The Pharma Innovation Journal 2022; 11(12): 310-313

www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277-7695
ISSN (P): 2349-8242
NAAS Rating: 5.23
TPI 2022; 11(12): 310-313
© 2022 TPI
www.thepharmajournal.com
Received: 13-10-2022
Accepted: 19-11-2022

Utsav Surati
1. ICAR-National Dairy Research Institute, Karnal, Haryana, India
2. ICAR-National Bureau of Animal Genetic Resources, Karnal, Haryana, India

Ymberzal Koul
1. ICAR-National Dairy Research Institute, Karnal, Haryana, India
2. ICAR-National Bureau of Animal Genetic Resources, Karnal, Haryana, India

Mohan M
1. ICAR-National Dairy Research Institute, Karnal, (Haryana) India
2. ICAR-National Bureau of Animal Genetic Resources, Karnal, Haryana, India

Gaurav Patel
1. ICAR-National Dairy Research Institute, Karnal, Haryana, India
2. ICAR-National Bureau of Animal Genetic Resources, Karnal, Haryana, India

Anmol
1. ICAR-National Dairy Research Institute, Karnal, Haryana, India
2. ICAR-National Bureau of Animal Genetic Resources, Karnal, Haryana, India

Saket K Niranjan
ICAR-National Bureau of Animal Genetic Resources, Karnal, Haryana, India

Corresponding Author:
Utsav Surati
1. ICAR-National Dairy Research Institute, Karnal, Haryana, India
2. ICAR-National Bureau of Animal Genetic Resources, Karnal, Haryana, India

Genome-wide mining and annotation of SNPs in *Bos indicus*

Utsav Surati, Ymberzal Koul, Mohan M, Gaurav Patel, Anmol and Saket K Niranjan

Abstract
The present study was aimed towards the identification of single nucleotide polymorphism in Sahiwal cattle using the reduced representation sequencing method. Blood samples were collected from ten unrelated Sahiwal cattle, DNA isolated and sequenced using ddRAD approach. After processing of sequenced data using various bioinformatics tools, a total of 279383 SNPs were identified in Sahiwal cattle genome with reference to the *Bos taurus* genome, and upon annotation, nearly half of the total variants were found to be in the intronic region (49.25%), followed by the intergenic region (41.25%). The SNPs identified in this study may serve as molecular markers for economically important traits and may be used in future breed improvement and conservation programs.

Keywords: Cattle**,** ddRAD, NGS, Sahiwal, SNPs

1. Introduction

Across the globe, cattle are economically important and provide a valuable source of food and draught power. There are hundreds of established breeds of cattle that display unique phenotypes, owing to the difference in genetic makeup due to forces of evolution. Understanding the polymorphism within cattle breeds is essential for breeding animals for improved productivity, growth rate, feed conversion efficiency, heat tolerance, and disease resistance. Single nucleotide polymorphisms (SNPs) are the most common type of polymorphism in the genome and have been used as genetic markers for marker-assisted selection in various species [9, 21, 20]. To identify potential SNPs related to economic traits, the genome should be scanned, followed by annotation of the identified SNPs to predict their function. SNPs found in one population may be completely monomorphic in another population [18], and hence genetic markers need to be identified for both *Bos taurus* and *Bos indicus* cattle breeds. However, bovine research is dominated by European taurine breeds and available SNP databases are biased toward taurine breeds [10]. Hence, for the improvement of native breeds, it is imperative to identify SNPs that are exclusive to Indicine breeds.

Advances in next-generation sequencing methods have made SNP discovery easier as reference genomes are now available for all the major livestock species. For genome-wide SNP identification, either the whole genome may be sequenced, or a sub-sampling method can be utilised. Sub-sampling methods or reduced representation methods are a good alternative to whole genome sequencing as they are cost-effective, computationally faster and allow a broad range of polymorphic loci testing with no reference sequence needed and no prior knowledge requirement. Restriction site-associated DNA sequencing (RADseq) is one such next-generation sequencing method that involves the use of restriction enzymes and molecular identifiers to study a fraction of the total genome [7]. RAD sequencing involves many techniques which are referred to as the 'RADseq family' [3]. One of the techniques of the RADseq family, known as double digest RAD sequencing (ddRAD), involves the use of a second restriction enzyme for the digestion of genomic DNA to reduce the cost and time for library preparation [17]. It also enables paired-end sequencing of identical loci across multiple samples, thus addressing a major shortcoming of the original RADseq method.

