

RESEARCH

Open Access



Genomic insights into six dog breeds owned in India: a ddRAD sequencing approach

Bhawanpreet Kaur¹, SK Mahajan² and CS Mukhopadhyay^{1*}

Abstract

The current study reports the maiden investigation aimed at molecular characterization of population structure and the discovery of tag-SNPs in the indigenous Gaddi and Mudhol Hound breeds, along with four other foreign breeds owned in India, using genome-wide SNP markers. Forty-six (46) dogs were genotyped using genotyping-by-sequencing double-digest restriction site-associated DNA (GBS-ddRAD), which identified 75,811 high-quality SNPs. Genetic analyses of the SNP data indicated significant variability within and between the experimental breeds, with indigenous breeds having a higher proportion of unique genetic variants. Linkage disequilibrium can indicate how genes are grouped on chromosomes and can help researchers understand genetic diversity and evolutionary relationships. A total of 2,033 tag-SNPs were identified across the six divergent dog breeds ($p < 10^{-6}$), representing the most informative markers within the dataset. These tag-SNPs were distributed across chromosomes 1–19, indicating regions of concentrated genetic variation, with the maximum (i.e., 242) number of tag-SNPs on chromosome 5 (Clu5: NC_051809.1). These findings echo the maiden report on tag-SNPs of Indian indigenous dog breeds (Gaddi dog and Mudhol Hound) and also provide comprehensive genomic insights. This knowledge base will further help researchers identify breed-specific molecular signatures and develop conservation strategies. It will also support future genome-wide association studies (GWAS) for trait and disease mapping in Indian dog breeds.

Keywords Gaddi Dog, Mudhol Hound, Tag-SNPs, Population diversity, Linkage disequilibrium

Introduction

Dog (*Canis lupus familiaris*) was the first animal domesticated by humans, with a shared evolutionary history dating back at least 10,000 years [1]. Breeding of domestic dogs has been practiced for a long time to perform a wide range of activities, such as sentinel guarding, hunting, and companionship, resulting in remarkable

phenotypic and behavioral variation [2, 3]. Animal genetics has advanced dramatically in recent years, with various molecular technologies allowing for deeper insights into the genome, a more refined understanding of genetic traits, disease susceptibility, and evolutionary patterns across species. Besides, dogs have become an important model for genetic studies. Recent advances in genomics have revolutionized canine genetic research, with single-nucleotide polymorphism (SNP) arrays emerging as a robust and cost-efficient tool for high-throughput genotyping in molecular breeding initiatives [4, 5]. A groundbreaking study examining 722 canine whole-genome sequences (WGS) has revealed remarkable insights into the genetic structure of pet dogs, documenting around 91 million SNPs and small insertion/deletion (INDEL)

*Correspondence:
CS Mukhopadhyay
csmbioinfo@gmail.com

¹Department of Bioinformatics, College of Animal Biotechnology, Ludhiana, India

²Department of Veterinary Surgery and Radiology, College of Veterinary Sciences, Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana, Punjab, India



variants [6, 7]. These studies emphasized that inbred organisms exhibit longer stretches of homozygosity, indicating reduced genetic polymorphism, stronger linkage disequilibrium (LD), and limited genetic exchange within specific regions of the genome. LD in animals refers to the non-random association of alleles at different genetic loci, resulting from their co-inheritance more often than would be expected by chance. Tag single-nucleotide polymorphisms (tag-SNPs) are specific SNPs chosen to represent larger sections of genetic variation across the genome. These representative marker-SNPs are useful in LD studies because they capture the genetic information of surrounding variants, making it easier to identify associations between genes and traits without analyzing every SNP. Relevant literature reports that the foreign canine genome exhibits extensive regions of high LD and reduced haplotype diversity, reflecting historical population bottlenecks and significant artificial selection pressures [8].

Recently, streamlined double-digest restriction site-associated DNA (ddRAD) techniques have attracted global attention for their cost-effective ability to identify genome-wide differences [9]. The Online Mendelian Inheritance in Animals (OMIA) database (as of November 2025) lists 994 dog traits (<https://www.omia.org/>) that may serve as analogs for human diseases [10]. The Dog10K consortium (with analysis of over 2,000 canid genomes) represents a collaborative effort to understand the genetic and phenotypic diversity of dogs.

In this study, two native Indian dog breeds, the Gaddi dog (from Himachal Pradesh) and the Mudhol Hound (from Karnataka), have been included. The Gaddi dog is recognized as a reliable means of protecting sheep and goats while grazing in mountainous terrains. The Mudhol Hound breed is popular as a hunting and guard dog. Limited reports are available on Indian dog breeds, especially on genomics and comparative studies of molecular markers across Indian and foreign breeds. A systematic and comprehensive study aimed at identifying SNPs and tag-SNPs in both native and foreign dog breeds has not yet been undertaken. It is worth noting that our lab has contributed to the molecular characterization and genome sequencing of the Gaddi dog [11–17].

To fill this knowledge gap, the present research was designed to explore genome-wide SNPs, identify patterns of LD and examine haplotype arrangements in divergent dog breeds maintained in India. Thus, this study has been conducted to provide comparative assessments of genomic variability and the evolutionary history of indigenous and foreign dog breeds, with important implications for conservation, management, breeding programs, and future genome-wide association studies.

Methods

Experimental design and DNA isolation

Peripheral blood samples were collected in vacutainer tubes containing anticoagulant, from 46 healthy dogs belonging to six divergent breeds (Labrador Retriever (Lab, $n=8$), Pug (Pug, $n=9$), German Shepherd (GS, $n=12$), Gaddi Dog (GD, $n=7$), Tibetan Mastiff (TM, $n=3$), and Mudhol Hound (MH, $n=7$) and corresponding to four states of India (Fig. 1; Supplementary Tables S1, S2). Blood collection was conducted in accordance with the guidelines and regulations ratified by the Institutional Animal Ethics Committee (IAEC) of Guru Angad Dev Veterinary and Animal Sciences University (GADVASU), Ludhiana, Punjab, India (registration number 497/GO/Re/SL/02/CPCSEA) (IAEC Application number: GADVASU/2021/IAEC/62/02). The phenol-chloroform method was used to extract genomic DNA from peripheral blood, followed by treatment with RNase and purification using the Qiaquick Nuclease Removal Kit (Qiagen, Valencia, CA) to eliminate RNA impurities. The quality of the extracted DNA was assessed using 1% agarose gel electrophoresis, and the DNA concentration was measured with a NanoDrop spectrophotometer (Thermo Scientific NanoDrop One, USA).

Library preparation and double-digest restriction site-associated DNA sequencing (ddRAD-seq)

The DNA extracted from 46 samples was sent to AgriGenome Laboratory Pvt. Ltd., Kerala, for ddRAD-genotyping by sequencing. Reduced representation sequencing (RAD-Seq) is a highly effective method for genotyping and gene mapping across organisms, even when a reference genome is not available. Each sample underwent double digestion with the restriction enzymes SphI and MluI (*Streptomyces phaeochromogenes* and *Micrococcus luteus*) during DNA library preparation [9]. After digestion, the fragments ends were ligated to specific P1 and P2 barcoded adapters using T4 DNA ligase. The ligation process was allowed to take place overnight (for more than 12 h). The mixture was maintained at room temperature (around 21 °C) and subsequently heat-inactivated at 63 °C for 10 min. To remove any unincorporated adapters and smaller DNA fragments, the ligation products were purified with AMPure magnetic beads.

