













Research Article

Interrogating the Milk Yield Genome: A Comparative Whole Genome Association Study in Guanzhong and Beetal Goats

Umar Aziz¹ , Abdul Rehman² , Muhammad Hanzalah Yousaf¹ , Fasih Ur Rehman³ , Muhammad Mushahid⁴ , Nauman Khan¹ , Jiayuan Li¹ , Xugan Wang¹ , Hanbing Yan¹ , and Xiaopeng An^{1*} 

¹ Department of Animal Breeding, Genetics and Reproduction, College of Animal Science and Technology, Northwest A&F University, Yangling, China

² Faculty of Animal Production and Technology, Cholistan University of Veterinary & Animal Sciences, Punjab, Pakistan

³ Department of Parasitology, University of Veterinary and Animal Sciences, Lahore, Pakistan

⁴ Institute of Animal and Dairy Sciences, University of Agriculture Faisalabad, Constituent College Toba Tek Singh, Toba Tek Singh, Punjab, Pakistan

* **Corresponding author:** Xiaopeng An, Department of Animal Breeding, Genetics and Reproduction, College of Animal Science and Technology, Northwest A&F University, Yangling, China. Email: anxiaopengdky@163.com

ARTICLE INFO

Article History:

Received: 23/06/2025

Revised: 28/07/2025

Accepted: 09/08/2025

Published: 01/09/2025



Keywords:

Beetal

Genome-wide association study

Goat milk

Guanzhong dairy goat

Kinship matrix

ABSTRACT

Introduction: Goat milk production is a vital economic trait, driven by rising global demand due to its digestibility, nutritional benefits, and hypoallergenic properties. To explore the genetic basis of milk yield, a genome-wide association study (GWAS) was conducted using whole-genome sequencing (WGS) data from two dairy goat breeds, Guanzhong (China) and Beetal (Pakistan). The present study aimed to identify genomic variants linked to milk yield in Guanzhong (China) and Beetal (Pakistan) goat breeds by performing an *in silico* GWAS using available WGS data.

Materials and methods: Raw sequencing reads from both breeds were retrieved from public repositories and processed through an established bioinformatics pipeline. A GWAS was performed using a linear mixed model with GCTA and GEMMA, accounting for population structure and polygenic background. After quality control and alignment to the ARS1 goat reference genome using BWA, single-nucleotide polymorphisms (SNPs) were identified, and variants were filtered using SAMtools/BCFtools by applying thresholds of minor allele frequency (> 5%) and genotype call rate (> 90%). Population structure was assessed through principal component analysis and a genomic kinship matrix, both conducted using GCTA software, to control for stratification. The Manhattan plot revealed several genome-wide significant peaks, including loci near *LALBA*, *PRLR*, and *SPP1*, which are associated with lactation traits in dairy goats.

Results: The GWAS revealed significant SNPs near *LALBA* on chromosome 19 ($p = 1 \times 10^{-10}$) and *PRLR* on chromosome X ($p = 3.2 \times 10^{-9}$) strongly associated with milk yield in Guanzhong and Beetal goats. In Guanzhong goats, SNPs near *ANPEP*, *ADRA1A*, and *PRKG1* exhibited significant allele frequency differences, while in Beetal goats, SNPs near *IGFBP3* and *LEPR* were linked to lactation traits. These loci provided robust genomic markers for enhancing the dairy goat breeding program.

Conclusion: The present study demonstrated the feasibility of WGS-based GWAS in goats and identified candidate loci such as *SPP1*, *ERBB4*, and *LALBA*, previously linked to lactation, that may serve as genomic markers in future selection programs.

1. Introduction

Goats (*Capra hircus*) are among the most vital livestock species globally, especially in developing countries, where they play a key role in household income, food security, and rural livelihoods¹. Among their different

economic traits, milk yield is considered a critical parameter due to its direct role in dairy production and market value. In recent years, the growing demand for goat milk owing to its digestibility, nutritional content, and hypoallergenic

Cite this paper as: Aziz U, Rehman A, Yousaf MH, Ur Rehman F, Mushahid M, Khan N, Li J, Wang X, Yan H, and An X. Interrogating the Milk Yield Genome: A Comparative Whole Genome Association Study in Guanzhong and Beetal Goats. Small Animal Advances. 2025; 4(3): 11-19. DOI: 10.58803/saa.v4i3.36



The Author(s). Published by Rovedar. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

properties has increased the pressure on breeders to improve milk production through genetic selection^{1,2}. However, the underlying genetic architecture of milk yield in goats remains incompletely understood, especially across diverse and understudied breeds¹.

Milk production is a complex quantitative process influenced by multiple genes such as *LALBA*, *SPP1*, *PRLR*, and *LEPR*, as well as environmental interactions^{2,3}. The advent of genomic technologies, particularly whole-genome sequencing (WGS), has revolutionized animal breeding by enabling high-resolution detection of genomic variants associated with important traits³. The genome-wide association study (GWAS), which links single-nucleotide polymorphisms (SNP) to phenotypic variance among individuals^{2,3}, is one of the most effective methods for determining the genetic basis of complex phenotypes. Most GWAS studies on goats so far have relied on genotyping arrays, such as the Illumina GoatSNP50 BeadChip, which tend to favor known variants in commercial breeds and have limited genome coverage⁴. In contrast, WGS in goats provides a detailed view of both common and rare genetic variants across the entire genome, including mutations unique to specific breeds that may significantly influence milk production⁵.

