

Systematic Approach for Compound Angus Populations Revealing Functional Genes and Improving Prediction Accuracy in Carcass Traits

Yuanyuan Yu

Southwest Minzu University

Yujiao Fu

Southwest Minzu University

Jiahong Zhao

Southwest Minzu University

Shiyu Wu

Southwest Minzu University

Shikai Wang

Southwest Minzu University

Zemin Li

Southwest Minzu University

Li Liu

Heilongjiang Provincial Academy of Agricultural Sciences

Fang Sun

Heilongjiang Provincial Academy of Agricultural Sciences

Jincheng Zhong

Southwest Minzu University

Jiabo Wang

Southwest Minzu University

Daoliang Lan

Southwest Minzu University

Kangzhu Yixi

kangzhu_smu@163.com

Southwest Minzu University

Research Article

Keywords: GWAS, GS, meta-analysis, carcass traits

Posted Date: March 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8747702/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Carcass traits, which reflect growth performance and muscle development, are economically important in beef cattle, yet their genetic determinants remain poorly characterized. Both single-population GWAS methods, such as BLINK, and cross-population meta-analysis approaches are widely used to identify genetic variants, yet their comparative performance in genomic prediction for complex traits in structured populations remains underexplored. Few studies have directly compared these methods in genomic prediction. To address this gap, this study aims to (i) identify functional genes associated with carcass traits and (ii) evaluate the context-dependent advantages of Covariate Adjustment (CA) and meta in genomic prediction. In this study, we analyzed carcass weight (CW), live weight (LW), and dressing percentage (DP) in 279 crossbred Angus cattle genotyped with the PHR0105_Bt140K_v1.0 SNP chip. Genome-wide association studies (GWAS) was performed on the full population using BLINK, and results from three subpopulations were combined via meta-analysis, with significance thresholds for both approaches determined by a shuffle-based method. Candidate genes located within ± 10 kb of significant SNPs were associated with different carcass traits, including STRIT1, SEL1L3, NOC4L and ANK1 for DP; SNCA and DNAH5 for CW; and GYPC, GPR158, and GUCY1A1 for LW. Prediction accuracy under MAS and MABLUP showed meta slightly outperformed BLINK in MAS, while BLINK was better with covariate adjustment; after incorporating kinship in MABLUP, meta achieved higher accuracy and population partitioning was negligible. Overall, MABLUP yielded the highest accuracy (0.52–0.79) versus MAS (0.37–0.54) in all traits. These findings provide a methodological basis for selecting appropriate GWAS strategies in structured populations and highlight candidate genes.

1. Introduction

With the continuous growth of the global population, beef consumption has increased steadily over recent decades[1]. Across the country, beef is the third most consumed meat after poultry and pork, highlighting its major role in overall meat supply[2]. Therefore, efficient beef production is crucial for maintaining a stable meat market and supporting the sustainability of the livestock sector. Given this context, carcass traits have become key indicators of beef production performance. Traits such as live weight, carcass weight, and dressing percentage directly determine beef yield and carcass quality. As classical quantitative traits controlled by multiple genes and shaped by environmental factors[3], identifying their genetic determinants is essential for improving breeding efficiency.

Genome-wide association studies (GWAS) leverage linkage disequilibrium to detect associations between genetic variants and phenotypic traits. By fitting multifactorial linear models, GWAS quantify genotype–phenotype relationships and identify significantly associated variants.[4, 5]. Despite these advances, GWAS power can diminish in structured or composite populations, where differences in allele frequencies, biased genetic backgrounds, and linkage disequilibrium (LD) patterns may obscure true signals and increase false positives. Thus, single-population GWAS may be insufficient to capture the full genetic architecture of complex traits across diverse cattle populations.

The population stratification is the major challenge in GWAS, especially from different data sources. In the traditional GWAS approach, the sub-population affiliation was considered as a covariance effect. This method was called the approach within model or Covariate Adjustment (CA). The estimated effect of the sub-population was a major variance component from phenotype values. Specifically, the output file derived from the CA procedure is uniformly designated as index through out this study. Another approach is the approach after results or meta-analysis, which uses the GWAS results from every single sub-population, then evaluates all markers' contribution under heterogeneity from different sources[6]. There is no exact conclusion of which method is the best, especially in beef conduction traits.

The Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK) method has been reported that it uses LD information to optimize covariates and Bayesian Information Criterion (BIC) to iteratively refine pseudo quantitative trait nucleotides (QTNs) within a fixed-effect framework, effectively controlling population structure and reducing false positives[7, 8]. While BLINK excels in analyzing individual GWAS datasets, integrating results across multiple populations requires a complementary approach. Meta-analysis originated in the 1970s and was first proposed by psychologist Gene V. Glass as a quantitative approach for synthesizing results from multiple independent studies[9]. With the development of GWAS, large-scale data from different populations have accumulated, making meta-analysis an important tool for integrating results, and identifying consistent signals. The METAL is specifically designed for GWAS meta-analysis, efficiently handling moderate to large-scale SNP data, supporting multiple input formats, and allowing sample-size weighting and allele coding adjustments to integrate data across cohorts[10].

To validate the effectiveness and utility of identified genetic markers, the most definitive approach is through gene knockout or targeted interference experiments. However, such methods are often impractical and ethically challenging in large animal studies. A more feasible alternative is to leverage marker-assisted selection (MAS)[11] or genomic selection (GS)[12] for predictive validation. In this context, the improvement in prediction accuracy contributed by significant markers in an independent validation population can serve as a key verification criterion. A truly associated and functional marker should explain a portion of the genetic variance and demonstrably enhance predictive performance in genomic selection models.

To overcome these limitations, we address two key research questions: i) identifying candidate genes associated with beef carcass traits; and ii) comparing and validating the predictive performance of SNPs detected by CA and meta-analysis in GS. The findings will not only evaluate their practical breeding utility but also provide methodological insights for optimizing multi-population GWAS strategies and enhancing genetic improvement of carcass traits.

2. Materials and Methods

2.1. Sample Collection and Genotyping

A total of 279 beef cattle were sampled from three locations in China: Sunwu County, Heilongjiang Province (48 Angus), Dachang Hui Autonomous County, Hebei Province (109 Angus × Wagyu crossbreds), and Qingyang City, Gansu Province (122 Angus). Phenotypic data for live weight (kg), carcass weight (kg), and dressing percentage (%) were collected following standardized carcass evaluation procedures, including electrical stimulation and inverted exsanguination. Live weight was measured after 6 hours fasting period. Carcass weight was recorded after removing the internal organs, and dressing percentage was calculated as carcass weight divided by live weight. Blood and muscle samples were collected, and information on age, ear tag number, and geographic origin was recorded for subsequent genotyping analyses. All individuals' informations were recorded in both of notes and Excel files as independent and verifiable backups.

Genomic DNA was extracted from muscle and blood samples, and genotyping was performed using the PHR0105_Bt140K_v1.0 SNP chip provided by Boready Biomarker Technologies (Shijiazhuang, China)[13]. The mensurable DNA (500ng) was sampled for fragmentation using restriction enzyme. The fractured DNA was connected to the A tail after terminal repair. The ligation products were added to sequencing primers with Barcodes and a high-fidelity PCR reaction system for PCR amplification. After the probe hybridization experiment, the bases of the hybridization sites were typed. All genotype and hybridization site information were saved in the HapMap file. The original data was updated to GitHub database (Data availability). Quality control was conducted in R 4.3.3, and SNPs were removed based on the following criteria: (1) missing rate > 0.05; (2) minor allele frequency (MAF) < 0.01; (3) presence of indels; and (4) multiple alleles. After filtering, genotype imputation was performed using Beagle (v4.1)[14], and high-quality SNPs were retained for subsequent GWAS and GS analyses.

2.2. Population Stratification

The Genome-Associated Prediction Integrated Tool (GAPIT)[15] implemented in R (v4.3.3) was employed to conduct a comprehensive set of genomic analyses, including linkage disequilibrium (LD) and heterozygosity assessments among markers, as well as population structure analyses through principal component analysis (PCA) and phylogenetic relationship reconstruction using a neighbor-joining (NJ) tree. Moreover, genome-wide LD decay curves were subsequently generated using PopLDdecay-3.42 to calculate LD and summarize its decay with increasing inter-marker distance[16].

2.3. Association Study

To establish an appropriate significance threshold, the shuffle method was applied to determine the genome-wide threshold line. Phenotypic values for the CW were randomly shuffled among individuals to break the true genotype–phenotype associations, and GWAS was repeated 50 times. The minimum P-value from each permutation was recorded, and the mean of these 50 values was used to define the final genome-wide significance threshold for all traits.