A unique dual-indexed barcode combination was applied to purified fragments through 14 cycles of PCR. The indexed PCR products were mixed in equal amounts and size-selected using Agencourt AMPure XP SPRI magnetic beads. The amplification procedure begins with an initial denaturation step at 95 °C for 3 min, followed by 25 cycles of denaturation at 95 °C for 30 s, annealing at 55 °C for 30 s, and extension at 72 °C for 30 s, culminating in a final extension at 72 °C for 5 min. To determine the effective concentration of each library, quantitative

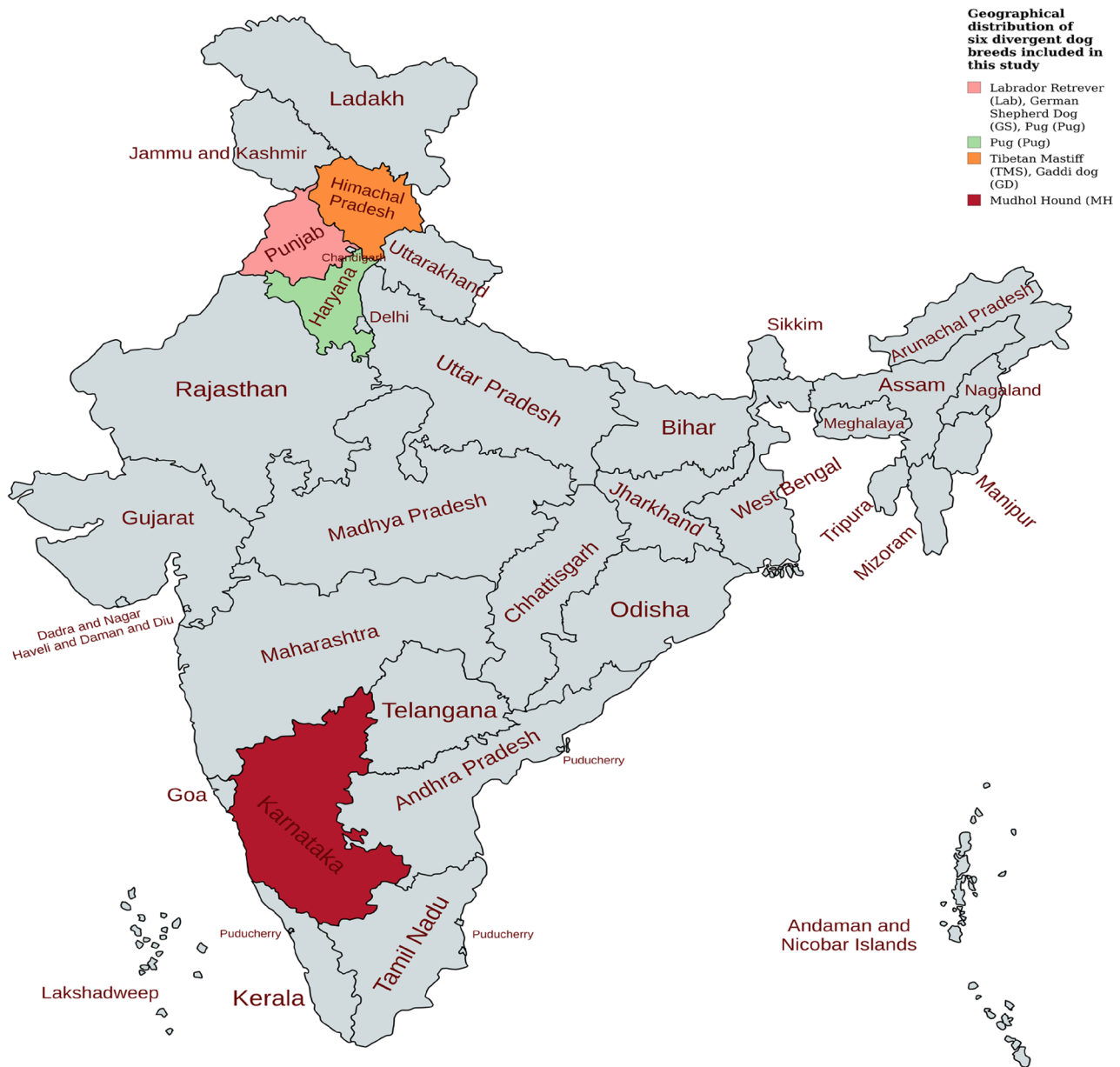


Fig. 1 Geographical distribution of the collected samples of the six dog breeds across the regions of India (Labrador Retriever, German Shepherd, Pug, Tibetan Mastiff, Gaddi Dog, and Mudhol Hound)

real-time PCR (qPCR) was performed. The libraries with acceptable insert sizes and effective concentrations exceeding 2nM were sequenced using Illumina 150 bp paired-end sequencing at AgriGenome Laboratory Private Limited, Kerala.

Quality Control (QC), read filtering, and SNP identification

The initial paired-end FASTQ files were assessed for quality using FASTQC. To ensure data integrity, reads with a quality score less than Q30 were removed from the dataset. Quality filtering was performed using dDocent (v.2.6.0) and Trimmomatic (v.0.38), which removed

low-quality bases (score < 20) and adapter sequences. A sliding window of 5 base pairs was utilized, trimming bases whenever the average quality score fell below 10. After the quality filtering process, the leftover reads were mapped to the reference genome utilizing BWA (version 0.7.8). The reference genome CanFam_GSD_1.0 (GCF_014441545.1) was used.

SNP calling summary

SNP detection utilized dDocent, which uses FreeBayes (v1.2.0) for calling variants. Variant calling was initially performed with two minimum read-depth thresholds

($RD \geq 5$ and $RD \geq 10$) using BCFtools (v1.6) to assess how read-depth filtering influences SNP reliability. The $RD \geq 5$ dataset served as an exploratory quality-control measure to detect low-confidence variants, whereas the stricter $RD \geq 10$ threshold was chosen to reduce false-positive SNP calls. Variants identified at $RD \geq 5$ but not preserved, at $RD \geq 10$ were regarded as potentially unreliable and were omitted. All subsequent analyses were conducted solely with the $RD \geq 10$ dataset. This dataset additional filtering with minor allele frequency (MAF) ≥ 0.05 and Hardy-Weinberg equilibrium (HWE) $p \geq 1 \times 10^{-6}$ to preserve high-confidence SNPs for analyses of population structure, LD and tag-SNPs.

Diversity and population structure analysis

VCFtools was utilized to compute both observed heterozygosity (H_o) and expected heterozygosity (H_e) for the six different dog breeds. The diversity metrics and phylogenetic relationships among dog populations were examined using Trait Analysis by Association, Evolution, and Linkage (TASSEL) software (v5.0) in R [18]. A Neighbor-Joining tree with a bootstrap value of 100 based on unweighted pair group method with arithmetic mean (UPGMA) was constructed using TASSEL (v5.0). Additionally, Principal Component Analysis (PCA) was performed on the trimmed dataset using TASSEL. PCA was conducted in R (v4.2) using the prcomp() function. The summary() function was used to extract the eigenvalues and percentage variance explained by each component. UPGMA method was employed to construct the phylogenetic tree using the R programming environment (with 100 bootstrap replicates).

Linkage disequilibrium (LD) mapping

LD analysis, tag-SNP identification and, the estimation of LD scores were claculated using SNP data through the LD Score Regression tool (v1.0.1). The initial analysis and generation of input files were executed using PLINK [4, 11]. Additionally, the TASSEL GBSv2 pipeline (v.5.2.82) was implemented to conduct LD analysis, providing further insights into the genetic structure of the examined breeds [19].

Haplotype identification

Haplotype identification was carried out using a range of libraries in the R programming environment via the Ubuntu Linux terminal. SNP genotype information was assessed through Haploview v4.2 [5] and TASSEL, applying two key parameters: linkage disequilibrium (D') and squared r (r^2), which gauges the non-random association between alleles at distinct loci on a chromosome, for initial analysis [20]. These parameters are crucial for understanding the extent to which alleles at various positions on a chromosome are inherited together, more or less

frequently than expected by chance. The overall count of haplotypes identified, along with their graphical depiction, was recorded. LOD (logarithm of odds) scores were computed using allele frequencies derived from the genotype data. The LOD score indicates the $\log_{10}$ likelihood ratio that assesses the hypothesis of linkage against the absence of linkage between the loci. Frequencies of alleles were calculated individually for each locus.

SNP tagging evaluation

SNP tagging analysis identifies the smallest set of SNPs that provides the highest level of informativeness, known as “tag-SNPs.” The dataset was analyzed using Bash and R scripts, and the selected tag-SNPs were assessed for statistical significance (p -value) after removal of ambiguous, multi-nucleotide polymorphisms (MNP) and SNPs that were not informative. Utilizing R and Python commands in a Linux terminal, the tag-SNPs were mapped to their corresponding chromosome locations.

Results

Quality Control (QC), alignment, and SNP calling

All 46 samples passed the filtering criteria during the quality checking. QC metrics, including percentage of reads passing filters (%QC), and the number of reads post-trimming, are provided in Supplementary Table S3. The number of clean reads obtained after processing ranged from 0.91272 to 0.95262 (i.e., 91–95% of total reads) per sample.

