

Advances in SNP Marker Development and Application on the Bovine Y Chromosome

Zhenyu Lei

*College of Animal Science and Technology, Northwest A&F University, Yangling, China
leizhenyu2023@nwfau.edu.cn*

Abstract. The male-specific region (MSY) of the bovine Y chromosome serves as an ideal tool for tracing paternal origins due to its paternal inheritance and absence of recombination. However, its high content of repetitive sequences has long restricted the accurate alignment of short reads and the complete assembly of reference sequences. In 2024, the completion of the telomere-to-telomere (T2T) gapless assembly of the bovine Y chromosome provided a full reference sequence for the systematic identification of Y-chromosomal single nucleotide polymorphisms (SNPs). Centering on the generational leaps in marker development technologies, this paper reviews the three-stage technological evolution from targeted amplification and next-generation sequencing to T2T assembly, analyzes the potential impacts of technical limitations on haplogroup distribution patterns and conclusions regarding paternal migration, sorts out the current application status and inconsistent findings of Y-SNPs in reproductive trait association studies and genomic selection, and finally outlines future directions including historical data calibration, genetic load assessment and breeding integration.

Keywords: Domestic cattle, Y chromosome, Single nucleotide polymorphism (SNP), Haplogroup, Paternal inheritance

1. Introduction

Domestic cattle comprise zebu (*Bos indicus*) and taurine cattle (*Bos taurus*). Early research on their domestication and dispersal history relied on mitochondrial DNA (mtDNA) and autosomal markers, and paternal genetic information long lagged behind owing to the difficulties in studying the Y chromosome. The bovine Y chromosome consists of the pseudoautosomal region (PAR) and the male-specific region (MSY): the PAR lies at both ends of the Y chromosome and recombines with the X chromosome during meiosis; the MSY accounts for over 95% of the Y chromosome length and does not participate in recombination during meiosis, thus can pass down paternal genetic information intact. This non-recombinant feature makes Y-chromosomal single nucleotide polymorphisms (SNPs) ideal markers for tracing paternal lineages, and can also be applied to monitor the paternal diversity of breeding populations. Skaletsky et al. [1] divided the MSY euchromatic sequences of the human Y chromosome into three categories: the X-transposed region, the X-degenerate region and the ampliconic region. This classification framework has since been widely adopted for mammals including domestic cattle. Nevertheless, the MSY is rich in repetitive

sequences and free of recombination, which poses two fundamental technical barriers to genome assembly: repetitive sequences prevent short reads from being uniquely aligned, and repetitive regions cannot be effectively spanned by short reads, leading to highly fragmented assemblies.

In 2005, Götherström et al. [2] classified the bovine Y chromosome into three major haplogroups, Y1, Y2 and Y3, based on early scattered SNP markers. In 2011, Edwards et al. [3] established the global distribution framework of these haplogroups through a systematic survey of 138 breeds worldwide. Since then, with the popularization of next-generation sequencing (NGS) technology, high-throughput SNP genotyping has facilitated the discovery of sub-haplogroups and the fine resolution of migration routes. In 2024, Olagunju et al. [4] completed the telomere-to-telomere (T2T) complete assembly of the bovine Y chromosome, obtaining the first gap-free sequence of the MSY from telomere to telomere, which marks the entry of the bovine Y chromosome reference sequence from the fragmentation stage to the complete assembly stage. However, prior to the completion of T2T, there was a lack of systematic examination of the SNP data bias detected from fragmented reference sequences, the difficulty of short read alignment under long segment duplications, and the way to integrate whole Y chromosome variation information into existing breeding systems. Taking the generational leap in reference genome assembly level as the clue, this paper systematically sorts out the technological evolution of Y-SNP marker development, the progress of haplogroup distribution patterns and its breeding application status, so as to point out directions for the key issues under whole Y chromosome analysis (Figure 1).