In this study, genome-wide association analysis was performed using the BLINK model implemented in the GAPIT (v3.0). BLINK is a multi-locus method that iteratively incorporates associated markers as

covariates while testing additional markers, thereby enhancing statistical power and reducing false positives[7]. The BLINK model can be expressed as follows:

$$Y = PCA + SNP_i + e$$

$$Y = PCA + pseudoQTN + SNP_i + e$$

where Y is the phenotype vector, and PCA denotes the top principal components fitted as fixed effects to account for population structure. $pseudoQTN$ represents the set of associated markers selected in previous iterations and included as covariates. SNP_i is the genotype of the marker being tested, and e is the residual error, assumed to follow a normal distribution.

BLINK employs two fixed-effect models: one for selecting pseudo QTNs and another for testing marker effects, which enhances both statistical power and computational efficiency. Model convergence is determined using the BIC, ensuring more accurate model selection compared with likelihood-based approaches. Manhattan and quantile–quantile plots were generated using GAPIT’s built-in functions to visualize the distribution and significance of association signals.

2.4 Meta-analysis

For the meta-analysis, we employed the z-score (sample size–based) method implemented in the METAL software[10]. This approach combines different parameters from individual studies, including sample size, P -values (p_i), and the direction of effects. For each variant, the method converts the p -value into a z -score using the formula.

$$Z = \frac{\sum_i Z_i w_i}{\sqrt{\sum_i w_i^2}}$$

where $w_i = \sqrt{N_i}$ is the square root of the sample size for study i , and

$$Z_i = \Phi^{-1}(1 - p_i/2) \times (\text{effect direction for study } i)$$

with Φ denoting the cumulative distribution function of the standard normal distribution and p_i the p -value of the i th study.

The overall P -value was calculated as

$$P = 2\Phi(-|Z|)$$

This method ensures that larger studies contribute more heavily to the combined test statistic, thereby weighting the different studies according to their sample size. The significance threshold for the meta-analysis was defined consistently with that used in the GWAS.

2.5 Gene annotation

Genes within the regions harboring significant association loci were annotated based on the bovine reference genome Bos_taurus.ARS-UCD2.0.114.gff3 (https://ftp.ensembl.org/pub/release-114/gff3/bos_taurus/). We converted genotype files in HapMap format to VCF format using TASSEL (v5.0)[17], which was then used for subsequent LD analysis. For each significant SNP, all SNPs within ± 500 kb upstream and downstream were extracted, and the corresponding genotype data were formatted as VCF files for input into LDBlockShow (v1.41)[18]. LDBlockShow calculated pairwise LD metrics, such as r^2 or D' , between SNPs and automatically defined LD blocks, providing a visualization of the linkage relationships and block structures within the candidate gene regions.

2.6 Genomic Selection

In this study, we conducted GS predictions using two approaches. First, MAS was implemented by utilizing only the significant SNPs identified from the GWAS, allowing us to evaluate the predictive power of top associated loci for the traits of interest. Second, a mixed-model approach incorporating both the kinship matrix and the significant SNPs—referred to as marker-assisted best linear unbiased prediction (MABLUP)—was applied. This method leverages the overall genetic relationships among individuals, captured by the kinship matrix, while also incorporating the effects of trait-associated markers, thereby improving the accuracy of genomic predictions. By using these approaches, we aimed to evaluate the effectiveness of SNPs discovered through AC and meta-analysis for genomic selection prediction.

For the MAS model, populations from Sunwu, Qingyang, and Dachang County were each subjected to five-fold cross-validation. In each fold, the selected training sets were combined, and the remaining individuals formed the test set. GWAS was conducted on the training set, with the population origin included as a covariate to identify significant SNPs. In the test set, the significant SNPs identified from the training set were included as fixed effects together with PCA components in a linear model (LM) to predict phenotypes and genomic estimated breeding values (gEBVs). Prediction accuracy was assessed by calculating the Pearson correlation between predicted values and observed phenotypes. If no significant SNPs were identified in the training set, only PCA was used as fixed effects.

In the meta part, GWAS was conducted separately for the three training sets without including the index as a covariate. The results were then combined in METAL to identify significant SNPs. Subsequent prediction in the test set followed the same procedure as in BLINK, using the significant SNPs together with PCA as fixed effects in the LM model to estimate phenotypes and gBV, and calculate prediction accuracy.

The prediction linear model can be expressed as:

$$Y = \mu + Xb + QTN\alpha + e$$

where Y is the vector of observed phenotypes, μ is the overall mean, Xb represents additional fixed effects, QTN contains the genotypes of the significant SNPs identified by GWAS, α is the vector of SNP effects, and e is the residual error.

For the MABLUP model, the division of training and test sets and the procedure for identifying significant SNPs were the same as in MAS. The significant SNPs identified by GWAS as fixed effects, while the kinship matrix was included as a random effect in MLM framework. The procedure for selecting significant SNPs was the same as in MAS. The model is formulated as:

$$Y = \mu + Xb + QTN\alpha + Zg + e$$

The parameter definitions are the same as in MAS, where (Zg) represents the random effects associated with the kinship matrix. Prediction accuracy was evaluated by the Pearson correlation between the gEBVs and the observed phenotypes.

In addition, we performed predictions using randomly divided populations into three equal subsets, following the same procedure for both MAS and MABLUP models. This experiment aimed to test whether the population structure adjustment can effectively improve prediction accuracy and to provide a finer comparison of CA and Meta prediction performance.

3.Result

3.1. Descriptive statistics and genetic parameters estimations

Phenotype Description

We analyzed three carcass traits in 279 cattle and summarized the descriptive statistics(mean and standard deviation). Most records of LW ranged from 450 kg to 980 kg. The mean value of LW, CW, and DP was 715.3 kg, 414.5 kg, and 57.9%. All traits exhibited normal distributions. Pearson's correlation coefficients were calculated for the three populations SW, QY and DC ((Fig. 1a-c)). Across these populations, a strong positive correlation was detected between LW and CW, a moderate positive correlation between CW and DP, and a negative correlation between LW and DP. To further examine inter-trait relationships and group-specific patterns, pairwise correlation analyses and density distributions were explored (Fig. 1d), yielding results consistent with those observed in the three individual populations. Group-specific density plots reveal distinct phenotypic distributions, particularly for DP, indicating possible population structure or differential selection.

Genotype conversion and view

Genotyping of 279 cattle was conducted using GBTS, targeting approximately 140K SNPs across the bovine genome. After quality control filtering to exclude SNPs containing indels, missing data, or multiple alleles, a total of 127,918 high-quality SNPs were retained for downstream analyses. For individual heterozygosity, the values were primarily clustered between 0.3 and 0.4, showing a distinct peak in this interval. In contrast, the heterozygosity of SNP markers was predominantly concentrated within 0.4–0.6,

with the highest frequency observed in this range. The MAF exhibited a distribution spanning 0.01 to 0.5, with frequencies gradually increasing across this range (Fig. S1).

We analyzed the correlation coefficient (R) and determination coefficient (R^2) between the neighbor SNPs pairs in the genotype matrix. The upper panels show the distribution of raw correlation coefficients (R), which remained a stable decline with increasing distance (Fig. S2). The lower panels display LD decay among markers, with R^2 decreasing as distance increased and reaching a plateau at approximately 8 kb with a value around 0.3. The overall LD decay patterns were highly similar among DC, SW, and QY, with only minor differences observed beyond 250 kb, suggesting comparable recombination rates and genetic diversity (Fig. S3)

3.2. Population structure and estimated heritability

Visualization of the first three principal components revealed three distinct clusters, with individuals within each cluster showing close genetic relatedness. Interestingly, cattle from DC and SW were genetically more similar, reflecting geographic proximity. The first three principal components accounted for 9.96%, 5.7%, and 2.24% of the total genetic variance, respectively, highlighting the moderate but discernible population differentiation (Fig. 2a). From the perspective of the NJ tree, the genetic structure of the populations revealed that QY formed an independent lineage, while DC and SW clustered on adjacent branches, consistent with partial shared ancestry (Fig. 2b).

For DP, the genetic proportion increased to 62.2%, with the residual contributing 37.8%. For CW, the genetic variance accounted for 45% of the total variance, while the residual variance explained 55%. For LW the genetic and residual components explained 51.7% and 48.3% of the total variance, respectively (Fig. S4). Here, the genetic variance represents variation attributable to heritable effects, reflecting the potential response to selection, whereas the residual variance includes environmental influences and unexplained variation.