The highest number of raw SNPs was observed in German Shepherds (1,962,853), followed by Pugs (1,686,939), Mudhol Hounds (1,589,552), Tibetan Mastiffs (1,619,515), Labrador Retrievers (1,404,371), and Gaddi Dogs (1,139,739). SNPs with a genotype quality (GQ) ≥ 30 were highest in German Shepherds (1,938,002), followed by Pugs (1,661,148), Mudhol Hounds (1,596,767), Tibetan Mastiffs (1,584,483), Labrador Retrievers (1,390,316), and Gaddi Dogs (1,125,220) (Table 1).

Table 1 The number of high-quality SNPs in each dog breed post series of filtering criteria

Breeds	Number of samples	Raw SNPs	SNP Genotype quality (GQ) ≥ 30
GS	12	19,62,853	19,38,002
Pug	9	16,86,939	16,61,148
MH	7	1,589,552	15,96,767
Lab	8	1,404,371	1,390,316
GD	7	11,39,739	11,25,220
TM	3	16,19,515	15,84,483
Total	46	94,02, 969	92,95, 936

* GS German Shepherd, Pug Pug, MH Mudhol Hound, Lab Labrador Retriever, GD Gaddi Dog, TM Tibetan Mastiff

After merging breed-wise datasets contained a total of 356,461 SNPs common to all analyzed breeds were retained. These overlapping SNPs were used for downstream comparative analyses. After initial quality filtering, 324,050 SNPs were observed at a read depth of 5 (RD 5), and further filtered at threshold read depth of 10 (RD 10), resulting in 313,694 high-confidence SNPs (Supplementary Table S4). Datasets with $RD \geq 10$ were selected to ensure analytical consistency and robustness. Additional filtering at RD 5 was conducted based on criteria like the proportion of missing genotypes, resulting in a final collection of 75,811 high-quality SNPs. The number of SNPs showed considerable variation among breeds following the filtering. Hardy–Weinberg Equilibrium (HWE) testing was utilized to evaluate differences in genotype frequencies across various loci. The histogram illustrating p -values supported this finding, showing concentrations at both the low ($p < 0.05$) and high ($p \approx 1.0$) ends of the distribution (Fig. 2). These findings suggest that while many loci conform to HWE expectations, a subset shows marked deviations, which may be attributed to biological influences such as selection, inbreeding, or population stratification, as well as to possible technical problems such as genotyping inaccuracies.

Nucleotide diversity and linkage disequilibrium analysis

Nucleotide diversity analysis, observed heterozygosity (H_o) and expected heterozygosity (H_e) for each of the six dog breeds were calculated using VCFtools and analyzed with TASSEL software (v. 5.0). The zygosity summary, based on RD10 genotype data and MAF threshold of 0.05 (Supplementary Tables S5 and S6), revealed SNPs

in strong LD ($r^2 > 0.5$). The UPGMA-based phylogenetic tree illustrated breed-specific clustering, with Labs, GS, and Pugs forming distinct and clearly separated clades, indicating significant genetic divergence and breed integrity. Conversely, the MH, GD, and TM groups are more closely aligned, suggesting a common evolutionary background or a stronger genetic link compared to the other breeds (Fig. 3). Overall, the tree emphasizes that while Labs, GS, and Pugs represent separate lineages, MH, GD, and TM appear more interrelated, suggesting a potential shared ancestry or breeding history. The bootstrap values for each node were nearly 100%, indicating the high reliability of the constructed tree.

PCA based on 2,033 tag-SNPs was used to evaluate the dataset's genetic architecture. The cumulative variance graph showed that the first 4 PCs accounted for most of the total variation, with more than 90% of the variance captured by the first 8–10 components. The first and the second PCs explained 38.5% and 22.1% of the total genetic variation, respectively (Fig. 4A, B, C; Table 2).

Linkage Disequilibrium (LD) analysis

LD referred to as gametic phase disequilibrium or gametic disequilibrium to the non-random association of alleles at various loci within a population. An initial examination of the ped and map files was performed using PLINK, followed by LD analysis conducted with TASSEL. The sequencing agency supplied input files in the HapMap format (*.hmp). LD calculations among SNPs were carried out using the r^2 statistic (Pearson product-moment correlation coefficient) [21]. TASSEL created a pairwise LD plot for marker sites, with

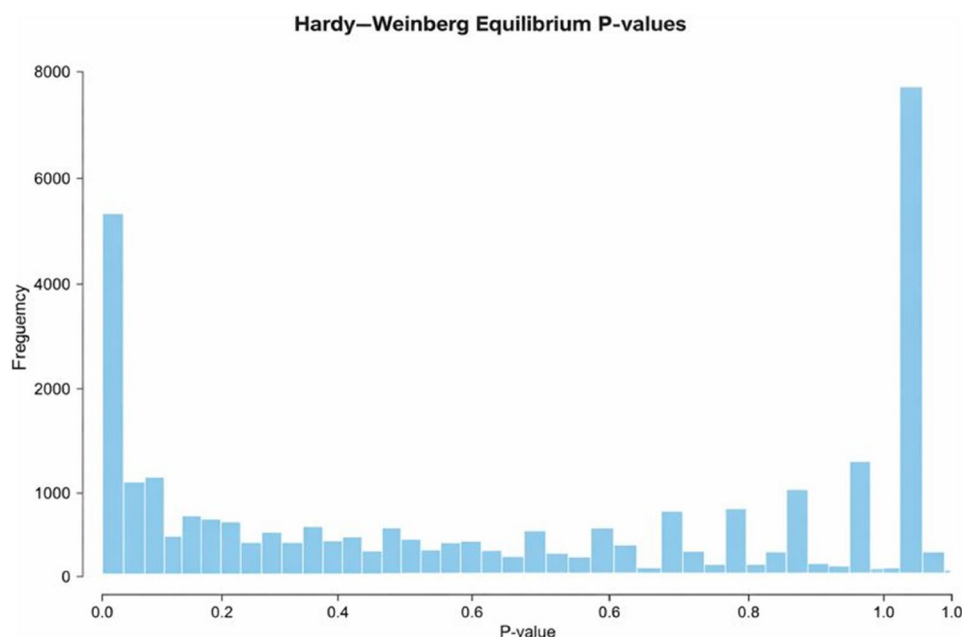


Fig. 2 Histogram of Hardy–Weinberg equilibrium (HWE) p -values across all loci

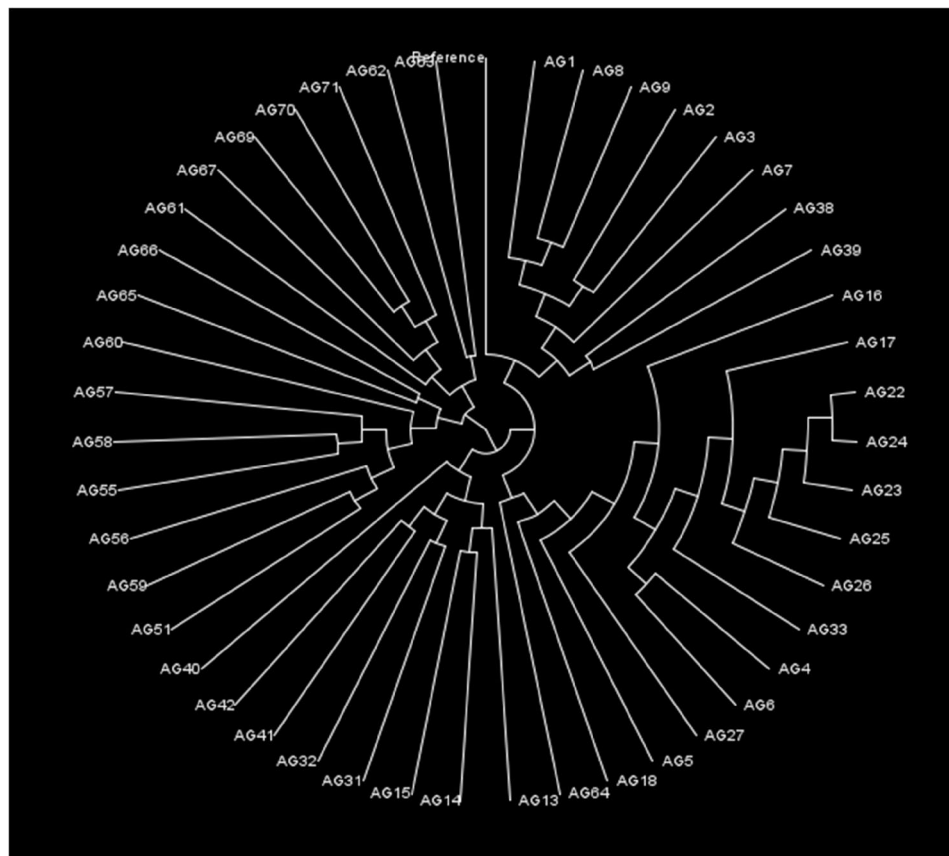


Fig. 3 Phylogenetic clustering of 46 dogs from six breeds was performed using the Neighbour-Joining method using TASSEL software

