Indian Journal of Animal Sciences **95** (2): 155–158, February 2025/Article
https://doi.org/10.56093/ijans.v95i2.140822

Genome assembly of indigenous Gaddi dog

KANWALJIT RANA[1] and C S MUKHOPADHYAY[1]✉

Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana, Punjab-141002

Received: 14 August 2023; Accepted: 13 February 2025

ABSTRACT

Dog is one of the oldest domesticated animals and the most familiar pet to mankind. The Gaddi dogs are robust and healthy and have been associated with herding and guarding for ancient times in the Himalayan vis-à-vis neighboring states like Himachal Pradesh, Uttarakhand, and Punjab. No information is available about the genetic architecture of these indigenous dogs. The present study was designed to sequence and assemble nuclear genome of Indigenous Gaddi dog for its genetic characterization. The blood samples were collected from five unrelated Gaddi dogs from the state of Punjab and Himachal Pradesh. The Illumina 150bp paired-end sequencing was done and 491.9 G raw data was generated with an error rate of 0.03% for all five samples. The quality of data was checked by using FastQC and Fastp tools. The Q30 of the processed data ranged from 90.5% to 93.3% and %GC from 40.1% to 41.5%. The genome assembly was done using Maryland Super Reads Celera Assembler (MaSuRCA). The number of contigs obtained for individual genome assemblies varied from 437250 to 455342, the N50 value ranged from 10028 to 10839 and the average contig length ranged from 5118 to 5344 bases. The raw sequence reads have been submitted to NCBI Bioproject id PRJNA843534 (SRA Accession: SRR22387066 to SRR22387070). The genetic tapestry of the Gaddi dogs has been revealed through this sequencing shot. Further, the enrichment of the genome assembly and annotation can have a deeper insight into the Gaddi dog's hidden potential and diverse adaptation behaviour.

Keywords: DNA, Gaddi dog, Genome assembly, Indigenous, Sequencing

Dog (*Canis lupus familiaris*) is the most familiar pet to man. There are about 400 breeds of domesticated dogs that exhibit considerable variations in morphological, physiological, and behavioral characteristics (Vaysse *et al.* 2011). The Gaddi dog breed (INDIA_DOG_0600_GADDI_19004) is robust and healthy and has been associated with herding and guarding for ancient times in the Himalayan vis-a-vis neighboring states like Punjab (Mukhopadhyay 2022). A healthy German Shepherd female genome was assembled by Field *et al.* (2020) as a reference genome to conduct disease and evolutionary studies in the future. The improved canid reference genome (CanFamGSD) was obtained using a combination of Pacific Bioscience, Oxford Nanopore, 10X Genomics, Bionano, and Hi-C technologies sequencing approaches (English *et al.* 2012). The issues associated with short-read assembly, the different types of data produced by second-generation sequencers, and the latest assembly algorithms designed for these data were summarized by Schatz *et al.* (2010). Zimin *et al.* (2013) evaluated the performance of Maryland Super Reads Celera Assembler (MaSuRCA) against two of the most widely used assemblers for Illumina data, Allpaths-LG and SOAPdenovo2, on two datasets from organisms for which high-quality assemblies are available: the bacterium *Rhodobacter sphaeroides* (Srikanth *et al.* 2022) and chromosome 16 of the mouse genome (Linacre 2021). Jung *et al.* (2020) summarised the approach genome assembly and annotation in time efficient manner. Leeb *et al.* (2022) highlighted the genetic factors related to skin disorders in dogs. Meadows *et al.* (2023) analyzed the genetic architecture and disease susceptibility (Kaur *et al*. 2023) in canine population through international Dog10K project. Sandhu *et al.* (2025) evaluated the microsatellite markers for breed identification in Gaddi dogs and other exotic dog breeds. Rana *et al.* (2025*)* uncovered the miRNAs from Gaddi dog genome.

Minimal information is available about the genetic architecture of indigenous dogs (Kalambhe *et al.* 2022, Tewari and Mukhopadhyay 2023). Sankhyan *et al.* (2022) documented the phenotypic characteristics of Gaddi dogs and highlighted the importance for Gaddi dog conservation and propagation. The indigenous Gaddi dog, being well adapted to the harsh terrains of the Himalayan region (Kaur *et al.* 2022a), has been the focus of this study. The genomics of these dogs have been unveiled for the first time through this effort.