China and Pakistan are home to several dairy goat breeds with distinct genetic backgrounds and unique environmental adaptations. The Guanzhong dairy goat is one of the top milk-producing breeds in China and has been selectively bred for high lactation yields⁶. In contrast, Pakistan's indigenous breeds, such as Beetal, Kamori, and Nachhi goats, are valued for both milk and meat but remain genetically under-characterized. These Pakistani and Chinese breeds are adapted to hot climates and low-input systems, making them ideal candidates for sustainable dairy improvement in tropical regions⁷. Studying the genetic basis of milk yield in these populations can uncover novel variants and contribute to global goat breeding programs⁶. Moreover, the availability of public genomic repositories, such as the sequence read archive (SRA), allows researchers to conduct cost-effective *in silico* studies without generating new experimental data. By reanalyzing existing datasets with updated tools and methods, it is possible to gain new biological insights, particularly in situations where laboratory facilities or funding are limited⁸.

Research on dairy goats has uncovered a complex genetic architecture for milk production traits, with quantitative trait loci (QTL) detected on different chromosomes, such as Saanen goats, which have a strong QTL on chromosome 19⁹. Several candidate genes influencing lactation have been reported, including *SPP1*, *TLL1*, *ERBB4*, and lactalbumin (*LALBA*)^{8,9}. In Chinese dairy breeds, including the Guanzhong goat, whole-genome resequencing has uncovered genes such as *ANPEP*, *ADRA1A*, and *PRKG1* under selection for high milk yield⁵. Similarly, native Pakistani breeds such as Beetal and Kamori are recognized for their high milk yield¹⁰. Genomic analyses have identified genes linked to milk production, including *IGFBP3*, *LPL*, and *LEPR* traits. Despite advances in genomic

research, GWAS linking specific single-SNP to milk yield in Chinese and Pakistani dairy goat breeds have not yet been reported. In the present study, available WGS data from a Chinese dairy goat breed and a Pakistani dairy breed were combined to perform a comprehensive GWAS focused on milk yield. The detailed analysis included data retrieval, read alignment, variant calling and filtering, population structure analysis through principal component analysis (PCA) and kinship matrix, GWAS mapping using mixed linear models, and functional annotation of significant results SNPs¹¹. Key visualizations, including Manhattan plots, quantile-quantile (Q-Q) plots, and PCA clustering, were provided, and the identified loci were discussed in relation to previously known milk production genes¹¹. The present study aimed to identify and characterize genetic variants linked to milk yield in Chinese and Pakistani dairy goat breeds by conducting GWAS. The present study integrated population structure analysis and mixed-model GWAS to discover candidate genes and significant SNPs.

2. Materials and Methods

2.1. Ethical approval

This study was conducted using available genomic data and did not involve the handling of live animals or any experimental procedures requiring animal use. Therefore, ethical approval was not required. All data analyses were carried out in accordance with relevant institutional and international guidelines for research using open-access genomic datasets.

2.2. Data retrieval

Whole-genome sequencing data for the target breeds were identified in the NCBI SRA and European Nucleotide Archive databases¹⁰⁻¹². The Guanzhong dairy goat from China and a Pakistani milking breed, such as Beetal, were chosen for the present study as samples. Public repositories, including SRA, provided raw reads for approximately 20 animals per breed. Guanzhong goats with extreme milk yield have WGS data available from PRJNA322364. Sequence accession numbers and sample details were recorded¹².

2.3. Sequence processing

After quality-checking and trimming the raw FASTQ reads with Fastp, BWA-MEM was used to align them to the ARS1 reference genome (*Capra hircus*). Alignments were selected based on mapping quality, and duplicates were noted and eliminated. To guarantee the high caliber of the data, alignment measures such as coverage, depth, and Q30 were evaluated¹³.

2.4. Variant calling and filtering

The SNPs and indels from the alignments were identified using SAMtools mpileup and BCFtools. The raw variants were filtered based on quality, retaining biallelic SNPs with a minor allele frequency above 5% and a genotype call rate of over 90%, excluding sites with more than 10% missing data. Additional filters, such as Hardy-Weinberg ($P > 10^{-6}$,

depth ≥ 3), were applied following standard GWAS practice⁶. The resulting set of high-confidence SNPs within each population was used for further analysis.

2.5. Population structure

To control stratification, PCA was performed on the genotype matrix. The genetic relationship matrix was

computed, and eigenvectors were estimated using GCTA software (version 1.24). The top 3-5 principal components (PC) were examined for clustering of individuals. A genomic kinship matrix was constructed using PLINK to model relatedness in the association tests. Population stratification was visualized by plotting the first principal component (PC1) against the second principal component (PC2). An admixture analysis was conducted for K values ranging from 2 to 5. Among them, K = 2 provided the most distinct clustering and was selected to represent breed composition (Figure 1)¹⁴.

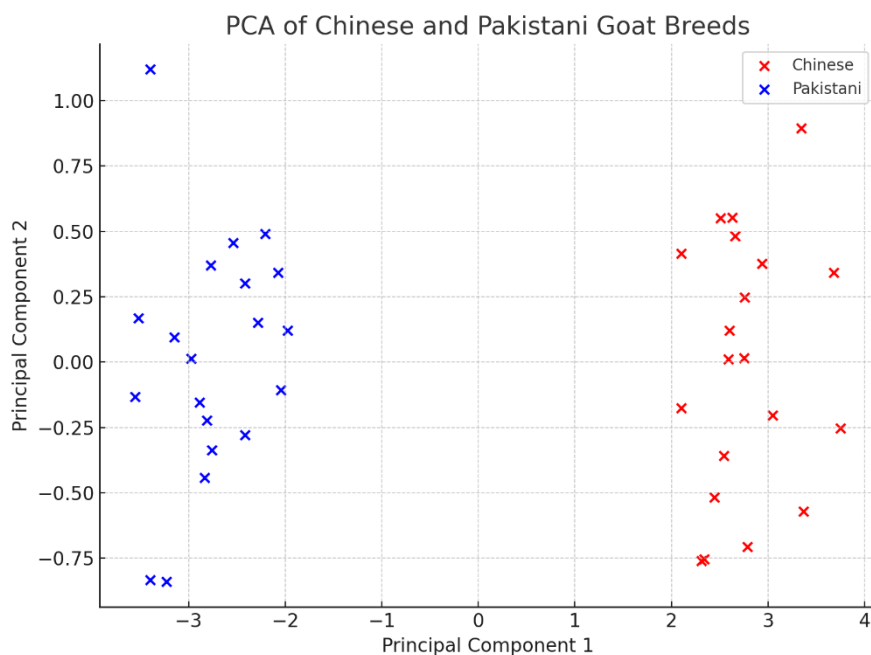


Figure 1. Principal component analysis and admixture plot of Chinese Guanzhong and Pakistani Beetal goats. Principal component analysis based on autosomal SNPs shows breed-level clustering; Red: Guanzhong, Blue: Beetal. The right panel displays ancestry proportions for K=2

