Genome-wide SNPs mining and annotation *in Bubalus bubalis*

Abstract

This study aimed to identify single nucleotide polymorphisms (SNPs) in the Murrah buffalo, often referred to as the "black gold of India," using the reduced representation sequencing technique.DNA was extracted and sequenced using the ddRAD technique from blood samples taken from 96 unrelated Murrah buffalo. After processing the sequenced genomic data with various bioinformatics tools, a total of 855,563 SNPs were identified in the Murrah buffalo genome, using the Bubalus bubalis genome as a reference. Annotation revealed that the majority of these variations were located in intronic regions (67.49%), followed by intergenic regions (20.15%). The SNPs identified in this study have the potential to serve as molecular markers for economically significant traits and could be valuable for future breed improvement and conservation efforts..

Keywords : Bioinformatics tools, Genetic variation, Murrah buffalo, Reduced representation sequencing, Single nucleotide polymorphisms (SNPs)

1. Introduction

Buffalo, a multi-purpose animal well adapted to tropical and sub-tropical climatic conditions,is a major dairy bovine in South Asian countries. At present, there are twenty registered breeds of buffalo in India, recognized by National Bureau of Animal Genetic Resources(NBAGR),exhibiting distinct phenotypes as a result of variations in their genetic composition, brought about by evolutionary factors. Enhancing buffalo productivity, growth rate, feed conversion efficiency, heat tolerance, and disease resistance necessitates a thorough understanding of polymorphisms among breeds. In various species, single nucleotide polymorphisms (SNPs)—the most prevalent type of genetic variation—have been widely used as markers for marker-assisted selection (MAS) [**9, 21, 20]**. To identify potential SNPs linked to economically important traits, the genome should be screened, and the detected SNPs should be annotated to predict their function.

SNPs found in one population may be entirely monomorphic in another.[18].

With the availability of reference genomes for all major livestock species, advances in next-generation sequencing techniques have made possible the detection of SNPs. One of two approaches can be used for genome-wide SNP identification: sub-sampling or whole genome sequencing. Reduced representation also known as sub-sampling techniques are an effective alternative for whole genome sequencing since they are less expensive, computationally faster, and provide testing of a wide variety of polymorphic loci without the need for a reference sequence or prior information.

One such next-generation sequencing technique that uses restriction enzymes and molecular identifiers to examine a portion of the whole genome is restriction site-associated DNA sequencing (RADseq) [7]. The "RADseq family" [3] is a collection of methods used in RAD sequencing. Double digest RAD sequencing (ddRAD), a method of the RADseq family, uses a second restriction enzyme to digest genomic DNA in order to cut down on the time and expense of library preparation [17]. Additionally, it addresses a significant flaw in the original RADseq approach by allowing paired-end sequencing of identical loci across many samples.

This study aimed to identify SNPs in the Murrah buffalo population using ddRAD sequencing data. Murrah is one of the most renowned buffalo breeds for high milk production and is widely utilized in various breed improvement programs across the country. Its superior germplasm has also been exported to several countries, including Egypt, Brazil, Bangladesh, Sri Lanka, Thailand, China, Nepal, and Vietnam, to enhance local buffalo populations.

Selective breeding and progeny testing programs have significantly improved the performance of native milch breeds. However, there remains substantial potential for genetic advancement through breeding programs that integrate genetic markers. Further research should focus on evaluating the impact of the identified polymorphic loci on economically significant traits to enhance genetic selection strategies.

2. Methods and Materials

Genomic DNA Analysis and Sample Collection

The genomic data of 96 Murrah (*Bubalus bubalis*) buffaloes was generated by ICAR-National Dairy Research Institute, Karnal and the bioinformatics analysis was conducted at ICAR- National Bureau of Animal Genetic Resources, Karnal.

The 96 Murrah buffaloes maintained at Livestock Research Centre (LRC), ICAR-National Dairy Research Institute, Karnal. Blood samples were taken according to the applicable guidelines and regulations, which were approved by the Institutional Animal Ethics Committee (IAEC) of the National Bureau of Animal Genetics Resources (ICAR-NBAGR), Karnal. DNA extraction was carried out after blood samples were collected, and the extracted DNA was then checked for quality, concentration, and purity in preparation for further analysis.

