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

Abstract
The present study aimed to determine single nucleotide polymorphisms (SNPs) in the Murrah buffalo, 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 of
sequenced genomic data using various bioinformatics tools, a total of 8,55,563 SNPs were identified in Murrah buffalo genome with reference to the *Bubalus bubalis* genome. Annotation revealed that over half of the total variations were in the intronic region (67.49%), followed by the intergenic region (20.15%).The SNPs identified in the present study may serve as molecular markers for economically important traits and may be used in future breed improvement and conservation initiatives.

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. Breeding buffaloes for increased productivity, growth rate, feed conversion efficiency, heat tolerance, and disease resistance requires an understanding of the polymorphism among breeds. In a variety of species, single nucleotide polymorphisms, the most common kind of polymorphisms in the genome, have been employed as genetic markers for marker-assisted selection **[9, 21, 20]**. To determine possible SNPs associated with economic traits, the genome should be screened, followed by annotation of the identified SNPs to predict their function.
SNPs discovered in a population could be absolutely monomorphic to 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.

The present study was aimed toward SNPs identification in Murrah buffalo population using ddRAD sequence data. Murrah is a best-known breed for milk production , widely employed for numerous breed improvement programs in the country. Germplasm of Murrah buffalo has been imported by the countries like Egypt, Brazil, Bangladesh, Srilanka, Thailand, China, Nepal and, Vietnam.
Native milch breed performance has significantly increased as a result of selective breeding and progeny testing programs. Nevertheless there is ample opportunity for genetic advancement through breeding programs which employ genetic markers. The impact of the identified polymorphic loci on economically significant traits may be further assessed.

2. Materials and Methods

2.1 Sample collection and genomic DNA analysis

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.

2.2 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].

2.3 Bioinformatics analyses of SNPs from ddRAD sequence data

2.3.1 Quality control and alignment

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 Variant calling and annotation

Using Samtools [14], the resultant SAM (Sequence Alignment Format) sequence alignment files were converted into BAM (Binary Alignment Format) files, which were then merged, mpileup, indexed, and sorted 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: Number of Transitions (Ts) and Transversions (Tv) in genomic sequence of Murrah buffalo

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

Table 3: Numbers of Base changes pattern (SNPs) in 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 current era of genetic improvement programs, a fundamental prerequisite for effective association studies,an initial understanding of genomic areas is a fundamental prerequisite for effective association studies, genomic selection, and fine

mapping of genes associated with complex phenotypes genomic selection, and fine mapping of genes associated to complex phenotypes is an initial understanding of genomic areas [8].

Therefore, using the ddRAD approach, a total of 252 million raw reads with a mean base-pair length of 151bp were collected. Following the initial quality control, and adapter trimming, the reads were aligned with buffalo reference genome sequence-Bubalus bubalis (UOA_WB_1), using Bowtie 2 tool.

On variant calling at read depths (RD) 10, a total of 8,14,919 SNPs were discovered when all Murrah buffalo genomes were combined. Similarly, 40,644 INDELS were discovere at RD 10 with a mapping quality threshold of 30 [1].

SNPs identified at RD10 were annotated using SnpEff for further processing. The results are given in Table 1 . Using SnpEff, a total of 8,55,563 variants were discovered in the sequence data with variant rate of one variant at every 3,038 bases. The intronic area included more than half of the overall variations (67.48%) followed by the intergenic region (20.17%). A total of 4,742 (0.211%) SNPs were found to be missense variants. With a transition/transversion (Ts/Tv) ratio of 2,5496 among the annotated SNPs, nucleotide transtition (12,750,309) outnumbered transversions (5,000,867) Table 1. The most frequent (1,47,426) substitutions in the base change pattern were cytosine to thymine (Table 3)

4. Conclusion

In order to identify SNPs in the indigenous buffalo, the study presents sequence alignment data from the 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. 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/fastqc; 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.