Identification of a Genome-Wide Significant SNP on Chromosome 5 Associated with Adult Body Weight in Nanyang Cattle via Whole-Genome Resequencing and GWAS

Pengpeng Zhang[1], Mengru Sheng[1], Xun Zhou[1], Dongtak Jeong[2*]
[1] College of Life Sciences, Xinyang Normal University, Xinyang, Henan, China
[2]Department of Medicinal & Life Science, College of Science and Convergence Technology, Hanyang University—ERICA, Ansan, Republic of Korea
**Corresponding Author*

Abstract: Adult body weight is an economically decisive trait in beef cattle production, directly governing carcass yield and the overall profitability of the production chain. Chinese indigenous cattle breeds, although well adapted to local environments, are characterized by relatively small body frames and consequently lower meat yields, which has contributed to a persistent national shortfall in beef supply. Marker-assisted selection offers a route to accelerate the genetic improvement of body weight in these populations, but its application depends on the discovery of reliable, breed-appropriate molecular markers. In this study, we performed whole-genome resequencing on a population of 285 healthy Nanyang cattle, generating approximately 30 Gb for each cattle. Then a genome-wide association study was conducted using a mixed linear model to dissect the genetic basis of adult body weight. After stringent quality control of sequencing reads, alignment to the ARS-UCD1.2 reference assembly, and variant filtering, association analysis identified a single nucleotide polymorphism located on chromosome 5 (Chr5:108852977) that reached the genome-wide significance threshold ($P = 4.67 \times 10^{-8}$). This locus carries a T/C allelic substitution with an estimated additive effect of 9.23 ± 1.36, with the T allele associated with significantly heavier adult body weight. A 189-bp genomic fragment flanking the variant (with the SNP at position 137) was characterized, a dedicated PCR primer pair was designed, and the genotype was validated by Sanger sequencing. The identified marker provides a novel, breed-specific molecular tool for the early selection of superior heavy-weight Nanyang cattle and offers a simple, rapid, and specific alternative to conventional genotyping approaches such as PCR-RFLP.

Keywords: Nanyang Cattle; Adult Body Weight; Genome-Wide Association Study; Single Nucleotide Polymorphism; Whole-Genome Resequencing.

1. Introduction

Beef consumption in China has climbed steadily with rising living standards, and it now takes up a growing share of the country's total meat intake. Yet domestic production has consistently fallen short of demand, making large-scale imports a yearly necessity. Much of this shortfall stems from the production characteristics of China's native cattle. These breeds were historically selected for draft power and farm work, not for meat, so they tend to have smaller frames and lower per-animal yields [1]. That is why increasing adult body weight through genetic selection has become a strategic goal—not just for the beef industry, but also for breeding programmes.

Adult body weight is a quantitative trait. It is shaped by many genes, each with a tiny effect, and it is also heavily influenced by feeding, housing, and other management practices. Conventional pedigree-based and quantitative genetic methods struggle to precisely localize the major effector genes underlying such complex traits. The advent of high-throughput single nucleotide polymorphism (SNP) genotyping has transformed the study of complex traits by enabling genome-wide association studies (GWAS). GWAS tests for statistical correlations between genetic variants and phenotypes across the entire genome and, relative to traditional linkage analysis, offers greater resolution and statistical efficiency in dissecting polygenic traits [2,3]. Markers

identified by GWAS can be incorporated directly into selection programmes, providing a rapid, accurate, and environment-independent means of accelerating genetic improvement. The general utility of GWAS for understanding the biology and architecture of complex traits has been extensively documented [4], and its specific value for livestock breeding has been reviewed in detail [5].

Two principal strategies are used to obtain the high-density genotype data that GWAS requires: SNP genotyping arrays and whole-genome resequencing. Most GWAS in cattle to date have relied on commercial SNP arrays, predominantly those developed by Illumina. While arrays are cost-effective, they suffer from intrinsic limitations: the panel of interrogated SNPs is fixed, so genuinely novel variants cannot be discovered, and the markers were largely ascertained in Western Bos taurus breeds. As a consequence, array-based panels often achieve poor SNP coverage and ascertainment in Chinese indigenous breeds, biasing or weakening association signals in these populations. Whole-genome resequencing (WGS) circumvents these constraints. By sequencing the entire genome of individuals or populations from a species with an existing reference assembly, WGS can detect novel SNPs as well as insertions/deletions, structural variants, and copy-number variants, thereby supplying a far richer and less biased set of polymorphic markers for genetic studies and molecular breeding.