3.3. GWAS, Meta-analysis, and Candidate Gene

As mentioned in the Methods, the GWAS was performed by two approaches: 1) Meta-analysis with each sub-population; 2) CA-analysis with the whole population. In the meta-analysis, 3 subpopulations were performed BLINK GWAS, respectively. All GWAS results in DP trait were used to draw the Manhattan and quantile-quantile (QQ) plots (Fig. 3). The dotted gray line indicates there are two methods to detect this common signal. The solid gray line indicates there are more than three methods to detect this common signal. Totally, there are 6 SNPs (there are 2 signals detected by CA and Meta in the same position) that were identified as significant. The relative functional gene names were marked in the Manhattan plot. The common position of significant signals was on chromosome 6, represented by rs37571 and rs37577 within *SEL1L3*, a protein-coding gene involved in ER-associated degradation (ERAD) and protein-folding quality control. The significant loci were mapped to several functional genes with biological relevance to muscle development and metabolism. rs5563 (Chr1) corresponds to *STRIT1*, a gene involved in potential regulation of muscle structural stability. rs127557 (Chr27) is located within *ANK1*, which plays a key role

in cytoskeletal stability and function. rs95061 (Chr17) is associated with *NOC4L*, a gene involved in fundamental cellular processes such as protein synthesis and cell growth.

For LW, four significant SNPs were identified: rs12171 (Chr2), rs69377 (Chr11), rs75952 (Chr13), and rs94988 (Chr17) (Fig. S5). Among these, rs94988 (Chr17) and rs12171 (Chr2) exhibited the highest proportions of variance explained (PVE = 20.96% and 91.86%, respectively), highlighting their major contributions to LW variation. rs94988 was consistently detected in the meta-analysis and combined population and is annotated to *GUCY1A1*, involved in smooth muscle relaxation and blood pressure regulation. rs12171 is annotated to *GYPC*, associated with red blood cell integrity and oxygen transport. Other significant loci, including *PSMB7* (Chr11), involved in muscle hypertrophy, and *GPR158* (Chr13), associated with endocrine regulation, showed lower PVE values.

For CW, three significant SNPs—rs106790 (Chr6), rs36257 (Chr20), and rs93045 (Chr17)—were detected, all exclusively in the combined population (Fig. S6). Among these, rs106790 and rs36257 exhibited negative effects, with PVEs of 2.19% and 24.83%, respectively, and are annotated to *SNCA* (Chr6) and *DNAH5* (Chr20). *SNCA* (Chr6) encodes α -synuclein, and *DNAH5* (Chr20) encodes a dynein heavy chain; neither gene is associated with slaughter traits. In contrast, rs93045 (Chr17) showed a positive effect with a high PVE of 52.50% and corresponds to an unannotated long non-coding RNA, suggesting a potential regulatory role. Given its large contribution to CW variation and chromosomal location, this locus on chromosome 17 represents a key candidate region for follow-up functional studies. All significant SNPs and candidate genes information were observed in Table 1, and their estimated genetic effects are in Table S1.

Table 1
Significant SNP and candidate gene information

SNP No. *	Chr	position	P.value	gene name	biotype	trait
rs5563	1	112371771	1.19E-09	STRIT1	protein_coding	DP
rs9962	2	34488282	4.01E-08	SLC4A10	protein_coding	DP
rs12183	2	79056259	2.67E-12	No_name	lncRNA	DP
rs30285	5	48843409	1.16E-07	No_name	lncRNA	DP
rs37571	6	45331116	1.10E-07	SEL1L3	protein_coding	DP
rs95061	17	44994156	2.95E-10	NOC4L	protein_coding	DP
rs99644	18	59327463	9.18E-07	No_name	Protein_coding	DP
rs116390	23	40500113	2.02E-09	No_name	LncRNA	DP
rs127557	27	36709501	1.50E-10	ANK1	protein_coding	DP
rs36257	6	34851001	3.84E-07	SNCA	protein_coding	CW
rs93045	17	3855180	8.18E-08	No_name	LncRNA	CW
rs106790	20	59342623	9.27E-09	DNAH5	protein_coding	CW
rs12171	2	78736403	1.55E-12	GYPC	protein_coding	LW
rs69377	11	95429986	2.44E-08	PSMB7	protein_coding	LW
rs75952	13	26167866	1.54E-07	GPR158	protein_coding	LW
rs94988	17	43621923	1.74E-10	GUCY1A1	protein_coding	LW
SNP No.* indicates the sequence number in the entire tag list.Chromosomes and locations refer to physical location information in genomic data. The gene names are annotated from the GTF file of the Bos_taurus.ARS-UCD2.0.114 reference genome.						

Through Fig. 3, the comparison of the GWAS results from individual subpopulation and whole population (whatever in CA and Meta) shows some common signals and different signals. To answer the reason for these differences, we compared the phenotype distribution of the genotypes of these significant markers(Fig. 4). The significant signal in such a population was marked in red. The rs5563 SNP was significant in the only whole population, while not significant in other subpopulations. That is because of there are enough samples in different genotypes in the whole population. In the subpopulation, the number of individuals with CT genotype is few or lacking. The same trend can be observed in the rs127557, and rs37571. The few or lack of the key genotype could also make fake positive (rs9962, rs30285, and rs99644). The same trend were also observed in LW and CW traits(Fig. S7 and S8).

3.4 The LD block between SNPs

To characterize the local LD patterns surrounding significant SNPs, we analyzed a 500 kb region upstream and downstream of each locus. The resulting LD heatmaps revealed distinct patterns among loci. For instance, strong LD blocks were observed around rs37571 and rs95061, indicating tight linkage with neighboring variants and highlighting their robustness as candidate loci for functional follow-up. In contrast, loci including rs5563, rs127557, rs9962, rs116390, rs12183, rs99644, and rs30285 exhibited weaker LD with adjacent markers, suggesting a more dispersed local haplotype structure. Pink-shaded regions in the heatmaps denote gene regions that either overlap with or are located near the significant SNPs (Fig. 5). Similarly, loci associated with LW and CW—such as rs69377, rs12171, rs75952, rs94988, rs36257, rs93045, and rs106790—generally showed modest LD with surrounding markers (Fig. S9 and S10). Overall, the diversity in LD patterns provides a valuable basis for prioritizing candidate loci for downstream functional validation.

3.5 Prediction with significant markers

In the individual population, predictive performance was used to evaluate the detection ability of the CA and meta methods. For all prediction models—including MAS and MABLUP—significant SNPs identified by each model were incorporated as covariates. Prediction accuracy was defined as the Pearson correlation between predicted gEBVs and observed phenotypes, as well as between predicted phenotypes and observed phenotypes. The former correlation assessed the predictability of gEBVs (reflecting all genetic factors), while the latter evaluated the predictability of phenotype (incorporating both genetic and non-genetic factors).

In comparisons of gEBVs accuracy, the meta approach consistently outperformed CA, with an average improvement of approximately 0.01. Notably, however, the CA model exhibited superior predictive performance in terms of phenotype accuracy. For the DP and CW traits, the improvement was 0.10, and for LW, it was approximately 0.04. This suggests that incorporating the covariance effects positively contributes to phenotype prediction accuracy (Fig. 6a and 6c).

To further investigate whether the enhanced predictive ability stems from the population indicator used in the CA approach, the entire population was randomly divided into three sub-populations. The GS and GWAS procedures were then reapplied using both the CA and meta methods, generating new indicators based on random allocation. This process was repeated 20 times. As a result, the prediction accuracy of CA and meta decreased substantially, dropping to around 0.10 (Fig. 6b and 6d).

When the kinship matrix was included in MABLUP, the meta method generally outperformed CA in prediction accuracy, with DP increasing by approximately 0.03 and LW and CW both increasing by about 0.1 (Fig. S10a). When the overall population was randomly partitioned, the prediction accuracy of both meta and CA showed little difference compared with that obtained from the original population, with discrepancies within 0.01 (Fig. S10b). In summary, the meta method demonstrated either superior or comparable performance overall, and the advantage associated with a predefined population indicator was diminished.

4. Discussion

4.1. Candidate Genes

In this study, we identified significant SNPs and candidate genes associated with carcass traits using BLINK and meta. Although only some loci were shared between the methods, their overlap provides mutual validation and highlights the polygenic and complex biological nature of carcass traits.