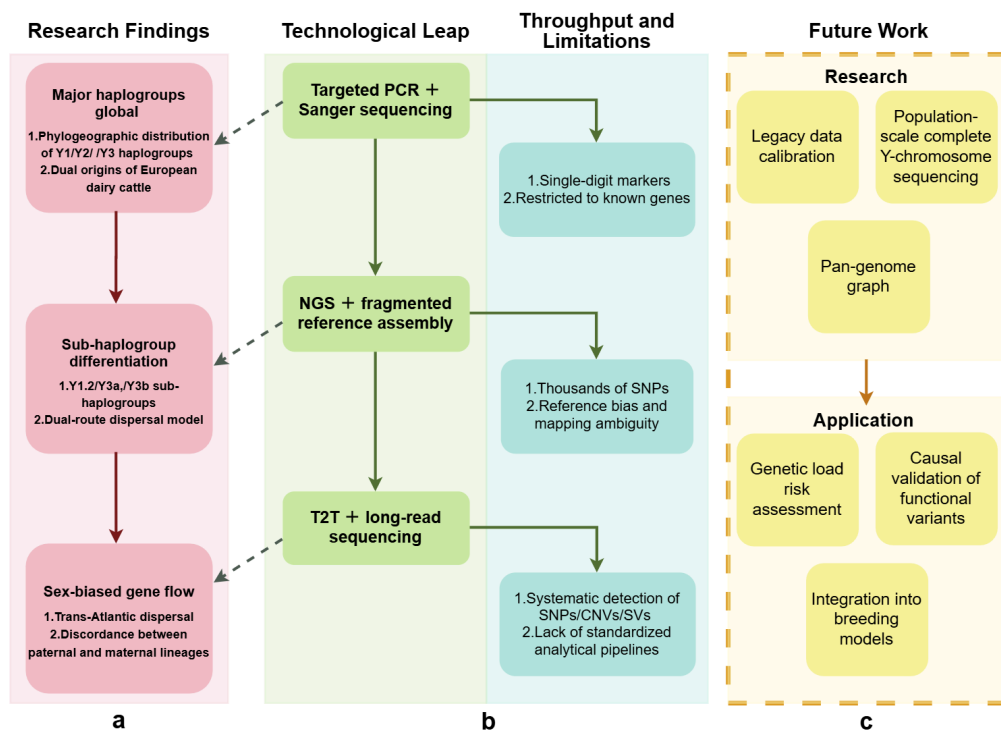


Figure 1. Overview of SNP marker development and application on the bovine Y chromosome (a) global haplogroup distribution and dispersal history, including dual origins, sub-haplogroup differentiation, and sex-biased gene flow (b) technological evolution from targeted PCR and next-generation sequencing (NGS) to telomere-to-telomere (T2T) assembly, with associated constraints such as reference bias and ambiguous mapping (c) future priorities for basic research (legacy data calibration, population-scale Y sequencing, pan-genome graph) and applied breeding (genetic load assessment, functional validation, integration into breeding models)

2. Technological evolution of Y-SNP marker development

The development methods and resolution of Y-SNP markers have always been constrained by sequencing technology and the assembly level of reference genomes. From targeted amplification to whole-chromosome analysis, this field has undergone three important leaps (Figure 1b).

2.1. Targeted amplification and sanger sequencing

When the reference genome only covered partial regions of the Y chromosome, marker development was mainly realized by targeted amplification of Y-chromosome specific genes via polymerase chain reaction (PCR), followed by Sanger sequencing of the amplified products to identify sequence variations. Using this strategy, Götherström et al. [2] sequenced the non-coding regions of *UTY*, *ZFY* and *DBY* genes, identified key SNP loci such as UTY19 (c.423 C>A) and ZFY10 (c.661 C>T), and accordingly divided the bovine Y chromosome into three haplogroups: Y1, Y2 and Y3. Bonfiglio et al. [5] further detected an 81 bp insertion site on the *USP9Y* gene, and developed a polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) genotyping method using the unique restriction site in the Y3 sequence. This strategy circumvents the problem of repetitive sequence alignment by selecting known Y-specific genes, but its throughput is extremely low. It can only detect a limited number of loci and hardly reflects the full landscape of the MSY.

