−0.13138	1.28E-19	0	MLW
14	14_81120608	81120608	A	0.085	−0.138198	5.47E-11	0	MLW
14	14_91314704	91314704	G	0.394	−0.0691414	8.78E-09	0	MLW
14	14_102530914	102530914	G	0.325	−0.0683913	1.04E-08	0	MLW
14	14_119494674	119494674	T	0.405	0.0668115	7.04E-09	0	MLW
16	16_15015108	15015108	A	0.447	2.42126	4.16E-08	0	NBA
16	16_27248646	27248646	A	0.448	0.0810732	3.44E-13	0	MLW
17	17_15821131	15821131	C	0.094	−1.92546	3.93E-15	0	LMD

from 0.0012 to 0.014. The highest correlation (0.014) was located between positions 52,910,665 and 53,779,338 bp. These shared loci are exclusively located between MLW and BFT, indicating that selection for improved MLW is unlikely to cause significant correlated changes in BFT within this specific genomic region. This minimal local genetic correlation suggests that the two traits can be optimized independently within this chromosomal segment.

TWAS revealed significant loci and genes linked to reproduction traits

Based on the expression data from 34 tissues, TWAS identified 2,143 significant genes associated with various traits (Table S8). We focused on seven tissues related to reproduction and hormone secretion, namely, embryo, hypothalamus, oocyte, ovary, placenta, pituitary, and uterus. For TNB, 31 genes were found among the 175 significant loci that are related to these seven tissues. The most significant TWAS association was for the *SCLT1* gene in the oocyte ($p=1\ .74\times10^{-18}$), which is also significant in the hypothalamus ($p=4.91\times10^{-11}$) and ovary ($p=1.18\times10^{-8}$). Interestingly, in the TWAS results for MLW, *CAPN9* and *SCLT1* were significant in the hypothalamus ($p=2.47\times10^{-25}$) and oocyte tissue ($p=1.80\times10^{-21}$), indicating that *CAPN9* plays a pivotal role in reproductive physiology by potentially regulating hormone synthesis and secretion in the hypothalamus, as well as influencing oocyte development and function. The consistent, high-significance of *SCLT1* across multiple reproductive tissues further underscores its potential as a key regulator of both TNB, NBA and MLW.

Annotation of shared local significant regions between MLW and BFT identified two genes: *COL9A1* and *B3GAT2*. *B3GAT2* exhibited marked significance in MLW-relevant tissues,with hypothalamus ($p=6.91\times10^{-6}$) and frontal cortex ($p=5.38\times10^{-9}$), suggesting that *B3GAT2* may contribute to the genetic regulation of maternal life weight by modulating neural mechanisms involved in growth and hormone signaling. *COL9A1* showed broad significant associations across MLW, NBA, BFT, and LMD, implicating its role in multiple traits. To further investigate the potential regulatory mechanisms underlying these shared significant regions between MLW and BFT, we analyzed the identified regulatory elements. We discovered 536 distinct regulatory elements on chromosome 1 and 3 on chromosome 13 within these shared regions (Table S11). Notably, within a

critical zone on chromosome 1 spanning 50.0 to 54.0 Mb, we identified 460 regulatory elements. Within the region associated with *COL9A1*, we found several strongly active enhancers and promoters, as well as ATAC islands, which are indicative of open chromatin and potential regulatory activity. The presence of these elements suggests a complex regulatory network that modulates gene expression through several mechanisms.

COLOC, SMR and integrated findings identified significant genes associated with reproduction traits

COLOC and SMR identified 127 and 1,050 significant genes, respectively, associated with various traits (Table S9, S10). Notably, COLOC revealed *RNF150* as a significant gene in blood tissues for both TNB and NBA. For MLW, the most significant gene in the hypothalamus was *CENPE* (PP4 = 0.941). SMR further implicated *SCLT1*, showing significance in the oocyte ($p = 9.75 \times 10^{-5}$) and hypothalamus ($p = 5.66 \times 10^{-5}$).

