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.

A research paper on Copy Number Variation Landscapes Underlying Breed-Specific Genome Structure in Nepalese buffaloes

Uddhav Paneru

pnruddhav.npl@gmail.com

NARC-National Buffalo Research Program (NBRP)

Gopal Ramdasji Gowane

ICAR-National Dairy Research Institute (NDRI)

Rani Alex

ICAR-National Dairy Research Institute (NDRI)

Badmus Kazeem Ajasa

Federal University

Vikas Vohra

ICAR-National Dairy Research Institute (NDRI)

Research Article

Keywords: Copy number variation, genomic hotspot, copy number variation region, high-frequency CNVR, structural variation, gene annotation

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10302210/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A research paper on

Copy Number Variation Landscapes Underlying Breed-Specific Genome Structure in Nepalese buffaloes

Uddhav Paneru[1], Gopal Ramdasji Gowane[2], Rani Alex[2], Badmus Kazeem Ajasa[3], and Vikas Vohra[2]

[1]NARC-National Buffalo Research Program (NBRP), Tarahara, Sunsari, Nepal @: pnruddhav.npl@gmail.com. https://orcid.org/0000-0003-1788-6269

[2]ICAR-National Dairy Research Institute (NDRI), Karnal, Haryana, India @: gopalgowane@gmail.com

[2]ICAR-National Dairy Research Institute (NDRI), Karnal, Haryana, India @: ranialex01vet@gmail.com

[3]Federal University, Gashua, Yobe State, Nigeria @: alqasiminternational@yahoo.com

[2]ICAR-National Dairy Research Institute (NDRI), Karnal, Haryana, India @: vohravikas@gmail.com. https://orcid.org/0000-0001-7581-4939

Corresponding author: pnruddhav.npl@gmail.com

Abstract

Copy number variations (CNVs) are large structural genomic variants characterized by the presence or absence of different copies of genomic segments relative to a reference genome, contributing substantially to phenotypic variation within populations. The present study aimed to identify genomic hotspots of CNVs in five Nepalese buffalo breeds using 64 whole-genome sequence datasets. A read-depth approach implemented in CNVnator with a bin size of 500 bp was used for CNV detection. Overlapping CNVs were merged into copy number variation regions (CNVRs). High-frequency CNVRs were defined as regions present in at least 80% of individuals within a breed. These regions were considered genomic hotspots of structural variation and were investigated for their potential association with disease resistance, environmental adaptation, and productive traits. Gene annotation was performed using ClusterProfiler and DAVID, followed by functional pathway enrichment analysis and literature mining. A total of 20,594 CNVs and 7,767 CNVRs were identified across the five breeds. Lime buffalo exhibited the highest number of CNVs (5,978), whereas Parkote buffalo showed the lowest number (2,964). Deletions outnumbered duplications in all breeds. Functional annotation revealed genes associated with mastitis resistance (CLDN34) in Lime buffalo, genes associated to metabolic substrates required for lactose synthesis during lactation (PDHA1), and immune-related genes including BTNL2 and neighboring BoLA loci implicated in tuberculosis and paratuberculosis susceptibility in Gaddi buffalo. The results provide important insights into the genomic divergence and adaptive differentiation of Nepalese buffalo breeds. This study demonstrates the potential utility of breed-specific CNVRs as genomic markers for disease resistance, productivity, and environmental adaptation in buffalo breeding programs.

Keywords: Copy number variation, genomic hotspot, copy number variation region, high-frequency CNVR, structural variation, gene annotation

Introduction

Buffaloes are among the most important livestock species in Nepal, contributing approximately 63% of national milk production and 52% of meat production [1]. Nearly half of Nepalese households keep buffaloes for milk, meat, manure, traction, and agricultural activities. Buffalo milk is preferred over cattle milk because of its higher fat and solid-not-fat content, attractive white color, and greater economic value [2]. Indigenous Nepalese buffaloes are broadly classified into Lime, Parkote, Gaddi, and Terai breeds, while Arnee represents the wild buffalo population found in Nepal. Lime, Parkote, and Gaddi are adapted to hilly regions above 1,000 meters above sea level, whereas Terai buffaloes are native to the hot and humid lowland plains.

Copy number variations (CNVs) are structural genomic variants, typically larger than 1 kb, that occur in variable copy numbers relative to a reference genome [3]. Unlike single nucleotide polymorphisms (SNPs) and small insertions/deletions, CNVs often encompass entire genes and regulatory regions, thereby influencing gene dosage, transcriptional regulation, and gene function [4]. CNVs can affect phenotypes through deletion or duplication of coding sequences, alteration of regulatory elements, exposure of recessive alleles, or disruption of gene structure at CNV breakpoints [4]. In addition, CNVs contribute substantially to tissue-specific transcriptomic variation and can influence the expression of genes located both within and outside the CNV boundaries [5].

Because of these functional effects, CNVs are increasingly recognized as important contributors to complex traits such as milk yield, fat percentage, protein percentage, and solid-not-fat content.

They also represent a major source of genetic variation involved in environmental adaptation, evolution, and disease resistance. Genome-wide studies in cattle have identified signatures of selection in CNV-overlapping genes associated with climatic adaptation and economically important traits, highlighting the biological significance of structural variation in livestock [6–8].

Historically, CNVs in livestock were detected using comparative genomic hybridization (CGH) arrays and SNP-based platforms. However, advances in next-generation sequencing have made whole-genome sequencing (WGS) the preferred approach because of its higher resolution, deeper coverage, and improved characterization of complex structural variants [9]. Recent studies have demonstrated that WGS can identify CNVs that remain undetected or unresolved by array-based methods [9].

Despite the growing interest in buffalo genomics, CNVs associated with major economic traits such as milk production, fat content, solid-not-fat percentage, and disease resistance remain largely unexplored in Nepalese buffalo populations. Therefore, the present study aimed to comprehensively characterize CNVs and CNVRs in five Nepalese buffalo breeds using whole-genome sequence data. Specifically, the objectives were to: (i) identify genome-wide CNVs and CNVRs, (ii) characterize breed-specific and high-frequency CNVRs, (iii) annotate genes overlapping these regions, and (iv) investigate functional pathways associated with adaptation, disease resistance, and productive traits. The findings provide a genomic framework for understanding structural variation in Nepalese buffaloes and may facilitate the development of marker-assisted selection strategies based on breed-specific CNVRs.