MATERIAL AND METHODS

Sample Collection and DNA extraction: The five

Present address: [1]Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana, Punjab-141002. ✉Corresponding author email: csmbioinfo@gmail.com

unrelated Gaddi dogs were used for sample collection. Permission from the concerned Institutional Animal Ethics Committee (IAEC) was obtained for the same. The blood samples were aseptically collected into 0.5 M EDTA vials and were further processed for DNA isolation through the standard PCI DNA extraction method (Sambrook and Russel 2001). The quality and quantity of the DNA were checked using gel electrophoresis and Nanodrop method and were further sent for sequencing.

Whole genome sequencing: The quality checked DNA samples were used for further sequencing.

Library preparation and quality control: The genomic DNA was fragmented randomly through sonication and was end polished, A-tailed followed by Illumina adapter ligation. Further PCR amplification was done using P5 and P7 index primers and final products were purified with AMPure XP systems. Final library constructs were analyzed by Agilent 2100 Bioanalyzer and quantified by real-time PCR.

Sequencing: The final libraries were loaded onto Novaseq 6000 sequencer for Illumina 150bp paired-end sequencing and the expected data volume was obtained. The raw data was obtained and subjected to further analysis.

Quality checking and processing of raw data: The raw data obtained through sequencing was further subjected to quality checking using the FastQC and Fastp tools onto the LINUX terminal. The quality of the raw data was further improved by trimming the adapter sequences and rejecting the low-quality reads.

Whole Genome Assembly: All 5 samples were individually assembled using MaSuRCA (Maryland Super Read Cabog Assembler) genome assembler. The Masurca toolkit uses QUORUM error corrector for Illumina data, Chromosome scaffolder, jellyfish mer counter, and MUMmer aligner. The MaSuRCA assembler combines the benefits of the de Bruijn graph and Overlap-Layout-Consensus assembly approaches for short Illumina reads (Lyu *et al.* 2021). Wang *et al.* (2019) had earlier sequenced and de novo assembled hog deer genome sequences using six different insert-size libraries.

Relevant information is provided using the following parameters: Number of Bases in the assembly 2. N50 3. Number of Contigs 4. Maximum Contig size. Marcais *et al.* (2015) compared QUORUM against several published error correctors and found it suitable for large data sets. Marcais and Kingsford (2011) proposed a fast and memory-efficient, Jellyfish having a k-mer counting algorithm and associated implementation. In a study by Chen *et al.* (2020), MaSuRCA performed better resulting in contiguous genomes.

Genome Assembly Assessment: The clean raw data was used to assemble the individual genomes of Gaddi dogs. Further, the quality of the genome assembly was assessed using the QUAST (Quality Assessment Tool) (Gurevich *et al.* 2013) (http://quast.sourceforge.net/quast), installed and used on the Linux terminal.

The tool evaluates genome assemblies by computing various metrics. The CanFam6 dog assembly available at NCBI was downloaded and used as a reference genome for QUAST analysis. The output files show comparative metrics i.e. GC content, N50, total nucleotide length, etc.

RESULTS AND DISCUSSION

Quality checking and filtration of Raw sequencing data: The raw data was obtained in Fastq format for the five whole genome samples. The quality of the raw data was checked using the Fastp tool.

The raw sequence reads were checked by using Fastp tool and about 99.40% clean reads were obtained for the **GKSn1** sample. The N-containing reads were minimal, 20,720, and low-quality reads were 3362 in number. The adapter content observed was 0.60%. Again, the 99.40% clean reads were obtained for **GKSp2** with minimal N containing and raw reads, 19666 and 3122 respectively with adapter content of 0.60%. The 99.43% clean reads were obtained in the **GKNh3** sample i.e. 705119160 reads with 0.60% adapter content and minimum N containing and low-quality reads, 21534 and 3576, respectively. The **GKG4** sample resulted in 99.36% clean reads with 0.64% adapter content and minimum low-quality reads. For sample number five, **GKC5**, 99.36 % of clean reads were generated, and an adapter content of 0.63%. The N-containing and low-quality reads were again minimal i.e. 20104 and 2926, respectively. The average Q30 (Phred score) for raw sequences of five samples ranged from 90.16% to 93.18%. The number of reads varied from 606 M to 709 M.

Raw data processing summary: The raw data was processed using the Fastp tool and all the low-quality reads along with other contaminants were removed for five samples. Clean reads were obtained and no adapter sequences were observed after the adapter sequences were trimmed (Table 1).