The Nanyang cattle (Bos taurus), one of China's five major indigenous breeds, is prized for its meat quality and environmental adaptability but, like other native breeds, would benefit from improvement in growth and body-weight traits. The objective of the present study was to exploit whole-genome resequencing to characterize genome-wide SNP variation in Nanyang cattle—thereby overcoming the low coverage of foreign SNP arrays in this breed—and to apply GWAS to identify molecular markers associated with adult body weight. The ultimate aim is to provide validated, breed-specific markers that can assist the effective and early selection of superior heavy-weight Nanyang cattle and accelerate the genetic improvement of Chinese indigenous beef cattle.

2. Materials and Methods

2.1 Experimental Population and Phenotype Recording

The experimental population comprised 285 healthy Nanyang cattle maintained at the Nanyang Cattle Original Breeding Farm in Nanyang City, Henan Province, China. Adult body weight was recorded for each animal and used as the phenotype for association analysis. Venous blood was collected from each animal into anticoagulant tubes and thoroughly mixed prior to DNA extraction.

2.2 Genomic DNA Extraction

Genomic DNA was isolated from whole blood specimens. Initially, 1 mL of blood was combined with an equal volume of PBS buffer. This mixture underwent gentle agitation for 15 min, followed by centrifugation at 3,500 g for 10 min at room temperature; subsequently, the supernatant was removed. This washing procedure was performed repeatedly until the supernatant appeared clear and the cell pellet became colorless. The resulting pellet was then resuspended in 1 mL of lysis buffer. After gentle mixing, the suspension was incubated in a 37 °C water bath for 1 h, followed by treatment with proteinase K and overnight incubation to ensure complete cell lysis and solution clarification. Once cooled to room temperature, the lysate was subjected to successive extractions: first with an equal volume of Tris-saturated phenol (mixed gently on ice for 20 min, then centrifuged at 12,000 g for 10 min), followed by a phenol/chloroform extraction (0.5 volume each), and finally a pure chloroform extraction. After each extraction, the upper aqueous phase was carefully transferred to a new sterile tube. DNA was precipitated by adding two volumes of pre-chilled absolute ethanol, followed by incubation at -20 °C for 30 min and subsequent recovery via centrifugation. The DNA pellet was rinsed twice using 1 mL of 70% ethanol, air-dried, and then resuspended with water, with storage at 4 °C to facilitate complete dissolution. Finally, the DNA concentration was quantified spectrophotometrically, and preserved at -80 °C.

2.3 Whole-Genome Resequencing and SNP Genotyping

We conducted paired-end whole-genome sequencing on the DNBSEQ-T7 platform, yielding approximately 30 Gb of raw data for each individual. Initial data processing involved quality assessment and trimming of raw FASTQ

files via fastp (v0.23.2) [6]. Subsequently, the clean reads were mapped to the bovine reference genome (ARS-UCD1.2) [7] using the BWA (v0.7.17) aligner [8], with only unique alignments being preserved for downstream analysis. Variant discovery was executed using the GATK (v4.3.0.0) pipeline, strictly adhering to the established best-practice guidelines [9,10]. To ensure high-quality variant calls, we implemented a hard-filtering strategy based on the parameters: QD less than 2, FS more thatn 60.0, and MQ less than40.0. Following this, further refinement was performed using VCFtools to filter out variants based on specific population and quality metrics: a minor allele frequency was set as equal or more than 0.05, a maximum missingness rate of 0.7, a read depth ranging from 4 to 1000, a minimum quality score of 30, and a genotype quality score of at least 80. Only biallelic sites were retained for subsequent association studies.

2.4 Genome-Wide Association Analysis

To investigate the potential associations between SNP genotypes and adult body weight, we employed a mixed linear model (MLM) framework, which was executed using the GEMMA software package in a Linux environment. The statistical model utilized is defined by the following equation（1）:

$$y=W\alpha+x\beta+u+\varepsilon \quad (1)$$

In this formulation, y represents the vector of phenotypic measurements for the n individuals; The matrix W accounts for fixed effects, incorporating a column of ones alongside the corresponding vector of coefficients α . The variable x denotes the genotype vector for the SNPs of interest, β represents the estimated effect size for each SNP; Additionally, u and ε denote the vectors for random effects and residual errors, The random effects are assumed to follow a multivariate normal distribution, SNPs surpassing the genome-wide significance threshold were considered to be significantly associated with adult body weight.