polymorphic sites represented on both the x-axis and y-axis. Above the diagonal, r^2 values were presented, while below it, the associated p -values derived from the rapid 1000-shuffle permutation test were displayed (Figs. 5 and 6). Each cell in the LD plot compared a pair of marker sites, employing color codes to signify levels of significant LD. A color gradient illustrated the significance threshold along both diagonals. Additionally, a genetic distance scale was provided for a theoretical tag-SNP, with darker zones indicating stronger LD and lighter colors indicating decreasing significance. Therefore, use of multiple LD metrics (D' , r^2 , and multi-loci measures) offers complementary insights into allele associations and recombination history. The identification of functional variants, precise haplotype block definition, and trustworthy genomic predictions for complex traits are all enhanced by this multi-metric approach [22–24].

Pairwise LD plots are crucial for evaluating LD patterns across numerous molecular markers. These plots are generally displayed as color-coded triangular diagrams, where notable pairwise LD levels—assessed through r^2 , p -value, and D' —aid in visualizing areas of strong LD (red blocks). Large red haplotype regions along the diagonal of the triangular plot indicate robust LD between loci within these regions, suggesting limited

or nonexistent recombination since the formation of the LD block. This indicates that the genetic material within these regions has largely remained stable over time.

The LD plots (Figs. 7 and 8) demonstrate these patterns, where lower p -values and higher r^2 values signify strong LD, making it easier to pinpoint loci clusters in high LD (red blocks). The r^2 and D' values range from 0 to 1, with 0 indicating linkage equilibrium (LE), which signifies statistical independence between loci, while 1 denotes complete LD. Strong block-like LD configurations are especially beneficial for association mapping, as they enhance the identification of genetic variants associated with complex traits.

Total LD score count

The overall count of LD scores was calculated in R, guided by the visualized LD plots. The strength of LD between SNP pairs was quantified using the squared correlation coefficient (r^2). A histogram (Fig. 9) illustrates the distribution of LD scores, measured using r^2 values. The majority of values are concentrated in the LD region, with values approaching 1, indicating strong LD. Conversely, the count is significantly lower in the linkage equilibrium region, where recombination has likely disrupted allele associations over time.

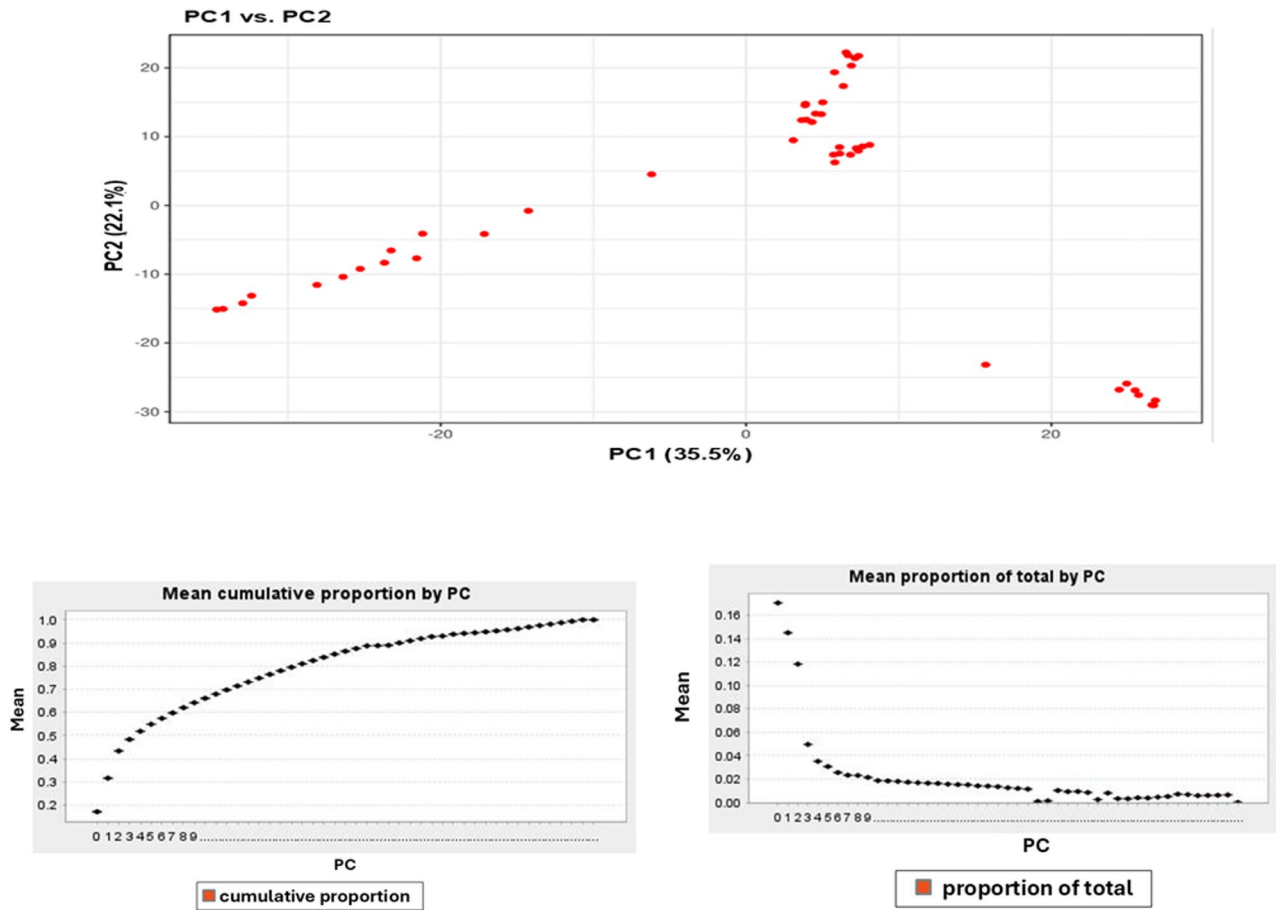


Fig. 4 Principal Component Analysis (PCA) of the dataset. **A** Bar plot showing the proportion of variance explained by each principal component. **B** Cumulative variance explained by successive principal components. **C** Scatter plot of individual samples projected along the first two principal components (PC1 and PC2), illustrating population structure among the analyzed dog breeds

Table 2 Proportion of variance (%) explained by the first and second principal components

Principal Component	Proportion of Variance (%)
PC1	38.5
PC2	22.1

Haplotype identification

Haplotypes consist of clusters of genetic variants that are inherited together on the same chromosome. Preliminary haplotype identification was conducted using Haploview v4.2, with analysis based on the LOD score [25]. As depicted in the histogram (Fig. 10), the graph showing the frequency distribution of LOD scores used in association mapping shows that most variants have low scores, with values less than 2, indicating no significant linkage between most loci. This suggests that most haplotypes are not linked.

Discovering the major SNP tagging

Genome-wide tag-SNP identification was performed using R and awk/sed scripts in a Linux environment. A comprehensive analysis of ddRAD sequencing data yielded 313,694 identified SNPs, which were subsequently sorted and filtered. Non-informative SNPs and MNPs, such as CTG<>GTG, were eliminated, resulting in a reduced SNP count of 70,268. The dataset was further refined by chromosome-wise sorting using R and bash scripts, resulting in 33,613 SNPs. The selection of tag-SNPs relied on *p*-values (0.000001) from LD scores, where lower *p*-values indicate greater significance (Supplementary Table S7). Ultimately, 2,033 tag-SNPs were recognized as the most significant. The *p*-value histogram for LD scores is shown in Fig. 11.

