Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Chromosome-Scale Genome Assemblies of Two Black Bengal Goat (Capra hircus) Individuals

Sharadindu Shil

dr.sharadindu@gmail.com

Genomics and Molecular Biology Laboratory, PBGSBS

Gouri Shankar Koner

Genomics and Molecular Biology Laboratory, PBGSBS

Manoranjan Roy

West Bengal University of Animal and Fishery Sciences

Biswajit Saha

Genomics and Molecular Biology Laboratory, PBGSBS

Diganta Pan

Kotulpur Buck Semen Station, Paschimbanga Go Sampad Bikash Sanstha (PBGSBS)

Jayanta Chowdhury

CLART, West Bengal Livestock Development Corporation Limited (WBLDCL)

Anusuya Debnath

Brainware University

Ayan Mukherjee

CLART, West Bengal Livestock Development Corporation Limited (WBLDCL)

Shankari Prasad Dutta

CLART, West Bengal Livestock Development Corporation Limited (WBLDCL)

Kasturi Mandal

Genomics and Molecular Biology Laboratory, PBGSBS

Ajoy Mandal

National Dairy Research Institute

Santanu Banik

National Dairy Research Institute

---

data-descriptor

Keywords:

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11141864/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

The Black Bengal Goat (Capra hircus), a small-bodied, prolific breed native to West Bengal, India and neighboring Bangladesh, is valued for its meat quality, disease resistance, and adaptability to marginal environments, yet lacks a dedicated reference-quality genome assembly. Here we report chromosome-scale genome assemblies for two Black Bengal Goat individuals — 803M (male) and D863F (female) — generated from PacBio HiFi long-read sequencing (~ 30× coverage) using hifiasm followed by reference-guided scaffolding against the Capra hircus ARS1.2 assembly. The resulting assemblies span 2.86 Gb (803M) and 2.78 Gb (D863F), with scaffold N50 values of 88 Mb and 92 Mb (L50 = 13 and 12), reflecting near-chromosome-level contiguity, and genome-mode BUSCO completeness of 99.5% and 99.3% (cetartiodactyla_odb10), respectively. Protein-coding genes were annotated using RNA-seq-evidence-based BRAKER1, yielding 23,563 (803M) and 47,030 (D863F) gene models after filtering for assembly-boundary artifacts. Demographic history reconstructed via PSMC revealed a shared Pleistocene population bottleneck followed by partial recovery, consistent between both individuals. These assemblies — deposited under BioProject PRJNA1502507 — provide the first chromosome-scale reference genome resource for Black Bengal Goat, supporting future work in comparative genomics, marker-assisted selection, and conservation genetics of this economically important South Asian breed.

Background & Summary

Goats (Capra hircus) are among the most widely distributed livestock species globally, valued for their adaptability to marginal environments and their contribution to meat, milk, and fiber production across smallholder and pastoral systems. The Black Bengal Goat, a small-bodied, prolific breed native to West Bengal, India, and neighboring Bangladesh, is of substantial economic and genetic importance in South Asia: it is prized for high-quality meat and pelt, early sexual maturity, and high litter size, and is widely reared by smallholder farmers across the region. Despite this importance, no dedicated, high-quality reference genome assembly for Black Bengal Goat has previously been available — prior molecular characterization of the breed has relied primarily on mitochondrial DNA and microsatellite markers, or on alignment to the general-purpose Capra hircus ARS1.2 reference assembly, which was built from a different breed (San Clemente goat) and does not capture breed-specific structural variation.

Here we report chromosome-scale, PacBio HiFi-based genome assemblies for two Black Bengal Goat individuals — 803M (male) and D863F (female) — together with RNA-seq-evidence-based gene annotation and population-genomic demographic reconstruction. These resources establish the first reference-quality genomic dataset specifically for this breed, supporting future comparative genomics, marker-assisted selection, and conservation genetics applications.