For DP, *ANK1*, a membrane–cytoskeleton adaptor, shows SNPs—including missense and promoter variants—associated with carcass weight, drip loss, and Warner–Bratzler shear force in yak, suggesting its influence on muscle architecture and meat quality[19, 20]. *STRIT1*, associated with sarcoplasmic reticulum calcium transport, is biologically connected to muscle contraction and growth[21]. *SEL1L3*, previously linked to milk protein content, was significant in both BLINK and meta analyses despite no known direct relation to dressing percentage[22]. Together, these three genes emerge as key candidate genes for carcass traits. Other detected genes, such as *NOC4L*, which encodes a nucleolar protein involved in ribosome biogenesis and pre-rRNA processing, potentially contributing to core cellular processes such as protein synthesis and cell growth, although functional evidence in cattle is limited[23]. *SLC4A10*, a regulator of neuronal function, was detected only in the DC population and has no known direct association with carcass traits[24].

For LW, the identified genes are primarily involved in oxygen transport, vascular regulation, and muscle protein turnover. *GYPC*, a red blood cell membrane protein, maintains cell shape and oxygen delivery, influencing muscle growth and carcass composition[25]. *GUCY1A1* regulates vascular function[26], while *PSMB7*, a proteasome-related gene, mediates protein turnover and muscle accretion, as supported by *MSTN*-dependent transcriptomic studies[27].

For CW, Francisco Alejandro Paredes-Sánchez et al. identified *SNCA* (α -synuclein) in Angus and Brangus cattle as associated with temperament[28]. Similarly, *DNAH5* is critical for proper cilia function and related physiological processes[29], but both genes appear to have minimal direct impact on carcass traits.

Moreover, several significant variants were located in long non-coding RNA regions, implying potential regulatory functions, although their biological roles in cattle remain largely unexplored. These findings collectively indicate that the identified loci may influence carcass traits through diverse mechanisms involving cytoskeletal stability, protein homeostasis, metabolic control.

4.2. Genomic Prediction Performance of CA and Meta

In addition to identifying candidate genes, it is crucial to assess the practical value of these genes or the genetic signals they represent in breeding. For this reason, we further compared the genomic predictive capabilities of SNPs discovered based on different GWAS strategies.

Beyond gene identification, our comparison of different genomic prediction frameworks provided important methodological insights. In MAS, when comparing gEBVs with observed phenotypes, meta markers showed slightly higher correlations. However, when comparing predicted values with observed values, CA performed markedly better, indicating that properly accounting for population structure can substantially improve prediction accuracy in mixed or stratified populations. The markedly reduced accuracy observed in randomly divided populations is likely due to unstable SNP effects and insufficient adjustment for population structure, further highlighting the important role of correcting for population structure in genomic selection. Interestingly, meta became superior again, whereas the index provided minimal additional benefit in MABLUP. This suggests that the kinship matrix can effectively adjust for population structure even under random grouping, while the effect of the index is relatively minor, likely because already absorb most of the structure-related variance, thereby reducing the incremental value of predefined groupings. Consistent with previous reports, we found that relying only on significant SNPs for prediction resulted in lower accuracy — likely because such a subset cannot capture all the genetic variation influencing complex traits[30]. The contrasting performance of CA and meta under different conditions highlights their complementary strengths and implies that no single method is universally optimal. Instead, a combined or context-specific strategy may be more effective for capturing both common and population-specific genetic signals related to carcass traits.

Our results indicate that for populations with obvious geographical or genetic stratification, explicitly correcting the population structure within the model can more effectively improve the prediction accuracy of MAS. Moreover, under the MABLUP model, integrating multi-population information based on meta-analysis can capture more stable genetic effects and is robust to the choice of initial population classification.

There are several limitations should be acknowledged. Functional validation of the identified candidate genes is still lacking, and several associated loci reside in non-coding regions where regulatory roles remain uncertain. Further transcriptomic, epigenomic, and gene-editing studies will be necessary to clarify the causal variants and their biological mechanisms. Additionally, the study population consisted of crossbred Angus cattle, and residual stratification or sampling bias cannot be fully ruled out despite the use of multiple correction strategies. Increasing sample size, expanding population diversity, and refining population structure modeling will be important directions for future work. Nevertheless, the integration of GWAS, meta-analysis, and genomic prediction offers a robust framework for detecting meaningful loci and evaluating their predictive value in breeding programs. The candidate genes and methodological comparisons presented here provide a valuable foundation for future molecular breeding efforts aimed at improving carcass traits and overall production efficiency in beef cattle.

5. Conclusions

In conclusion, this study integrated GWAS, meta-analysis, and genomic selection methodologies to uncover the genetic basis of key carcass traits in beef cattle. Some significant SNPs and biologically relevant candidate genes were identified, pointing to pathways involved in cytoskeletal organization,

protein turnover and metabolic regulation. Methodologically, the comparative analyses demonstrated that meta provides improved genomic prediction under standard GS frameworks, whereas the CA approach enhances accuracy, highlighting the complementary strengths of the two strategies. Together, these findings improve our understanding of the genetic architecture of carcass traits and offer practical insights for optimizing genomic prediction models in structured populations. The candidate genes and predictive approaches identified provide a valuable foundation for future functional validation and for advancing molecular breeding programs aimed at improving carcass performance and production efficiency in beef cattle.

Abbreviations

SNP
Single nucleotide polymorphism
GWAS
Genome-wide association study
GS
genomic selection
CA
Covariate Adjustment
BIC
Bayesian Information Criterion
QTNs
pseudo quantitative trait nucleotides
BLINK
Bayesian-information and linkage-disequilibrium iteratively nested keyway
CW
Carcass weight
LW
Live weight
DP
Dressing percentage
MLM
Mixed linear model
BLINK
Bayesian-information and linkage-disequilibrium iteratively nested keyway GS:Genomic selection
gEBVs
Genomic estimated breeding values
MAF
Minor allele frequency
PCA

Principal component analysis
NJ
Neighbor-joining
GAPIT
Genome association and prediction integrated tool
MAS
Marker-Assisted Selection
MABLUP
Marker-Assisted Best Linear Unbiased Prediction
LM
Linear model

Declarations

Ethics Statement

All animal procedures were approved by the Animal Ethics and Welfare Association of Southwest Minzu University (Approval No. 16053). All experimental protocols adhered to relevant institutional and national guidelines, and reporting complies with the ARRIVE guidelines.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding:

This work was supported by the Heilongjiang Province Key Research and Development Project, China (2022ZX02B09); the Key R&D Project of Tibet Autonomous Region: "Breeding and Application Research of New Yak Strains in Tibet" (XZ202501ZY0031); the Qinghai Science and Technology Program, China (2022-NK-110); the Fundamental Research Funds for the Central Universities, China (Southwest Minzu University, ZYN2024050); and the Program of Chinese National Beef Cattle and Yak Industrial Technology System, China (CARS-37).

Author Contribution

YYY wrote the first draft of the manuscript. DLL and JBW conceived the study and revised the manuscript. JHZ, YJF, SWu, SWang, ZL,LL, FS, and JCZ contributed to data collection and analysis. JBW and YX KJ provided critical revisions and approved the final version. All authors read and approved the final manuscript.

Data Availability

All raw sequencing outputs were deposited in the NCBI Sequence Read Archive (SRA) under Bioproject accession PRJNA1418350. The datasets generated and/or analyzed during the current study are publicly available at: <https://github.com/yy0587023-rgb/Systematic-Approach-for-Compound-Angus-Populations>.

References

1. Jadeja R, Teng XM, Mohan A, Duggirala K. Value-Added Utilization of Beef by-Products and Low-Value Comminuted Beef: Challenges and Opportunities. *Curr Opin Food Sci*. 2022;48:100938. 10.1016/j.cofs.2022.100938.
2. OECD; Food and Agriculture Organization of the United Nations. *OECD-FAO Agricultural Outlook 2022–2031*; OECD-FAO Agricultural Outlook; OECD Publishing, 2022; ISBN 978-92-64-58870-7.
3. Lee S-H, Park B-H, Sharma A, Dang C-G, Lee S-S, Choi T-J, Choy Y-H, Kim H-C, Jeon K-J, Kim S-D, et al. Hanwoo Cattle: Origin, Domestication, Breeding Strategies and Genomic Selection. *J Anim Sci Technol*. 2014;56(2). 10.1186/2055-0391-56-2.
4. Jiang J, Ma L, Prakapenka D, VanRaden PM, Cole JB, Da YA. Large-Scale Genome-Wide Association Study in U.S. Holstein Cattle. *Front Genet*. 2019;10:412. 10.3389/fgene.2019.00412.
5. Zhu H, Li X, Zhang M, Liu S, Zhang Y, Zheng Y, Wei Z, Han M, Huang H, Fu T, et al. Detection of Selection Signatures and Genome-Wide Association Analysis of Body Weight Traits in Xianan Cattle. *Genes*. 2025;16:682. 10.3390/genes16060682.
6. Zeggini E, Ioannidis JP. Meta-Analysis in Genome-Wide Association Studies. *Pharmacogenomics* 2009, 10, 191–201. 10.2217/14622416.10.2.191
7. Huang M, Liu X, Zhou Y, Summers RM, Zhang ZBLINK. A Package for the next Level of Genome-Wide Association Studies with Both Individuals and Markers in the Millions. *GigaScience* 2019, 8, giy154. 10.1093/gigascience/giy154
8. Neupane B. Systematic Comparison of GWAS Methods in Wheat: Balancing Statistical Power. False Positive Control, and Computational Efficiency; 2025.
9. Shadish WR. Introduction to the Special Issue on the Origins of Modern Meta-analysis. *Res Synthesis Methods*. 2015;6:219–20. 10.1002/jrsm.1148.