Materials and Methods

Sampling and whole-genome resequencing

Whole-genome sequencing (WGS) data from 64 Nepalese buffaloes (*Bubalus bubalis*) representing five breeds were retrieved from the NCBI Sequence Read Archive (SRA) under BioProject accessions PRJNA1057008 and PRJNA1101683 [10]. The dataset comprised 21 Lime, 15 Gaddi, 14 Terai, 7 Parkote, and 7 Arnee buffaloes that has average sequencing depth of 10x.

Read quality control and sequence alignment

Raw sequencing reads were initially assessed for quality using FastQC v0.12.1 [11]. Adapter sequences and low-quality bases were removed using fastp with default parameters. High-quality reads were aligned against the water buffalo reference genome (NDDB_SH_1, released September 2021) using BWA-MEM v0.7.12 [12]. The resulting Sequence Alignment/Map (SAM) files were retained for downstream processing.

Post-alignment processing

SAM files were converted to Binary Alignment/Map (BAM) format using SAMtools v1.17 [13]. BAM files were subsequently validated, sorted according to genomic coordinates, and assigned read-group information using Picard Tools v2.25.1 implemented within the Genome Analysis Toolkit (GATK) v4.0.2.0 [14]. PCR duplicates were identified and marked using the MarkDuplicates module to minimize amplification bias during CNV detection.

Population genetic analysis

To investigate the genetic relationship among Nepalese buffalo breeds, principal component analysis (PCA) was performed using PLINK v1.9 [15]. PCA was used to evaluate breed differentiation and validate the genetic distinctiveness of the sampled populations.

Identification of copy number variations and copy number variation regions

Copy number variations (CNVs) were detected using the read-depth approach implemented in CNVnator v0.4.1 [16]. CNVnator partitions the genome into non-overlapping bins and estimates read-depth (RD) signals to identify genomic regions with abnormal copy numbers. PCR duplicate-free BAM files were used to generate read-depth histograms. The optimal bin size was determined according to the CNVnator recommendations by evaluating the ratio of mean read depth to read-depth variance, with values between 4 and 5 considered appropriate. A bin size of 500 bp was selected for all samples. Following GC-content correction and read-depth normalization, CNVs were identified using the call function. Candidate CNVs were retained using the following filtering criteria: CNV size between 1 kb and 5 Mb and statistical significance ($p < 0.01$). Copy number variation regions (CNVRs) were generated by merging overlapping CNVs with at least 1 bp overlap using BEDTools v2.31.0 [17]. CNVRs were subsequently classified as loss, gain, or mixed (regions containing both deletions and duplications) using CNVRanger v1.18.0 [18]. Chromosomal distributions of CNVRs were visualized using the RIdeogram package in R.

Identification of high-frequency CNVRs and functional annotation

To identify breed-specific structural variation hotspots, CNVRs were filtered according to their occurrence frequency within each breed. High-frequency CNVRs were defined as regions present in at least 80% of individuals belonging to the same breed. These regions were considered

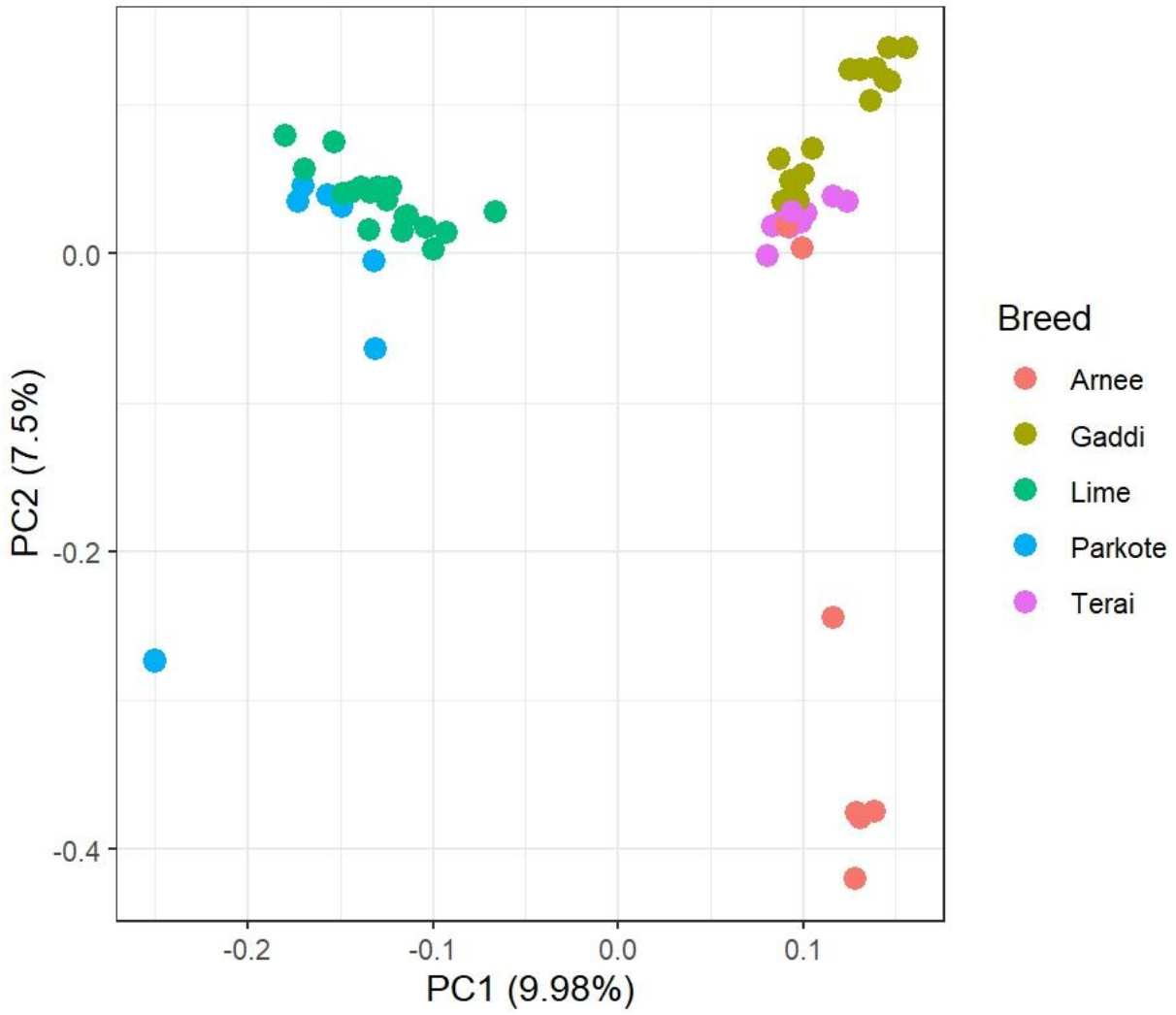
candidate structural variants potentially associated with breed-specific adaptation and economically important traits.