The DNA extraction for this study was performed using the phenol-chloroform method, a widely used technique for isolating high-quality genomic DNA. This method involves cell lysis, followed by organic phase separation using phenol and chloroform, and precipitation of DNA with ethanol or isopropanol. The extracted DNA was then quantified using a NanoDrop spectrophotometer and assessed for integrity through agarose gel electrophoresis before proceeding with downstream applications such as ddRAD sequencing.

2.1 Library preparation and ddRAD sequencing

Following the initial genomic DNA quality and quantity evaluation, the standard RAD sequencing protocol was used [17]. The restriction enzymes *Sph I* and *MluC* were used to double digest the extracted DNA. In order to prepare the library, digested products were barcoded using adapters on the 5' and 3' ends of DNA using both an inline barcode and an Illumina index. Following size selection and pooling, samples were sequenced using Illumina HiSeq 2000, which produced short, unique product sizes up to 150 bp in length. Following the first genomic DNA quality and quantity check, the standard RAD sequencing protocol was used [17, 23].

2.3 Bioinformatics analysis of SNPs derived from ddRAD sequencing data

2.3.1 Alignment and quality control

Raw sequence FastQC was used to quality-check FASTQ files [2]. Using PRINSEQ, adapters and barcode sequences were trimmed across restriction enzymes [19]. Low-quality sequences were eliminated using STACKS [5] based on a PHRED score less than 15. Again, using Bowtie2, quality passed sequences were aligned with the reference genome of Mediterranean buffalo (UOA_WB_1) [13].

2.3.2 Annotation and variant calling

Using Samtools [14], the resulting SAM (Sequence Alignment Format) files were converted into BAM (Binary Alignment Format) files, which were then merged, indexed, sorted, and processed with mpileup to generate a single BCF file. Using vcftools, variant calling was carried out with a quality score of at least 30 and read depths (RD) of 2, 5, and 10 . SnpEff tool was used to annotate the SNPs obtained at RD 10 [6].

Table 1: Various genomic variants identified by SnpEff

Type of Variants	**SnpEff**	
	Count	**%**
3 prime_UTR variant	15,767	0.703
5_prime_UTR_variant	4,104	0.183
Downstream_gene_variant	1,14,179	5.115
Intergenic_variant	4,52,179	20.149
Intron_variant	1,514,544	67.488
Missense_variant	4,742	0.211
Non_coding_transcript_exon_variant	6,793	0.303
Splice_acceptor_variant	72	0.003
Splice_donor_variant	49	0.002
Splice_region_variant	2,299	0.102
Stop_gained	29	0.001
Start_lost	10	0.001
Synonymous_variant	9,103	0.406
Upstream_gene_variant	1,12,810	5.027

Table 2: Count of transitions (Ts) and transversions (Tv) in the genomic sequence of Murrah buffalo

Type of change	Number
Transitions	12,750,309
Transversions	5,000,867
Ts/Tv ratio	2.5496

Table 3: Pattern of base changes (SNPs) in the genomic sequence of Murrah buffalo

BASE	A	C	G	T
A	0	34,715	1,47,354	21,249
C	33,317	0	34,798	1,39,969
G	1,41,017	34,674	0	33,164
T	21,064	1,47,426	34,681	0

3. Results and Discussion

In the era of advanced genetic improvement programs, a fundamental prerequisite for conducting effective association studies, genomic selection, and fine mapping of genes linked to complex phenotypes is a comprehensive understanding of genomic regions. Such knowledge is essential for identifying key genetic markers that influence traits of economic and biological significance in livestock species, including buffalo.