10. Willer CJ, Li Y, Abecasis GR. Fast and Efficient Meta-Analysis of Genomewide Association Scans. *Bioinformatics*. 2010;26:2190–1. 10.1093/bioinformatics/btq340.
11. Wakchaure R, Ganguly S. Marker Assisted Selection (MAS) in Animal Breeding: A Review. *J Drug Metab Toxicol*. 2015;06. 10.4172/2157-7609.1000e127.
12. Singh BD, Singh AK. Genomic Selection. In *Marker-Assisted Plant Breeding: Principles and Practices*; Springer India: New Delhi, 2015; pp. 295–312 ISBN 978-81-322-2315-3.
13. Xu Y, Yang Q-N, Zheng H-J, et al. Genotyping by Target Sequencing (GBTS) and Its Applications. *Sci Agric Sin*. 2020;53:2983–3004. 10.3864/j.issn.0578-1752.2020.15.001.
14. Browning SR, Browning BL, Rapid. *Am J Hum Genet*. 2007;81:1084–97. 10.1086/521987. and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering.
15. Wang J, Zhang ZGAPIT. Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. *Genom Proteom Bioinform*. 2021;19:629–40. 10.1016/j.gpb.2021.08.005.
16. Zhang C, Dong S-S, Xu J-Y, He W-M, Yang T-L, PopLDdecay. A Fast and Effective Tool for Linkage Disequilibrium Decay Analysis Based on Variant Call Format Files. *Bioinformatics*. 2019;35:1786–8. 10.1093/bioinformatics/bty875.
17. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: Software for Association Mapping of Complex Traits in Diverse Samples. *Bioinformatics*. 2007;23:2633–5. 10.1093/bioinformatics/btm308.
18. LDBlockShow. A Fast and Convenient Tool for Visualizing Linkage Disequilibrium and Haplotype Blocks Based on Variant Call Format Files.
19. Giacomello E, Quarta M, Paolini C, Squecco R, Fusco P, Toniolo L, Blaauw B, Formoso L, Rossi D, Birkenmeier C, et al. Deletion of Small Ankyrin 1 (sAnk1) Isoforms Results in Structural and Functional Alterations in Aging Skeletal Muscle Fibers. *Am J Physiology-Cell Physiol*. 2015;308:C123–38. 10.1152/ajpcell.00090.2014.
20. Hu J, Gao X, Shi B, Chen H, Zhao Z, Wang J, Liu X, Li S, Luo Y. Sequence and Haplotypes of Ankyrin 1 Gene (ANK1) and Their Association with Carcass and Meat Quality Traits in Yak. *Mamm Genome*. 2021;32:104–14. 10.1007/s00335-021-09861-9.
21. Veluri R, Pollin G, Wagenknecht JB, Urrutia R, Zimmermann MT. An Integrative Multitiered Computational Analysis for Better Understanding the Structure and Function of 85 Miniproteins 2025.
22. Sanchez M-P, Ramayo-Caldas Y, Wolf V, Laithier C, El Jabri M, Michenet A, Boussaha M, Taussat S, Fritz S, Delacroix-Buchet A, Sequence-Based GWAS, et al. Network and Pathway Analyses Reveal Genes Co-Associated with Milk Cheese-Making Properties and Milk Composition in Montbéliarde Cows. *Genet Sel Evol*. 2019;51. 10.1186/s12711-019-0473-7.
23. Zhu X, Zhang W, Guo J, Zhang X, Li L, Wang T, Yan J, Zhang F, Hou B, Gao N, et al. Noc4L-Mediated Ribosome Biogenesis Controls Activation of Regulatory and Conventional T Cells. *Cell Rep*. 2019;27:1205–e12204. 10.1016/j.celrep.2019.03.083.

24. Ding W, Gong W, Bou T, Shi L, Lin Y, Shi X, Li Z, Wu H, Dugarjaviin M, Bai D. Analysis of ROH Characteristics Across Generations in Grassland-Thoroughbred Horses and Identification of Loci Associated with Athletic Traits. *Animals* 2025, 15, 2068. 10.3390/ani15142068
25. Jiang A, Zhang J, Yuan C-B, Xiang B-J, Huang D-J, Gao L-F, Guang-Xin E. Whole-Genome Scanning for the Heat-Resistance-Associated Genes in the Droughtmaster Breed (*Bos Taurus*). *3 Biotech* 2019, 9, 95. 10.1007/s13205-019-1620-0
26. Childers KC, Garcin ED. Structure/Function of the Soluble Guanylyl Cyclase Catalytic Domain. Nitric Oxide. 2018;77:53–64. 10.1016/j.niox.2018.04.008.
27. Sadkowski T, Jank M, Zwierzchowski L, Siadkowska E, Oprządek J, Motyl T. Gene Expression Profiling in Skeletal Muscle of Holstein-Friesian Bulls with Single-Nucleotide Polymorphism in the Myostatin Gene 5'-Flanking Region. *J Appl Genet.* 2008;49:237–50. 10.1007/BF03195620.
28. Paredes-Sánchez FA, Sifuentes-Rincón AM, Lara-Ramírez EE, Casas E, Rodríguez-Almeida FA, Herrera-Mayorga EV, Randel RD. Identificación de Genes Candidatos y SNP Relacionados Con El Temperamento Del Ganado Utilizando Un Análisis GWAS Junto Con Un Análisis de Redes Interactuantes. *Rev Mex Cienc Pecu.* 2022;14:1–22. 10.22319/rmcp.v14i1.6077.
29. Dementieva NV, Dysin AP, Shcherbakov YS, Nikitkina EV, Musidray AA, Petrova AV, Mitrofanova OV, Plemyashov KV, Azovtseva AI, Griffin DK, et al. Risk of Sperm Disorders and Impaired Fertility in Frozen–Thawed Bull Semen: A Genome-Wide Association Study. *Animals.* 2024;14:251. 10.3390/ani14020251.
30. Kamprasert N, Aliloo H, Van Der Werf JHJ, Duff CJ, Clark SA. Effect of Using Preselected Markers from Imputed Whole-Genome Sequence for Genomic Prediction in Angus Cattle. *Genet Sel Evol.* 2025;57:52. 10.1186/s12711-025-00999-7.