Genes overlapping high-frequency CNVRs were identified using BEDTools v2.31.0 based on the NDDB_SH_1.gtf genome annotation file [17]. Functional enrichment analyses were performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) to identify significantly enriched Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. KEGG pathways with $p < 0.001$ were further validated using the clusterProfiler package in R. Only pathways consistently identified by both approaches were considered for biological interpretation. Owing to the limited functional annotation available for buffalo, the Bos taurus genome annotation was used as the background reference for enrichment analysis.

3. Results

3.1 Genome-wide identification of CNVs and population structure

Principal component analysis (PCA) based on genome-wide CNV data clearly separated the five buffalo breeds (Fig. 1). The first two principal components explained 9.98% (PC1) and 7.50% (PC2) of the total genetic variation, respectively. Lime and Parkote buffaloes clustered closely, whereas Gaddi and Terai buffaloes formed a separate domestic cluster. Arnee buffaloes were distinctly separated from the domestic breeds, indicating clear genetic differentiation among the studied populations. There were only seven Parkote buffaloes were available for whole-genome sequencing and PCA indicated substantial genetic similarity between Lime and Parkote buffaloes, Parkote was excluded from downstream functional annotation analyses.



147

148 **Fig. 1** Principal component analysis (PCA) of five Nepalese buffalo breeds based on genome-wide
149 CNV data.

150

151 **3.2 Genome-wide distribution of CNVs**

152 Genome-wide CNVs were detected using the read-depth approach implemented in CNVnator. A
153 bin size of 500 bp produced an optimal mean read depth to variance ratio of approximately 4,
154 satisfying the recommended criteria for reliable CNV detection. The total number of CNVs

differed among breeds (Fig. 2). Lime buffalo exhibited the highest number of CNVs (5,978), whereas Parkote buffalo had the lowest (2,964). In all breeds, deletions were more frequent than duplications and mixed CNVs.

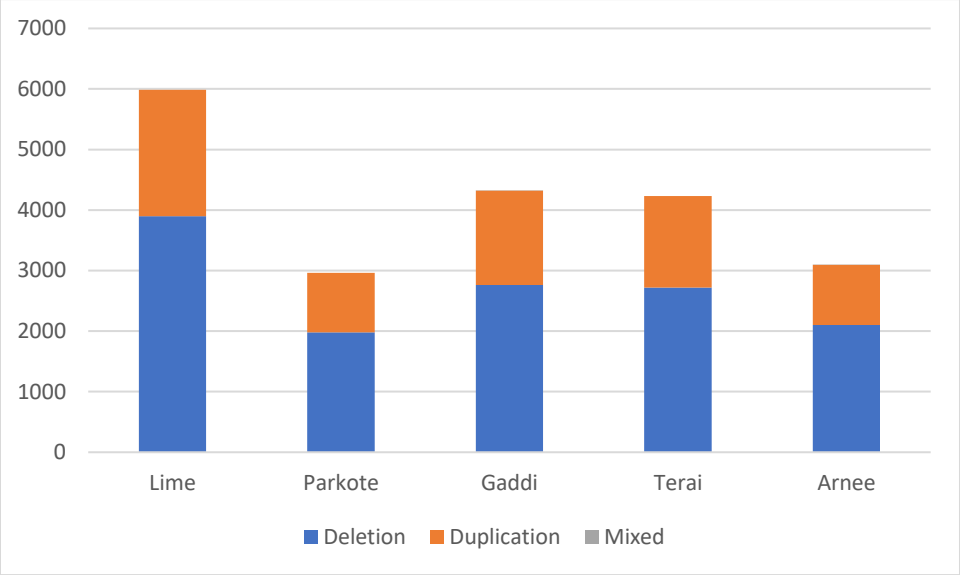


Fig. 2 Total number of CNVs identified in each Nepalese buffalo breed.

Total number of CNVR observed in 5 breeds of buffaloes are presented in Fig 3. Total number of CNVR observed in Lime, Parkote, Gaddi, Terai and Arnee were 1753, 1411, 1557, 1518 and 1528, respectively. Out of different type of CNVR, deletion type was higher than duplication and mixed type of CNVR in all breeds.

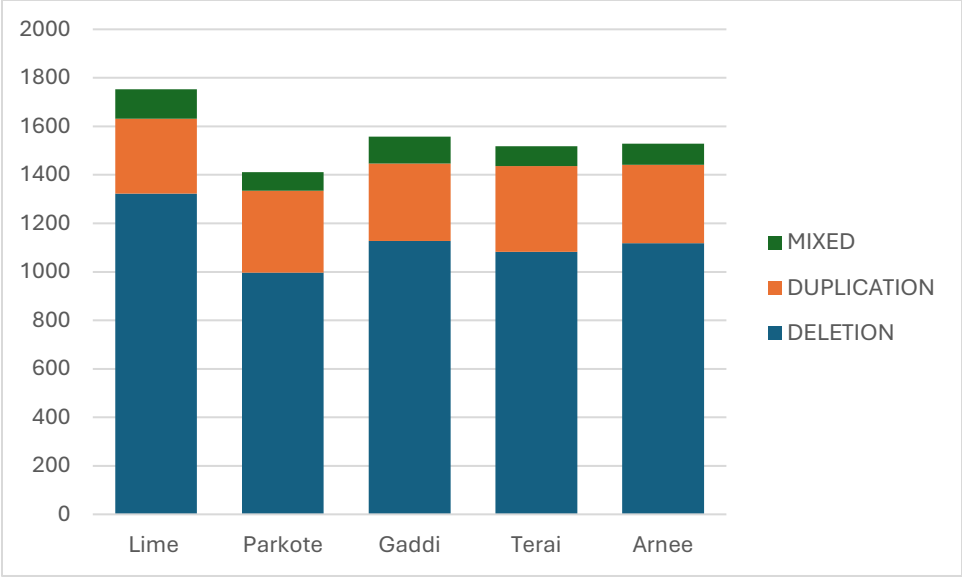


Fig. 3 Total number of CNVR per breed

The total number of CNVs identified in the present study exceeded that previously reported in six Indian buffalo breeds. Likewise, the number of CNVRs observed in Nepalese buffaloes was comparable to that reported in Ethiopian indigenous cattle populations.