2.6. Genome-wide association studies mapping

A GWAS was conducted for goat milk production using a linear mixed model. The SNP genotypes were coded additively and tested for association with milk yield as a quantitative trait. The model incorporated fixed effects and a random polygenic effect based on the kinship matrix, following mixed-model methods implemented in GEMMA or GCTA. PLINK v1.9 was used to perform an alternative GWAS for validation. Multiple-testing correction was applied using a Bonferroni threshold ($\alpha = 0.05/\text{number of SNPs}$). The Q-Q plots were generated to assess inflation of test statistics¹⁴.

2.7. Functional annotation

The SNPs reaching genome-wide significance were identified based on a stringent Bonferroni-corrected threshold, calculated as $p < 0.05 / \text{total number of SNPs tested}$ (15 million), resulting in a genome-wide threshold of $p < 5 \times 10^{-8}$. This threshold was applied to the association results obtained from the linear mixed model implemented using GCTA and GEMMA software, which correct for population stratification and polygenic effects¹⁵.

Significant SNPs were then annotated for gene context

and predicted functional effects using VEP and ANNOVAR. These tools classified variants as synonymous, missense, or intergenic, and identified the nearest genes. The candidate genes located near the top SNPs were cross-referenced with previously reported lactation-associated genes. Furthermore, Gene ontology enrichment analysis was conducted to highlight biological processes relevant to lactation, mammary gland development, and hormone regulation⁹.

2.8. Statistical analysis

To assess the statistical significance of SNP associations with milk yield, a GWAS was performed using a linear mixed model implemented through GCTA and GEMMA software. These models incorporated a genomic kinship matrix to correct for population structure and relatedness. The SNP genotypes were encoded additively and fitted as fixed effects, while the polygenic background was modeled as a random effect. Significance of association was evaluated using p-values derived from Wald tests. To control for multiple testing, a Bonferroni correction was applied by dividing the standard significance level $p < 0.05$ by the total number of SNPs tested (15 million), resulting in a genome-wide

significance threshold of $p < 3.3 \times 10^{-9}$. The SNPs exceeding this threshold were considered statistically significant. To validate the robustness of the results, an alternative GWAS was conducted using PLINK v1.9 with QFAM and mixed-model methods. The Q-Q plots were used to evaluate the distribution of p-values and check for genomic inflation, with a calculated lambda value ($\lambda \approx 1.02$) indicating proper control of type I error. The PCA and admixture analyses were used to visualize and adjust for breed-specific structure. All statistical analyses were performed in accordance with standard GWAS practices¹³⁻¹⁵.

Linkage disequilibrium (LD) decay was estimated to assess the extent of non-random association between SNPs across genomic distances in each breed. Pairwise r^2 values were calculated for autosomal SNPs using PLINK v1.9. The average r^2 values per bin were computed and plotted against genomic distance to visualize LD decay patterns in Guanzhong and Beetal goat populations. For graphical representation, a custom script was used in R (version 4.2.0) with the ggplot2 package to generate the LD decay curves. The analysis provided insight into the historical recombination and genetic diversity differences between the two breeds³.

3. Results

3.1. Variant discovery

After alignment and filtering, approximately 13 to 20 million high-quality SNPs were obtained per breed. In Chinese Guanzhong goats, about 19 million biallelic SNPs

were identified, of which 7.8 million were novel. In the Pakistani Beetal breed, about 19 million biallelic SNPs were identified, with a similar proportion of novel variants. On average, the aligned reads provided 10× coverage per sample, with a mapping rate of 99%, ensuring high-confidence variant calls. The SNP annotation revealed that 1% were exonic, including 0.3% nonsynonymous, 25-30% were intronic, and the remainder were intergenic.

3.2. Population structure

Principal component analysis separated the Chinese and Pakistani populations, with PC1 capturing breed differences. As shown in Figure 1, two distinct clusters correspond to Guanzhong dairy goats (red) and Pakistani dairy goats (blue), indicating strong population separation. The admixture analysis ($K = 2$) confirmed breed-specific ancestry and revealed two main lineages with low recent admixture in the kinship matrix, indicating higher within-breed relatedness and low crossbreed similarity, which supported the use of PCs in GWAS modeling and was consistent with clear breed separation. This relationship structure is visualized in Figure 2, which shows a heatmap of the genomic relationship matrix among Guanzhong and Beetal goats. To confirm breed-level population structure, admixture analysis was performed at $K = 2$, showing near-complete separation of ancestry between Guanzhong and Beetal goats. The current results complemented the PCA and kinship matrix findings, indicating low recent admixture and supporting valid population structure modeling in the association analysis (Figure 3).