Table 1. Raw data quality check (QC) summary obtained after whole genome sequencing

Sample ID	Reads	Yield (Bases)	Q30
GKSn1	707.1 M	106.07 G	92.55%
GKSp2	648.7 M	97.3 G	92.97%
GKNh3	709.1 M	106.3 G	93.18%
GKG4	608.2 M	91.2 G	90.16%
GKC5	606.2 M	90.9 G	90.61%

The total clean reads obtained after data processing range from 596.2 M to 700.7 M reads. The GC content for five samples ranges from 40.1% to 41.5%, which was ideally near the theoretical GC content (Table 2).

Whole Genome Assembly: The raw data was filtered and again the quality of the raw data was checked using the FastQC tool. Clean reads without any contaminating sequences were selected for further analysis. The genome assembly was done using the Maryland Super Reads

Table 2. Data quality check statistics after raw data processing

Parameter	GKSn1	GKSp2	GKNh3	GKG4	GKC5
Total_reads	697.6 M	640.7 M	700.7 M	597.5 M	596.2 M
Total_bases	97.5 G	89.6 G	98.03 G	83.5 G	83.4 G
Q20_bases	95.1 G	87.5 G	95.7 G	80.5 G	80.5 G
Q30_bases	90.5 G	83.5 G	91.5 G	75.7 G	75.8 G
Q20_rate	97.4%	97.6%	97.7%	96.3%	96.5%
Q30_rate	92.7%	93.1%	93.3%	90.5%	90.9%
GC_content	40.1%	40.1%	40.2%	41.0%	41.5%

Celera Assembler (MaSuRCA). The de novo approach was used to assemble the individual Gaddi dog genomes. The Masurca assembler produced a contig-level assembly. Five individual genome assemblies were obtained.

Genome Assembly: The MaSuRCA assembler combines the benefits of the de Bruijn graph and Overlap-Layout-Consensus assembly approaches for short Illumina reads. The Masurca toolkit uses QUORUM error corrector for Illumina data, Chromosome scaffolder, jellyfish mer counter, and MUMmer aligner.

Summary of five assembled genome:

The five individual assemblies of contig level were obtained through the MaSuRCA assembler. Contig level assemblies were obtained with approximately 0.4 million contigs per assembly and N50 ranges 10028 to 10839 (Table 3).

Genome Assembly Assessment: The assembled Gaddi genome was assessed for its assembly quality using the Quality Assessment Tool (QUAST). The reference dog genome CanFam6 was downloaded from the ensemble and the comparative output was obtained from the QUAST output. The QUAST was installed locally on the Linux system and the Gaddi dog assembly was assessed along with the CanFam6 dog assembly.

The total genome length observed in the case of the Gaddi dog was 2337014036 while it was 231280198 for CanFam6. The total size of the Gaddi dog genome obtained was 2.2G while 2.31G for the reference genome. The %GC content observed in Gaddi dog assembly was 40.70% whereas it was 41.26% for reference dog genome assembly.

The indigenous canine breeds have not been characterized at the molecular level. The genomic architecture of indigenous dogs has not been explored. No information was available about the genetic features of Indigenous dogs. Exotic canine breeds like German shepherd, Labrador retriever, Boxer, Basenji, etc had been well-characterized, assemblies are available on NCBI since the first genome availability of canine in 2004.

The new GSD assembly was approximately 80 times as contiguous as the assembled canid reference genome and contained far lesser gaps (306 vs 23,876) and scaffolds (429 vs 3,310) than the CanFamv3.1. Using 10X Genomics linked reads, genome sequence was assembled resulting in 2.72 Gb in length (contig N50, 66.04 Kb and scaffold N50, 20.55 Mb), in which expected genes were detected up to 94.5%. The annotation of 22,473 protein-coding genes, 37,019 tRNAs, and 1,058 Mb repeated sequences were completed. Low-cost high-quality reference genomes for the African wild dog (*Lycaonpictus*) were generated (Armstrong *et al.* 2019). Kaur *et al.* (2023b) cariied out genetic diversity analysis through SNPs genome-wide association in exotic dog breeds and Gaddi dogs.

Dogs have been associated with humans for ages and served the humans in guarding and hunting activities. They have also well adapted to different geographical terrains resulting in diverse genomic architecture. In the present study, the Gaddi dog genome sequence was assembled de novo to the contig level, size 2.2Gb using Illumina short reads 150bp paired-end data. It had a genome size of 2.2 Gb which was comparable to the indigenous dog, Gaddi's genome has been assembled for the first time. The already available exotic dog breed's genome size varies from 2.3 to 2.4 Gb and assemblies up to chromosome level. The raw sequence reads for five whole genome samples have been submitted to NCBI-SRA (Bioproject id: PRJNA843534) with SRA Accession Nos. SRR22387066 to SRR22387070. Various aspects of the Gaddi dog genome could be explored to study the evolutionary relationship and also disease association studies. Further validation studies for the predicted sequences can pave the way to getting a deeper insight into the Indigenous Gaddi dog genome.