3.3 Genomic distribution of CNVRs

The chromosomal distribution of CNVRs is presented in Table 2. Overall, CNVRs covered 1.56% of the autosomal genome and 1.90% of the entire buffalo genome, including the sex chromosomes. CNVR coverage varied among chromosomes. The X chromosome exhibited the highest proportion of CNVR coverage, whereas chromosome 12 showed the lowest genomic coverage.

3.4 High-frequency CNVRs

To investigate breed-specific structural variation, CNVRs were classified according to their frequency within each breed into high-frequency (≥80% of animals), medium-frequency (20–80%), and rare (<20%) categories. Only high-frequency CNVRs from Lime, Gaddi, and Terai buffaloes were included in downstream functional annotation because each breed contained at least 13 whole-genome sequenced individuals. A large proportion of genes overlapping these high-frequency CNVRs corresponded to predicted or uncharacterized LOC genes. Distribution of CNVR according to chromosome distribution is presented in Fig 4.

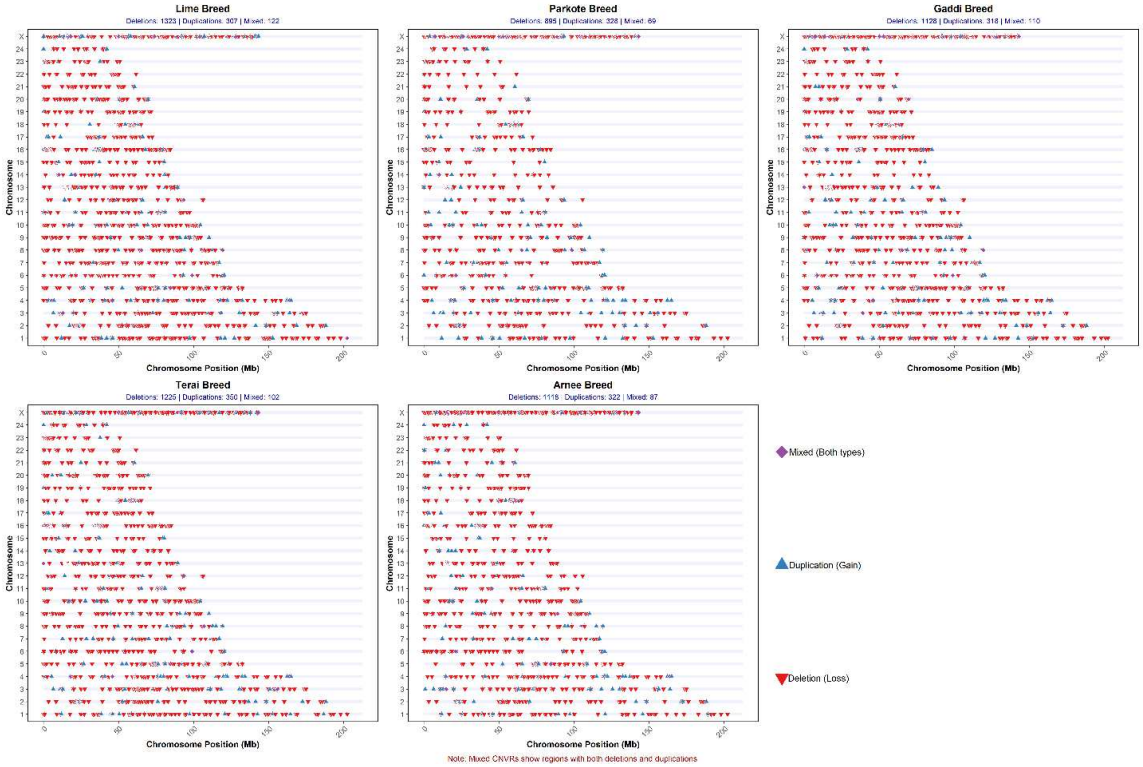


Fig. 4 Ideogram depicting size of buffalo chromosome

4. Discussion

Genome-wide copy number variation (CNV) analysis was performed using whole-genome sequencing data from 64 Nepalese buffaloes representing five indigenous breeds. A total of 20,594 CNVs were identified and subsequently merged into 7,767 copy number variation regions (CNVRs). These CNVRs collectively spanned 53.13 Mb, representing approximately 1.90% of the buffalo reference genome. Among the identified CNVRs, 748 regions were common to all five breeds. Genome-wide CNV analysis demonstrated extensive structural genomic variation among Nepalese buffalo breeds. The identification of 20,594 CNVs and 7,767 CNVRs, together covering 1.90% of the buffalo genome, indicates that CNVs represent an important source of genetic diversity in indigenous buffalo populations. Similar genome coverage has been reported in previous buffalo whole-genome sequencing studies, supporting the reliability of the present CNV detection pipeline [23]. Furthermore, the observed genomic coverage falls within the range reported for ruminant species using next-generation sequencing approaches [24].

Principal component analysis clearly distinguished the five buffalo breeds, demonstrating substantial genetic differentiation that agrees with previous population genomic studies of buffalo [19,20]. The close clustering of Lime and Parkote buffaloes suggests either a recent common ancestry or historical gene flow between these breeds. In contrast, the distinct separation of Arnee buffaloes reflects their wild origin and confirms their considerable genetic divergence from domestic buffalo populations. These findings provide confidence that breed-specific CNVR analyses capture biologically meaningful structural variation.

The present study detected a larger number of CNVs than previous investigations conducted in Indian buffalo populations [21]. Several factors may explain this difference, including the larger sample size, differences in sequencing depth, higher genetic diversity among Nepalese buffaloes,

and variations in CNV detection algorithms and filtering criteria. Such methodological differences have previously been shown to substantially influence CNV discovery across livestock populations. Deletions were consistently more abundant than duplications in all five breeds, a pattern that has been widely reported in buffalo and cattle CNV studies [21,22]. This predominance may reflect both biological processes underlying structural variation and the greater sensitivity of read-depth methods for detecting deletions relative to duplications.