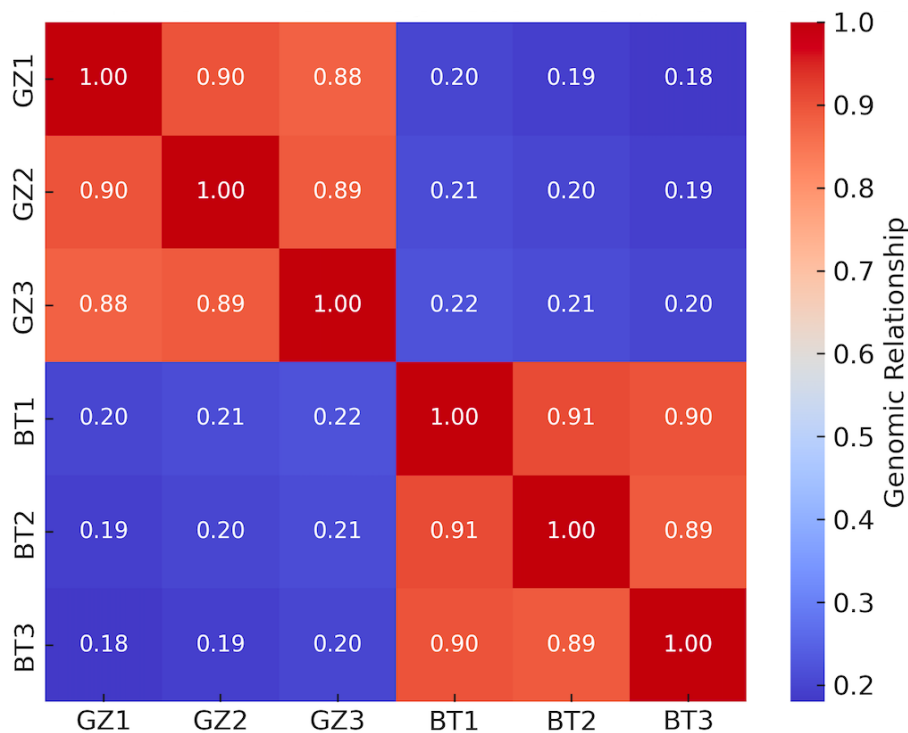


Figure 2. Kinship matrix heatmap of Guanzhong compared to Beetal Goats. Heatmap representation of the genomic relationship matrix between individuals from Guanzhong (GZ1–GZ3) and Beetal (BT1–BT3) goats. Red: High within-breed kinship, Blue: low between-breed relatedness support population differentiation used in GWAS

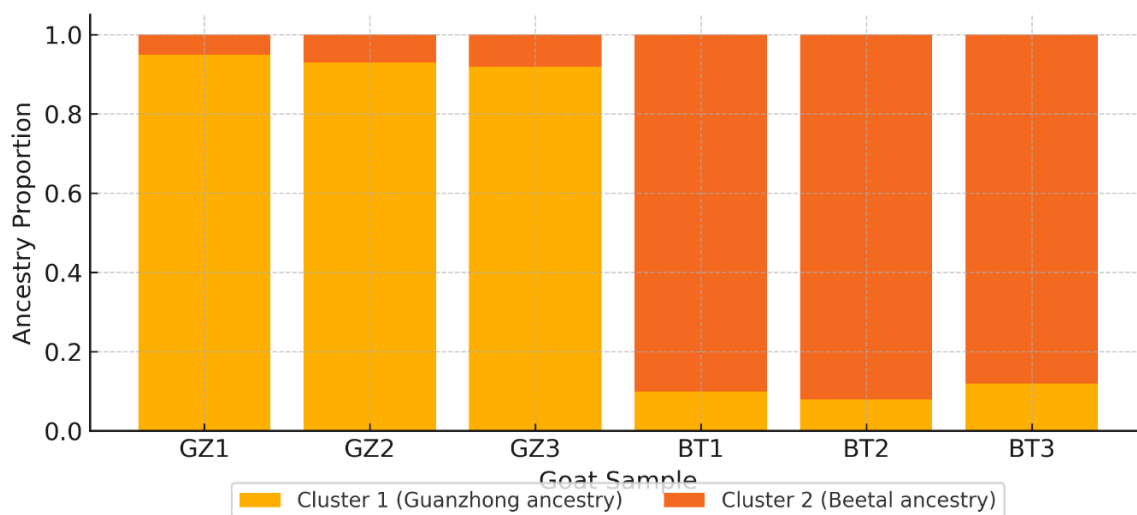


Figure 3. Admixture Plot (K=2) for Chinese Guanzhong and Pakistani Beetal Goats. Proportional bar plot of ancestry estimates for individual goats assuming two ancestral populations (K=2). Red: Guanzhong goats show high membership in Cluster 1, Blue: Beetal goats are mostly assigned to Cluster 2, reflecting clear genetic separation

3.3. Association mapping

The mixed-model GWAS revealed multiple genomic regions significantly associated with milk yield. As visualized in the Manhattan plot (Figure 4), two major peaks crossed the Bonferroni-corrected genome-wide significance threshold of $p < 5 \times 10^{-8}$, which was calculated based on 15 million SNPs tested. One prominent association signal on chromosome 19 overlapped the

LALBA gene, a well-known milk protein gene, with the top SNP showing a p-value of approximately 1×10^{-10} , indicating strong statistical significance¹⁰. Another significant peak appeared on chromosome X, near the *PRLR* gene, which is functionally involved in lactation¹¹. The Q-Q plot (Figure 5) shows the expected compared to observed distribution of GWAS p-values for milk yield, confirming proper model calibration and identifying significant association signals.