To achieve this objective, we employed the double digest Restriction-site Associated DNA (ddRAD) sequencing approach, which enabled the generation of a total of 252 million raw reads with an average read length of 151 base pairs (bp). After stringent quality control measures, including adapter trimming and filtering of low-quality reads, the processed reads were aligned to the Bubalus bubalis reference genome (UOA_WB_1) using Bowtie 2. This alignment facilitated the identification of high-confidence genetic variants across the sequenced Murrah buffalo genomes.

Variant calling, performed at a read depth (RD) threshold of 10, revealed a total of 814,919 single nucleotide polymorphisms (SNPs) when all Murrah buffalo genomes were combined. In addition, 40,644 insertions and deletions (INDELs) were identified, with a mapping quality threshold set at 30 to ensure accuracy. These findings highlight the genetic diversity within the analysed population and provide a valuable resource for further functional genomic studies.

For annotation, SNPs identified at RD10 were processed using SnpEff, which enabled the classification of 855,563 total variants with an estimated variant rate of one per 3,038 bases. Among these variants, the majority were located within intronic regions (67.48%), followed by intergenic regions (20.17%), indicating that most identified polymorphisms do not directly impact protein-coding sequences. However, 4,742 SNPs (0.211%) were classified as missense variants, which could potentially alter protein function and be of functional importance in trait-associated studies.

Analysis of nucleotide substitution patterns revealed a transition/transversion (Ts/Tv) ratio of 2.5496, reflecting a predominance of transitions over transversions, as commonly observed in mammalian genomes. Specifically, 12,750,309 transition events were identified, compared to 5,000,867 transversions. The most frequently occurring base substitution was cytosine (C) to thymine (T) transitions, with a total count of 147,426 (Table 3). This bias towards C>T transitions is consistent with the expected mutational patterns arising from spontaneous deamination of methylated cytosine residues.

The findings of this study provide a foundational dataset for exploring genetic diversity and marker-assisted selection in Murrah buffalo. The high density of SNPs identified through ddRAD sequencing underscores the potential for genome-wide association studies (GWAS) and genomic selection programs aimed at improving economically significant traits. The discovery of missense variants further opens avenues for investigating candidate genes influencing traits such as milk production, disease resistance, and growth performance. Future research should focus on validating these SNPs in larger buffalo populations and integrating functional genomics approaches to uncover their phenotypic effects.

4. Conclusion

To identify SNPs in indigenous buffalo, this study analyzes sequence alignment data from Murrah buffalo using a reference genome. Furthermore, it will be easier to understand the domestication pattern, environmental adaption, and population mixing of indigenous buffalo attributed to the SNPs discovered in this study. In order to create a low density chip for genomic selection, polymorphic loci identified in the Murrah genome could also be related economically.

5. Acknowledgements

Authors are thankful to Directors of ICAR-National Bureau of Animal Genetic Resources and National Dairy Research Institute and Head of Division of Animal Genetics and Breeding, National Dairy Research Institute, for providing the necessary facilities.

6. Conflict of Interest

The authors have declared no conflict of interests exist.

ETHICAL APPROVAL:

Blood samples were taken according to the applicable guidelines and regulations, which were approved by the Institutional Animal Ethics Committee (IAEC) of the National Bureau of Animal Genetics Resources (ICAR-NBAGR), Karnal.

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

Option 2:

Author(s) hereby declare that generative AI technologies such as Large Language Models, etc. have been used during the writing or editing of manuscripts. This explanation will include the name, version, model, and source of the generative AI technology and as well as all input prompts provided to the generative AI technology

Details of the AI usage are given below:

1.

2.

3.

6. References

1. Altmann A, Weber P, Bader D, Preuß M, Binder EB, Müller-Myhsok B. A beginners guide to SNP calling from high-throughput DNA-sequencing data. Human genetics. 2012 Oct;131(10):1541-54.

2. Andrews S. Babraham bioinformatics-FastQC a quality control tool for high throughput sequence data.URL:https://www.bioinformatics.babraham.ac.uk/projects/fast qc; c2010 Feb.

3. Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics. 2016 Feb;17(2):81-92.