One limitation of the present study is the absence of phenotypic records and estimated breeding values (EBVs) for the sequenced animals. Consequently, the identified CNVRs could not be directly associated with economically important traits through genotype–phenotype association analyses. In addition, Parkote buffaloes were excluded from downstream functional annotation because of the limited number of sequenced individuals and their close genetic relationship with Lime buffaloes. Future studies integrating whole-genome sequencing with phenotypic, transcriptomic, and association analyses will be necessary to validate the biological significance of the identified CNVRs and their contribution to adaptation and productive performance.

4.3 Genomic distribution of CNVRs

The genome-wide distribution of CNVRs demonstrated that structural variants occupied approximately 1.90% of the buffalo genome, indicating that copy number variation represents a substantial source of genomic diversity in Nepalese buffalo. This estimate is highly consistent with the 2.2% genome coverage reported by Liu et al. in water buffalo using whole-genome sequencing [23]. Although Ayalew et al. [22] reported a lower genome coverage (0.7%) in Ethiopian cattle, differences in species, sample size, sequencing depth, and CNV detection methodologies likely account for these discrepancies. Moreover, the observed genome coverage falls within the range

(0.13–8.96%) reported across ruminant species, indicating that the present estimates are comparable with previous livestock genomic studies [24].

The uneven chromosomal distribution of CNVRs, with the highest coverage observed on the X chromosome and the lowest on chromosome 12, suggests that structural variation is not uniformly distributed throughout the buffalo genome. Similar non-random chromosomal distributions have been reported in cattle and buffalo and may reflect differences in recombination rates, repetitive DNA content, mutation frequency, and selective pressures acting on individual chromosomes.

4.4 Biological significance of high-frequency CNVRs

High-frequency CNVRs were selected for functional annotation because structural variants occurring in a large proportion of individuals within a breed are more likely to represent stable genomic features than rare polymorphisms. Such regions may have accumulated through natural or artificial selection and therefore represent candidate loci associated with environmental adaptation, disease resistance, and economically important production traits. Nevertheless, because the present study did not include phenotypic records or gene expression data, these regions should be regarded as candidate genomic regions requiring further functional validation.

A considerable proportion of genes located within high-frequency CNVRs were annotated as predicted or uncharacterized LOC genes. This finding is consistent with previous buffalo genome studies and reflects the current limitations of functional annotation in the buffalo reference genome. Copy number variants frequently occur within rapidly evolving genomic regions enriched for duplicated sequences and lineage-specific genes, many of which remain poorly characterized. Continued improvement of buffalo genome annotation, together with transcriptomic and

comparative genomic analyses, will facilitate the biological interpretation of these candidate structural variants.

Gene Annotation

Lime buffalo

A total of 384 genes were identified within high-frequency CNVRs in Lime buffalo, although a substantial proportion corresponded to uncharacterized *LOC* genes, reflecting the incomplete annotation of the buffalo genome. The annotated genes were primarily enriched in pathways related to immunity, reproduction, muscle development, and metabolism, suggesting that structural variation may contribute to adaptation and productive performance in this indigenous hill breed.

Among the immune-associated genes, *CXCR3* encodes a chemokine receptor responsible for recruiting activated T lymphocytes and natural killer (NK) cells to inflammatory sites, thereby regulating Th1-mediated immune responses and antiviral defense [25]. *PLCG2*, an essential component of B-cell receptor and mast-cell signaling, has been implicated in both innate and adaptive immunity and may influence resistance to intramammary infections. Likewise, *CLDN34*, a member of the tight junction protein family, contributes to maintaining epithelial barrier integrity within the mammary gland. Tight junction stability is essential for preventing pathogen invasion and preserving milk composition, suggesting that CNVRs overlapping *CLDN34* may influence susceptibility to mastitis [26].

Several metabolic genes were also identified within high-frequency CNVRs. *PDHA1* links glycolysis to the tricarboxylic acid cycle and is fundamental to cellular energy production and lactose synthesis during lactation. *MAP3K15*, previously associated with stress-response pathways and pancreatic islet function, may participate in cellular adaptation to oxidative and environmental

stress [27]. Although its function has not been investigated in buffalo, structural variation affecting this gene could contribute to metabolic homeostasis under the fluctuating climatic conditions of Nepal's mid-hill environment.

Reproductive function was represented by *PGR*, which regulates ovarian cyclicity, uterine receptivity, and pregnancy maintenance [28]. In addition, TSPYL2, an X-linked member of the nucleosome assembly protein family, has been implicated in gonadal development and reproductive physiology [29]. The occurrence of these genes within high-frequency CNVRs is consistent with the well-recognized reproductive fitness and adaptability of indigenous Lime buffaloes maintained under extensive production systems.

Genes associated with muscle growth and carcass characteristics were also detected. *HDAC4* regulates skeletal muscle differentiation through epigenetic mechanisms and influences lipid metabolism [30]. *ARHGAP42* contributes to vascular smooth muscle regulation and may influence muscle perfusion and metabolic efficiency [31], whereas *SHROOM2* participates in cytoskeletal organization and tissue morphogenesis. *PYCR2*, which catalyzes proline biosynthesis, contributes to collagen synthesis and connective tissue development, potentially influencing muscle integrity and meat quality.

Overall, the high-frequency CNVRs identified in Lime buffalo encompass genes involved in immune competence, reproductive performance, metabolic regulation, and muscle development. These findings suggest that structural variation may contribute to the remarkable adaptability and productive resilience of this indigenous hill breed.

Gaddi buffalo

High-frequency CNVRs identified in Gaddi buffalo overlapped several genes involved in immunity, metabolism, growth, and reproduction, indicating that structural variation may contribute to adaptation in this high-altitude breed.

PDHA1 plays a central role in oxidative metabolism by linking glycolysis with the tricarboxylic acid cycle and provides metabolic substrates required for lactose synthesis during lactation [32,33]. Structural variation affecting this gene may therefore influence metabolic efficiency and milk production.