For MLW, integrated analysis (TWAS, COLOC, SMR) identified *CENPE* as candidate gene (Fig. 3b); *CENPE* showed strong evidence of association ($p = 1.59 \times 10^{-5}$, Fig. 3c) and robust colocalization (PP4 = 0.941) in the hypothalamus (Fig. 3d), with further support from TWAS in the pituitary ($p = 6.54 \times 10^{-6}$). These findings align with the known function of *CENPE*, a kinetochore protein that plays a pivotal role in chromosome segregation and cell cycle regulation. In humans, where *CENPE* expression is notably higher in EBV-transformed lymphocytes compared to other tissues and is higher in males than in females (Fig. 3a). While these data originate from humans, given the conserved role of *CENPE* in the cell cycle, it is plausible to infer that in Large White pigs, *CENPE* may also exhibit tissue-specific expression, particularly in reproductively relevant tissues. *CENPE* region contains multiple regulatory elements, including a strongly active promoter/transcript and several enhancers and quiescent elements, indicating a sophisticated regulation of *CENPE* expression.

Role of significant SNPs in enhancing predictive accuracy for genomic selection

To explore the application of GWAS findings in breeding, we used the GFBLUP model to observe the impact of significant SNPs on improving the accuracy of genomic selection. Recognizing the low heritability of reproduction traits, which necessitates an expanded feature set to capture more of the genetic variation, we selected the top 0.2% of SNPs based on the GWAS *p* values and included them as additional genetic relationship matrices in the GFBLUP model (Fig. 4a-b). This approach was evaluated against the standard GBLUP model using a fivefold cross-validation. Notably, the inclusion of these significant SNPs led to substantial improvements in predictive accuracy. Predictive accuracy for MLW increased from 0.0168 with GBLUP to 0.0242 with GFBLUP. Similar improvements were observed for NBA (0.0901 to 0.0967), TNB (0.0735 to 0.0905), and WeightCV (0.0018 to 0.0217).

Discussion

In this study, we conducted a GWAS to identify genetic variants linked to production and reproduction traits in Yorkshire pigs. We detected 277,120 significant loci for these traits, providing insights into their genetic architecture. The large number of loci suggests that these traits are influenced by a complex genetic framework, consistent with findings in other livestock species that demonstrate their polygenic nature [43]. Notably, reproduction traits exhibited a higher number of significant loci despite their lower heritability estimates. This can be attributed to their polygenic nature, where numerous genetic variants with small individual effects contribute to trait variation. The heritability of these traits is influenced not only by their genetic architecture but also by environmental factors such as nutrition, management practices, and physiological conditions, which may explain the lower heritability estimates observed. The identification of a larger number of significant loci, particularly at a lower significance threshold, reflects the presence of many small-effect variants that collectively contribute to these traits.

To enhance the resolution and power of our GWAS, we utilized the PGRP panel for genotype imputation to millions of SNPs, which significantly increased the marker density in our study. This high-density genotyping is especially valuable for traits governed by many small-effect variants, as it improves our ability to detect significant associations. These findings align with previous research in pigs [44], supporting the idea that reproduction traits are shaped by a combination of numerous genetic and environmental factors. The complex genetic architecture of these traits highlights the need for a more nuanced selection strategy, considering both genetic and environmental influences.

The results indicate that reproduction and production traits are interrelated, with these relationships varying across different parities. The differing genetic correlations suggest that breeding strategies should consider both reproduction and production traits to optimize improvements in both areas. Specifically, the negative correlations between production traits, such as ADG, and reproduction traits, like MLW, in the third parity emphasize the necessity of carefully managing selection pressures to avoid adverse effects on reproductive performance. Our computational analysis also revealed positive relationships among several production and reproduction traits.