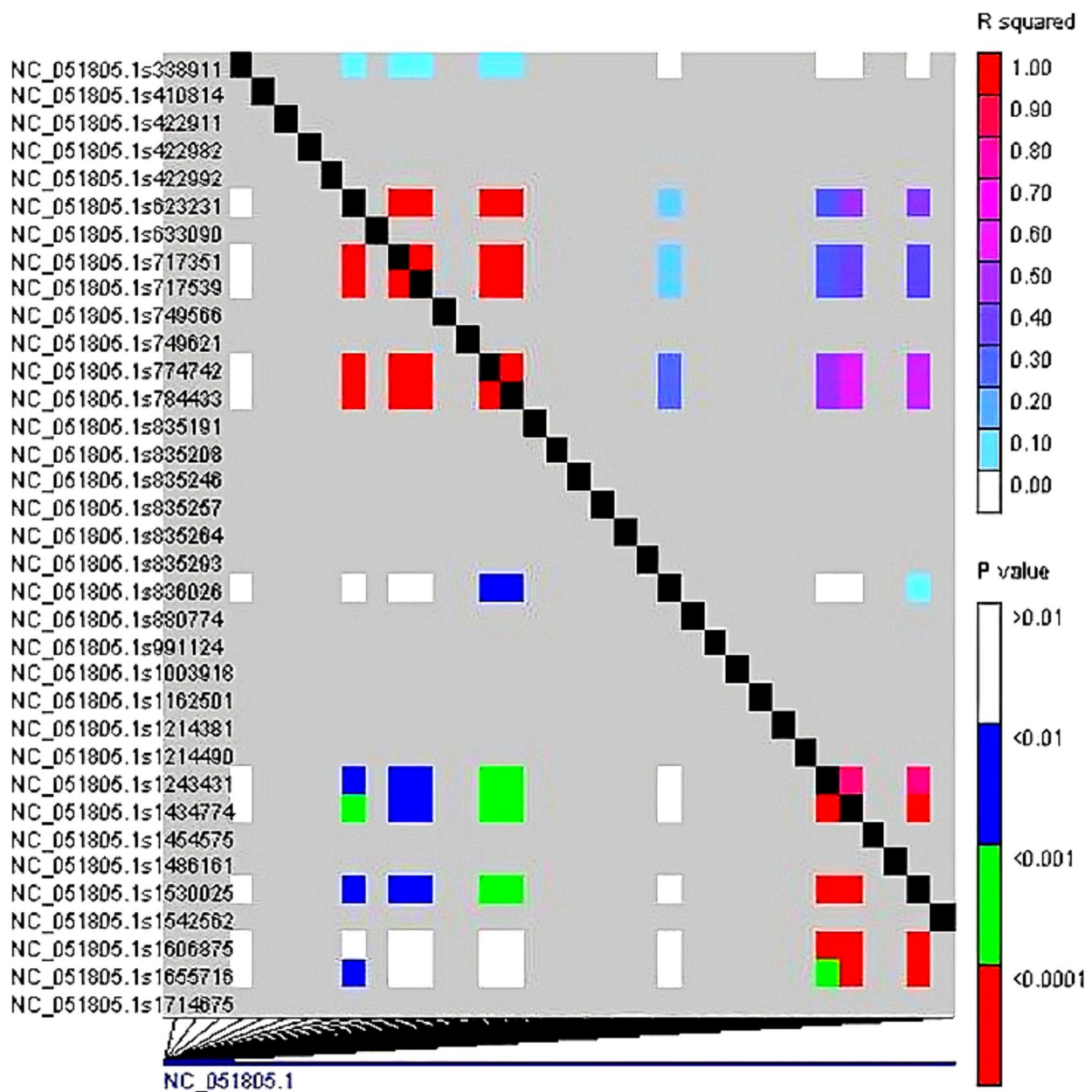


Fig. 5 LD Heatmap for Chromosome 1 (Clu1) shows detailed visualization of LD strength and significance among SNPs with strong LD blocks marked in red

Associated tag-SNPs chromosome-wise

Chromosome-wise sorting of identified tag-SNPs in dogs revealed their distribution across specific chromosomes. Tag-SNPs were detected on chromosomes 1–19, while no tag-SNPs were found on the remaining chromosomes (Supplementary Table S8). The majority of tag-SNPs were concentrated on larger chromosomes, indicating genomic regions with high LD. This clustering suggests that these tag-SNPs represent groups of SNPs strongly associated within these genomic regions [19].

Discussion

Genetic variation among native dog breeds serves as a crucial marker of their evolutionary background, adaptability, and enduring viability. This research utilized genome-wide SNP data to examine population structure, LD, and tag-SNP distribution in two native Indian dog breeds, the Gaddi dog and Mudhol Hound, alongside specific foreign breeds kept in India. The integrated examination of population structure and LD offers an understanding of breed divergence and genomic arrangement.

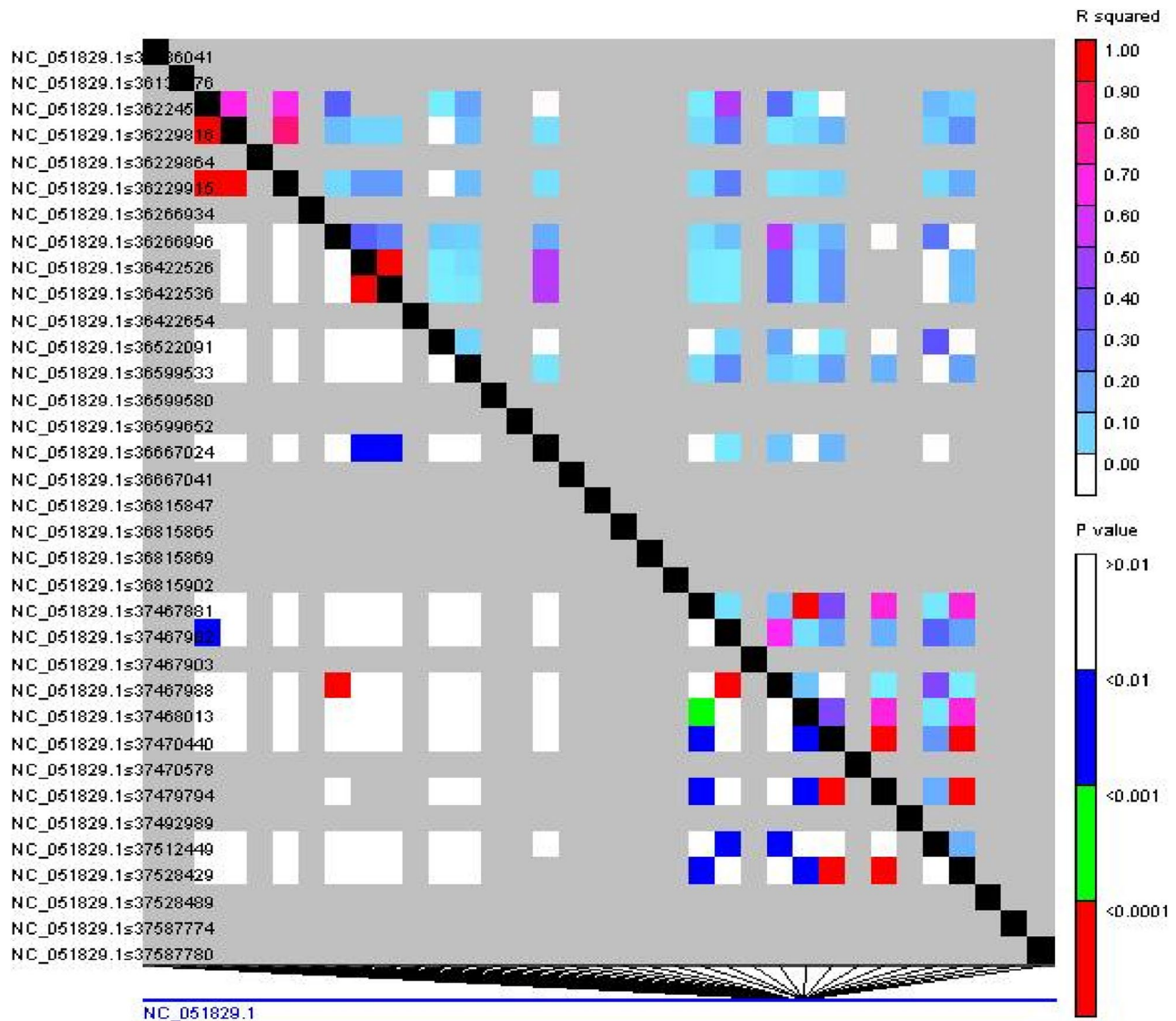


Fig. 6 LD Heatmap for Chromosome 2 (Clu2) shows the degree of LD across SNPs, useful to find recombination cold spots or evolutionary selective sweeps

Analyses of population structure, such as PCA and phylogenetic reconstruction, showed evident genetic grouping among the six examined breeds, suggesting separate evolutionary lineages and minimal recent interbreeding. Indigenous breeds created distinct clusters apart from foreign breeds, reinforcing breed integrity and the existence of unique genetic lineages. These results align with international canine genomic research, including the Dog10K initiative, which noted significant phylogeographic organization among dog populations globally [3].