2.5 Characterization of the Associated Region And Primer Design

The flanking nucleotide sequence surrounding the genome-wide significant locus was extracted from the reference genome using SAMtools (v1.16.1), yielding a 189-bp fragment in which the candidate variant is located at position 137. A primer pair was designed against this fragment for targeted amplification and validation: forward primer F:5′-GACACAGAAGAAAGGACAAAG-3′; reverse primer R: 5′-CATGAAAGTGAAAAGTGAAAG-3′.

2.6 PCR Amplification, Purification, and Genotype Validation

Targeted genotyping results were verified using PCR amplification followed by Sanger sequencing. The 20 μL PCR reaction mixture contained 10 μL of PCR Mix, 0.5 μmol/L of both forward and reverse primers, 50 ng of genomic DNA template, and nuclease-free water. The thermal cycling parameters were set as follows: an initial denaturation at 94 °C for 4 min; 35 cycles of 20 s denaturation at 94 °C, 20 s annealing at 58 °C, and 20 s extension at 72 °C. The specificity of the amplification was evaluated via agarose gel electrophoresis. After purification with a commercial kit according to the manufacturer's instructions, the products were subjected to Sanger sequencing using the reverse primer. Finally, the resulting chromatograms were inspected via SnapGene (v6.02) to assess the genotype at position 137, with heterozygosity characterized by the presence of double peaks at the variant locus.

3. Results

3.1 Genome-Wide Significant Association on Chromosome 5

Following the analytical pipeline outlined in Figure 1, whole-genome resequencing of the 285 Nanyang cattle, followed by alignment and stringent variant filtering, produced a high-quality genome-wide SNP set suitable for association analysis. Applying the GEMMA mixed linear model to adult body weight identified a single SNP that reached genome-wide significance, as visualized in the Manhattan plot (Figure 2). The associated variant is located on bovine chromosome 5 at genomic coordinate Chr5:108852977 and carries a T/C allelic substitution. The association statistics for this locus are summarized in Table 1.

The pipeline proceeds from venous blood sampling and genomic DNA extraction, through paired-end resequencing variant calling and filtering (GATK/VCFtools), and adult body-weight phenotyping, to mixed-linear-model association analysis in GEMMA and selection of the significant marker.

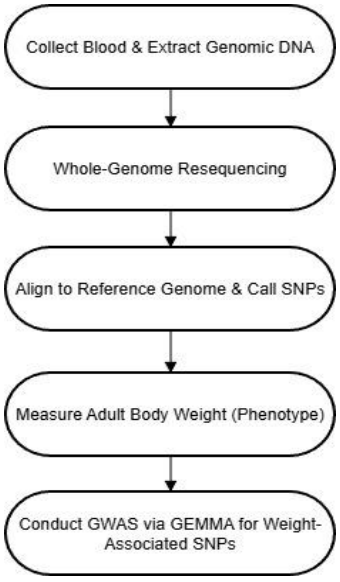

Figure 1. Workflow for Identifying the SNP Associated with Adult Body Weight in Nanyang Cattle by GWAS.

The horizontal axis represents the genomic position of SNPs across chromosomes and the vertical axis represents $-\log_{10}(P)$. The marker indicated by the black circle and arrow.

The P-value of 4.67×10^{-8} is well below the conventional genome-wide significance threshold, indicating a robust statistical association rather than a chance finding. Our estimate for the T allele was positive and sizeable (9.23 ± 1.36), meaning that animals carrying this allele were significantly heavier than those with the C allele. We therefore consider this marker a strong candidate for breeding programmes targeting adult body weight in Nanyang cattle

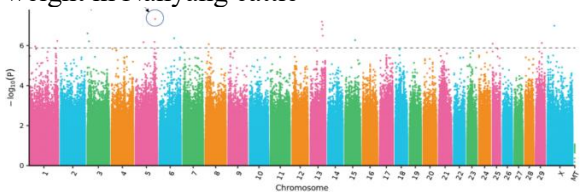

Figure 2. Manhattan Plot of the Genome-Wide Association Analysis for Adult Body Weight in Nanyang Cattle..