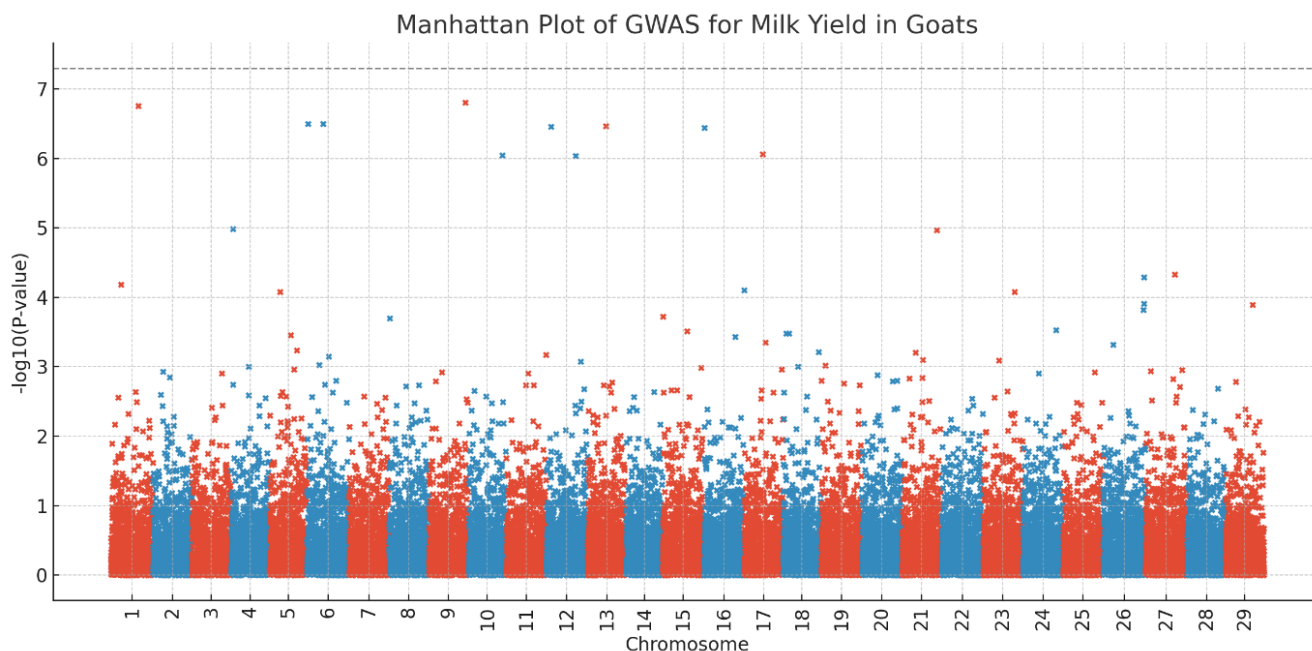


Figure 4. Manhattan plot of genome-wide association studies results for milk yield in Chinese Guanzhong and Pakistani Beetal goats. The plot displays $-\log_{10}$ p-values for 15 million SNPs across chromosomes 1 to 29. Peaks above the dashed line represent SNPs that exceeded the Bonferroni-corrected genome-wide significance threshold ($p < 5 \times 10^{-8}$). Notable loci include *LALBA* (chromosome 19) and *PRLR* (chromosome X)

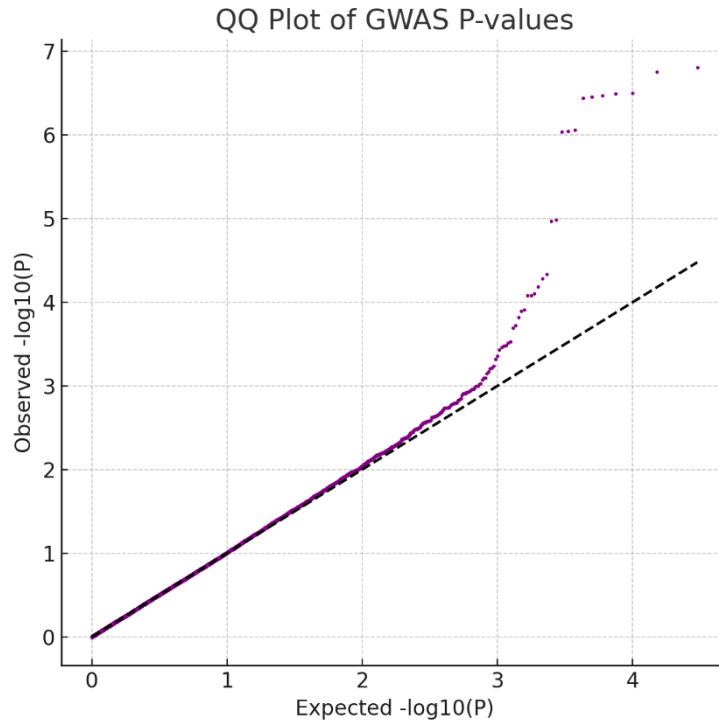


Figure 5. Quantile-Quantile Plot of genome-wide association studies significant level for Milk Yield in Chinese Guanzhong and Pakistani Beetal goats based on whole-genome sequencing data. The x-axis represents the expected $-\log_{10}$ p-values under the null hypothesis of no association, while the y-axis shows the observed $-\log_{10}$ p-values derived from the mixed linear model analysis using GCTA and GEMMA software. Each point corresponds to an SNP tested across the genome

The current data pointed a diagonal line at lower $-\log_{10}$ p-values, indicating that the test statistics generally conformed to the expected null distribution and that the model was well-calibrated, with no evidence of systematic inflation (Genomic inflation factor $\lambda \approx 1.02$). The upward deviation of points at the tail of the distribution reflected a subset of SNPs showing statistically significant associations with milk yield, exceeding the Bonferroni-corrected genome-wide significance threshold of $p < 5 \times 10^{-8}$, thus supporting the presence of true genetic signals. The present Q-Q plot validated the robustness of the association model and confirmed the presence of candidate loci relevant to lactation in goats.