The discovery of numerous high-quality SNPs after rigorous quality control demonstrates the effectiveness of the genotyping and filtering approach utilized. SNP retention variability is anticipated in reduced-representation sequencing methods, as it relies on sequencing depth, marker density, and filtering thresholds [26–28]. The collection of tag-SNPs found in this research signifies genomic areas with significant LD and offers a compact but informative marker set for subsequent analyses.

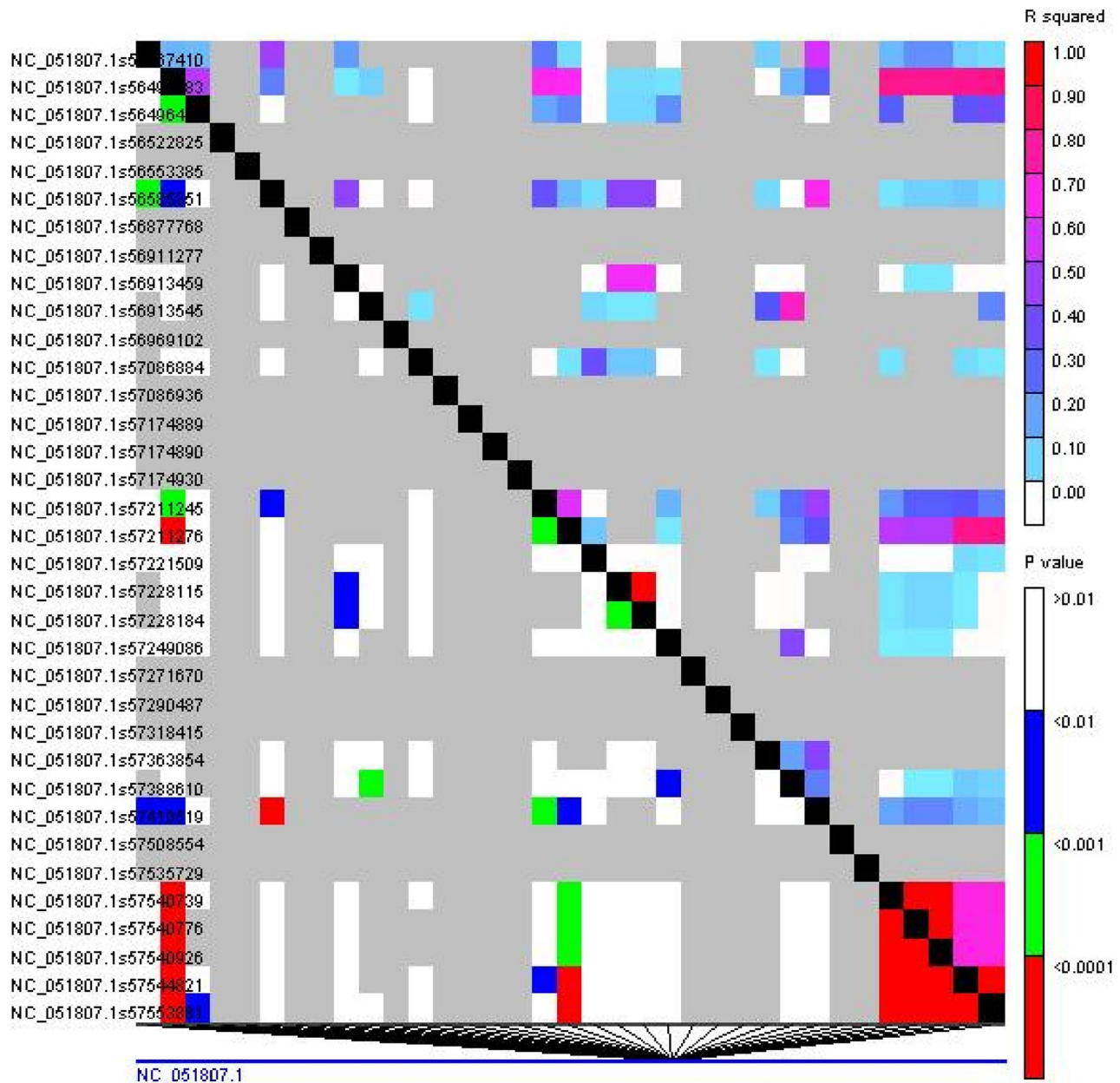


Fig. 7 A colour-coded illustration of LD patterns between SNP locations on Chromosome 3 (Clu3), emphasising clusters of co-inherited variations

Analysis of Hardy–Weinberg equilibrium revealed that although the majority of loci matched equilibrium predictions, a portion exhibited notable deviations. These variations can occur due to biological elements like selection, inbreeding, or non-random mating, especially in controlled or isolated populations, along with technical aspects related to high-throughput sequencing datasets.

Breed-specific LD patterns were noted, with native breeds typically exhibiting quicker LD decay and shorter LD blocks than foreign breeds. Typically, shorter LD blocks

and increased haplotype diversity are linked to larger effective population sizes or less intense artificial selection.

In summary, the unique genetic signatures and LD features found in native breeds highlight their importance as exclusive genetic resources influenced by local adaptation and conventional breeding methods [29]. The tag-SNPs discovered in this study provide a valuable resource for conservation genomics, breed characterization, and forthcoming genome-wide association studies in Indian dog populations [30–32].