2.2. NGS resequencing and conventional reference genomes

The advent of NGS technology has broken the throughput bottleneck, yet the fragmented state of reference sequences limits the completeness of variant detection. Whole-genome resequencing combined with MSY region variant extraction has lifted the throughput of Y-SNP markers from single digits to thousands. On this basis, Xia et al. [6] achieved joint analysis of Y-SNPs and Y-chromosomal short tandem repeats (Y-STRs); Ma et al. [7] identified three (sub)haplogroups Y1, Y2a and Y2b in Qaidam yellow cattle; Wu et al. [8] revealed remarkable differences in paternal genetic structure among local breeds through analysis of Weining cattle and Sinan cattle.

Nonetheless, the early fragmented reference sequences brought two problems to MSY analysis of NGS data. The first is reference bias: variants in regions missing from the assembly cannot be detected for lack of a reference. The second is ambiguous alignment: long segment duplications in the MSY make it impossible for short reads to align uniquely. For multi-mapping reads, whether the analysis pipeline adopts random allocation or direct discarding, it may introduce SNP genotyping errors or cause systematic underestimation of duplicated regions. Webster et al. [9] systematically described the mechanism by which sequence homology of sex chromosomes leads to short read misalignment and affects variant detection, and this principle also applies to segmental duplication regions within the Y chromosome. Therefore, these limitations could not be fundamentally resolved before the emergence of complete reference sequences.

2.3. T2T complete assembly and whole-chromosome profiling

Aligning NGS or third-generation sequencing data to the T2T complete reference sequence enables systematic identification of various variants including SNPs, structural variations (SVs) and copy number variations (CNVs) (Figure 1b), which fundamentally solves the problems of incomplete reference genomes and difficult alignment of repetitive sequences. From targeted amplification to whole-genome resequencing, and from fragmented reference sequences to T2T complete assembly,

the coordinated progress of sequencing technology and reference genomes has driven this field from genotyping of a handful of markers to the stage of fine whole-chromosome profiling [4].

3. Geographical distribution of Y-chromosomal haplogroups: paternal migration of global domestic cattle

The geographical distribution of haplogroups uncovered by Y-SNP markers provides core evidence for reconstructing the paternal domestication and migration history of domestic cattle. The evolution of this understanding has always kept pace with the improvement of marker resolution: from distinguishing basic haplogroups to resolving internal substructures, and then to capturing sex-biased gene flow, each technological breakthrough has promoted more accurate reconstruction of migration routes.

3.1. Basic haplogroups and global distribution

In the Sanger sequencing era, Götherström et al. [2] divided the bovine Y chromosome into three major haplogroups: Y1, Y2 and Y3. On this basis, Edwards et al. [3] established their spatial distribution framework via a systematic survey of 138 breeds worldwide: Y1 is concentrated in Northern and Northwestern Europe, Y2 dominates Central Europe and the Mediterranean region and extends to Northern Asia, and Y3 prevails in South Asia, East Africa and Southern China. The clear geographical separation of Y1 and Y2 across Europe supports the dual-origin hypothesis of European dairy cattle (Figure 1a), namely that the two may have entered Europe independently along the Balkan-Central European route and the Mediterranean-Iberian route respectively.

3.2. Subtype differentiation and refinement of migration routes

In the era of whole-genome resequencing application, although the reference genome remained incomplete, the leap in marker throughput allowed researchers to resolve the internal structure of haplogroups in greater detail. The application of high-throughput SNP genotyping has revealed the substructures within the three major haplogroups, further revising the initial migration model. The novel Y1.2 haplogroup identified by Pelayo et al. [10] in native Spanish cattle is exclusive to the Iberian Peninsula, indicating that this region has undergone an independent paternal evolutionary history, consistent with its position as the terminal of the southern dispersal route of taurine cattle. For zebu cattle, Pérez-Pardal et al. [11] further subdivided the Y3 haplogroup into three subclades, Y3A, Y3B and Y3C, through a systematic survey of Y chromosomes from South Asian and African zebu: Y3A is distributed in East Africa and South Asia, Y3B is mainly found in West Africa, and Y3C is concentrated in northern South Asia. This fine substructure demonstrates the multi-pathway spread of zebu from the South Asian domestication center to the African continent. The distinct subclade detected by Pérez-Pardal et al. [12] in African taurine cattle further suggests that the African continent may preserve older paternal variants of taurine cattle than Europe.