4. Brouard JS, Boyle B, Ibeagha-Awemu EM, Bissonnette N. Low-depth genotyping-by-sequencing (GBS) in a bovine population: strategies to maximize the selection of high quality genotypes and the accuracy of imputation. BMC genetics. 2017 Dec;18(1):1-4.

5. Catchen JM, Amores A, Hohenlohe P, Cresko W. Postlethwait JH. Stacks: building and genotyping loci de novo from short-read sequences. G3: Genes, Genomes, Genetics. 2011;3:171-82.

6. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012 Apr 1;6(2):80-92.

7. Davey JW, Blaxter ML. RADSeq: next-generation population genetics. Briefings in functional genomics. 2010 Dec 1;9(5-6):416-23.

8. Gurgul A, Semik E, Pawlina K, Szmatoła T, Jasielczuk I,Bugno-Poniewierska M. The application of genome-wide SNP genotyping methods in studies on livestock genomes. Journal of applied genetics. 2014 May;55(2):197-208.

9. He Y, Zhou X, Zheng R, Jiang Y, Yao Z, Wang X, *et al*. The Association of an SNP in the EXOC4 gene and reproductive traits suggests its use as a breeding marker in pigs. Animals. 2021 Feb 17;11(2):521.

10. Iqbal N, Liu X, Yang T, Huang Z, Hanif Q, Asif M, Khan QM, *et al*. Genomic variants identified from whole genome resequencing of indicine cattle breeds from Pakistan. PLoS One. 2019 Apr 11;14(4):e0215065.

11. Joshi BK, Singh A, Gandhi RS. Performance evaluation, conservation and improvement of Sahiwal cattle in India. Animal Genetic Resources Information. 2001 Apr;31:43- 54.

12. Keller I, Bensasson D, Nichols RA. Transition transversion bias is not universal: a counter example from grasshopper pseudogenes. PLoS genetics. 2007 Feb;3(2):e22.

13. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357-9.

14. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov 1;27(21):2987-93.

15. Malik AA, Sharma R, Ahlawat S, Deb R, Negi MS, Tripathi SB. Analysis of genetic relatedness among Indian cattle (*Bos indicus*) using genotyping-by-sequencing markers. Animal genetics. 2018 Jun;49(3):242-5.

16. Patel AB, Subramanian RB, Padh H, Shah TM, Mohapatra A, Reddy B, *et al*. Identification of single nucleotide polymorphism from Indian Bubalus bubalis through targeted sequence capture. Current Science. 2017 Mar 25:1230-9.

17. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One. 2012;7(5):e37135.

18. Sanchez JJ, Phillips C, Børsting C, Balogh K, Bogus M, Fondevila M, *et al*. A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis. 2006 May;27(9):1713-24.

19. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011 Mar 15;27(6):863-4.

20. Sebastiani C, Arcangeli C, Torricelli M, Ciullo M, D'avino N, Cinti G, *et al*. Marker-assisted selection of dairy cows for β-casein gene A2 variant. Italian Journal of Food Science. 2022 Apr 2;34(2):21-7.

21. Tao L, He XY, Wang FY, Pan LX, Wang XY, Gan SQ, *et al*. Identification of genes associated with litter size combining genomic approaches in Luzhong mutton sheep. Animal Genetics. 2021 Aug;52(4):545-9. 22. Yang F, Chen F, Li L, Yan L, Badri T, Lv C, *et al*. GWAS using 2b-RAD sequencing identified three mastitis important SNPs via two-stage association analysis in Chinese Holstein cows. bioRxiv. 2018 Jan 1:434340.

22. Surati U, Mohan M, Jayakumar S, Verma A, Niranjan SK. Genome-wide in silico analysis leads to identification of deleterious L290V mutation in RBBP5 gene in *Bos indicus*. Anim Biotechnol. 2023;34(9):4851-4859. doi: 10.1080/10495398.2023.2199502.

23. Utsav Surati, Ymberzal Koul, Mohan M, Gaurav Patel, Anmol and Saket K Niranjan. Genome-wide mining and annotation of SNPs in Bos Indicus. The Pharma Innovation Journal 2022; 11(12): 310-313