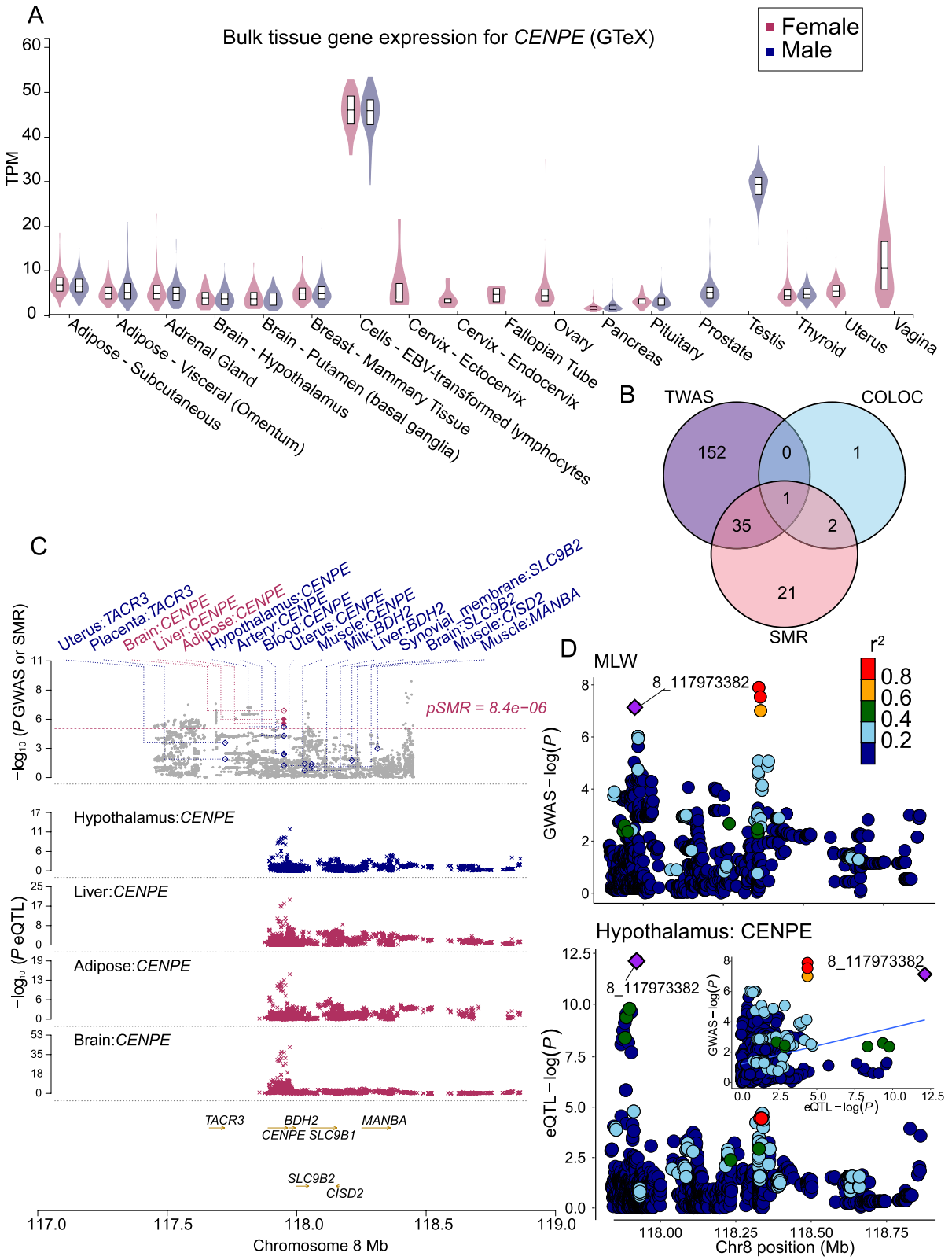

Fig. 3 The post-GWAS analysis of MLW. **A** The expression of *CENPE* gene across various human tissues and between sexes from the GTEx database. **B** Venn diagram showing the significant genes associated with MLW detected by COLOC, SMR, and TWAS approached. **C** The results of SMR of *CENPE*. **D** The results of COLOC of *CENPE*