Immune-related genes were strongly represented. *BTNL2* regulates T-cell activation and immune homeostasis, and polymorphisms within *BTNL2* and neighboring BoLA loci have previously been associated with susceptibility to bovine tuberculosis and paratuberculosis [34–36]. *PLCG2,* another immune-related candidate, mediates B-cell receptor signaling and has been associated with immune responsiveness and mastitis susceptibility [37,38]. These observations suggest that CNVRs encompassing immune-related genes may contribute to disease resilience in Gaddi buffaloes raised under extensive grazing systems.

Genes associated with muscle development were also identified. *FHL1* regulates skeletal muscle differentiation, whereas *HDAC4* functions as an epigenetic regulator of muscle growth, skeletal development, and lipid metabolism [39–41]. Reproductive performance may additionally be influenced by structural variation affecting *PGR*, a key regulator of progesterone signaling and pregnancy establishment [42].

Collectively, these candidate genes indicate that structural variation in Gaddi buffalo may influence immune function, metabolic efficiency, muscle development, and reproductive

performance. Nevertheless, transcriptomic and genotype–phenotype association studies are required to establish the functional consequences of these CNVRs.

Terai buffalo

Terai buffaloes inhabit the hot and humid lowland plains of Nepal, where animals are continuously exposed to elevated temperature, humidity, and increased pathogen pressure. High-frequency CNVRs identified in this breed contained genes primarily associated with immune regulation, metabolism, cellular stress responses, and reproduction.

HDAC4 functions as an epigenetic regulator of gene expression involved in skeletal muscle development, metabolism, and cellular responses to environmental stress [30]. Epigenetic regulation is increasingly recognized as an important mechanism underlying adaptation to climatic stress in livestock. *MGAM*, which encodes maltase-glucoamylase, plays an essential role in carbohydrate digestion and glucose utilization, potentially contributing to efficient energy metabolism under tropical production conditions.

Several immune-related genes were also identified within high-frequency CNVRs. *BTNL2* regulates adaptive immune responses through modulation of T-cell activation [43,44]. *CD55* and *CFH* are major regulators of the complement cascade that protect host tissues from excessive complement activation while maintaining effective innate immunity [45,46]. *PLCG2* mediates B-cell receptor signaling and innate immune activation, whereas *WDFY4* participates in antigen presentation and adaptive immune responses [47]. The enrichment of these immune-associated genes may represent genomic adaptations that enhance disease resilience under the hot and humid conditions characteristic of the Terai region.

Unlike the previous version, mitochondrial genes were not interpreted because copy number estimates for mitochondrial DNA generated from whole-genome sequencing are highly influenced by tissue-specific mitochondrial abundance and sequencing depth. Consequently, mitochondrial CNVs detected by read-depth methods such as CNVnator should be interpreted cautiously and are generally not considered robust candidates for adaptive CNVR analyses.

Overall, the high-frequency CNVRs identified in Terai buffalo highlight genes involved in immune regulation (*BTNL2, CD55, CFH, PLCG2, WDFY4*), metabolic regulation (MGAM), and cellular stress responses (*HDAC4*). These candidate genes provide plausible genomic mechanisms underlying environmental adaptation and disease resilience in buffaloes inhabiting the subtropical lowlands of Nepal. However, further transcriptomic and functional validation studies are necessary to determine whether these structural variants directly influence economically important traits or heat tolerance.

Conclusion

A large number of structural variations were observed in this study for five major breeds of buffalo of Nepal, with reasonable number of breed-specific and common variants. A variation in CNV numbers among the different breeds was observed, with Lime possessing the highest number of CNVs and Parkote the lowest, and deletions outnumbering duplications in all breeds. Annotation of the high frequency CNVR revealed genes for mastitis resistance, milk production, fertility, muscle growth and development. One of the major gene, *CLDN34* observed in Lime buffalo, produce a tight junction protein that contribute to epithelial barrier integrity in the mammary gland. Maintenance of tight junction structure is crucial for preventing pathogen invasion and preserving milk composition. Similarly, *PHDA1* genes observed in Gaddi buffalo help in linking glycolysis to the tricarboxylic acid cycle, is central to oxidative metabolism and provides substrates for

lactose synthesis, and milk production from the mammary glands. All annotated genes in the study are practically relevant because these are derived from breed-specific high frequency CNVR (≥80), which is recent polymorphism in majority of the population for particular breed. Results indicate that CNVs and CNVRs govern candidate genes related to important functional traits and adaptability, and can be used for the selection of buffalo for important traits based on breed-specific variants.

Declarations

Ethics approval and consent to participate

Secondary data was used in this by downloading data from NCBI. Therefore, no ethical approval is required.

Availability of the data and materials

Whole-genome sequencing (WGS) data used in this study were derived from NCBI Sequence Read Archive (SRA) under BioProject accessions PRJNA1057008 and PRJNA1101683. Further processed dataset such as BAM files and CNVs file can be provided if requested to the corresponding author.

Conflict of interest

Authors express no conflict of interest.

Funding Source of Research

No specific funding was received for this paper.

Author Contributions

UP, VV and GRG are involved in the conceptualization of the study; UP involved in main Analysis, writing and preparation of original draft; BKA assisted in gene annotation and gene mining, VV, GRG and RA supervised the study and involved in critical review and editing. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

I would like to thank NDRI Karnal, Haryana, India for providing computation facilities and other research supports. Nepal Agricultural Research Council, Nepal for the study leave and Department of Science and Technology (DST), India for the scholarship for the duration of stay in India.

References

1. Ministry of Agriculture and Livestock Development (MoALD) (2022) *Agri-Business Promotion and Statistical Division*. Singha Durbar, Kathmandu, Nepal.

2. Paudel LN (2017) Breed improvement through intensive local selection of indigenous buffaloes for food and nutritional security and livelihood improvement in mid-hills of Nepal. In: *Proceedings of the International Buffalo Symposium*, Chitwan, Nepal, p 134.

3. Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7(2):85–97. https://doi.org/10.1038/nrg1767

4. Zhang F, Gu W, Hurles ME, Lupski JR (2009) Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 10:451–481. https://doi.org/10.1146/annurev.genom.9.081307.164217

5. Henrichsen CN, Vinckenbosch N, Zöllner S, Chaignat E, Pradervand S, Schütz F, Ruedi M, Kaessmann H, Reymond A (2009) Segmental copy number variation shapes tissue transcriptomes. Nat Genet 41(4):424–429. https://doi.org/10.1038/ng.345

6. Fadista J, Thomsen B, Holm LE, Bendixen C (2010) Copy number variation in the bovine genome. BMC Genomics 11:284. https://doi.org/10.1186/1471-2164-11-284

7. Zhang Y, Hu Y, Wang X, Jiang Q, Zhao H, Wang J, Ju Z, Yang L, Gao Y, Wei X, Bai J, Zhou Y, Huang J (2020) Population structure and selection signatures underlying high-altitude adaptation inferred from genome-wide copy number variations in Chinese indigenous cattle. Front Genet 10:1404. https://doi.org/10.3389/fgene.2019.01404

8. Chen N, Cai Y, Chen Q, Li R, Wang K, Huang Y, Hu S, Huang S, Zhang H, Zheng Z, Song W, Ma Z, Ma Y, Dang R, Zhang Z, Xu L, Jia Y, Liu S, Yue X, Deng W, Zhang X, Sun Z, Lan X, Han J, Chen H, Bradley DG, Jiang Y, Lei C (2018) Whole-genome resequencing reveals

worldwide ancestry and adaptive introgression events of domesticated cattle. Nat Commun 9:2337. https://doi.org/10.1038/s41467-018-04737-0

9. Bowman P, Grimes H, Dallosso AR, Berry I, Mullin S, Rankin J, Low KJ (2025) Whole genome sequencing for copy number variant detection to improve diagnosis and management of rare diseases. Dev Med Child Neurol 67(1):126–131. https://doi.org/10.1111/dmcn.15985

10. Dhakal A, Si J, Sapkota S, Pauciullo A, Han J, Gorkhali NA, Zhao X, Zhang Y (2024) Whole-genome sequencing reveals genetic structure and adaptive genes in Nepalese buffalo breeds. BMC Genomics 25:1082. https://doi.org/10.1186/s12864-024-10993-0

11. Andrews S (2015) *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Babraham Bioinformatics, Babraham Institute, Cambridge, UK. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

12. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324

13. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H (2021) Twelve years of SAMtools and BCFtools. GigaScience 10(2):giab008. https://doi.org/10.1093/gigascience/giab008

14. Van der Auwera GA, O'Connor BD (2020) *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. 1st edn. O'Reilly Media, Sebastopol.

15. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575. https://doi.org/10.1086/519795

16. Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 21(6):974–984. https://doi.org/10.1101/gr.114876.110

17. Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842. https://doi.org/10.1093/bioinformatics/btq033

18. da Silva V, Ramos M, Groenen MAM, Crooijmans RPMA, Johansson AM, Regitano LCA, Coutinho LL, Zimmer R, Waldron L, Geistlinger L (2020) CNVRanger: Association analysis of CNVs with gene expression and quantitative phenotypes. Bioinformatics 36(3):972–973. https://doi.org/10.1093/bioinformatics/btz632

19. Deng TX, Ma XY, Lu XR, Duan AQ, Shokrollahi B, Shang JH (2022) Signatures of selection reveal candidate genes involved in production traits in Chinese crossbred buffaloes. J Dairy Sci 105(2):1327–1337. https://doi.org/10.3168/jds.2021-21102

20. Deng TX, Ma XY, Duan A, Lu XR, Abdel-Shafy H (2024) Genome-wide copy number variant analysis reveals candidate genes associated with milk production traits in water buffalo (*Bubalus bubalis*). J Dairy Sci 107(9):7022–7037. https://doi.org/10.3168/jds.2023-24614

21. Ahmad SF, Chandrababu Shailaja C, Vaishnav S, Kumar A, Gaur GK, Janga SC, Ahmad SM, Malla WA, Dutt T (2023) Read-depth based approach on whole genome resequencing data reveals important insights into the copy number variation (CNV) map of major global buffalo breeds. BMC Genomics 24:680. https://doi.org/10.1186/s12864-023-09720-8

22. Ayalew W, Wang X, Tarekegn GM, Tessema TS, Chu M, Liang C, Naboulsi R, Van Damme R, Bongcam-Rudloff E, Ping Y (2024) Whole-genome sequencing-based copy number variation analysis in Ethiopian cattle reveals adaptations to diverse environments. BMC Genomics 25:1088. https://doi.org/10.1186/s12864-024-10936-5

23. Liu M, Zhou Y, Rosen BD, Van Tassell CP, Stella A, Tosser-Klopp G, et al (2019) Diversity of copy number variation in the worldwide goat population. Heredity 122(5):636–646. https://doi.org/10.1038/s41437-018-0150-6

24. Celus CS, Ahmad SF, Gangwar M, Kumar S, Kumar A (2025) Deciphering new insights into copy number variations as drivers of genomic diversity and adaptation in farm animal species. Gene 939:149159. https://doi.org/10.1016/j.gene.2024.149159

25. Groom JR, Luster AD (2011) CXCR3 in T cell function. Nat Immunol 12(7):589–594. https://doi.org/10.1038/ni.2042

26. Kobayashi K, Kumagai H, Kameyama A (2013) Tight junction proteins in mammary gland development and mastitis resistance. J Dairy Sci 96(9):5397–5409. https://doi.org/10.3168/jds.2013-6736

27. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, Payne AJ, Steinthorsdottir V, Scott RA, Grarup N, et al (2018) Fine-mapping type 2 diabetes loci to

single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet 50(11):1505–1513. https://doi.org/10.1038/s41588-018-0241-6

28. Robker RL, Akison LK, Russell DL (2009) Control of oocyte release by progesterone receptor-regulated gene expression. Nucl Recept Signal 7:e012. https://doi.org/10.1621/nrs.07012

29. Lau YFC, Li Y, Kido T (2011) The TSPY gene family: Role in gonad development and spermatogenesis. Int J Dev Biol 55(7–9):1029–1037. https://doi.org/10.1387/ijdb.103225yl

30. Zhang X, Dong J, Zhao Y, et al (2018) HDAC4 regulates skeletal muscle fiber type and lipid metabolism. J Biol Chem 293(11):4304–4317. https://doi.org/10.1074/jbc.RA117.000987

31. Luo Y, Wang Y, Liu L, Wang J, Li Y, Liu H, et al (2017) ARHGAP42 regulates vascular smooth muscle contractility and blood pressure. Nat Commun 8:15208. https://doi.org/10.1038/ncomms15208