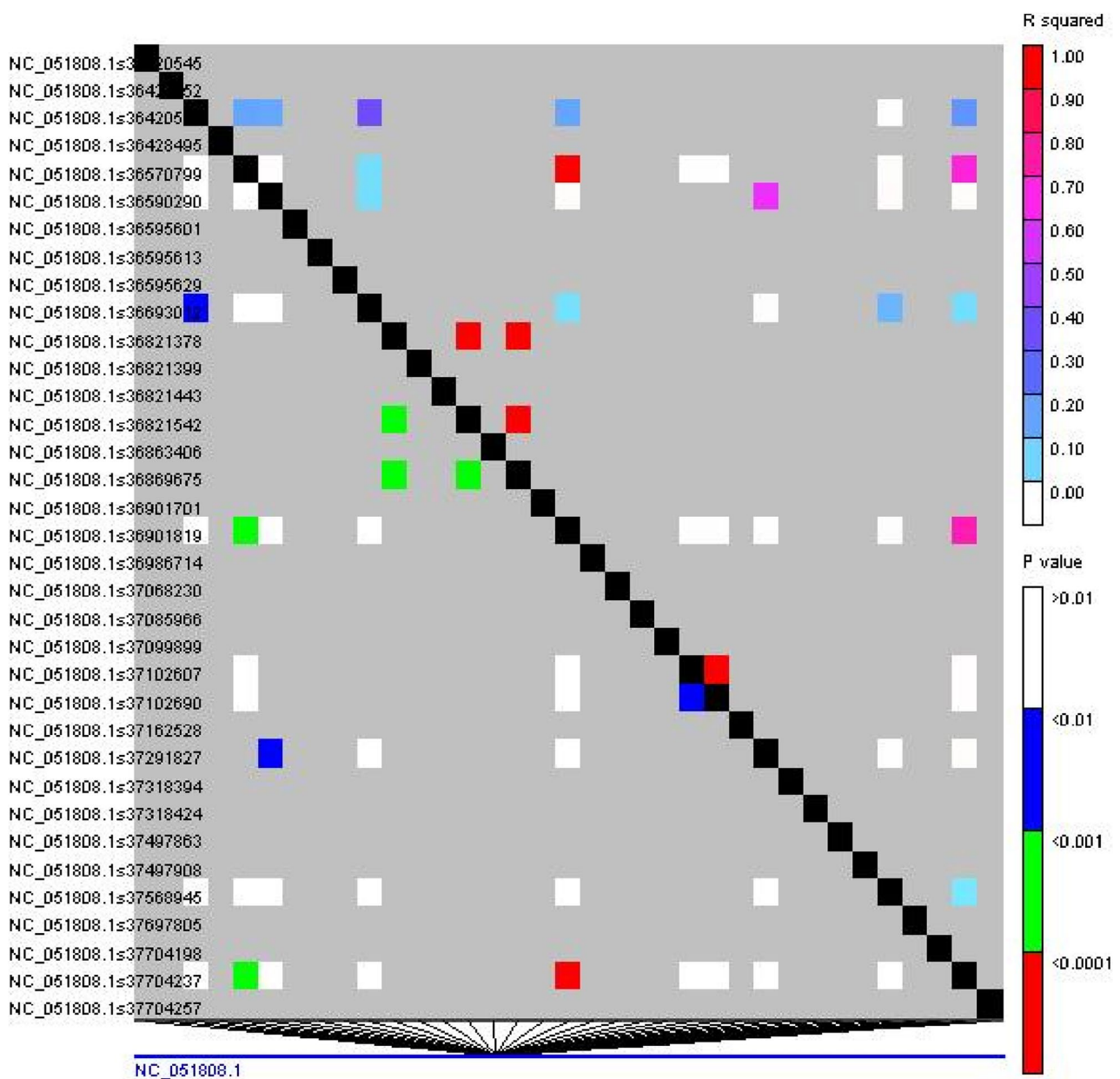


Fig. 8 Pairwise LD matrix displaying r^2 values and significance (p -values) for SNPs on Chromosome 4 (Clu4), color gradients represent LD strength and significance levels

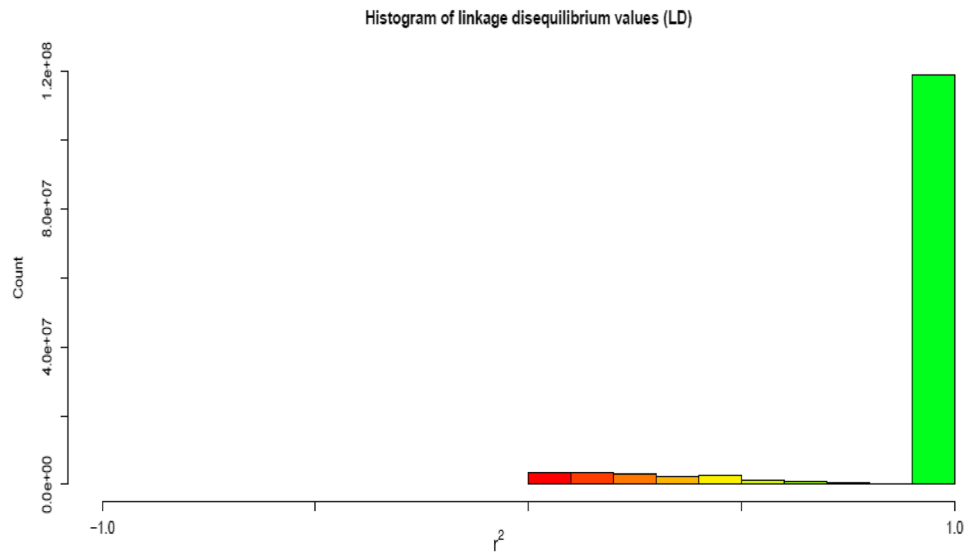


Fig. 9 Distribution of Linkage Disequilibrium (LD) values (r^2) Across SNP Pairs demonstrates the high prevalence of SNP pairs with strong LD (r^2 near 1), supporting extensive genomic linkage blocks within the dataset

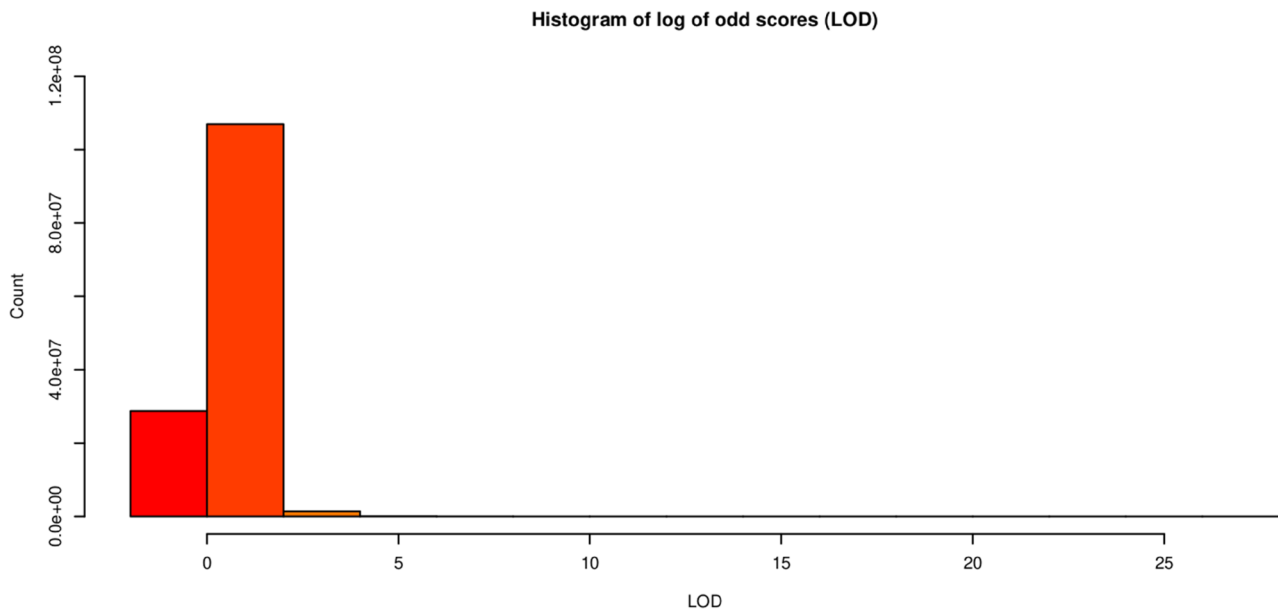


Fig. 10 Histogram of haplotypes represented as Logarithm of Odds (LOD) scores across the genome, most variants show low scores with values that were found to be less than 2, indicating no significant linkage between most loci

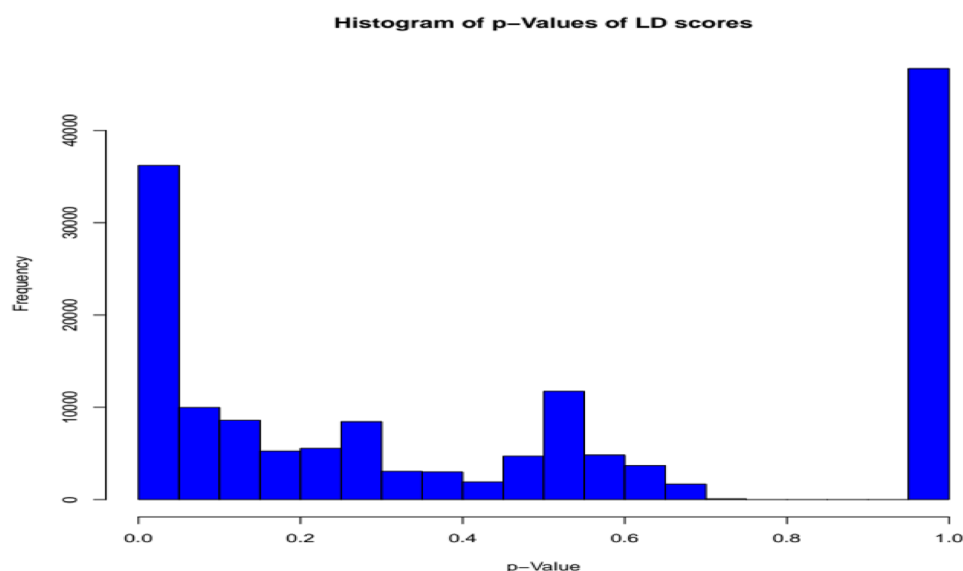


Fig. 11 Histogram of *p*-Values Derived from Linkage Disequilibrium (LD) Score Analysis

Conclusion

This report is the first of its kind as it serves pioneering testament to the population structure, SNPs, and tag-SNP LD status in the well-known Indian dog breeds, specifically the Gaddi dog and Mudhol Hound. It is worth noting that the Indian Council of Agricultural Research-National Bureau of Animal Genetic Resources (ICAR-NBAGR) officially recognized the Gaddi dog as a breed in January 2025. The findings offer significant insights into the molecular framework of the population structure determined by LD and tag-SNPs. Through the examination of genome-wide SNP data obtained from ddRAD sequencing, the study enhances our comprehension of the genetic makeup of the varied canine germplasm raised in India by pet owners.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-026-12758-z>.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

The authors thank Mr. Anil Jamwal (Integrated farmer and Tibetan Mastiff breeder, Palampur), Mr. Newton Sidhu (Director, PHG-CTBI, Mohali), Dr. Yathish HM (Assistant Professor), and various pet owners for providing samples.