Figures

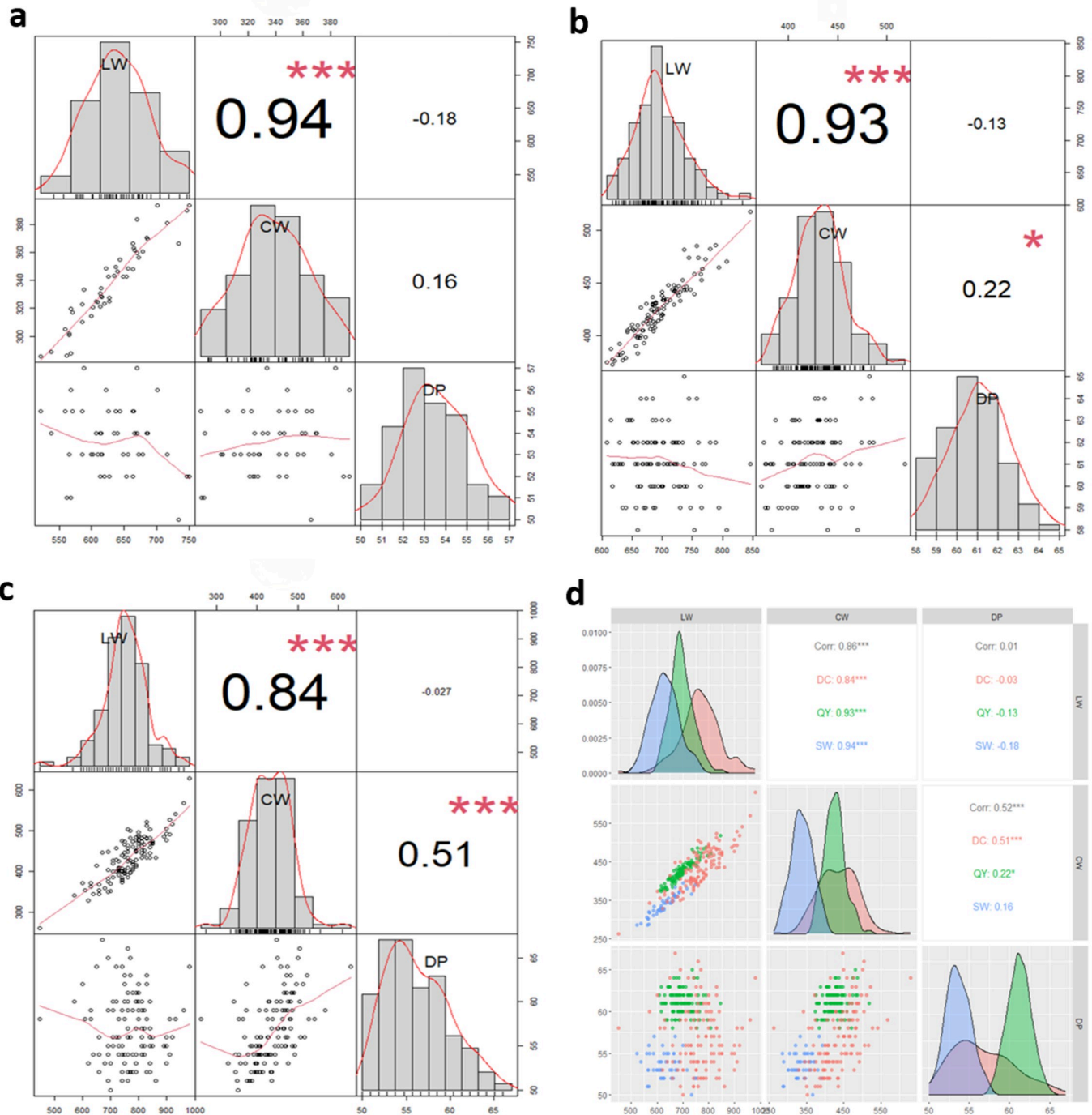


Figure 1

Correlation and distribution analysis of three carcass traits: live weight (LW), carcass weight (CW), and dressing percentage (DP). Pairwise phenotypic correlations were assessed using Pearson's correlation coefficients within three distinct cattle populations: Sunwu (SW) (a), Qingyang (QY) (b), and Dachang (DC) (c). Each panel displays the kernel density plots along the diagonal, scatterplots in the lower triangle, and correlation coefficients in the upper triangle. Significance levels are denoted as *P < 0.05, **P < 0.01, and ***P < 0.001 (a–c). Comparative distribution of LW, CW, and DP across the three

populations (SW in blue, QY in green, and DC in red). Diagonal panels show trait-wise kernel density estimates, while the lower triangle displays pairwise scatterplots colored by population. The upper triangle summarizes Pearson's correlation coefficients between traits within each population(d).

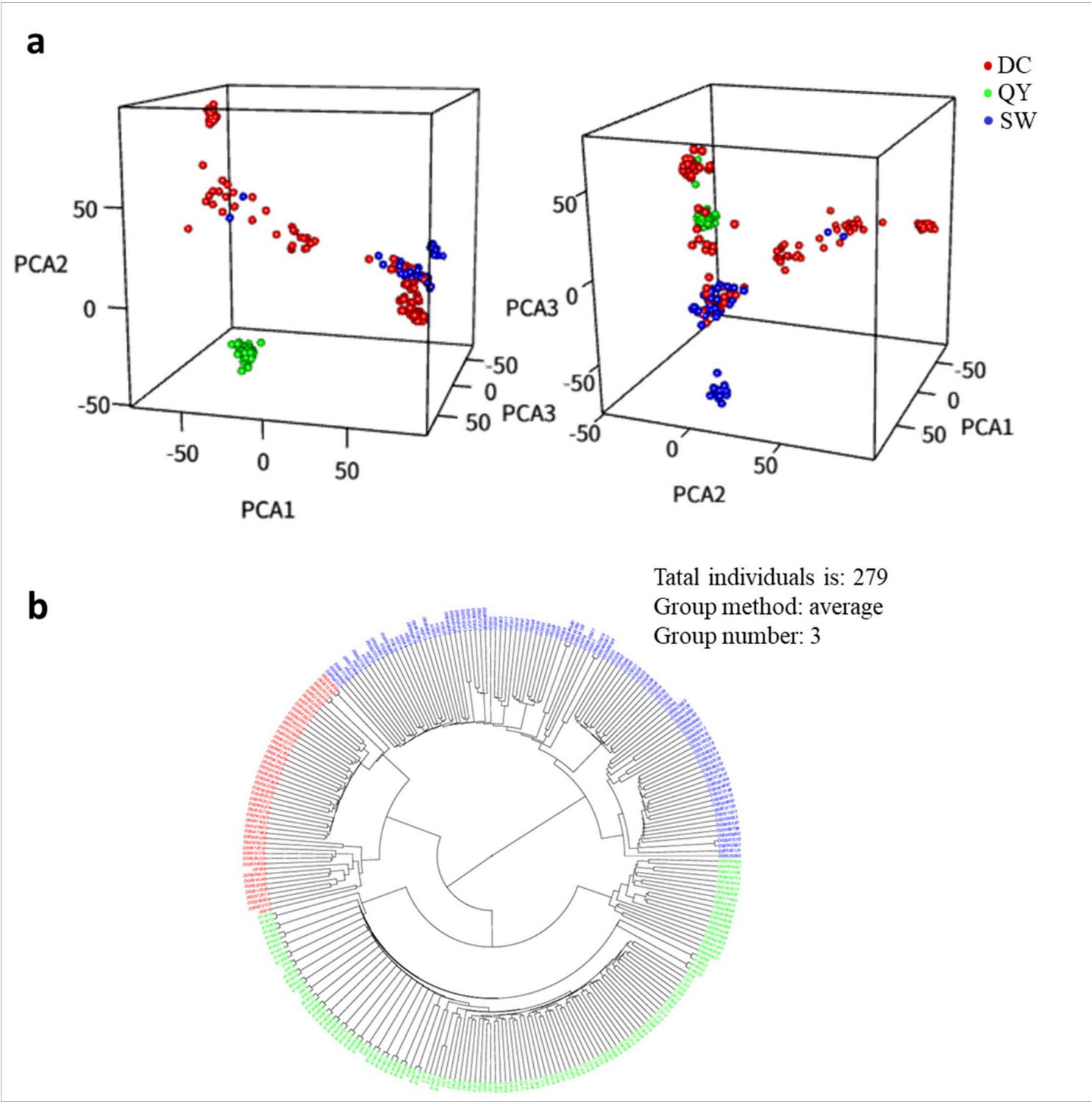


Figure 2

Genetic structure and phylogenetic relationships among individuals from Dachang (DC), Qingyang (QY), and Sunwu (SW). (a) PCA of genome-wide SNP genotypes, shown from two viewing angles derived from

a three-dimensional PCA space. Each point represents an individual, with DC in red, QY in green, and SW in blue. The PCA reveals a distinct clustering of QY individuals, while DC and SW show partially overlapping distributions, indicating a higher degree of genetic affinity. (b) Neighbour-joining (NJ) tree constructed using pairwise genetic distances. Tip colors represent clustering results assigned by GAPIT based solely on genetic relatedness, and therefore do not correspond directly to the color scheme used in panel (a). The tree topology recapitulates the PCA pattern, showing QY as an independent lineage, while DC and SW cluster on adjacent branches, suggesting that they may share parts of an ancestral genetic background.

