Indian J Dairy Sci 77(1): 71-75
https://doi.org/10.33785/IJDS.2024.v77i01.010

RESEARCH ARTICLE

Genome-wide SNP identification and annotation from high coverage whole genome sequenced data of Bhadawari buffalo

Ameya Santhosh, Vikas Vohra*, Rani Alex and Gopal Gowane

Received: 30 August 2023 / Accepted: 11 September 2023 / Published online: 23 February 2024

Abstract: Eukaryotic genomes are rich in Single Nucloetide Polymorphisms(SNPs) which have gained immense importance as the genetic marker of choice especially for the production traits in livestock species. The high throughput sequencing technologies, which offer quick and greater coverage genome sequencing data has improved the accuracy of SNP discovery through various bioinformatic pipelines. This is the preliminary study to uncover the SNPs from high coverage (45x) sequenced data of the unique Bhadawari buffalo breed of India. Blood was collected from two pure bred animals from the breeding tract and DNA was sequenced using Illumina Highseq technology. The sequences were aligned to the buffalo reference genome and variant calling was done.The variant count (SNPs, Insertions, and deletions) discoverable at six ascending read depths (2,5,10, 20, 30 and 40) have been analysed. The Transition to Transversion ratio (Ti/Tv ratio) is found to be nearly 2.2. Chromosome wise distribution of SNPs showed that in all read depths maximum number of SNPs were found to be in the 1st chromosome of Bhadawari buffalo genome.

Keywords: Bhadawari buffalo, Whole genome sequencing, SNP, Read depth

Introduction

Buffaloes are hardy animals maintained as a source of income in the households of marginal farmers of India. The technical barriers for scientific upgrading of indigenous buffaloes have been greatly overcome in the last few years (Kumar et al. 2019). Buffalo genome, especially that of the Indian buffaloes other than Murrah buffalo remain unexplored to a great extent.Considering the decline in the population of pure indigenous breeds due to intense cross breeding, we are at the verge of losing the diversity of the indigenous buffalo genomes shortly. Especially for the breeds with peculiar properties like Bhadawari. Bhadawari buffaloes are reported to be the Indian buffalo breed with highest milk fat percentage. These animals are reported to be economical for maintaining due to their hardy nature and have less calf mortality rate. Their populations are mainly concentrated in the Indian states of Uttar Pradesh and Madhya Pradesh and scattered in northern parts of India (Pundir et al. 1996). The population of this breed is declining and has reached few thousands in recent years. Recent studies have evaluated the mixing of Murrah genome to Bhadawari due to cross breeding (Tyagi et al. 2021).

The upliftment of a particular breed with a declining population like that of Bhadawari should be done in an intensive manner. Even though selective breeding and improvement can be an option, the time and cost required for such programs limits its usage in comparison to the genomic selection (Choi et al. 2013). Among the various technologies for unveiling the valuable genomic areas of the large animal genome, Whole genome sequencing give benefits like uniformity of read coverage and detecting the polymorphisms outside the coding regions (Meynert et al. 2014). Single Nucleotide Polymorphisms are emphasized now a days for analysing the variations in the genome due to their abundance in the genome (Schultz et al. 2020). Detection of structural variants in livestock carries many challenges and bias due to the quality of the reference assembly, false positives in annotation and difficulty in detecting karyotype errors (Bickhart and Liu, 2014). Most research rely on low coverage data due to the high expense of whole genome sequencing. However, this may result in an error rate more than 15%, affecting further downstream analysis such as Runs of homozygosity (ROH) area detection, emphasising the significance of high coverage data (> 30x) to yield more accurate results. (Ceballos et al. 2018). Recent studies suggest that accuracy and number of structural variants per bovine sample increased as the coverage of the short read whole genome sequences data increased from 15x to 47x (Lee et al. 2023).