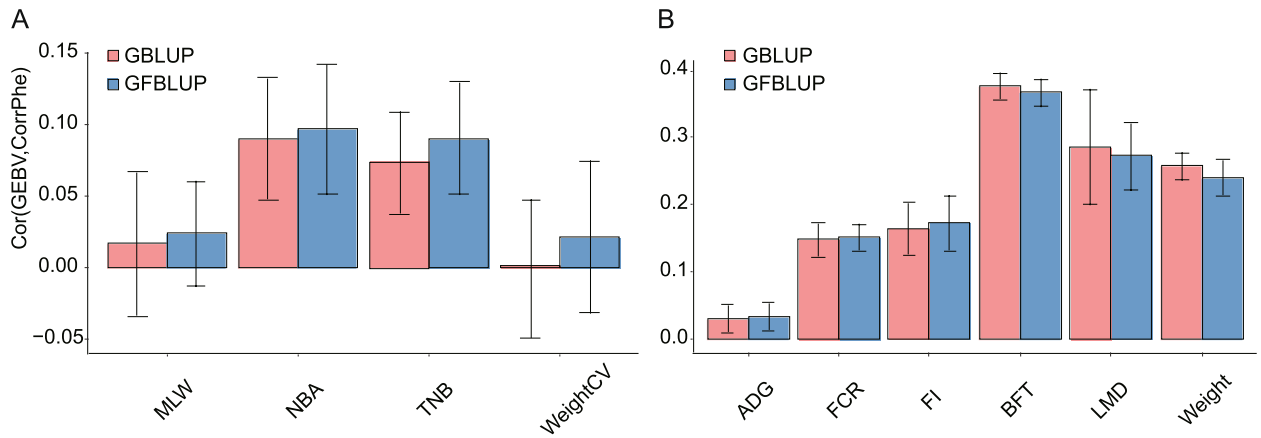

Fig. 4 Impact of top SNPs on genomic prediction accuracy. Correlation between genomic estimated breeding values (GEBV) and corrected phenotypes for (**A**) reproduction traits and (**B**) production traits. The figure compares the predictive accuracy of standard GBLUP and GFBLUP models incorporating the top 0.2% of SNPs. Error bars represent standard errors

While a negative correlation between these categories has traditionally been assumed, posing challenges for balanced selection, our findings suggest a more nuanced relationship. The long-standing emphasis on production traits has inadvertently contributed to a decline in reproductive performance. However, recent years have seen a growing recognition of the importance of reproduction traits. Despite this, progress in breeding for both traits has been hindered by limited knowledge of their genetic backgrounds. Our study's identification of positively correlated traits (e.g., ADG and WeightCV, with a correlation coefficient of 0.2925 in the first parity) provides valuable insights to guide breeding efforts. Notably, the correlations between reproduction and production traits appeared relatively consistent across parities, although minor variations suggest potential influences of parity on these relationships. This study may have limitations, such as sample size and the specific traits chosen. Further research with larger datasets and additional traits could offer more comprehensive insights. These results highlight the importance of targeted selective breeding strategies to improve both productive and reproductive efficiency across different parities, ultimately enhancing overall productivity and economic returns in livestock breeding programs.

The annotation of GWAS results indicated that a substantial number of identified significant genes have been previously associated with related traits in the literature, reinforcing the credibility of our findings [44–46]. For instance, genes such as *EGR2*, *BMPR1B*, and *FSHR*, previously linked to reproduction traits, were also found to be significantly associated with MLW in our analysis. By employing post-GWAS analyses, we uncovered several promising loci and genes warranting further investigation. For example, the *CENPE* gene, identified within a significant locus for MLW, also demonstrated significance in TWAS, COLOC, and SMR analyses. *CENPE*, a protein-coding gene, plays a crucial role in regulating chromosome segregation, cell division, and mitosis. Given its expression in the hypothalamus and uterus, *CENPE* may influence cell proliferation and, consequently, MLW through its involvement in the mitotic process. Furthermore, *CENPE*, located at Chr8:117973382, exhibited high expression in human cultured fibroblasts and Epstein-Barr virus (EBV)-transformed lymphocytes, with expression levels higher in females than in males (Fig. 3a). Since both cell types are involved in cell division and proliferation, this elevated expression suggests a unique function for this gene. Within the *CENPE* gene region, several regulatory elements were identified, likely interacting in a complex network to control *CENPE* expression. Quiescent elements may silence the gene under certain conditions, while a strongly active promoter drives its expression in others. Weakly transcribed regions could contribute to tissue-specific or inducible expression. Additionally, accessible ATAC islands and flanking transcription start sites (TSSs) may facilitate transcription, while enhancers, including medium and poised enhancers, could amplify or modulate gene expression. The interplay of these regulatory elements likely contributes to the spatiotemporal regulation of *CENPE* expression, ensuring appropriate expression in different cell types and developmental stages. Further studies are needed to fully elucidate the intricate interactions and regulatory dynamics within the *CENPE* gene region. It is hypothesized that this gene may regulate cell proliferation and influence trait expression through yet-to-be-elucidated mechanisms.

By integrating insights from multiple GWAS results, we identified a set of loci that significantly affect both reproduction and production traits, with several shared loci located on chromosomes 1 and 13. Among these loci, *COL9A1*, a gene encoding Collagen Type IX Alpha 1 Chain, was particularly noteworthy. Our analysis indicated that *COL9A1* may influence both BFT and MLW, though the direction of effect varies between traits. The local genetic correlation at the shared loci on chromosome 1 is relatively low because it is estimated by considering the entire genomic region rather than individual genes within that region. Since *COL9A1* represents only a small portion of this region, its effect does not necessarily reflect the overall genetic correlation between BFT and MLW. The estimated correlation tends to regress toward the regional mean, and the contribution of a single gene within this region may be independent of the broader correlation pattern. Despite this, *COL9A1* still appears to play a role in the shared region, with its TWAS results showing significant associations with both traits. Mutations in *COL9A1* have been associated with Epiphyseal Dysplasia and Stickler Syndrome, which further supports its potential involvement in trait variation in Yorkshire pigs. LD analysis of this region identified several LD blocks, indicating that SNPs within these blocks could exert stronger effects on the traits. For instance, at locus Chr1:50072259 (Fig. 5a-c), the TT genotype ($n=1298$) was associated with the highest BFT, while the trend was reversed for MLW, where CC ($n=606$) genotypes exhibited higher values. This suggests that *COL9A1* may influence these traits through different mechanisms. The presence of many regulatory elements in this region further complicates the interpretation of the gene's role.