Methods

Animal sampling and ethics statement

The two Black Bengal Goat individuals used for genome assembly, 803M (male) and D863F (female), were drawn from the institutional herd maintained at ICAR-National Dairy Research Institute, Eastern Regional Station (NDRI-ERS), Kalyani, Nadia, West Bengal, India. Blood samples (approximately 0.4 mL per animal) were collected by a single jugular venipuncture, performed by a registered veterinarian (Dr. Ajoy Mandal, Principal Scientist, NDRI-ERS) under NDRI-ERS's institutional research mandate and associated farm activities. Sampling followed standard veterinary practice and the provisions of the Prevention of Cruelty to Animals Act, 1960. The procedure caused no more than transient discomfort comparable to a routine needle prick; no sedation, experimental intervention, or euthanasia was performed, and both animals were returned to the institutional herd immediately and unharmed. As a minimal-risk, non-experimental diagnostic/breed-characterization procedure performed on institutionally maintained herd animals by a registered veterinarian under NDRI-ERS's institutional mandate, formal review by an Institutional Animal Ethics Committee (IAEC) under the CPCSEA Breeding of and Experiments on Animals (Control and Supervision) Rules, 1998 was not required, consistent with accepted guidelines for routine vertebrate sampling.

Genome sequencing and assembly

Two Black Bengal Goat individuals — 803M (male) and D863F (female) — were sequenced on the PacBio HiFi platform (~ 30× coverage) by MedGenome Labs Ltd. Genome assembly was performed with hifiasm v0.25.0-r726 (Cheng et al., 2021) in HiFi-only mode. Contigs were scaffolded against the Capra hircus ARS1.2 reference assembly (GCF_001704415.2) using RagTag v2.1.0 (Alonge et al., 2022). The resulting assemblies comprised 1,073 scaffolds (803M) and 1,816 scaffolds (D863F), with total lengths of 2.86 Gb and 2.78 Gb, scaffold N50 values of 88 Mb (L50 = 13) and 92 Mb (L50 = 12), and GC content of 43.14% and 42.70%, respectively. Assembly completeness was assessed with BUSCO v6.1.0 (Manni et al., 2021) in genome mode (euk_genome_min setting, miniprot gene predictor) against the cetartiodactyla_odb10 lineage dataset (n = 13,335 BUSCOs). Base-level assembly accuracy was assessed with Merqury (Rhie et al., 2020) using individual-specific Illumina-derived k-mer databases (k = 21, built independently for each animal from its own whole-genome Illumina reads, at near-identical depth: 2.89 Gb input bases for 803M, 2.84 Gb for D863F).

Decontamination

Scaffolded assemblies were screened for adaptor contamination using NCBI's FCS-adaptor tool and for foreign/contaminant sequence using FCS-GX (Astashyn et al., 2024), run via usegalaxy.org owing to local memory constraints. Flagged contaminant contigs were removed prior to submission: 2 Anaplasma-derived contigs from 803M and 49 Theileria-derived contigs from D863F. Sequence headers were renamed to NCBI-compliant conventions (scaffold_01–scaffold_29, scaffold_MT, unplaced_scaffold_####) prior to submission, as RagTag-embedded reference accession numbers in headers are not permitted by NCBI.

Genome annotation

Protein-coding genes were annotated using BRAKER1 (Hoff et al., 2019) in RNA-seq-evidence-only mode. RNA-seq evidence was drawn from BioProject PRJEB23196 (Muriuki et al., 2019; 8 tissues, 58 samples), quality-trimmed with fastp (Chen et al., 2018) and aligned with STAR (Dobin et al., 2013) in two-pass mode. Annotation completeness was assessed with BUSCO (artiodactyla_odb12 lineage dataset) in protein mode. Gene models were filtered to remove likely assembly-boundary artifacts, excluding models located within 1,000 bp of a scaffold edge and occupying < 30% of that scaffold's length (14 genes excluded for 803M, 51 for D863F). Final GFF3 files were generated with gffread (Pertea & Pertea, 2020), applying --keep-genes and correcting transcript feature types to mRNA. Submission-format validation was performed with table2asn v1.28.1094.