Department of Animal Genetics and Breeding, National Dairy Research Institute(ICAR-NDRI), Karnal, India, Haryana

Ameya Santhosh: ameyasanthosh1128@gmail.com, ORCID 0000-0003-2689-7613
Vikas Vohra:vohravikas@gmail.com , ORCID 0000-0001-7581-4939
Rani Alex: ranialex01vet@gmail.com , ORCID 0000-0001-9163-3724
Gopal Gowane: gopalgowane@gmail.com,ORCID 0000-0001-6535-7818

Vikas Vohra(✉)
Department of Animal Genetics and Breeding, National Dairy Research Institute(ICAR-NDRI), Karnal, India, Haryana

This is a preliminary and first study to uncover the Variants (SNPs and InDels) from high coverage whole genome sequenced data from Bhadawari buffalo breed. The results of this study pave light to the comparison of potential variants in this buffalo breed's genome along with the assessment of counts of SNPs at different filtration level which can only be done using high coverage sequenced data.

Materials and methods

DNA isolation and sequencing

Genomic DNA was isolated by Phenol Chloroform method (Sambrook and Russel, 2004) from the blood collected from two healthy Bhadawari Buffaloes from the breeding tract in Etawah district of Uttar Pradesh. Illumina HighSeq 2000 technology was used to sequence the 150bp paired end libraries to obtain high quality sequenced data.

Pre-processing of data and Mapping

The quality of the whole genome sequenced data was evaluated by FastQC v0.12.1 and quality control was performed by Trimmomatic version 0.39. The quality-controlled reads were mapped to reference genome (NDDB_SH_1) of Murrah buffalo using BWA mem (BWA v 0.7.17).

Variant calling and annotation

Genome wide variants were identified using Samtools (samtools v 0.1.19) and BCFtools (bcftools v 0.1.19). Samtools was used for the conversion of sam file to bam format followed by sorting, indexing, and merging the bam files. The conversion to vcf format was done using Bcftools. The variants were filtered using Vcftools (vcftools v 0.1.16) with a PHRED >30 and keeping maximum read depth 500 and minimum as 2,5,10,20,30,40.

Screening of SNPs annotated with genes related to Fat percentage in buffalo

Candidate genes associated with Milk fat percentage was searched from the literature (Vohra et al. 2021, Deng et al. 2016, Ferritas et al. 2016, Dubey et al. 2015, Tanpure et al. 2012) and the the SNPs detected from Bhadawari buffalo genome annotated with these genes were enumerated.

Results and Discussion

Detection of Variants (SNP and InDels)

45X coverage data was remaining after quality control and adapter removal of the 54 X coverage raw sequenced data**.** Leading and trailing bases with a quality score less than 5 and reads with a length less than 50 bp and PHRED score less than 33 was removed in this process. The quality-controlled reads of the samples on alignment with *Bubalus bubalis* reference genome (NDDB_SH_1) showed 99.86% and 99.88% mapping rate. The number of SNPs annotated at 6 different filtration levels from Bhadawari buffalo genome are shown in Table 1. The total number of SNPs and InDels were in near range upon filtration at read depth 2, 5, 10 and 20. The number of SNPs annotated was 12,104,222; 10,607,561 and 6,724,029 at read depth 20, 30 and 40. The number of InDels annotated were 1,135,489; 982,999 and 614,616 at read depth 20, 30 and 40 read depth.

In a similar study conducted on a large scale with 71 buffaloes, about 28,347,965 SNPs were detected from 12x coverage data with less than 2 read depth (Chen et al. 2023). In all read depths, largest number of SNPs were detected from chromosome number 1 (Table 2) of the genome followed by chromosome number 2. Majority of the SNPs were in the Intronic region just like the results studies in buffalo (Chen et al. 2023) and other large eukaryotes (Hedayat-Evrigh et al. 2020). The percentage of variants in intronic and intergenic regions were 67.4% and 19.145% respectively (Figure 1). Nearly similar percentages of 61.9 % and17.71% are reported from ddRAD data of sixty-three Murrah buffaloes (Mishra et al. 2020). In other bovine species including cattle, 6 million SNPs were detected from 15x coverage were data of a single animal at read depth less than 5 (Kawahara-Miki et al. 2011). These results clearly suggest the increase in SNP counts at higher coverage. When the number of purebred animals is less the accuracy and efficiency of detection of variants should be improved by a higher coverage data.