Our study not only unearthed a wealth of candidate loci and genes but also demonstrated the practical utility of these findings in improving the accuracy of genomic selection. By incorporating these discoveries as features into GFBLUP, we achieved an enhancement in selection accuracy compared to conventional GBLUP. However, the prediction accuracy for reproduction traits remained relatively low (Fig. 4b), which can be attributed to their complex genetic architecture

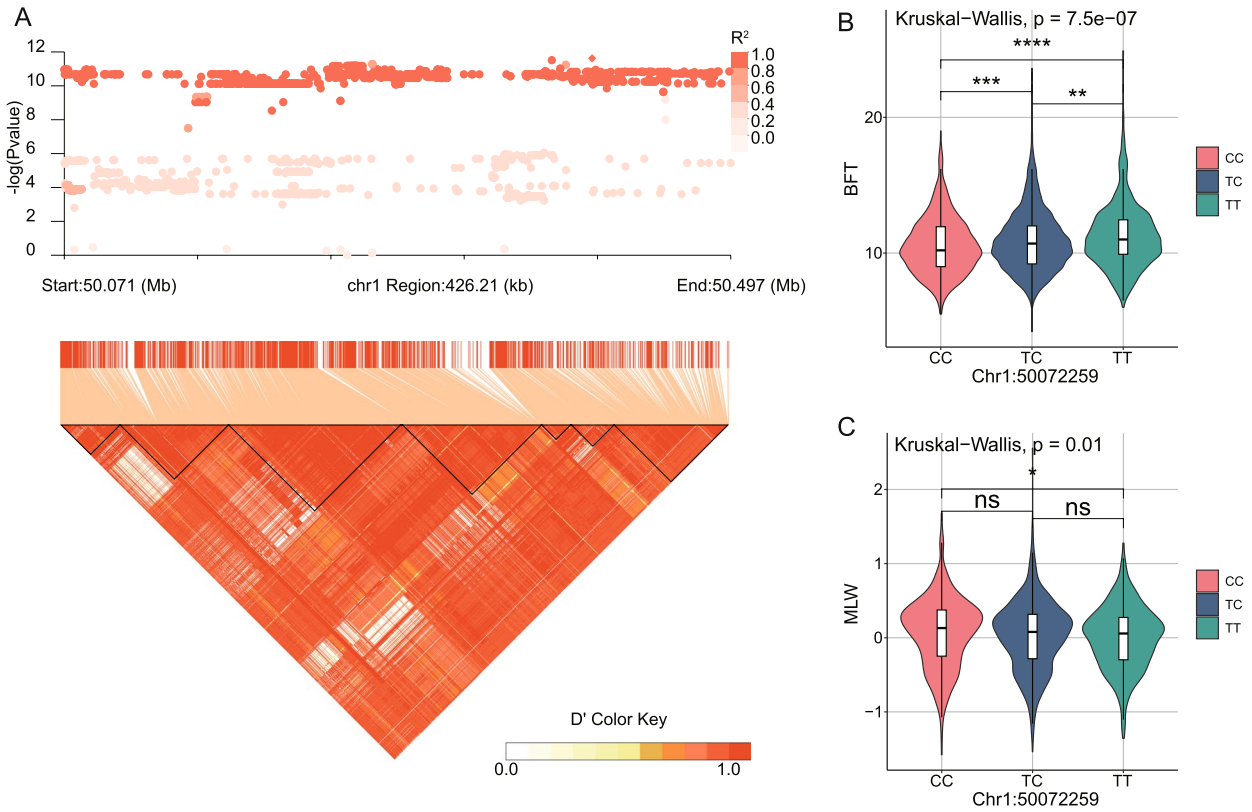

Fig. 5 Shared associated regions between BFT and MLW. **A** The linkage disequilibrium (LD) block spanning the chromosomal region Chr1:50.071Mb-50.497Mb, including the corresponding R-squared (R^2) values. **B** Genotype distribution of Chr1:50,072,259 locus and the association between different genotypes and phenotypes. The difference of phenotypes of different genotypes was detected by t-test. The number of * represents the degree of significance level. * means the *p* value is less than 0.05, ** means the *p* value is less than 0.01, *** means the *p* value is less than 0.001. **C** Genotype distribution of Chr1:50,072,259 locus and the association between different genotypes and DEBV