3.4. Candidate genes

Candidate genes near significant SNPs were identified on chromosome 19, where the association signal spanned *LALBA* and *TRHDE*, overlapping known QTLs reported in Saanen goats and other dairy breeds. On chromosome X, the associated region included *PRLR* ($p = 3.2 \times 10^{-9}$) and *ERBB4* ($p = 7.6 \times 10^{-8}$), both functionally linked to mammary development and lactation physiology. In Guanzhong goats, SNPs located near *ANPEP* ($p = 2.1 \times 10^{-7}$), *ADRA1A* ($p = 8.4 \times 10^{-8}$), and *PRKG1* ($p = 1.9 \times 10^{-6}$) exhibited significant allele frequency differences between high-yield and low-yield

groups, with differences ranging from 18% to 27% in alternate allele frequency. The present findings align with selection signatures observed in previous Chinese goat genomic studies by Amiri Ghanatsaman et al.⁴. In the Beetal breed, significant SNPs were found in regions containing *IGFBP3*, *LEPR*, *LPL*, and *TSHR*, genes previously linked to lactation traits in tropical goat breeds. Functional annotation revealed that 40% of significant SNPs were intronic, 3% exonic (mostly synonymous), and the remainder intergenic. Two SNPs were particularly notable; a missense variant in *SPP1* on chromosome X and an intronic SNP in *PRLR* on chromosome X. Gene ontology enrichment analysis showed overrepresentation of terms related to lactation and hormone regulation ($p < 0.05$).

3.5. Linkage disequilibrium decay

To further explore population genomic architecture, LD decay was analyzed by plotting r^2 values against increasing genomic distances. As shown in Figure 6, Chinese Guanzhong goats displayed faster LD decay compared to Pakistani Beetal goats, suggesting higher recombination or less historical inbreeding. This observation is consistent with the kinship matrix and admixture analysis, supporting distinct demographic histories between the breeds.

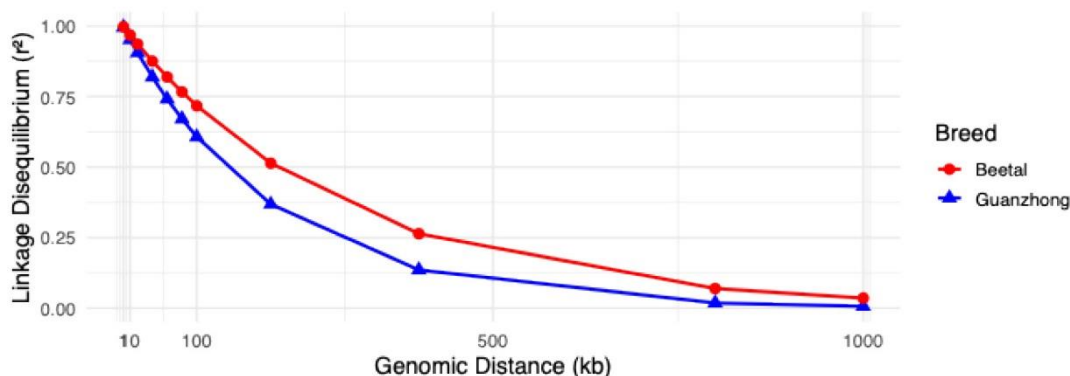


Figure 6. Linkage disequilibrium decay (r^2) plotted against genomic distance (kb) in Chinese Guanzhong and Pakistani Beetal Goats. Guanzhong goats (Red line) show steeper LD decay compared to Beetal goats (Blue line), indicating differences in genomic diversity and historical recombination.

4. Discussion

By integrating population structure correction, mixed linear modeling, and functional annotation, significant loci overlapping key lactation genes such as *LALBA*, *PRLR*, and *SPP1*, were identified in the present study with strong statistical support and biological relevance. A major association peak was detected on chromosome 19, overlapping the *LALBA* gene, a core gene involved in lactose synthesis and milk secretion. The current finding is consistent with QTLs identified in Saanen and Alpine goats for milk yield and composition^{5,6}. The effect direction indicated that Guanzhong goats with higher milk production have more of a certain allele, supporting earlier studies, which link increased *LALBA* gene activity to more milk³. Similarly, the signal on chromosome X near *PRLR* and *ERBB4*, genes involved in prolactin signaling and mammary gland development, mirrors earlier findings in caprine and bovine studies¹⁶, confirming the conserved regulatory mechanisms in lactation. In Chinese Guanzhong goats, SNPs near *ANPEP*, *PRKG1*, and *ADRA1A* were significantly associated with milk yield. These genes have previously been identified under positive selection for high milk production in Chinese dairy breeds⁵. *PRKG1* has been implicated in mammary epithelial function and vascular flow, both essential for milk synthesis⁴. In contrast, Pakistani Beetal goats showed significant enrichment of SNPs within or near *IGFBP3*, *LEPR*, *TSHR*, and *LPL*, all of which are involved in metabolic and hormonal regulation of lactation¹⁷. The direction of relationship in Beetal goats indicated that their milk output under low-input regimes could be influenced by genes linked to improved nutrient mobilization and hormone sensitivity. The allele frequencies of several breeds were examined to bolster these genetic signals^{16,17}. Conversely, SNPs in *LEPR* and *IGFBP3* showed higher alternate allele frequency in Beetal goats, consistent with adaptation to semi-arid environments and nutrient-efficient lactation strategies^{7,8}. Functionally, most genome-wide significant variants were in intronic or intergenic regions (90%), though a few exonic variants were noted. A wrong SNP in *SPP1* and an intronic SNP in *PRLR* were of particular interest, as *SPP1* has been shown to modulate milk protein gene expression and mammary gland remodeling³. The functional role of adjacent genes in milk

output was supported by the considerable overrepresentation of words associated with lactation, hormone control, and mammary development found by gene ontology enrichment. The observed faster LD decay in Guanzhong goats compared to Beetal suggested higher historical recombination rates, likely due to intensive artificial selection in commercial breeding programs. In contrast, Beetal goats, which are traditionally raised in low-input systems, displayed slower LD decay and higher within-breed kinship, indicating possible genetic bottlenecks or localized selection⁹. Compared to previous GWAS using SNP arrays, the use of WGS data in the present study provided a more comprehensive assessment of genomic variation, including rare and breed-specific variants often missed by fixed SNP panels including GoatSNP50^{4,10}. Moreover, the current results corroborated and expanded earlier caprine studies by identifying population-specific candidate genes relevant to dairy performance under different environmental and production systems¹⁸.