Authors' contributions

Bhawanpreet Kaur did the benchwork presented in the paper. CS Mukhopadhyay designed the study and supervised the work. SK Mahajan provided the samples. All authors read and approved the manuscript. All authors contributed to the manuscript revision and read and approved the submitted version.

Funding

The authors acknowledge funding from the Department of Biotechnology, Government of India, through the collaborative research project "Parentage Determination and Cytogenetic Profiling in Dogs (DBT-19I)".

Data availability

The NCBI Bioproject accession number is PRJNA843534. The data is freely available at <https://www.ncbi.nlm.nih.gov/biosample/34074032>.

Declarations

Ethics approval and consent to participate

The study was approved by the Institutional Animal Ethics Committee (IAEC) of Guru Angad Dev Veterinary and Animal Sciences University (GADVASU), Ludhiana, Punjab, India (registration number 497/GO/Re/SL/02/CPCSEA) (IAEC Application number: GADVASU/2021/IAEC/62/02). Informed consent was obtained from the owners of the animals included in the study. All experimental research was conducted in accordance with the guidelines set by the IAEC, GADVASU.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 27 June 2025 / Accepted: 13 March 2026

Published online: 29 March 2026

References

1. Freedman AH, Gronau I, Schweizer RM, Vecchy O-D, Han D, Silva E, Novembre PM, J. Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet*. 2014;10(1):e1004016.
2. Dutrow EV, Serpell JA, Ostrander EA. Domestic dog lineages reveal genetic drivers of behavioral diversification. *Cell*. 2022;185(25):4737–55.
3. Zhou T, Pu SY, Zhang SJ, Zhou QJ, Zeng M, Lu JS, Wang GD. Dog10K: an integrated Dog10K database summarizing canine multi-omics. *Nucleic Acids Res*. 2025;53(D1):D939–47.
4. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559–75.

5. Barrett JC, Fry B, Maller JDMJ, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005;21(2):263–5.
6. Lander R. The Canine Genome: Discoveries, Applications, and Future Potential. 2016.
7. Plassais J, Kim J, Davis BW, Karyadi DM, Hogan AN, Harris AC, Ostrander EA. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nat Commun*. 2019;10(1):1489.
8. Serres-Armero A, Davis BW, Povolotskaya IS, Morcillo-Suarez C, Plassais J, Juan D, Marques-Bonet T. Copy number variation underlies complex phenotypes in domestic dog breeds and other canids. *Genome Res*. 2021;31(5):762–74.
9. Liu C, Chen H, Yang X, Zhang C, Ren Z. Exploring the genomic resources of seven domestic Bactrian camel populations in China through restriction site-associated DNA sequencing. *PLoS ONE*. 2021;16(4):e0250168.
10. Online Mendelian Inheritance in Animals (OMIA). OMIA: Online Mendelian Inheritance in Animals – A comparative knowledgebase of genetic disorders and traits in animals. Faculty of Veterinary Science, University of Sydney. 2025. Retrieved June 29, 2025, from <https://www.omia.org/home/>.
11. Kaur J, Mohan M, Singh B, Sethi RS, Narang D, Kaur S, Mukhopadhyay CS. In-vitro transcriptomic profiling of indigenous Gaddi vis-à-vis exotic Labrador dogs: Insights from systems biology. *Front Veterinary Sci*. 2025;12:1489905.
12. Rana K, Randhawa SS, Mohindroo J, Sethi RS, Mukhopadhyay CS. Biocomputational identification of microRNAs from indigenous Gaddi dog genome. *Gene Rep*. 2025;39:102167.
13. Rana K, Mukhopadhyay CS. Genome assembly of indigenous Gaddi dog. *Indian J Anim Sci*. 2025;95(2):155–8.
14. Tewari S, Mukhopadhyay CS. In silico mining of protein-coding and non-coding RNA (ncRNA) specific genes in exotic versus indigenous Gaddi dogs. *Curr Biotechnol*. 2023;12(3):190–202.
15. Kaur B, Kaur J, Kashyap N, Arora JS, Mukhopadhyay CS. A comprehensive review of genomic perspectives of canine diseases as a model to study human disorders. *Can J Vet Res*. 2023;87(1):3–8.
16. Mukhopadhyay CS, Kaur B. Applications of tag-SNPs in quantitative trait loci (QTL) identification. In: Sobti RC, Mukesh M, Sobti A, editors. *Genomic, proteomics, and Biotechnology*. 1st ed. CRC; 2022. pp. 89–100. <https://doi.org/10.1201/9781003220831>.
17. Sandhu Y, Mahajan S, Sethi RS, Arora JS, Mukhopadhyay CS. Differential karyotype profiling of three popular breeds of dogs in India. *Indian J Anim Sci*. 2020;90(11):1488–90.
18. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*. 2007;23(19):2633–5. <https://doi.org/10.1093/bioinformatics/btm308>.
19. Jannink JL, Walsh B. Association mapping in plant populations. *Quant Genet Genomics Plant Breed*. 2002;59–68. <https://doi.org/10.1079/9780851996011.005>.
20. Meisner J, Benros ME, Rasmussen S. Leveraging haplotype information in heritability estimation and polygenic prediction. *Nat Commun*. 2025;16(1):126.
21. VanLiere JM, Rosenberg NA. Mathematical properties of the r^2 measure of linkage disequilibrium. *Theor Popul Biol*. 2008;74(1):130–7.
22. Karlsson EK, Sigurdsson S, Ivansson E et al. Extent of linkage disequilibrium in large-breed dogs: chromosomal and breed variation. *Mamm Genome*. 2013;24(9–10):486–98. <https://doi.org/10.1007/s00335-013-9469-7>. Epub 2013 Oct 21. PMID: 24062056.
23. Hill WG, Robertson A. Linkage disequilibrium in finite populations. *Theor Appl Genet*. 1968;38(6):226–31.
24. Axelsson E, Ratnakumar A, Arendt M, L, Maqbool K, Webster M, T, Perloski M, ... Lindblad-Toh K. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature*. 2013;495(7441):360–364.
25. Kong A, Cox NJ. Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet*. 1997;61(5):1179–88.
26. Boyko AR, Quignon P, Li L, et al. A simple genetic architecture underlies morphological variation in dogs. *PLoS Biol*. 2009;7(8):e1000451. <https://doi.org/10.1371/journal.pbio.1000451>.
27. Kaur B, Yathish HM, Kashyap N, Mukhopadhyay CS. Phylogeographic and genetic diversity analysis through genome-wide SNPs in indigenous and exotic canine breeds owned in India. *Discover Biotechnol*. 2025;2(1):5.
28. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double digest RAD-seq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*. 2012;7(5):e37135.
29. Wittenburg D, Doschoris M, Klosa J. Grouping of genomic markers in populations with family structure. *BMC Bioinformatics*. 2021;22(1):1–12.
30. Mueller JC. Linkage disequilibrium for different scales and applications. *Brief Bioinform*. 2004;5(4):355–64.
31. Huang X, Zhu TN, Liu YC, Qi GA, Zhang JN, Chen GB. Efficient estimation for large-scale linkage disequilibrium patterns of the human genome. *Elife*. 2023;12:RP90636.
32. Zhang K, Deng M, Chen T, Waterman MS, Sun F. A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci*. 2002;99(11):7335–9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.