3.3. Sex-biased gene flow and transoceanic dispersal

Higher-resolution markers not only refine migration routes, but also reveal the sex asymmetry of gene flow. Chinese yellow cattle exhibits a north-south differentiation pattern of Y2 (north) and Y3 (south) [6, 13], which is highly consistent with mtDNA evidence, confirming the dual-origin model that taurine cattle and zebu were introduced along the Eurasian steppe and southeast coastal routes respectively and converged in the Central Plains. In Central Asia, however, Yamanaka et al. [14]

found that all Y chromosomes of Kazakhstani native cattle are of taurine type, while their mtDNA contains zebu lineages. Such paternal-maternal genetic decoupling strongly indicates that zebu genes can introgress into taurine populations through female mediation without leaving paternal traces. Similarly, analysis of Peruvian Creole cattle by Yalta-Macedo et al. [15] showed that their Y chromosomes originate from the Iberian Peninsula, while the African zebu Y chromosome signals detected in some American cattle populations record the genetic input from the transatlantic slave trade. Together, these findings demonstrate that the temporal resolution of Y-SNP markers is sufficient to resolve both millennium-scale domestication dispersal and century-scale recent transmission events (Figure 1a).

4. Applications and challenges

Y-SNP markers are not only tools for tracing paternal history; the genes located in the MSY are also closely related to male reproductive function, which opens up the possibility of applying Y chromosome variation in breeding improvement. Nevertheless, from marker development to breeding practice, the field still faces dual challenges of unstable association validation and imperfect modeling strategies.

4.1. Expression features and functions of MSY genes

Chang et al. [16] conducted a systematic transcriptomic analysis of the bovine MSY and detected 28 protein-coding genes/families in total, with a total copy number as high as 1,274. Among them, 16 genes/families are unique to Bovidae, which revises the previous view that the Y chromosome has scarce genes and low transcriptional activity. Importantly, most MSY ampliconic genes are mainly or exclusively expressed in testicular tissue, suggesting that variations in this region may affect spermatogenesis and male reproductive function. In terms of variant types, currently reported functional evidence related to reproductive traits mostly focuses on copy number variations. For instance, the copy numbers of *HSFY* and *ZNF280BY* are associated with testis size and conception rate in Holstein bulls [17], and the copy number of *ZNF280AY* is negatively correlated with normal sperm percentage and sperm concentration [18]. As for Y-SNPs, direct evidence of functional association remains limited. However, differences in reproductive traits among different Y chromosome haplogroups provide indirect support for the potential functions of Y-SNPs: bulls carrying the zebu-origin Y3 haplogroup generally show higher sperm motility under heat stress than those carrying taurine-origin Y1 or Y2 haplogroups, which indirectly implies that Y chromosome haplogroups may be associated with heat stress tolerance.

4.2. Discrepant results in reproductive trait association studies

The association between Y chromosome genetic variation and reproductive traits has not been consistently verified across all studies. Pacheco et al. [19] performed association analysis of sex chromosome markers on 5,014 US Holstein bulls and found that none of the 24 Y-chromosome specific SNPs were significantly associated with bull conception rate, whereas multiple markers in the X-chromosome specific region showed significant associations. In contrast, the aforementioned copy number variation studies on *HSFY*, *ZNF280BY* and *ZNF280AY* all detected significant associations. Such inconsistency may arise from the following reasons: first, the Y chromosome has a small effective population size and low genetic variability, and the sample size of most studies is insufficient to capture minor effects; second, loci detected by conventional SNP genotyping may

only be passenger markers, while the true causal variants are likely copy number variations or large structural variations that are systematically omitted in standard pipelines; third, Y chromosome genes may have epistatic interactions with autosomes or the mitochondrial genome, leading to biased effect estimates in marginal association analysis.