32. Zhao FQ, Keating AF (2014) Lactose synthesis in the mammary gland: Insights from transgenic mouse models. Asian-Australas J Anim Sci 27(7):919–927. https://doi.org/10.5713/ajas.2014.14122

33. Bionaz M, Loor JJ (2008) Gene networks driving bovine milk fat synthesis during the lactation cycle. BMC Genomics 9:366. https://doi.org/10.1186/1471-2164-9-366

34. Minozzi G, Buggiotti L, Stella A, Strozzi F, Luini M, Williams JL (2010) Genetic loci involved in antibody response to *Mycobacterium avium* subsp. *paratuberculosis* in cattle. PLoS ONE 5(6): e11117. https://doi.org/10.1371/journal.pone.0011117

35. Pant SD, Schenkel FS, Verschoor CP, You Q, Kelton DF, Moore SS, Karrow NA (2010) A principal component regression-based genome-wide analysis approach reveals the presence of a novel QTL on BTA7 for MAP resistance in Holstein cattle. Genomics 95(3):176–182. https://doi.org/10.1016/j.ygeno.2010.01.001

36. Bhide MR, Mucha R, Mikula I, Kisova L, Skrabana R, Novak M, Mikula I (2009) Novel mutations in TLR genes cause hyporesponsiveness to *Mycobacterium avium* subsp. *paratuberculosis* infection. BMC Genet 10:21. https://doi.org/10.1186/1471-2156-10-21

37. Wang X, Wu J, Zhang Y, Wang Y, Li Y, Zhang Q, et al (2017) A missense mutation in the phospholipase C gamma 2 (PLCG2) gene is associated with exfoliative ichthyosis in Hereford cattle. BMC Genomics 18:453. https://doi.org/10.1186/s12864-017-3818-6

38. Jiang P, Wang M, Zhang C, Li X, Wang Z, Guo Y, Zhang J, Yan W, Xi X (2025) PLCG2 across human disease: Genetic variants, signaling mechanisms, and clinical implications. J Transl Med 23:110. https://doi.org/10.1186/s12967-025-07632-7

544 39. Lee SH, Park BH, Sharma A, Dang CG, Lee SS, Choi TJ, Kim HC, Yeon SH, Kang HS, Lee
545 HK (2012) Genome-wide association study identifies major loci affecting carcass weight,
546 backfat thickness and marbling score in Hanwoo cattle. PLoS ONE 7(7):e40853.
547 https://doi.org/10.1371/journal.pone.0040853
548
549 40. Wang X, Zhang Y, Li Y, Wang Y, Gao X, Chen Y (2015) Genome-wide association study in
550 Chinese Simmental cattle identified a potential candidate gene FHL3 for meat production
551 traits. Anim Genet 46(2):211–212. https://doi.org/10.1111/age.12264
552
553 41. Bongiorni S, Gruber CEM, Chillemi G, Bueno S, Failla S, Moioli B, Ferrè F, Valentini A (2016)
554 Skeletal muscle transcriptional profiles in two Italian beef breeds, Chianina and
555 Maremmana, reveal breed-specific variation. Mol Biol Rep 43(4):253–268.
556 https://doi.org/10.1007/s11033-016-3957-3
557
558 42. Spencer TE, Bazer FW, Johnson GA (2004) Conceptus signals for establishment and
559 maintenance of pregnancy. Anim Reprod Sci 82–83:537–550.
560 https://doi.org/10.1016/j.anireprosci.2004.05.014
561
562 43. Lebrero-Fernández C, Wenzel UA, Akeus P, et al (2016) Butyrophilin-like proteins regulate γδ
563 T-cell activation and epithelial immune responses. Front Immunol 7:315.
564 https://doi.org/10.3389/fimmu.2016.00315
565

566 44. Franke A, Hampe J, Rosenstiel P, et al (2008) Systematic association mapping identifies
567 NELL1 as a novel inflammatory bowel disease gene. PLoS ONE 3(2):e691.
568 https://doi.org/10.1371/journal.pone.0000691
569
570 45. Rainard P, Riollet C (2006) Innate immunity of the bovine mammary gland. Vet Res 37(3):369–
571 400. https://doi.org/10.1051/vetres:2006007
572
573 46. Zipfel PF, Skerka C (2009) Complement regulators and inhibitory proteins. Nat Rev Immunol
574 9(10):729–740. https://doi.org/10.1038/nri2620
575
576 47. Yang W, Tang H, Zhang Y, Tang X, Zhang J, Sun L, Yang J, Cui Y, Zhang L, Hirankarn N, Lau
577 CS, Hao F, Wen Y, Li X, Wong RW, Sheng Y, Hu Y, Pan HF, Ye DQ (2018) Meta-analysis
578 of genome-wide association studies identifies genetic variants in WDFY4 associated with
579 immune-mediated diseases. Nat Commun 9:4972. https://doi.org/10.1038/s41467-018-
580 07497-6
581

582 **Table 2.** Chromosomal distribution and genomic coverage of CNVRs in Nepalese buffalo.

Chromosome	CNVR region (bp)	Buffalo reference genome (bp)	Percentage
NC_059181.1	14575000	143192433	10.18
NC_059163.1	1263000	116997125	1.08
NC_059164.1	1606500	119318788	1.35

NC_059165.1	1946000	110262714	1.76
NC_059173.1	616000	72603365	0.85
NC_059172.1	3210000	85115218	3.77
NC_059166.1	1261000	104551495	1.21
NC_059167.1	2602500	102416932	2.54
NC_059160.1	3195500	164971372	1.94
NC_059161.1	3170000	132500020	2.39
NC_059179.1	1008500	52095120	1.94
NC_059178.1	319500	61771150	0.52
NC_059168.1	504500	106386409	0.47
NC_059169.1	4153000	89364132	4.65
NC_059157.1	2266000	202348575	1.12
NC_059159.1	2306000	174872328	1.32
NC_059158.1	2275000	188164321	1.21
NC_059171.1	465000	81832314	0.57
NC_059170.1	565000	83479656	0.68
NC_059177.1	321000	60624307	0.53

NC_059176.1	1538000	69611172	2.21
NC_059175.1	526500	71582194	0.74
NC_059174.1	2000000	65791760	3.04
NC_059180.1	236000	42172481	0.56
NC_059162.1	1203000	120418900	1.00
Mean coverage	2125300	104897771.2	1.90

583

584

585

586