and lower heritability. These traits are influenced by numerous small-effect loci, making it more challenging to capture their genetic variation effectively using genomic prediction models. While GFBLUP outperformed GBLUP for some production traits, it showed lower prediction accuracy for others. This may be due to the fact that production traits often have fewer major-effect loci, which GFBLUP can better leverage, whereas reproduction traits rely on a larger number of minor-effect variants. Additionally, GBLUP, as a well-established method, may provide more stable performance for traits with highly polygenic architectures. Since GFBLUP is a more recent approach, further optimization may be needed to improve its predictive power, particularly for traits with lower heritability.

Our study has laid a foundation for understanding the genetic architecture of reproduction and production traits in Yorkshire pigs. The next crucial step involves subjecting these identified candidate genes and loci to rigorous experimental validation. This process will enable us to confirm their functional significance and establish a direct link between genetic variation and phenotypic traits. Furthermore, we envision the integration of additional technical tools and knowledge to further illuminate the genetic mechanisms underlying these traits. Pan-genomic analyses, which delve into the structural variation (SV) landscape, offer a powerful avenue for exploring the role of large genomic variants in shaping trait variation [47]. By combining these approaches, we can paint a comprehensive picture of the genetic landscape governing reproduction and production traits in Yorkshire pigs, paving the way for targeted breeding strategies that enhance both reproductive performance and growth efficiency in this important pig breed.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11416-0.

Supplementary Material 1: Figure S1 Manhattan plot showing GWAS results for reproduction traits in Yorkshire pigs. Traits included are TNB (A), NBA (B), WeightCV (C), and MLW (D). Significance threshold is indicated by the horizontal dashed line.

Supplementary Material 2: Figure S2 Manhattan plot showing GWAS results for production traits in Yorkshire pigs. Traits included are BW (A), BFT (B), LMD (C), ADFI (D) , ADG (E), and FCR (F). Significance threshold is indicated by the horizontal dashed line.

Supplementary Material 3: Figure S3 Quantile-quantile (QQ) plots of GWAS results for reproduction and production traits. Traits included are TNB (A), NBA (B), WeightCV (C), MLW (D), BW (E), BFT (F), LMD (G), ADFI (H), ADG (I), and FCR (J).

Supplementary Material 4: Table S1. Genetic correlations between reproduction and production traits in Yorkshire pigs across three parities.

Supplementary Material 5: Table S2. Summary of 10 phenotypes.

Supplementary Material 6: Table S3. Genes searched around significant loci for TNB.

Supplementary Material 7: Table S4. Genes searched around significant loci for NBA.

Supplementary Material 8: Table S5. Genes searched around significant loci for MLW.

Supplementary Material 9: Table S6. Genes searched around significant loci for WeightCV.

Supplementary Material 10: Table S7. Genes searched around significant loci for production traits.

Supplementary Material 11: Table S8. The result of TWAS.

Supplementary Material 12: Table S9. The result of COLOC.

Supplementary Material 13: Table S10. The result of SMR.

Supplementary Material 14: Table S11. The element of important region.

Supplementary Material 15: Table S12. Phenotypic correlations between all traits in Yorkshire pigs.

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

Authors' contributions
YP and ZZ conceived and designed the study, obtained funding, and reviewed and revised the manuscript. RW participated in the design of the study, performed the statistical analysis, drafted and revised the manuscript. ZYZ processed the raw genotype data. HH helped organize the productive phenotypes. JM revised the manuscript. PY carried out the RNA-seq data analysis. HC helped organize the reproductive phenotypes. WZ played crucial roles in the collaboration between Zhejiang University and SciGene Biotechnology Co., Ltd. XH was responsible for collecting the phenotypic data. JW participated in the collection of phenotypic data. YH participated in the collection of phenotypic data. YF was responsible for the collaboration between SciGene Biotechnology Co., Ltd. and our laboratory. ZW conceived and designed the study. QW was involved in the funding application. All authors read and approved the final manuscript.