Population genomics

Effective population size (Ne) was estimated using GONE (Santiago et al., 2020) and SNeP (Barbato et al., 2015). Historical demographic trajectories were reconstructed using PSMC (Li & Durbin, 2011), assuming a per-generation mutation rate of 2.5×10^{-8} and a generation time of 2 years, following the precedent established for goat/sheep PSMC analyses by Alberto et al. (2018).

Data Records

Genome assemblies, together with annotation files (GTF, protein FASTA, CDS FASTA), have been deposited in GenBank under BioProject PRJNA1502507 (accession numbers JCDURR000000000 for 803M and JCDURS000000000 for D863F). These records are held under a confidentiality hold and will be released publicly upon publication of this manuscript or by March 2027, whichever occurs first. RNA-seq evidence used for annotation is publicly available under BioProject PRJEB23196 ((Muriuki et al., 2019).

Technical Validation

Genome assembly, annotation, and completeness statistics for both individuals are summarized in Table 1.

Table 1
Genome assembly and completeness statistics for two PacBio HiFi-sequenced Black Bengal Goat individuals.

Metric	803M (male)	D863F (female)
Total assembly length	2.86 Gb	2.78 Gb
Number of scaffolds	1,073	1,816
Scaffold N50 / L50	88 Mb / 13	92 Mb / 12
Contig N50	75 Mb	17 Mb
GC content	43.14%	42.70%
BUSCO, genome mode (cetartiodactyla_odb10, n = 13,335)	C:99.5% [S:98.0%, D:1.5%], F:0.1%, M:0.4%	C:99.3% [S:96.9%, D:2.4%], F:0.2%, M:0.5%
Merqury consensus QV	57.3	30.6
Protein-coding genes (BRAKER1)	23,563	47,030
BUSCO, annotation mode (artiodactyla_odb12, protein)	41.7% Complete	64.5% Complete

Both assemblies achieved near-chromosome-scale contiguity, with scaffold N50 values of 88 Mb (803M) and 92 Mb (D863F) and L50 values of 13 and 12 scaffolds, respectively (Table 1) — consistent with the expected karyotype of *Capra hircus* (2n = 60; 29 autosomes plus the X chromosome). Genome-mode BUSCO completeness was high for both assemblies (803M: 99.5% Complete; D863F: 99.3% Complete; cetartiodactyla_odb10, n = 13,335), with low duplication (1.5% and 2.4%) and minimal fragmentation or missingness.

Merqury consensus QV differed substantially between the two assemblies (803M: QV = 57.3; D863F: QV = 30.6), despite comparable genome-mode BUSCO completeness (99.5% vs. 99.3%) and assembly contiguity (scaffold N50 88 Mb vs. 92 Mb). Both k-mer databases were independently constructed and validated against their respective animal's own Illumina reads at near-identical depth (2.89 Gb vs. 2.84 Gb input bases), ruling out short-read coverage as an explanation (Fig. 1). The source of this discrepancy was not further resolved in this study and is noted as a limitation.

Population genomic analysis via PSMC revealed concordant demographic trajectories between 803M and D863F: a Pleistocene population bottleneck (Ne trough of approximately 12,000–15,000, occurring roughly 50,000–80,000 years ago), an ancient effective population size of 110,000–125,000, and partial recovery to approximately 35,000–40,000 individuals by roughly 10,000 years ago (Fig. 2).

Submission-format validation of the annotated assemblies with table2asn (v1.28.1094) returned zero errors for both individuals (803M: 0 errors, 8,842 warnings; D863F: 0 errors, 4,281 warnings); remaining warnings were reviewed and did not indicate structural problems with the assemblies or gene models.

Usage Notes

These assemblies are provided as reference-quality genomic resources for Black Bengal Goat and related caprine research. Users conducting fine-scale variant calling against the D863F assembly should be aware of its comparatively lower Merqury consensus QV relative to 803M (see Technical Validation) and may wish to apply additional quality filtering. Gene annotations reflect RNA-seq-evidence-only BRAKER1 predictions and should be considered a first-pass annotation; protein-mode BUSCO completeness (41.7% for 803M, 64.5% for D863F) indicates annotation completeness lags assembly completeness, likely reflecting tissue/expression coverage limitations in the RNA-seq evidence base (PRJEB23196) rather than assembly quality.