The missense (28.4%) to silent (71.3%) ratio was 0.3 ratio 986 (Table 3). The transition to transversion ratio (Ti/Tv ratio) indicating the rate of substitution mutations (Wang et al. 2014) is also considered as a quality parameter for high throughput sequencing studies (Durbin et al. 2010). The transitions (10,904,964) and transversion (4,897,340) ratio was found to be 2.2 in the current study. A ratio of 2.0-2.1 is mentioned for WGS

Table 1: Number of Single Nucleotide Polymorphisms and Insertion- Deletions (SNPs & InDELs) remaining after filtration at read depth (RD) 2,5,10, 20,30 and 40 from the whole genome sequenced data

Species	*Bubalus Bubalis* (Bhadawari)					
Filteration level	RD2	RD5	RD10	RD20	RD30	RD40
SNPs	12,361,170	12,351,047	12,323,806	12,104,222	10,607,561	6,724,029
Ins	557,662	557304	555441	540512	468734	296098
Del	611,371	611045	609529	594977	514265	318518
Total variants	13,530,203	13519396	13488776	13239711	11590560	7338645

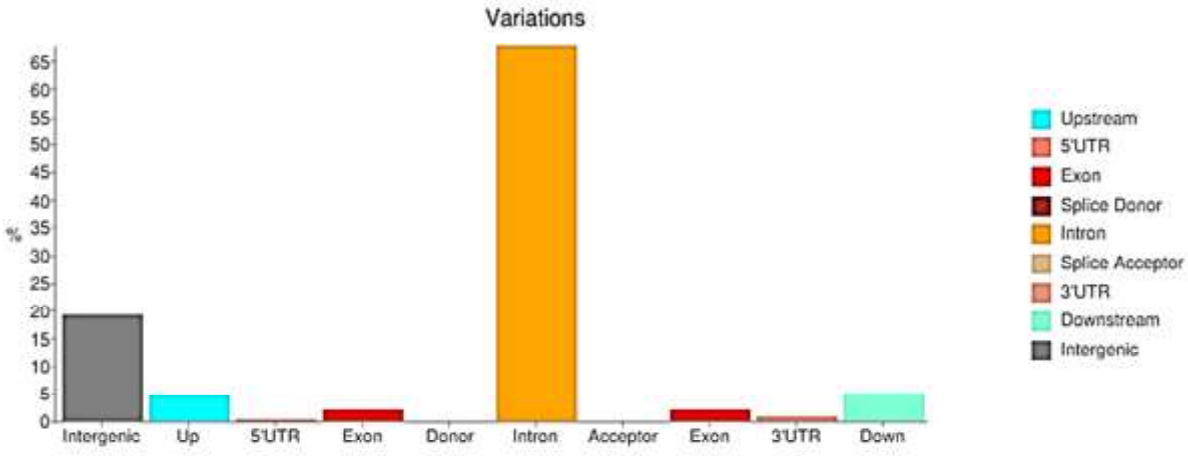

Fig. 1 Percentage of total variants region wise

Table 2: Chromosome wise distribution of SNPs

Chromosome number	Number of SNPs detected at different read depth		
	20	30	40
1	968746	849557	536281
2	858202	753298	479808
3	773600	676048	423107
4	770277	676700	433880
5	616289	540124	342834
6	542018	473920	296750
7	565283	497268	315029
8	555100	487101	307756
9	521672	456142	286810
10	497119	438000	281151
11	473976	415015	263430
12	470116	410797	255327
13	540852	476746	323856
14	374621	326247	201428
15	392421	344207	215738
16	427299	370366	232427
17	353463	309368	194473
18	294834	257556	161822
19	367546	322419	202586
20	324520	282285	175991
21	276071	240628	149460
22	308793	270124	167966
23	257032	225038	140742
24	198184	172223	104650
X	376141	336337	230680
Mitochondrial	47	47	47
Total	12104222	10607561	6724029

data by the International Genome Sample Resource (http://www.1000genomes.org/).