Funding
This work was supported by the National Key Research and Development Program of China (2023YFF1001100), Zhejiang Science and Technology Major Program on Agricultural New Variety Breeding (grant nos. 2021C02068-5); and National Natural Science Foundation of China (grant nos. 32272832).

Data availability
Data supporting this study are publicly available in the European Variation Archive (EVA) under the project accession number PRJEB83454. The specific analyses associated with this study are accessioned as ERZ24985230 and ERZ24985231. The data can be accessed via the following link: https://www.ebi.ac.uk/eva/?eva-study=PRJEB83454. Phenotype data are available upon request and subject to agreement with the breeding organization.

Declarations

Ethics approval and consent to participate
Ethical permission to samples from pigs was approved by the Institutional Animal Care and Use Committee of Zhejiang University. All procedures in which pigs were involved were per the agreement of the Institutional Animal Care and Use Committee of Zhejiang University.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]College of Animal Sciences, Zhejiang University, Hangzhou 310058, China. [2]Hainan Institute, Zhejiang University, Yongyou Industry Park, Yazhou Bay Sci-Tech City, Sanya 572000, China. [3]SciGene Biotechnology Co., Ltd, Hefei 230022, China.

Received: 12 November 2024 Accepted: 28 February 2025
Published online: 29 March 2025