4.3. Instrumental value for paternal diversity management

Beyond genetic association with reproductive traits, another application direction of Y-SNP markers is to monitor the paternal genetic diversity of breeding populations. Against the background of widespread artificial insemination, the excessive use of a small number of elite breeding bulls may lead to a sharp narrowing of the paternal genetic base of populations. Incorporating Y haplogroup information into mating schemes helps optimize bull utilization strategies and reduce the risk of genetic vulnerability. Panigrahi et al. [20] also noted in their cross-species review that Y chromosome genes are gaining increasing attention in breeding programs and can be used to improve male reproductive performance and disease resistance.

5. Summary and prospects

5.1. Re-evaluation of historical data and population sequencing

Following the completion of T2T assembly, the field faces a dual task: calibrating legacy data and advancing new sequencing efforts. Before T2T, Y-SNP studies based on fragmented reference sequences generally suffered from two types of systematic bias: reference bias and ambiguous alignment [4, 9], which need to be evaluated through re-alignment and re-genotyping of historical samples using the T2T sequence as the reference. Meanwhile, moving from T2T assembly of a few individuals to population-scale whole Y chromosome sequencing can unlock the application potential of the new reference sequence. A feasible strategy is to proceed in phases: initially, select 5–10 bulls carrying different known haplotypes from each representative global cattle breed, perform long-read sequencing to assemble the entire Y chromosome, generate a "bovine Y chromosome pan-genome map", and systematically document population-specific structural variations; on this basis, design targeted genotyping marker panels for the identified variants to reduce the detection cost for large-scale populations (Figure 1c).

5.2. From genetic load assessment to breeding model integration

The non-recombinant nature of the MSY brings a long-overlooked breeding risk: the entire region is inherited as an indivisible super-haplotype block, with favorable and deleterious variants bound together. In the breeding system dominated by artificial insemination, high-frequency selection for dominant haplotypes may unconsciously enrich their linked deleterious structural variants. Therefore, complete sequencing of dominant Y haplotypes to assess genetic load should become a necessary preliminary step before breeding application.

However, current mainstream commercial SNP chips (such as the Illumina BovineHD BeadChip) contain only a small number of Y-chromosome specific loci, and these loci are often filtered out in routine genetic evaluation pipelines due to difficult genotyping and extremely low variation; GWAS analyses also generally exclude sex chromosomes, making it hard to integrate Y chromosome variation into existing breeding value estimation systems. Genomic prediction of sex chromosome markers by Pacheco et al. [19] showed that incorporating bovine X chromosome (BTAX) markers

can improve prediction accuracy, but the independent contribution of bovine Y chromosome (BTAY) markers remains unclear due to limited variation in the studied population. The SNP prioritization strategy proposed by Hay [21] indicates that marker optimization for specific populations and traits can enhance prediction accuracy, providing a methodological reference for the breeding application of Y-SNPs.

Transforming Y chromosome variation from a tracing tool into a breeding marker also requires a leap from association to causation. One feasible strategy is to perform single-cell multi-omics sequencing (scMulti-omics) on testicular tissues of bulls carrying different Y haplotypes, identify functional variants driving expression differences at the cell type and developmental stage levels, and establish a regulatory network from variation to phenotype. In terms of breeding model integration, given the low polymorphism of the Y chromosome and its non-recombination with autosomes, directly assigning independent breeding values to it faces difficulties such as unstable variance component estimation and difficulty in separating additive effects. By contrast, two approaches are more operable: first, incorporate validated functional Y haplotypes as fixed effects into the breeding value estimation model to directly estimate their fixed offset on target traits; second, in mating optimization, take the genetic load assessment results of each haplotype together with known phenotypic effects as constraints to avoid passive enrichment of deleterious variants (Figure 1c).