Table 1. Genome-wide Significant SNP Associated with Adult Body Weight in Nanyang Cattle.

Chromosome	Position	Reference allele	Alternative	P-val	Effect (± SE)
5	108852977	T	C	4.67×10^{-8}	9.23 ± 1.36

3.2 PCR Amplification and Genotype Validation

The genomic context around Chr5:108852977 was extracted, yielding a 189-bp fragment that contains the T/C polymorphism at position 137. This sequence was then used to design a locus-specific primer pair for downstream genotyping. PCR amplification with the designed primers produced a single, clean band of the expected length on an agarose gel (Figure 3A), confirming primer specificity. The genotype at position 137 was reliably resolved by Sanger sequencing of the purified amplicon using the reverse primer. As shown in the sequencing chromatogram (Figure 3B), heterozygous (T/C) individuals could be unambiguously identified by the overlapping T and C peaks at the variant position, demonstrating that all three possible genotypes are accurately distinguished by the assay. Taken together, these results confirm that the marker can be genotyped in a straightforward in-vitro sequencing format.

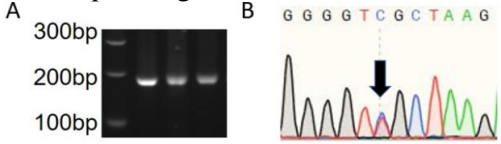

Figure 3. PCR Amplification and Sequencing Validation of the SNP Position.

(A) PCR products were analyzed by agarose gel electrophoresis. A 100–300 bp DNA ladder was loaded in the leftmost lane, while the remaining three lanes exhibited clear, specific amplification bands of the expected size for the target fragment.

(B) A representative Sanger sequencing chromatogram is shown. The target polymorphic site, indicated by a black arrow, displayed a heterozygous T/C double peak, confirming the presence of both alleles at this marker.

4. Discussion

A resequencing based GWAS was conducted in this study, through which a genome wide significant SNP on chromosome 5 (Chr5:108852977) was identified in association with adult body weight in Nanyang cattle. The association strength ($P = 4.67 \times 10^{-8}$), together with the sizeable additive effect estimated for the T allele, suggests that this locus marks a genomic region of biological importance for growth and body size in this native breed.

A key methodological feature of this work is the use of whole-genome resequencing rather than a fixed SNP array. Commercial bovine arrays were largely designed on the basis of variation segregating in Western Bos taurus breeds, and their marker density and ascertainment are

frequently suboptimal in Chinese indigenous populations. By sequencing each animal to an average depth of roughly 30 Gb and calling variants directly against the ARS-UCD1.2 assembly, we were able to interrogate breed-specific variation that array platforms cannot capture, thereby reducing ascertainment bias and improving the resolution of the association scan [7]. The stringent, multi-layered filtering applied at both the variant and site levels—including hard filters on quality-by-depth, strand bias, and mapping quality, together with thresholds on allele frequency, missingness, depth, and genotype quality—was important for limiting false-positive associations that can arise from low-quality genotype calls.

The choice of a mixed linear model implemented in GEMMA is well suited to this design. Body weight in cattle is highly polygenic, and populations such as a single breeding-farm herd typically contain substantial relatedness and possible cryptic population structure. By incorporating a genomic kinship matrix as a random effect, the MLM accounts for these confounders and controls the inflation of test statistics that would otherwise produce spurious associations. That a single locus nonetheless surpassed genome-wide significance is consistent with the expectation that, even for highly polygenic traits, individual variants of comparatively larger effect can be detected when genotyping is dense and phenotyping is accurate.

According to previous studies, a large number of loci had been identified in bovine Chromosome 5 which significantly associated with body size [8]. Further, Whole-genome-sequence-based analyses also confirmed new candidate genes in nearby regions [9,10].

It should be noted that, the current study only pinpointed molecular markers and did not conduct functional experiments to confirm the true causal genes. To distinguish tag SNPs from true causal variants and unravel the molecular mechanisms underlying weight regulation, additional experimental work is required. Follow-up studies should include fine-mapping of this locus, transcriptomic profiling of nearby genes, and comprehensive functional assays for all genes adjacent to the Chr5:1088529777 site. These complementary approaches will allow us to move past marker identification and establish how the underlying variant modulates animal body weight. In the future, functional validation via transgenic or gene-edited animal models that carry the SNP, followed by phenotypic measurement of carcass performance, can supply straightforward proof of the SNP functional importance.