References

1. Jiang Y, Tang S, Xiao W, Yun P, Ding X. A genome-wide association study of reproduction traits in four pig populations with different genetic backgrounds. Asian-Australas J Anim Sci. 2020;33:1400–10.
2. Chang Wu Z, Wang Y, Huang X, Wu S, Bao W. A genome-wide association study of important reproduction traits in large white pigs. Gene. 2022;838:146702.
3. Song H, Dong T, Yan X, Wang W, Tian Z, Sun A, et al. Genomic selection and its research progress in aquaculture breeding. Rev Aquac. 2023;15:274–91.
4. do Nascimento AV, Romero ÂR da S, Utsunomiya YT, Utsunomiya ATH, Cardoso DF, Neves HHR, et al. Genome-wide association study using haplotype alleles for the evaluation of reproductive traits in Nelore cattle. PLoS One. 2018;13:e0201876.
5. Desta ZA, Ortiz R. Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. 2014;19:592–601.
6. Gutierrez-Reinoso MA, Aponte PM, Garcia-Herreros M. genomic analysis, progress and future perspectives in dairy cattle selection: a review. Animals. 2021;11:599.
7. Zhang Y, Zhang J, Gong H, Cui L, Zhang W, Ma J, et al. Genetic correlation of fatty acid composition with growth, carcass, fat deposition and meat quality traits based on GWAS data in six pig populations. Meat Sci. 2019;150:47–55.
8. Zhang H, Shen L-Y, Xu Z-C, Kramer LM, Yu J-Q, Zhang X-Y, et al. Haplotype-based genome-wide association studies for carcass and growth traits in chicken. Poult Sci. 2020;99:2349–61.
9. Freebern E, Santos DJA, Fang L, Jiang J, Parker Gaddis KL, Liu GE, et al. GWAS and fine-mapping of livability and six disease traits in Holstein cattle. BMC Genomics. 2020;21:41.
10. Jiang J, Cao Y, Shan H, Wu J, Song X, Jiang Y. The GWAS analysis of body size and population verification of related SNPs in Hu sheep. Front Genet. 2021;12:642552.
11. 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.
12. Xu P, Ni L, Tao Y, Ma Z, Hu T, Zhao X, et al. Genome-wide association study for growth and fatness traits in Chinese Sujiang pigs. Anim Genet. 2020;51:314–8.
13. Shi L, Wang L, Fang L, Li M, Tian J, Wang L, et al. Integrating genome-wide association studies and population genomics analysis reveals the genetic architecture of growth and backfat traits in pigs. Front Genet. 2022;13.
14. Xue Y, Li C, Duan D, Wang M, Han X, Wang K, et al. Genome-wide association studies for growth-related traits in a crossbreed pig population. Anim Genet. 2021;52:217–22.
15. Höglund J, Rafati N, Rask-Andersen M, Enroth S, Karlsson T, Ek WE, et al. Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. Sci Rep. 2019;9:16844.
16. 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.
17. Ding R, Savegnago R, Liu J, Long N, Tan C, Cai G, et al. The SWine IMputation (SWIM) haplotype reference panel enables nucleotide resolution genetic mapping in pigs. Commun Biol. 2023;6:1–10.
18. Zhang K, Liang J, Fu Y, Chu J, Fu L, Wang Y, et al. AGIDB: a versatile database for genotype imputation and variant decoding across species. Nucleic Acids Res. 2023;52:D835–49.
19. 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.
20. Li MJ, Yan B, Sham PC, Wang J. Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. Brief Bioinform. 2015;16:393–412.
21. Berry DP, Wall E, Pryce JE. Genetics and genomics of reproductive performance in dairy and beef cattle. Animal. 2014;8:105–21.
22. Pedersen LD, Kargo M, Berg P, Voergaard J, Buch LH, Sørensen AC. Genomic selection strategies in dairy cattle breeding programmes: Sexed semen cannot replace multiple ovulation and embryo transfer as superior reproductive technology. J Animal Breed Genet. 2012;129:152–63.
23. Bereskin B, Frobish LT. Carcass and related traits in duroc and yorkshire pigs selected for sow productivity and pig performance1. J Anim Sci. 1982;55:554–64.
24. Lopez BIM, Song C, Seo K. Genetic parameters and trends for production traits and their relationship with litter traits in Landrace and Yorkshire pigs. Anim Sci J. 2018;89:1381–8.
25. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. Am J Hum Genet. 2018;103:338–48.
26. 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.
27. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. Imeta. 2023;2:e107.
28. 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.
29. Xing Y, Li G, Wang Z, Feng B, Song Z, Wu C. GTZ: a fast compression and cloud transmission tool optimized for FASTQ files. BMC Bioinformatics. 2017;18:549.
30. Rubinacci S, Hofmeister RJ, Sousa da Mota B, Delaneau O. Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. Nat Genet. 2023;55:1088–90.
31. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.
32. Madsen P, Su G, Labouriau R, Christensen O. DMU – A Package for Analyzing Multivariate Mixed Models. the proceedings of the 8th World Congress on Genetics Applied to Livestock Production; Brasil. 2006.
33. Kassahun D, Taye M, Kebede D, Tilahun M, Tesfa A, Bitew A, et al. Phenotypic and genetic parameter estimates for early growth, growth rate and growth efficiency-related traits of Fogera cattle in Ethiopia. Veter Med Sci. 2022;8:387–97.
34. Garrick DJ, Taylor JF, Fernando RL. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol. 2009;41:55.
35. Parker CC, Gopalakrishnan S, Carbonetto P, Gonzales NM, Leung E, Park YJ, et al. Genome-wide association study of behavioral, physiological and gene expression traits in outbred CFW mice. Nat Genet. 2016;48:919–26.
36. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48:245–52.
37. Zhang Z, Chen Z, Teng J, Liu S, Lin Q, Wu J, et al. FarmGTEx TWAS-server: an interactive web server for customized TWAS analysis. Genomics Proteomics Bioinform. 2025:qzaf006.
38. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. J R Stat Soc Ser B Stat Methodol. 2020;82:1273–300.
39. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48:481–7.
40. Bulik-Sullivan B. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47.
41. Zhang Y, Lu Q, Ye Y, Huang K, Liu W, Wu Y, et al. SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. Genome Biol. 2021;22:262.
42. Yin L, Zhang H, Tang Z, Yin D, Fu Y, Yuan X, et al. HIBLUP: an integration of statistical models on the BLUP framework for efficient genetic evaluation using big genomic data. Nucleic Acids Res. 2023;51:3501–12.
43. Vanvanhossou SFU, Scheper C, Dossa LH, Yin T, Brügemann K, König S. A multi-breed GWAS for morphometric traits in four Beninese indigenous cattle breeds reveals loci associated with conformation, carcass and adaptive traits. BMC Genomics. 2020;21:783.
44. Yang Y, Gan M, Yang X, Zhu P, Luo Y, Liu B, et al. Estimation of genetic parameters of pig reproductive traits. Front Vet Sci. 2023;10.

45. Kwon SG, Hwang JH, Park DH, Kim TW, Kang DG, Kang KH, et al. Identification of differentially expressed genes associated with litter size in berkshire pig placenta. PLOS ONE. 2016;11:e0153311.
46. Ma X, Yi H. BMP15 regulates FSHR through TGF-β receptor II and SMAD4 signaling in prepubertal ovary of Rongchang pigs. Res Vet Sci. 2022;143:66–73.
47. Li R, Gong M, Zhang X, Wang F, Liu Z, Zhang L, et al. A sheep pangenome reveals the spectrum of structural variations and their effects on tail phenotypes. Genome Res. 2023;33:463–77.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.