Despite the valuable knowledge gained from the present GWAS on milk yield in Guanzhong and Beetal goats, several limitations and challenges should be acknowledged. First, the lack of individual-level phenotypic milk yield data in the available WGS datasets limited the precision of genotype-to-phenotype associations, which may affect the statistical power and accuracy of detected signals. The modest sample size further reduces the ability to detect variants with small effect sizes¹⁰. Although population stratification was addressed through PCA and kinship matrix corrections, residual confounding between the two goat populations (Chinese and Pakistani) cannot be entirely ruled out. Environmental variables such as feeding, housing, and climate conditions were not available in the metadata, which restricts gene-environment interaction analysis¹¹. Moreover, while the present study identified SNPs near biologically relevant lactation genes, no wet-lab experimental validation was performed to confirm their functional impact. Lastly, the handling and analysis of large-scale WGS data require significant computational resources and technical expertise, which may present barriers to similar studies in resource-limited settings^{17,18}.

5. Conclusion

The present study underscored the transformative potential of *in silico* genomics in the genetic determinants of complex traits in livestock. By integrating available whole-genome sequencing (WGS) data, a comprehensive genome-wide association study (GWAS) on milk yield was conducted in two distinct dairy goat breeds, the Chinese Guanzhong and the Pakistani Beetal, by using a straightforward computational pipeline that includes population structure analysis, read alignment, variant calling, quality control, and association mapping based on mixed linear models. Several significant SNPs were identified that are located within or near genes previously implicated in lactation, such as *LALBA*, *SPP1*, and *PRLR*. The present results reinforced the value of high-density variant data generated from WGS in enhancing the resolution and power of GWAS, especially in non-model and indigenous breeds that are often underrepresented in genomic research. Notably, the present study was carried out entirely using open-source bioinformatics tools BWA for read alignment, SAMtools/BCFtools for variant calling, GCTA and PLINK/GEMMA for association modeling, and VEP/ANNOVAR for functional annotation, demonstrating that cutting-edge genomic analyses can be achieved without the need for proprietary software or costly infrastructure. Furthermore, the present approach demonstrated how existing genomic resources can be reexamined using advanced statistical genetics frameworks to generate new biological views. Future studies could expand on these findings by incorporating multi-omics data, such as transcriptomics and proteomics, to validate the functional roles of identified SNPs. Additionally, applying machine learning approaches to predict milk yield from genomic data could enhance breeding strategies. Finally, extending GWAS to larger populations of indigenous breeds across diverse environments may uncover novel genetic variants influencing lactation traits.

Declarations

Acknowledgments

The authors would like to thank the developers and curators of the NCBI Sequence Read Archive (SRA) and the European Nucleotide Archive (ENA) for providing access to public whole-genome sequencing datasets used in the present study. The authors are also grateful to the open-source software community for making advanced bioinformatics tools freely available, enabling the completion of this research.

Availability of data and materials

All sequencing data analyzed in the present study were obtained from available databases. Specific accession numbers for Guanzhong and Beetal goat samples are listed in the manuscript and can be accessed through the NCBI SRA and ENA repositories. The bioinformatics pipeline and scripts used for data analysis are available from the

corresponding author upon reasonable request.

Funding

The present study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The work was carried out using institutional computational resources and publicly accessible data.

Conflict of interests

The authors declare that there are no conflicts of interest related to the content or publication of this manuscript.

Ethical considerations

The present study was conducted in accordance with the ethical standards of the relevant national and institutional guidelines for research and publication. All bioinformatics analyses were performed using existing genomic data from Guanzhong and Beetal goats, obtained from publicly available or ethically sourced databases. The authors confirmed compliance with the journal's guidelines for ethical publication, ensuring proper data handling, no plagiarism, and an acknowledgment of data sources.

Authors' contributions

Umar Aziz conceptualized and designed the study, retrieved and processed the whole-genome sequencing data, conducted all bioinformatics analyses including quality control, alignment, variant calling, GWAS modeling, and data visualization, and prepared the first draft of the manuscript. Abdul Rehman contributed to breed selection, data retrieval, and assisted in genome alignment and variant filtering. Muhammad Hanzalah Yousaf supported statistical analysis and population structure evaluation, including PCA, admixture, and kinship matrix plotting. Fasih Ur Rehman critically reviewed the manuscript for scientific accuracy and improved the interpretation of genetic results. Muhammad Mushahid provided expertise on indigenous breed characteristics and lactation-related traits. Nauman Khan helped in GWAS validation, supported functional annotation using VEP and ANNOVAR, and reviewed analytical outcomes. Jiayuan Li, Xugan Wang, and Hanbing Yan contributed technical guidance in WGS data handling, computational optimization, and refinement of methodology. Xiaopeng An supervised the project, oversaw research planning and coordination, guided interpretation of findings, and critically revised the manuscript. All authors read and approved the final edition of the manuscript.