ACKNOWLEDGMENTS

We are thankful to the Department of Biotechnology, Government of India (DBT-GOI), for providing funding for the research through the scheme DBT-19-I "Parentage determination and cytogenetic profiling in dogs." We are also grateful to the pet owners for providing the sample collection support.

Table 3. Assembly statistics for five whole genome sequencing (WGS) samples

Stats	GKSn1	GKSp2	GKNh3	GKG4	GKC5
Contigs	437250	455342	449629	453466	443189
Bases	2337014036	2330501545	2341260274	2322267954	2320758305
Avg	5344.80	5118.13	5207.09	5121.15	5236.50
Max	96344	110282	119030	105899	129985
N50	10839	10141	10781	10028	10279
N60	8599	8053	8509	7954	8143

REFERENCES

Armstrong E E, Taylor R W, Prost S, Blinston P, Meer Van Der E, Madzikanda H, Mufute O, Mandisodza-Chikerema R, Stuelpnagel J, Sillero-Zubiri C and Petrov D. 2019. Cost-effective assembly of the African wild dog (*Lycaonpictus*) genome using linked reads. *GigaScience* **8**(2), giy124. https://doi.org/10.1093/gigascience/giy124

Chen Z, Erickson D L and Meng J. 2020. Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing. *BMC genomics* **21**(1): 631. https://doi.org/10.1186/s12864-020-07041-8

English A C, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny D M, Reid J G, Worley K C and Gibbs R A. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PloS one* **7**(11): e47768.

Field M A, Rosen B D, Dudchenko O, Chan E K, Minoche A E, Edwards R J, Barton K, Lyons R J, Tuipulotu D E, Hayes V M, Omer A D, Colaric Z, Keilwagen J, Skvortsova K, Bogdanovic O, Smith M A, Aiden E L, Smith T P L, Zammit R A and Ballard J W O. 2020. Canfam_GSD: *De novo* chromosome-length genome assembly of the German Shepherd Dog (*Canis lupus familiaris*) using a combination of long reads, optical mapping, and Hi-C. *GigaScience*: **9**(4). https://doi.org/10.1093/gigascience/giaa027

Gurevich A, Saveliev V, Vyahhi N and Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**(8): 1072–5. https://doi.org/10.1093/bioinformatics/btt086

Jung H, Ventura T, Chung J S, Kim W J, Nam B H, Kong H J, Kim Y O, Jeon M S and Eyun S I. 2020. Twelve quick steps for genome assembly and annotation in the classroom. *PLoS computational biology* **16**(11): e1008325. https://doi.org/10.1371/journal.pcbi.1008325.

Kalambhe D, Bhanot R, Malik H, Mukhopadhyay C S and Gill J P S. 2022. The indigenous dogs of india: the forgotten treasure. *Project Monitoring Unit- DBT-GADVASU-CRC, Directorate of Research, Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana*. ISBN: 978-93-5659-734-1

Kaur B, Kaur J, Kashyap N, Arora J S and Mukhopadhyay C S. 2023a. A comprehensive review of genomic perspectives of canine diseases as a model to study human disorders. *Canadian Journal of Veterinary Research* **87**(1): 3-8.

Kaur B, Rana K and Mukhopadhyay C S. 2022. Ancestry, temperament, facts, and grooming of Indigenous 'Gaddi'dog breed. *Veterinary alumnus* **44**(2): 86-8.

Kaur B, Yathish H M, Kashyap N and Mukhopadhyay C S. 2023b. Phylogeographic and genetic-diversity analysis through genome-wide snps in indigenous and exotic canine breeds. http://dx.doi.org/10.2139/ssrn.4615672

Leeb T, Roosje P and Welle M. 2022. Genetics of inherited skin disorders in dogs. *Veterinary Journal* **279**: 105782. doi: 10.1016/j.tvjl.2021.105782

Linacre A. 2021. Animal Forensic Genetics. *Genes (Basel)* **12**(4): 515. doi: 10.3390/genes12040515

Lyu G, Feng C, Zhu S, Ren S, Dang W, Irwin D M, Wang Z and Zhang S. 2021. Whole Genome Sequencing Reveals Signatures for Artificial Selection for Different Sizes in Japanese Primitive Dog. Breeds. *Frontiers in Genetics* **12**: 671686. doi: 10.3389/fgene.2021.671686.