Declarations

Code Availability

Custom scripts used for assembly header renaming, gene-model edge/length quality filtering, and pipeline orchestration are available at https://github.com/sshil1/bbgoat-assembly-scripts, archived on Zenodo (DOI: 10.5281/zenodo.22893527). All other software used is publicly available and cited above with version numbers; standard parameters were used except where noted.

Author Contributions

Sharadindu Shil: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing.

Gouri Shankar Koner: Investigation, Resources.

Manoranjan Roy: Resources, Validation.

Biswajit Saha: Investigation, Resources.

Diganta Pan: Investigation, Resources.

Jayanta Chowdhury: Project administration, Resources.

Anusuya Debnath: Writing – review & editing.

Ayan Mukherjee: Investigation, Resources.

Shankari Prasad Dutta: Investigation, Resources.

Kasturi Mandal: Data curation, Investigation.

Ajoy Mandal: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing.

Santanu Banik: Validation, Writing – review & editing.

Competing Interests

The authors declare no competing interests.

Acknowledgements

The authors thankfully acknowledge DAH&VS, WB, Director, NDRI and MD, WBLDCL.

References

1. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods, 18(2), 170–175.
2. Alonge, M. et al. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biology, 23, 258.
3. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Molecular Biology and Evolution, 38(10), 4647–4654.
4. Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biology, 21, 245.
5. Astashyn, A. et al. (2024). Rapid and sensitive detection of genome contamination at scale with FCS-GX. Genome Biology, 25, 60.
6. Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). Whole-genome annotation with BRAKER. Methods in Molecular Biology, 1962, 65–95.
7. Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884–i890.
8. Dobin, A. et al. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15–21.
9. Muriuki C, Bush SJ, Salavati M, McCulloch MEB, Lisowski ZM, Agaba M, Djikeng A, Hume DA and Clark EL (2019) A Mini-Atlas of Gene Expression for the Domestic Goat (Capra hircus). Front. Genet. 10:1080. doi: 10.3389/fgene.2019.01080
10. Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. F1000Research, 9, 304.
11. Santiago, E. et al. (2020). Recent demographic history inferred by high-resolution analysis of linkage disequilibrium. Molecular Biology and Evolution, 37(12), 3642–3653.

12. Barbato, M., Orozco-terWengel, P., Tapio, M., & Bruford, M. W. (2015). SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. Frontiers in Genetics, 6, 109.
13. Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature, 475(7357), 493–496.
14. Alberto, F. J. et al. (2018). Convergent genomic signatures of domestication in sheep and goats. Nature Communications, 9, 813.
15. Shil, S. et al. sshil1/bbgoat-assembly-scripts: assembly header renaming, gene-model QC filtering, and pipeline orchestration scripts. Zenodo. https://doi.org/10.5281/zenodo.22893527