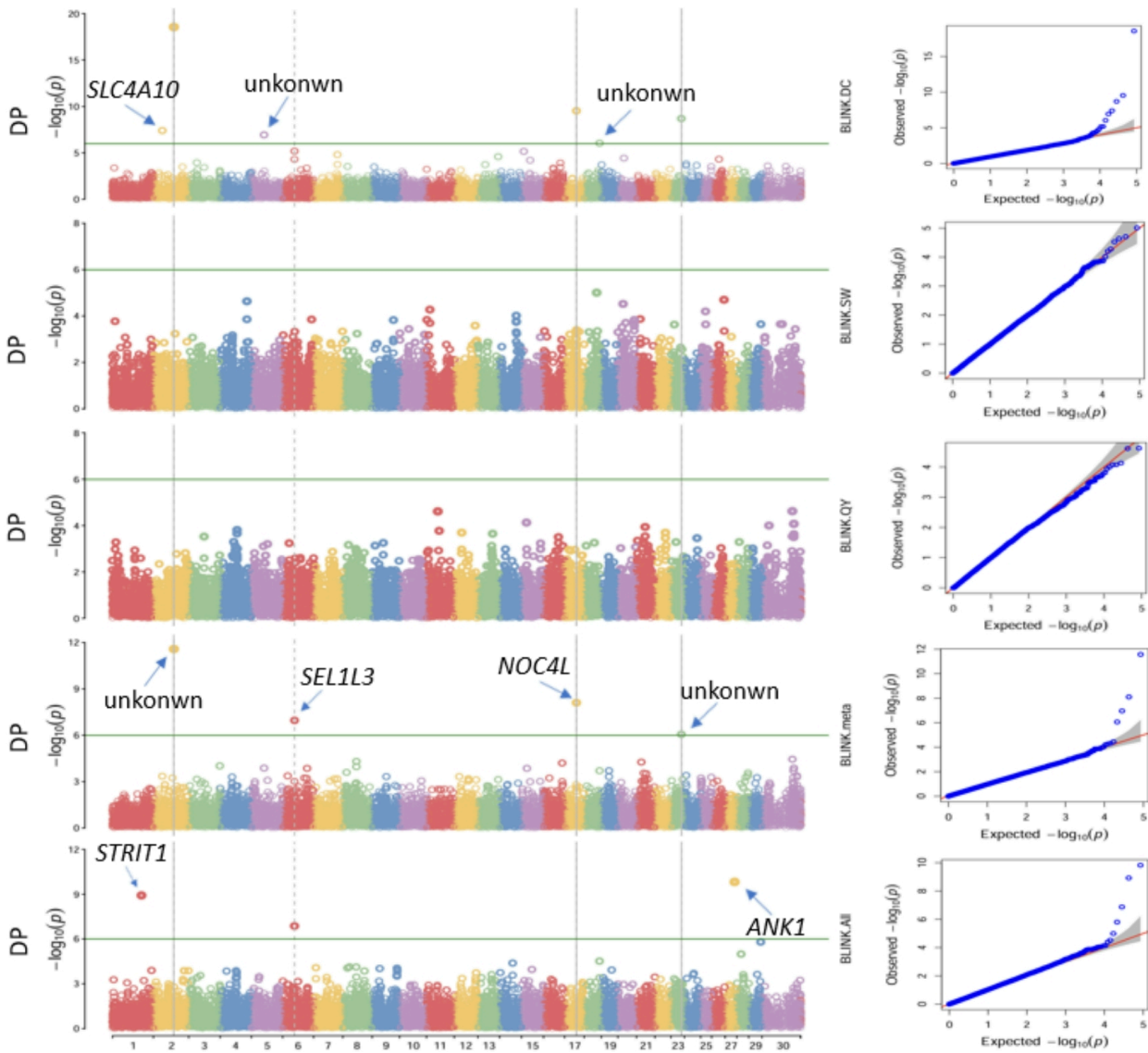


Figure 3

Manhattan and quantile–quantile (QQ) plots of the p-values for the genome-wide association study of the DP trait in cattle, performed separately for the Sunwu, Dachang, and Qingyang populations, for the combined population, and for the meta-analysis, based on the BLINK method—the horizontal line of significance threshold ($p\text{-value} < 1.01 \times 10^{-6}$) was used to distinguish significantly associated loci, and the different colors to distinguish different chromosomes.

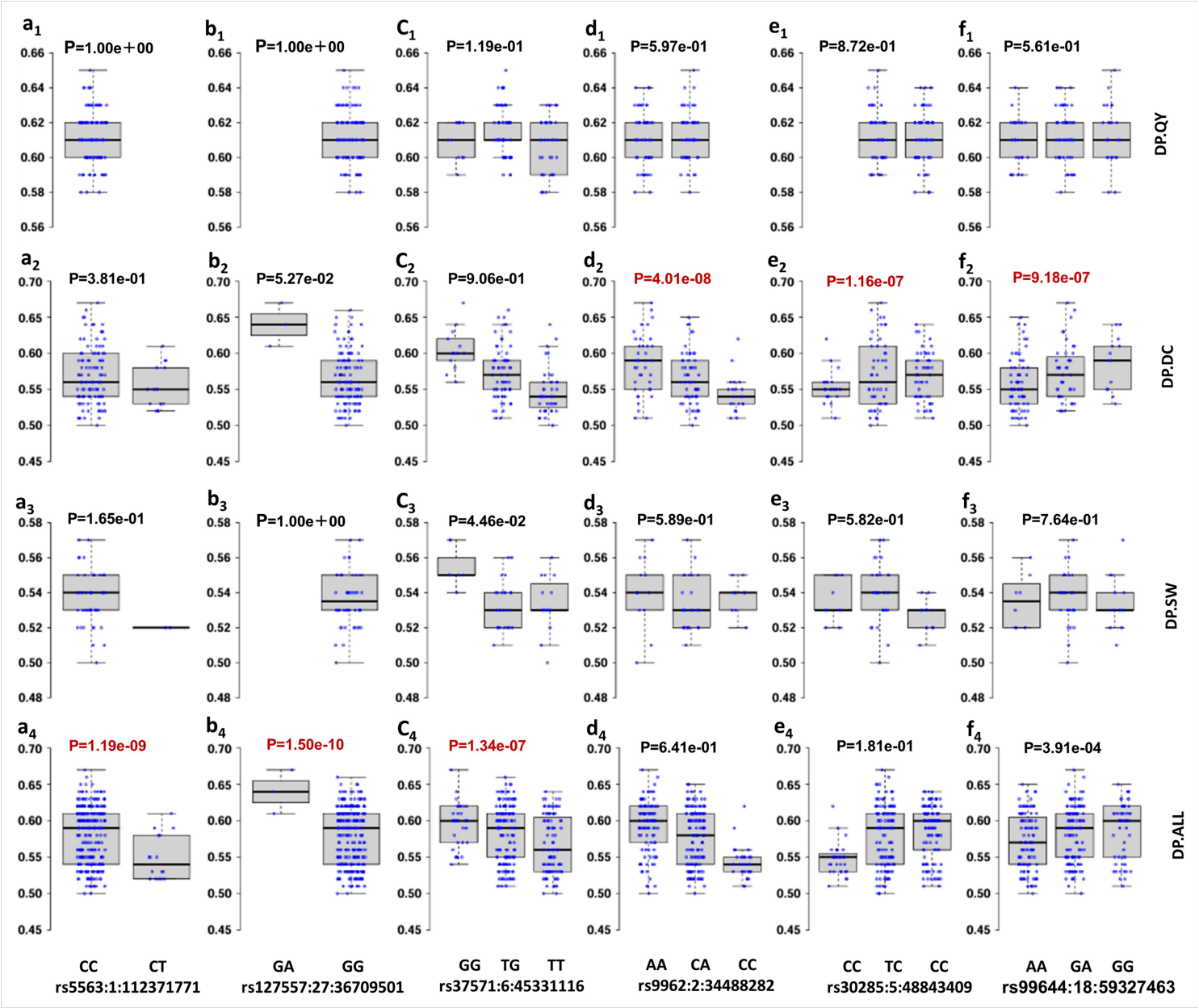


Figure 4

Boxplots showing the distribution of DP phenotypic values for different genotypes at each significant SNP. Each row represents a distinct population: Qingyang (QY), Dachang (DC), Sunwu (SW) and ALL (the combined dataset of all populations). Genotype classes are shown on the x-axis, and the observed DP values (%) are displayed on the y-axis. For each significant SNP, the p-values for that SNP in different populations are shown in red at the top of the panel.