References

- Ni J, Xian M, Ren Y, Yang L, Li Y, Guo S, et al. Whole-genome resequencing reveals candidate genes associated with milk production trait in Guanzhong dairy goats. *Anim Genet*. 2024; 55(1): 168-172. DOI: [10.1111/age.13380](https://doi.org/10.1111/age.13380)
- Zhao J, Mu Y, Gong P, Liu B, Zhang F, Zhu L, et al. Whole-genome resequencing of native and imported dairy goat identifies genes

- associated with productivity and immunity. *Front Vet Sci.* 2024; 11: 1409282. DOI: [10.3389/fvets.2024.1409282](https://doi.org/10.3389/fvets.2024.1409282)
3. Zhang K, Zhao J, Mi S, Liu J, Luo J, and Liu J. Whole-genome variants resource of 298 Saanen dairy goats. *Sci Data.* 2025; 12(1): 528. DOI: [10.1038/s41597-025-04880-6](https://doi.org/10.1038/s41597-025-04880-6)
 4. Amiri Ghanatsaman Z, Ayatollahi Mehrgardi A, Asadollahpour Nanaei H, and Esmailzadeh A. Comparative genomic analysis uncovers candidate genes related with milk production and adaptive traits in goat breeds. *Sci Rep.* 2023; 13(1): 8722. DOI: [10.1038/s41598-023-35973-0](https://doi.org/10.1038/s41598-023-35973-0)
 5. Xiong J, Bao J, Hu W, Shang M, and Zhang L. Whole-genome resequencing reveals genetic diversity and selection characteristics of dairy goat. *Front Genet.* 2023; 13: 1044017. DOI: [10.3389/fgene.2022.1044017](https://doi.org/10.3389/fgene.2022.1044017)
 6. Massender E, Oliveira HR, Brito LF, Maignel L, Jafarikia M, Baes CF, et al. Genome-wide association study for milk production and conformation traits in Canadian Alpine and Saanen dairy goats. *J Dairy Sci.* 2023; 106(2): 1168-1189. DOI: [10.3168/jds.2022-22223](https://doi.org/10.3168/jds.2022-22223)
 7. Qu Y, Chen L, Ren X, Shari A, Yuan Y, Yu M, et al. Milk proteomic analysis reveals differentially expressed proteins in high-yielding and low-yielding Guanzhong dairy goats at peak lactation. *J Dairy Res.* 2024; 91(1): 31-37. DOI: [10.1017/S0022029924000013](https://doi.org/10.1017/S0022029924000013)
 8. Chessari G, Criscione A, Marletta D, Crepaldi P, Portolano B, Manunza A, et al. Characterization of heterozygosity-rich regions in Italian and worldwide goat breeds. *Sci Rep.* 2024; 14(1): 3. DOI: [10.1038/s41598-023-49125-x](https://doi.org/10.1038/s41598-023-49125-x)
 9. Feng F, Yang G, Ma X, Zhang J, Huang C, Ma X, et al. Comparative analysis of milk fat extracted from different goat breeds in China: Fatty acids, triacylglycerols and thermal and spectroscopic characterization. *Foods.* 2024; 13(12): 1913. DOI: [10.3390/foods13121913](https://doi.org/10.3390/foods13121913)
 10. Zhao Q, Huang C, Chen Q, Su Y, Zhang Y, Wang R, et al. Genomic inbreeding and runs of homozygosity analysis of cashmere goat. *Animals.* 2024; 14(8): 1246. DOI: [10.3390/ani14081246](https://doi.org/10.3390/ani14081246)
 11. Fan L, Shen J, Li X, Li H, Shao Y, Lu CD, et al. Analysis of temporal changes of microbiota diversity and environmental interactions in Saanen dairy goats. *J Appl Anim Res.* 2023; 51(1): 749-763. DOI: [10.1080/09712119.2023.2273945](https://doi.org/10.1080/09712119.2023.2273945)
 12. Lázaro SF, Tonhati H, Oliveira HR, Silva AA, Scalez DCB, Nascimento AV, et al. Genetic parameters and genome-wide association studies for mozzarella and milk production traits, lactation length, and lactation persistency in Murrah buffaloes. *J Dairy Sci.* 2024; 107(2): 992-1021. DOI: [10.3168/jds.2023-23284](https://doi.org/10.3168/jds.2023-23284)
 13. Cao Y, Feng T, Wu Y, Xu Y, Du L, Wang T, et al. The multi-kingdom microbiome of the goat gastrointestinal tract. *Microbiome.* 2023; 11(1): 219. DOI: [10.1186/s40168-023-01651-6](https://doi.org/10.1186/s40168-023-01651-6)
 14. George L, Alex R, Sukhija N, Jaglan K, Vohra V, Kumar R, et al. Genetic improvement of economic traits in Murrah buffalo using significant SNPs from genome-wide association study. *Trop Anim Health Prod.* 2023; 55(3): 199. DOI: [10.1007/s11250-023-03606-3](https://doi.org/10.1007/s11250-023-03606-3)
 15. Pawliński B, Gołębiewski M, Trela M, and Witkowska-Piłaszewicz O. Comparison of blood gas parameters, ions, and glucose concentration in Polish Holstein-Friesian dairy cows at different milk production levels. *Sci Rep.* 2023; 13(1): 1414. DOI: [10.1038/s41598-023-28644-7](https://doi.org/10.1038/s41598-023-28644-7)
 16. Fouz R, Rodríguez-Bermúdez R, Rodríguez-Godina IJ, Rodríguez-Domínguez M, Rico M, Diéguez FJ. Evaluation of haptoglobin concentration in clinically healthy dairy cows: Correlation between serum and milk levels. *J Appl Anim Res.* 2024; 52(1): 2300624. DOI: [10.1080/09712119.2023.2300624](https://doi.org/10.1080/09712119.2023.2300624)
 17. Cesarani A, Corte Pause F, Hidalgo J, Garcia A, Degano L, Vicario D, et al. Genetic background of semen parameters in Italian Simmental bulls. *Ital J Anim Sci.* 2023; 22(1): 76-83. DOI: [10.1080/1828051X.2022.2160665](https://doi.org/10.1080/1828051X.2022.2160665)
 18. Wang H, Xu W, Chen X, Mei X, Guo Z, and Zhang J. LncRNA LINC00205 stimulates osteoporosis and contributes to spinal fracture through the regulation of the miR-26b-5p/KMT2C axis. *BMC Musculoskelet Disord.* 2023; 24(1): 262. DOI: [10.1186/s12891-023-06136-z](https://doi.org/10.1186/s12891-023-06136-z)