Variant annotation

Table 3: Number of SNPs annotated along with the genes related to milk fat percentage from the Bhadawari buffalo genome

Sl No.	Reported gene IDs associated with fat percentage in buffalo	Gene name	The dairy performance pathways and processes associated with the gene	No. of SNPs in Bhadawari genome annotated for the gene
1	CAMTA1	Calmodulin Binding Transcription Activator 1	Fatty acid metabolism (Vohra et al.2021)	3603
2	CCSER1	Coiled-coil serine rich protein 1	Growth and feed efficiency in beef cattle (Abo-Ismail et al. 2018)	7044
3	DGAT1	Type I Diacylglycerol O-acyltransferase	Glycerol 3 phosphate pathway (Khan et al. 2021)	72
4	LEP	Leptin gene	Regulation of bone remod elling (Haruna et al. 2021)	86
5	MC4R	Melanocortin 4 Receptor	Energy metabolism and regulation of feeding behaviour and metabolism (Deng et al.2016)	91
6	SCD	Stearoyl-CoA Desaturase	Saturated fatty acid metabolism (Rincon et al. 2012)	213
7	SREBF1	Sterol regulatory element binding protein 1	SREBF1 pathway, De novo synthesis of saturated fatty acids (Rincon et al. 2012)	60
8	STAT1	Signal transducer and activator of transcription 1	Tri glyceride synthesis (Thaller et al.2003)	188
9	TG	Thyroglobulin	Lipid metabolism (Kaczor etal.2017)	13067
10	ETS2	ETS proto-oncogene 2	Expression of other genes in mammary epithelial cells (Anderson et al. 2007)	97
11	ROR1	Receptor Tyrosine Kinase Like Orphan Receptor1	Mammary gland development (Dillon et al. 2002)	1708
12	CACNG6	Calcium voltage -gated channel auxiliary subunit gamma 6	Calcium channel stabilization (Lee et al. 2010)	119
13	SH3BP5L	SH3 binding domain protein 5	Membrane transport of lactose (Lopdell et al.2017)	59
14	ZNF672	Zinc Finger Protein 672	Energy metabolism (Zhou et al. 2022)	22

SNVs annotated with the genes reported to be related to milk fat percentage were enumerated in Table 3. Among them largest number of SNPs were found to be affecting the gene Thyroglobulin (TG) followed byCCSER1 and CAMTA1. The genomic regions associated with the milk fat related genes ETS2, ROR1, CACNG6, SH3BP5L, ZNF672, CAMTA1, CCSER1 and DGAT1were found from Genome wide association study of Murrah buffalo (Vohra et al. 2021 & Ferritas et al. 2016). Similarly, the SNPs associated with LEP gene was annotated in earlier studies from Mehsana (Tanpure et al. 2012), MC4R from Chineese buffalo, SCD from Niliravi buffalo, SERBF1 from Niliravi× Guanxi buffaloes (Deng et al. 2016) and TG from Mehsana and Niliravi buffaloes (Dubey et al. 2015).

Conclusions

This was the first study which explored the Bhadawari buffalo genome to identify the genetic variants through high coverage next generation sequencing technology. The variants were screened at higher read depths of 20, 30 and 40 to get an SNP count of 12, 10.6 and 6.7 million SNPs from 45x coverage data.

Chromosome number 1 harbours most of the SNPs in the Bhadawari genome. Our study suggests to incorporate higher coverage data for variant calling for keeping the accuracy of downstream processing. Further studies with a large number of Bhadawari animals is recommended in future for the genome wide association of the variant positions that are annotated in this study.