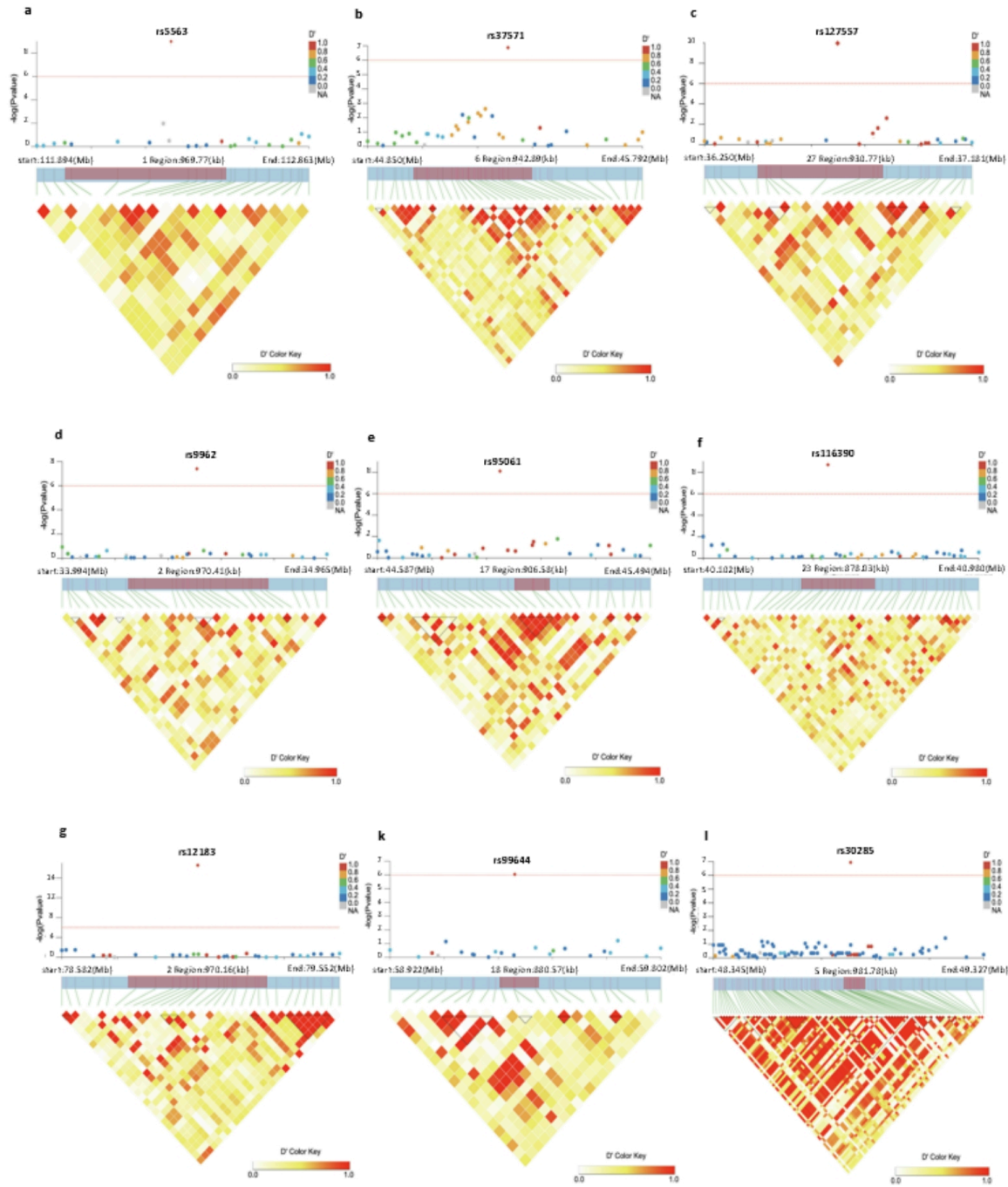


Figure 5

Local Manhattan plots (top), annotated gene regions (middle), and linkage disequilibrium (LD) heatmaps (bottom) for SNPs significantly associated with DP trait. Each panel shows the genomic region surrounding the lead SNP, with window sizes adjusted according to local SNP density and LD patterns. In the Manhattan plots, the red horizontal line represents the genome-wide significance threshold ($p < 1.01 \times 10^{-6}$). The middle panel displays the annotated gene structures within the targeted genomic interval;

the pink shaded blocks indicate the gene regions overlapping or adjacent to the lead SNP, highlighting potential candidate genes. The LD heatmaps illustrate pairwise D' values among SNPs in the region, where darker red indicates stronger LD.

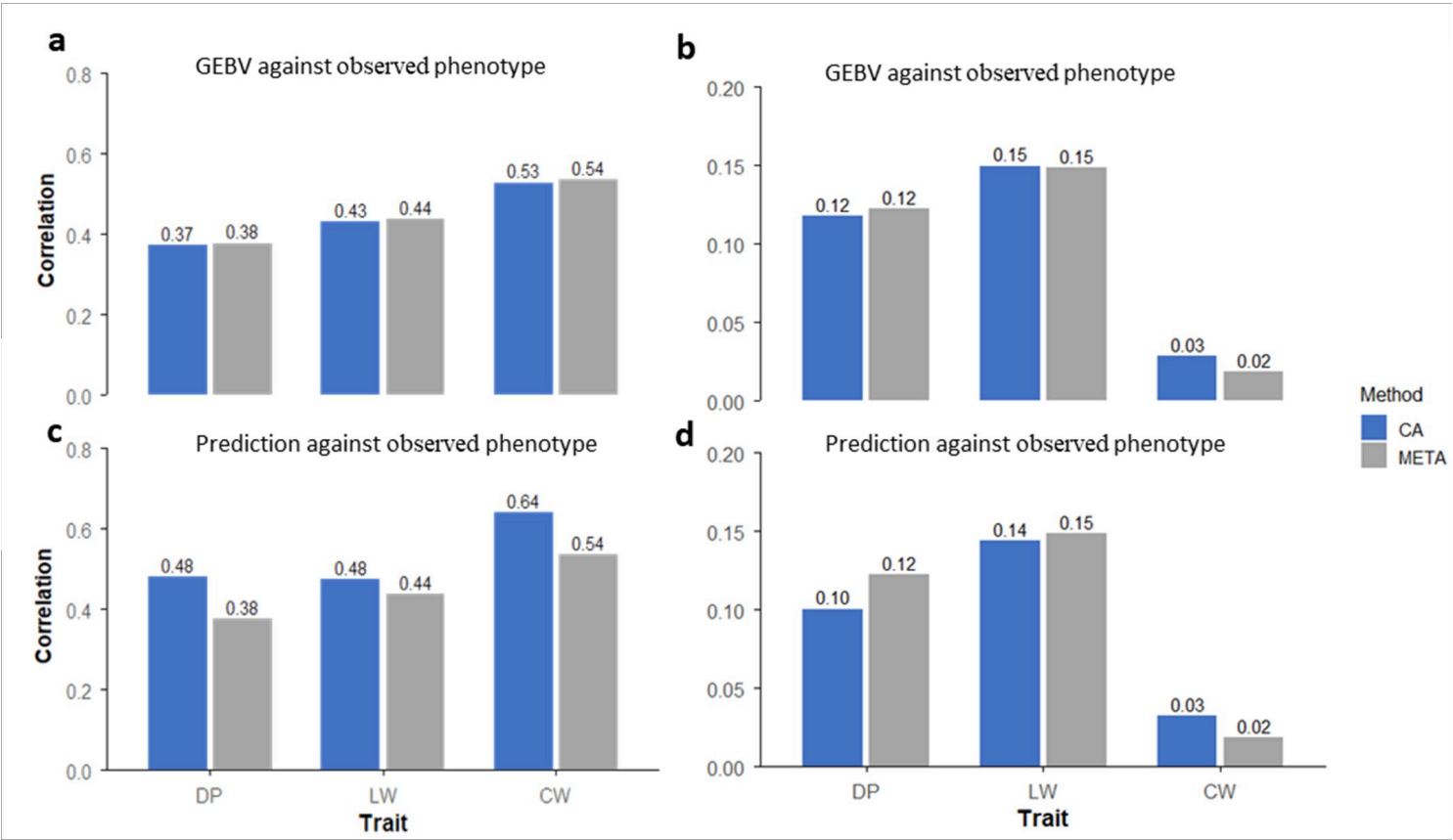


Figure 6

Comparison of genomic prediction performance between the CA and META models across three slaughter traits (DP, LW, CW) under two population-division strategies. Genomic selection (GS) was performed using a marker-assisted selection (MAS) framework. Panel (a/c) shows results based on the biologically structured (normal) population division, whereas panel (b/d) shows results obtained from a random population division. GEBV against observed phenotype, representing the correlation between genomic estimated breeding values (gEBVs) and observed phenotypes; Prediction against observed phenotype, representing the cross-validated correlation between predicted and observed phenotypes; Blue and grey bars represent CA and META, with correlation values labeled above each bar.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFiguretable.docx](#)