The present study was aimed toward SNP discovery in Sahiwal population using ddRAD sequence data. Sahiwal cattle is a prominent milch breed of India, famous for its endurance in hot subtropical climates, disease resistance, and low maintenance cost. The average milk yield of Sahiwal cattle in organized farms ranges between 1500 to 2500 kg and can go beyond 4500 kg in well-managed herds.

The average age at first calving, calving interval, and lactation length are 36 months, 420 days, and 315 days respectively. The fat and solids-non-fat (SNF) percent range from 4.6 to 5.2 and 8.9 to 9.3 percent, respectively [11]. Sahiwal cattle have been utilised to develop crossbred strains of dairy cattle like Karan Swiss, Karan Fries, Frieswal, Jamaica Hope and Australian Milking Zebu. Selective breeding and progeny testing schemes have greatly improved the performance of native milch breeds. However, there is still scope for genetic improvement via the employment of genetic markers in breeding programs. The identified polymorphic loci may further be evaluated for their impact on economically important traits.

2. Materials and Methods

2.1 Sample collection and genomic DNA analysis

The samples were collected from ten unrelated Sahiwal (*Bos indicus*) cattle 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. After collecting blood samples, DNA extraction was performed, followed by quality checking, concentration and purity checking of extracted DNA for future analyses.

2.2 Library preparation and ddRAD sequencing

After the initial quality and quantity check of genomic DNA, the standard RAD sequencing protocol was followed [17]. Double digestion of isolated DNA was carried out using restriction enzymes *Sph* I and *Mlu*C. Digested products were barcoded with adapters on 5 ' and 3 ' ends of DNA using both Illumina index and an inline barcode for library preparation. After pooling and size selection, sequencing of samples was done using Illumina HiSeq 2000 which generated short and unique product size up to 150 bp length.

2.3 Bioinformatics analyses of ddRAD sequence data

2.3.1 Quality control and alignment of sequences

Raw sequence FASTQ files were quality-checked using FastQC [2]. Trimming of adapters and barcode sequences over restriction enzymes was done using PRINSEQ [19]. The sequences which were low in quality were discarded using STACKS [5] on the basis of a PHRED score < 15. Alignment of quality passed sequences with *Bos taurus* (Ensembl's ARS-UCD1.2) reference genome was done using Bowtie2 [13].

2.3.2 Variant calling and annotation

The resulting SAM (Sequence Alignment Format) sequence alignment files were converted into BAM (Binary Alignment Format) files using Samtools [14] which was further sorted, indexed, merged and mpileup to get a single BCF file. Variant calling was done using vcftools at read depth (RD) of 2, 5, and 10 with quality score≥30 (Fig. 1). Annotation of the SNPs obtained at RD 10 was done using SnpEff [6].