This SNP marker holds clear practical value for livestock breeding. Carriers of the T allele exhibit greater adult body weight, meaning breeders can genotype young candidate animals at this locus. Genomic pre-selection based on genotypes like this is the core advantage of marker-assisted selection, as it effectively shortens the generation interval while simultaneously improving overall selection precision.

To facilitate routine detection in breeding programs, we designed a simple and robust genotyping protocol. A single pair of primers amplifies the 189 bp target fragment, and genotypes are directly resolved via Sanger sequencing, with heterozygous individuals easily identifiable by distinct dual sequencing peaks. Unlike traditional PCR-RFLP methods, which require a unique restriction enzyme cutting site and complete enzymatic digestion to distinguish variants, direct sequencing avoids such constraints. This sequencing approach delivers superior performance in terms of simplicity, speed, sensitivity and specificity. The protocol only requires standard PCR and Sanger sequencing equipment, making it readily implementable within routine livestock breeding laboratories.

Despite these limitations, our work yields a concrete outcome: we have developed a novel breed-specific molecular marker for adult body weight in Nanyang cattle, together with a validated, ready-to-use genotyping method. This combination directly tackles the persistent issue that foreign SNP arrays perform poorly in Chinese indigenous breeds, and it provides breeders with a practical tool to expedite genetic gains in body weigh.

5. Conclusion

Through whole-genome resequencing of 285 Nanyang cattle and a GEMMA mixed-linear-model GWAS, we identified a genome-wide significant SNP on chromosome 5 (Chr5:108852977; T/C; $P = 4.67 \times 10^{-8}$) associated with adult body weight, in which the T allele confers significantly heavier weight (effect 9.23 ± 1.36). We characterized a 189-bp marker region, designed a locus-specific primer pair, and validated genotyping by PCR and

Sanger sequencing. This marker provides a simple, rapid, sensitive, and specific tool for the early molecular selection of superior heavy-weight Nanyang cattle and constitutes a new molecular resource for indigenous beef-cattle breeding. Future work should focus on independent validation across cohorts and breeds and on functional dissection of the underlying causal gene.

Acknowledgments

This work was supported in part by the Key Scientific Research Project of Higher Education Institutions in Henan Province (24A230016).

References

[1] Zhang Y, Liu X, Liu J, et al. Genomic Insights into Genetic Diversity and Adaptation of Nanyang Cattle: Implications for Conservation and Breeding. Animals, 2025, 15(20):3033.

[2] Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. The American Journal of Human Genetics, 2017, 101(1):5-22.

[3] Ogunbawo AR, Hidalgo J, Mulim HAA, et al. Applying the Algorithm for Proven and Young in GWAS Reveals High Polygenicity for Key Traits in Nellore Cattle. Frontiers in Genetics, 2025, 16:1549284.

[4] Bai F, Cai Y, Qiu M, et al. LCORL and STC2 Variants Increase Body Size and Growth Rate in Cattle and Other Animals. Genomics, Proteomics & Bioinformatics, 2025, 23(3):qzaf025.

[5] Sharma A, Lee JS, Dang CG, et al. Stories and challenges of genome-wide association studies in livestock—a review. Asian-Australasian Journal of Animal Sciences, 2015, 28(10):1371-1379.

[6] Chen S, Zhou Y, Chen Y, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 2018, 34(17):i884-i890.

[7] Mafolo KS, MacNeil MD, Neser FWC, et al. Preliminary Evaluation of Blending, Tuning, and Scaling Parameters in ssGBLUP for Genomic Prediction Accuracy in South African Holstein Cattle. Animals, 2025, 15(19):2866.

[8] Bouwman AC, Daetwyler HD, Chamberlain AJ, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nature Genetics, 2018, 50(3):362-367.

[9] Purfield DC, Evans RD, Berry DP. Reaffirmation of known major genes and identification of novel candidate genes associated with carcass-related metrics based on whole-genome sequence within a large multi-breed cattle population. BMC Genomics, 2019, 20(1):720.

[10] An B, Xu L, Xia J, et al. Multiple association analysis of loci and candidate genes that regulate body size at three growth stages in Simmental beef cattle. BMC Genetic, 2020, 21(1): 32.