Acknowledgement

We thank the Director, ICAR-NDRI, Karnal for providing facilities for conducting the research work

References

Abo-Ismail MK, Lansink N, Akanno E, Karisa BK, Crowley JJ, Moore SS, Bork E, Stothard P, Basarab JA, Plastow GS (2018) Development and validation of a small SNP panel for feed efficiency in beef cattle. J Anim Sci 96:375-397.

Anderson SM, Rudolph MC, McManaman JL (2007) Key stages in mammary gland development, Secretory activation in the mammary gland: it is not just about milk protein synthesis. Breast Cancer Res 9:204

Bickhart DM, Liu GE (2014) The challenges and importance of structural variation detection in livestock. Front Genet 5:37

Ceballos FC, Hazelhurst S, Ramsay M (2018) Assessing runs of Homozygosity: a comparison of SNP Array and whole genome sequence low coverage data. BMC genomics 19:1-12

Chen Z, Zhu M, Wu Q, Lu H, Lei C, Ahmed Z, Sun J (2023) Analysis of genetic diversity and selection characteristics using the whole genome sequencing data of five buffaloes, including Xilin buffalo, in Guangxi, China. Front genet 13:1084824

Choi JW, Choi BH, Lee SH, Lee SS, Kim HC, Yu D, Chung WH, Lee KT, Chai HH, Cho YM, Lim D (2015) Whole-genome resequencing analysis of Hanwoo and Yanbian cattle to identify genome-wide SNPs and signatures of selection. Molecules cells 38(5):466

Deng, T.X., Pang, C.Y., Liu, M.Q., Zhang, C, Liang, X.W (2016) Synonymous single nucleotide polymorphisms in the MC4R gene that are significantly associated with milk production traits in water buffaloes. Genet Mol Res 15:1-8

Deng T, Pang C, Ma X, Lu X, Duan A, Zhu P, Liang X (2016) Four novel polymorphisms of buffalo INSIG2 gene are associated with milk production traits in Chinese buffaloes. Mol Cell Probes 30:294-299

Dillon C (2002) *Intracellular trafficking, and function of receptor tyrosine kinases in mammary gland development*. University of London, University College London (United Kingdom).

Dubey PK, Goyal S, Mishra SK, Yadav AK, Kathiravan P, Arora R, Malik R, Kataria RS (2015) Association analysis of polymorphism in thyroglobulin gene promoter with milk production traits in riverine buffalo (Bubalus bubalis). Meta gen 5:157-161

Freitas AC, De Camargo GMF, Aspilcueta-Borquis RR, Stafuzza NB, Venturini GC, Tanamati F, Hurtado-Lugo NA, Barros CC, Tonhati H (2016) Polymorphism in the A2M gene associated with high-quality milk in Murrah buffaloes (*Bubalus bubalis*) Genet Mol Res 15

Haruna IL, Zhou H, Hickford JG (2021) Variation in bovine leptin gene affects milk fatty acid composition in New Zealand Holstein Friesian× Jersey dairy cows. Arch Anim Breed, 64:245-256

Hedayat-Evrigh N, Khalkhali-Evrigh R, Bakhtiarizadeh MR (2020) Genome-wide identification and analysis of variants in domestic and wild bactrian camels using whole-genome sequencing data. Int J Genomics 2020

Kaczor U, Famielec, M, Dudziak P, Kaczor A, Kucharski M, Mandecki A (2017) Fatty acid binding protein 4 (FABP4) and thyreoglobulin (TG) polymorphisms in relation to milk performance traits in the Holstein-Friesian cattle. Acta Scientiarum Polonorum Zootechnica 16

Kawahara-MikiR, Tsuda K, ShiwaY, Arai-Kichise Y, Matsumoto T, Kanesaki Y, Oda SI, Ebihara S, Yajima S, Yoshikawa H, Kono T (2011) Whole-genome resequencing shows numerous genes with nonsynonymous SNPs in the Japanese native cattle Kuchinoshima-Ushi. BMC Genomics 12:1-8