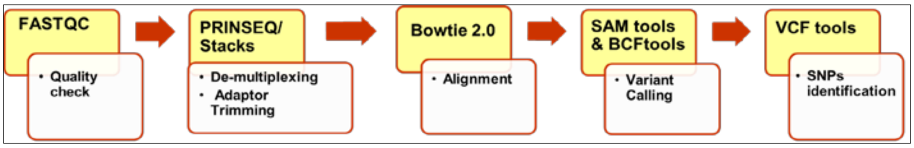

Fig 1: Bioinformatics Pipeline For the identification of SNPs from raw reads

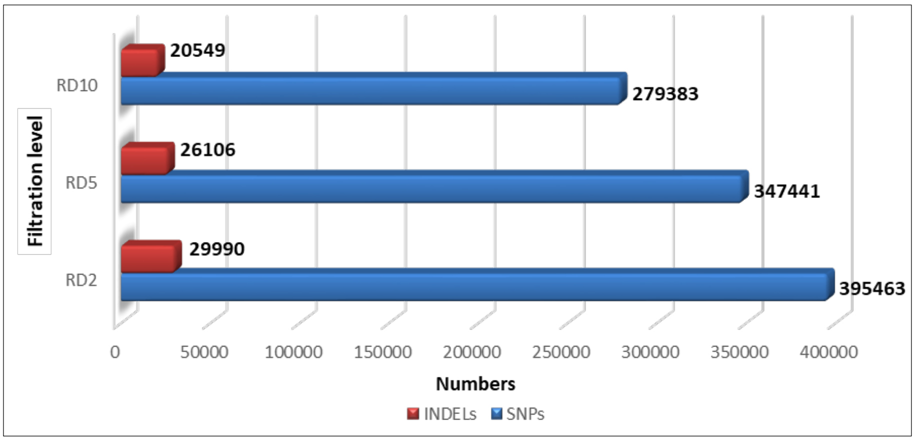

Fig 2: Number of SNPs and INDELs identified at different filtration level

Table 1: Various genomic variants identified by SnpEff

VARIANTS	SnpEff	
	numbers	%
3_prime_UTR_variant	966	0.22
5_prime_UTR_variant	464	0.11
downstream_gene_variant	18,952	4.27
intergenic_variant	1,83,053	41.25
intron_variant	2,18,595	49.25
missense_variant	692	0.16
non_coding_transcript_exon_variant	254	0.06
splice_acceptor_variant	2	0
splice_donor_variant	3	0
splice_region_variant	355	0.08
start_lost	2	0
stop_gained	9	0
stop_retained_variant	1	0
synonymous_variant	1,607	0.36
upstream_gene_variant	18,855	4.25

Table 2: Number of Transitions (Ts) and Transversions (Tv)

Transitions	13,89,995
Transversions	5,38,289
Ts/Tv ratio	2.5822

Table 3: Numbers of Base changes pattern (SNPs) in Sahiwal genomic sequences

BASE	A	C	G	T
A	0	10,577	46,374	6,723
C	10,936	0	11,386	54,018
G	54,133	11,368	0	11,110
T	6,797	46,263	10,635	0

3. Results and Discussion

The preliminary knowledge of genomic regions is a crucial requirement for effective association studies, genomic selection, and fine mapping of genes associated with complex phenotypes in the current era of genetic improvement programmes [8]. Therefore, using the ddRAD approach, the current experiment was conducted on Sahiwal cattle and 279383 SNPs specific to Sahiwal were identified.

The ddRAD sequence data was generated for ten Sahiwal cattle. A total of 31.98 million raw reads were obtained after ddRAD sequencing with mean base pair length of 151bp. After the initial quality control consisting of adapter trimming and removal of poor-quality reads, a total of 29.64 million reads were obtained, with an average 95.05% of the total raw reads. In the current study, STACKS was used to examine the mean quality score across windows, whereas PRINSEQ software was used to calculate the mean quality score throughout the read, which could result in the loss of more reads. After assessing quality control initially, the retained reads were aligned with *Bos taurus* (ARS-UCD1.2) of Ensembl (http://ftp.ensembl.org/pub/release-104/fasta/bos_taurus/dna/) using Bowtie2 as it is comparatively faster and sensitive than BWA [13].

The variants discovered in Sahiwal cattle were filtered for a quality score ≥ 30 and a depth of 2, 5, or 10 reads. A total of 395463, 347441 and 279383 SNPs were identified at read depths (RD) 2, 5 and 10 respectively in Sahiwal cattle. Likewise a total of 29990, 26106 and 20549 INDELs were identified at read depths 2, 5 and 10, respectively (Fig. 2). The SNPs identified in the current study were similar to earlier reports. By using a GBS method, Malik *et al.* (2018) [15] found 107488 SNPs in 24 animals from seven different Indian cattle breeds. Yang *et al.* (2018) [22] used modified RAD sequencing and found 1058 SNPs in 40 dairy cows. Brouard *et al.* (2017) [4] found 272,103 SNPs in 48 dairy cows using double enzyme GBS approach.

SNPs obtained at RD10 were further annotated with SnpEff (Table 1) as reported in the previous study [1]. Among total variants, nearly half of the variants were found to be in the intronic region (49.25%), followed by the intergenic (41.25%) region. A total of 1607 (0.36%) SNPs were synonymous, while 692 (0.16%) were missense/nonsynonymous. Nine stop gained (non-sense) SNPs were also found. Variant rate was estimated to be one variant at every 9,408 bases. More transitions (13,89,995) were found in annotated SNPs than transversions (5,38,289), with a ratio (transition/transversion, Ts/Tv) of 2.58 (Table 2). The ratio is a little higher than usual. However, the ratio varies between species and even within individuals from region to region [12]. It can also be higher than the average ratio when target-based approaches are used on the genome [16]. Numbers of base changes pattern in SNPs are shown in Table 3.

4. Conclusion

The study reports sequence alignment data from the Sahiwal cattle using a reference genome to find SNPs in indigenous cattle of India. Further, the SNPs found in this work will make it easier to understand the domestication pattern, adaptation to local environment, and population admixture of indigenous cattle. Polymorphic loci present in Sahiwal genome may further be associated with economically important traits so that a low density chip can be developed for genomic selection.

5. Acknowledgements

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

6. Conflict of Interest

The authors have declared no conflict of interests exist.

7. 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.