Figures

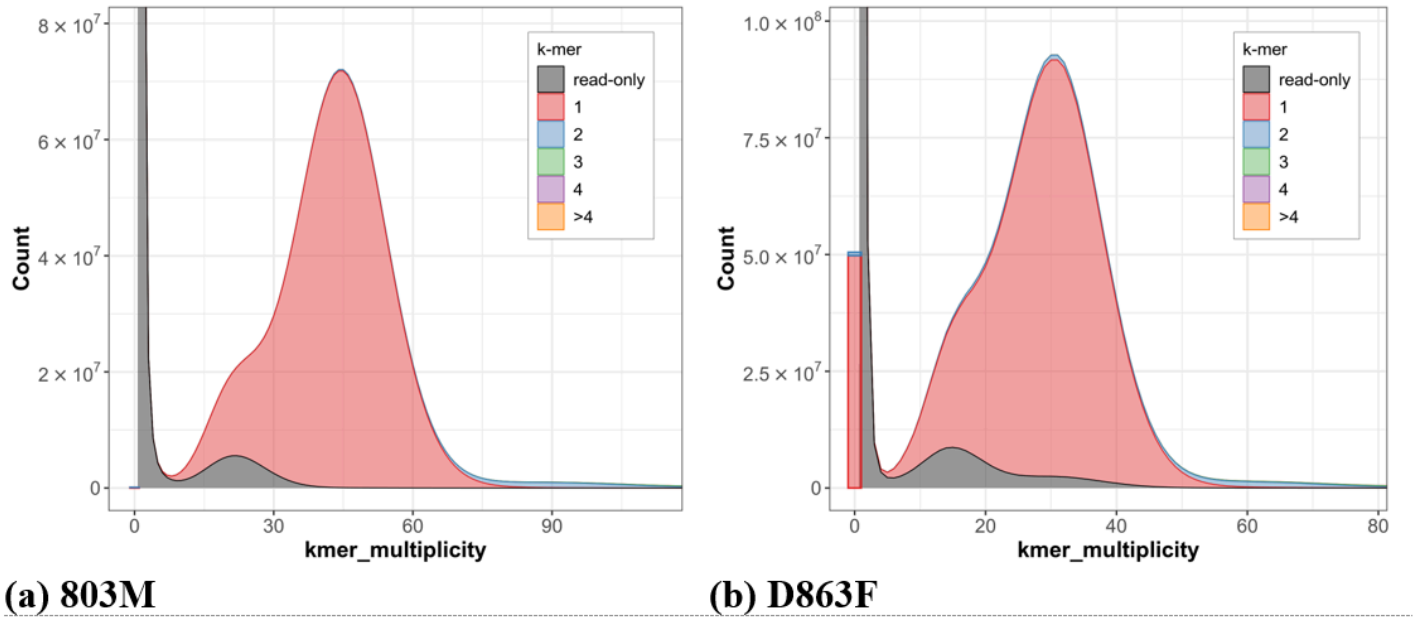


Figure 1

Merqury k-mer copy-number spectra plots for (a) 803M and (b) D863F, generated from independently built, individual-specific Illumina k-mer databases (k = 21). Colors indicate the copy number at which each k-mer is found in the assembly (0 = read-only, i.e. absent from the assembly; 1–4 and >4 = present at that copy number); the height and shape of the copy-number-1 peak reflect assembly base-level accuracy and underlie the Merqury consensus QV values reported in the text.

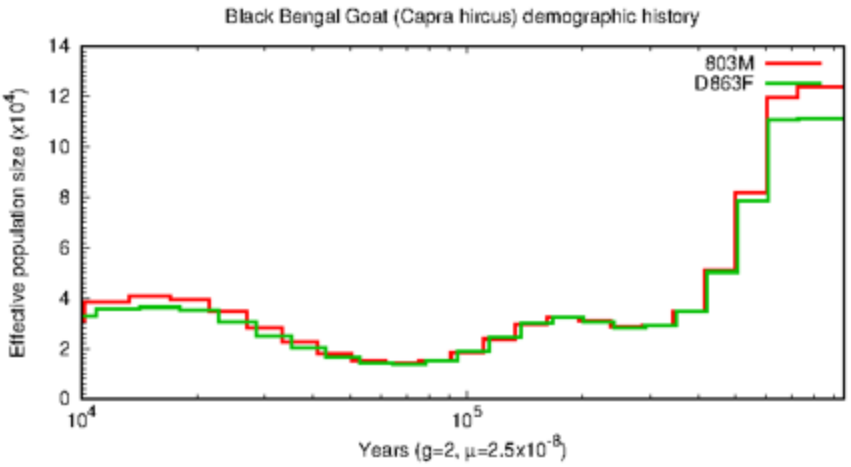


Figure 2

PSMC-inferred demographic history of Black Bengal Goat, showing effective population size (Ne, ×10⁴) through time for 803M (red) and D863F (green), assuming a per-generation mutation rate of 2.5×10⁻⁸ and a generation time of 2 years. Both individuals show a concordant Pleistocene bottleneck followed by partial recovery in effective population size.