Khan MZ, Ma Y, Ma J, Xiao J, Liu Y, Liu S, Khan A, Khan IM, Cao Z (2021) Association of DGAT1 with cattle, buffalo, goat, and sheep milk and meat production traits. Front Vet Sci 8: 712470

Kumar M, Dahiya SP, Ratwan P, Kumar S, Chitra A (2019) Status, constraints, and future prospects of Murrah buffaloes in India. Indian J Anim Sci 89:1291-1302

Lee YL, Bosse M, Takeda Moreira GCM, Karim L, Druet T, Oget-Ebrad C, Coppieters W, Veerkamp RF, Groenen MA, Georges M (2022) High-resolution structural variation catalogue in a large-scale whole genome sequenced bovine family cohort data (preprint available in research square)

Lee JS, Kim JH, Bae JS (2010) Association of *CACNG6* polymorphisms with aspirin-intolerance asthmatics in a Korean population. BMC Med Genet 11: 138

Lopdell TJ, Tiplady K, Struchalin M, Johnson TJJ, Keehan M, Sherlock R (2017) DNA and RNA-sequence based GWAS highlights membrane-transport genes as key modulators of milk lactose content. BMC Genomics 18:968

Meynert AM, Ansari M, FitzPatrick DR, Taylor MS (2014) Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinform 15:1-11

Mishra DC, Sikka P, Yadav S, Bhati J, Paul, SS, Jerome A, Singh I, Nath A, Budhlakoti N, Rao AR, Rai A (2020) Identification and characterization of trait-specific SNPs using ddRAD sequencing in water buffalo. Genomics 112:3571-3578

Pundir RK, Vij PK, Singh RV, Nivsarkar AE (1996) Bhadawari buffaloes in India. Anim Genet Resour 17:101-113

Rincon G, Islas-Trejo A, Castillo AR, Bauman DE, German BJ, Medrano JF (2012) Polymorphisms in genes in the SREBP1 signalling pathway and SCD are associated with milk fatty acid composition in Holstein cattle. J Dairy Res 79:66-75

RM Durbin (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

Sambrook J, Russell DW (2001) Molecular Cloning-Sambrook & Russel-Vol. 1, 2, 3 Cold Springs Harbor Lab Press: Long Island, NY, USA

Schultz B, Serão N,Ross JW (2020) Genetic improvement of livestock, from conventional breeding to biotechnological approaches. Anim Agric 393–405

Tanpure T, Dubey PK, Singh KP, Kathiravan P, Mishra BP, Niranjan SK, Kataria RS (2012) PCR-SSCP analysis of leptin gene and its association with milk production traits in river buffalo (Bubalus bubalis). Trop Anim Health Prod 44:1587-1592

Thaller G, Kühn C, Winter A, Ewald G, Bellmann O, Wegner J, Zühlke H, Fries R (2003) DGAT1, a new positional and functional candidate gene for intramuscular fat deposition in cattle. Animal Genet 34:354-357

Tyagi SK, Mehrotra A, Singh A, Kumar A, Dutt T, Mishra BP, PandeyAK (2021) Comparative signatures of selection analyses identify loci under positive selection in the Murrah Buffalo of India. Front Genet 12:673697

Vohra V, Chhotaray S, Gowane G, AlexR, Mukherjee A, Verma A, Deb SM (2021) Genome-wide association studies in Indian Buffalo revealed genomic regions for lactation and fertility. Front Genet 12:696109

Wang J, Raskin L, Samuels DC, Shyr Y, Guo Y (2015) Genome measures used for quality control are dependent on gene function and ancestry. Bioinformatics 31:318-323

Zhou Y, Wang Y (2022) Prognostic implication of an energy metabolism related 11 gene signature in lung cancer. J. Biochem. Mol. Toxicol 36:23171