References

- [1] Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., et al. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423, 825–837.
- [2] Götherström, A., Anderung, C., Hellborg, L., et al. (2005). Cattle domestication in the Near East was followed by hybridization with aurochs bulls in Europe. *Proceedings of the Royal Society B: Biological Sciences*, 272, 2345–2351.
- [3] Edwards, C. J., Ginja, C., Kantanen, J., et al. (2011). Dual origins of dairy cattle farming—evidence from a comprehensive survey of European Y-chromosomal variation. *PloS One*, 6, e15922.
- [4] Olagunju, T. A., Rosen, B. D., Neibergs, H. L., et al. (2024). Telomere-to-telomere assemblies of cattle and sheep Y-chromosomes uncover divergent structure and gene content. *Nature Communications*, 15, 8277.
- [5] Bonfiglio, S., De Gaetano, A., Tesfaye, K., et al. (2012). A novel USP9Y polymorphism allowing a rapid and unambiguous classification of *Bos taurus* Y chromosomes into haplogroups. *Animal Genetics*, 43, 611–613.
- [6] Xia, X., Yao, Y., Li, C., et al. (2019). Genetic diversity of Chinese cattle revealed by Y-SNP and Y-STR markers. *Animal Genetics*, 50, 64–69.
- [7] Ma, Z. J., Lei, C. Z., Chen, N. B., et al. (2025). Genome-wide genetic diversity and functional gene exploration in Qaidam yellow cattle.
- [8] Wu, D. Y., Wang, N., Yan, H. X., et al. (2025). Study on Y-chromosomal genomic genetic diversity and paternal origin of Weining cattle and Sinan cattle. *Chinese Journal of Animal Science*, 61, 238–243.
- [9] Webster, T. H., Couse, M., Grande, B. M., et al. (2019). Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. *GigaScience*, 8, giz074.
- [10] Pelayo, R., Penedo, M. C. T., Valera, M., et al. (2017). Identification of a new Y chromosome haplogroup in Spanish native cattle. *Animal Genetics*, 48, 450–454.
- [11] Pérez-Pardal, L., Sánchez-Gracia, A., Álvarez, I., et al. (2018). Legacies of domestication, trade and herder mobility shape extant male zebu cattle diversity in South Asia and Africa. *Scientific Reports*, 8, 18027.
- [12] Pérez-Pardal, L., Royo, L. J., Beja-Pereira, A., et al. (2010). Y-specific microsatellites reveal an African subfamily in taurine (*Bos taurus*) cattle. *Animal Genetics*, 41, 232–241.
- [13] Li, R., Zhang, X. M., Campana, M. G., et al. (2013). Paternal origins of Chinese cattle. *Animal Genetics*, 44, 446–449.
- [14] Yamanaka, H., Murata, K., Tabata, R., et al. (2019). Kazakhstani native cattle reveal highly divergent mtDNA from *Bos taurus* and *Bos indicus* lineages with an absence of *Bos indicus* Y chromosome. *Animal Science Journal = Nihon Chikusan Gakkaiho*, 90, 29–34.

- [15] Yalta-Macedo, C. E., Veli, E. A., Diaz, G. R., et al. (2021). Paternal ancestry of Peruvian creole cattle inferred from Y-chromosome analysis. *Livestock Science*, 244, 104376.
- [16] Chang, T.-C., Yang, Y., Retzel, E. F., et al. (2013). Male-specific region of the bovine Y chromosome is gene rich with a high transcriptomic activity in testis development. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 12373–12378.
- [17] Yue, X.-P., Dechow, C., Chang, T.-C., et al. (2014). Copy number variations of the extensively amplified Y-linked genes, HSFY and ZNF280BY, in cattle and their association with male reproductive traits in Holstein bulls. *BMC Genomics*, 15, 113.
- [18] Liu, R., Low, W. Y., Tearle, R., et al. (2019). New insights into mammalian sex chromosome structure and evolution using high-quality sequences from bovine X and Y chromosomes. *BMC Genomics*, 20, 1000.
- [19] Pacheco, H. A., Rezende, F. M., & Peñagaricano, F. (2020). Gene mapping and genomic prediction of bull fertility using sex chromosome markers. *Journal of Dairy Science*, 103, 3304–3311.
- [20] Panigrahi, M., Rajawat, D., Nayak, S. S., et al. (2025). Genetic diversity of the Y-chromosome in domestic animals: From evolutionary insights to functional implications. *Molecular Biology Reports*, 52, 1031.
- [21] Hay, E. H. (2026). Prioritization of SNP markers for genomic prediction in closed beef cattle populations. *Translational Animal Science*, 10, txaf166.