Marcais G, Yorke J A and Zimin A. 2015. QuorUM: An Error Corrector for Illumina Reads. *PLoS ONE* **10**(6): e0130821. https://doi.org/10.1371/journal.pone.0130821

Marcais G and Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**(6): 764–70. https://doi.org/10.1093/bioinformatics/btr011

Meadows J R S, Kidd J M, Wang G D, Parker H G, Schall P Z, Bianchi M, Christmas M J, Bougiouri K, Buckley R M, Hitte C, Nguyen A K, Wang C, Jagannathan V, Niskanen J E, Frantz L A F, Arumilli M, Hundi S, Lindblad-Toh K, Ginja C, Agustina K K, and Ostrander, E A 2023. Genome sequencing of 2000 canids by the Dog10K consortium advances the understanding of demography, genome function and architecture. *Genome biology*, **24**(1), 187. https://doi.org/10.1186/s13059-023-03023-7.

Mukhopadhyay C S. 2022. Genomic Characterization of Gaddi breed of Dog to Envisage Molecular Signatures. (**Editorial**). *Acta Scientific Veterinary Sciences* **4**(10): 1-2. DOI: 10.31080/ASVS.2022.04.0507

Rana K, Randhawa S S, Mahindroo J, Sethi R S and Mukhopadhyay C S. 2025. Biocomputational identification of microRNAs from indigenous Gaddi dog genome. *Gene Reports* **39**: 102167. https://doi.org/10.1016/j.genrep.2025.102167

Sambrook J and Russell D W. 2001. Molecular Cloning: A Laboratory Manual. *Cold Spring Harbor Laboratory Press, New York* **3**(1): 2344.

Sankhyan V, Thakur R, Dogra P K and Thakur A. 2022. Phenotypic characterization and documentation of Gaddi dog of western Himalayan region of India. *The Indian Journal of Animal Sciences* **92**(10): 1189-93.

Schatz M C, Delcher A L and Salzberg S L. 2010. Assembly of large genomes using second-generation sequencing. *Genome research* **20**(9): 1165–73. https://doi.org/10.1101/gr.101360.109

Sandhu Y, Kaur B, Kaur M, Yathish H M and Mukhopadhyay C S. 2025. Microsatellite DNA Analysis of Genetic Diversity and Parentage Testing in Popular Dog Breeds in India. *Indian Journal of Animal Research* **B-5477**: 1-8. doi 10.18805/IJAR.B-5477

Srikanth K, von Pfeil D J F, Stanley B J, Griffitts C and Huson H J. 2022. Genome-Wide Association Study with Imputed Whole Genome Sequence Data Identifies a 431 kb Risk Haplotype on CFA18 for Congenital Laryngeal Paralysis in Alaskan Sled Dogs. *Genes (Basel)* **13**(10): 1808. doi: 10.3390/genes13101808

Tewari S and Mukhopadhyay C S. 2023. *In silico* Mining of Protein-coding and Non-coding RNA (ncRNA) Specific Genes in Exotic versus Indigenous Gaddi Dogs. *Current Biotechnology* **12**(3): 190-202.

Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Pielberg G R, Sigurdsson S, Fall T, H. Seppälä E H, Hansen M S T, Lawley C T, Karlsson E K, The LUPA Consortium, Bannasch D, Vilà C, Lohi H, Galibert F, Fredholm M, Häggström J, Hedhammar A, André C, Lindblad-Toh K, Hitte C and Webster M T. 2011. Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS genetics* **7**(10): e1002316.

Wang W, Yan H J, Chen S Y, Li Z Z, Yi J, Niu L L, Deng J P, Chen W G, Pu Y, Jia X, Qu Y, Chen A, Zhong Y, Yu X M, Pang S, Huang W L, Han Y, Liu G J and Yu J Q. 2019. The sequence and de novo assembly of hog deer genome. *Scientific data*, **6**, 180305. https://doi.org/10.1038/sdata.2018.305

Zimin A V, Marçais G, Puiu D, Roberts M, Salzberg S L and Yorke J A. 2013. The MaSuRCA genome assembler. *Bioinformatics* **29**(21): 2669–77. https://doi.org/10.1093/bioinformatics/btt476