B-4934
[1-6]

RESEARCH ARTICLE Indian Journal of Animal Research

Identification of Genome-wide Single Nucleotide Polymorphisms in Indigenous Cattle Breeds of Tamil Nadu

L. Mahalakshmi, R. Thiagarajan, S.M.K. Karthickeyan, Sabiha Hayath Basha 10.18805/IJAR.B-4934

ABSTRACT

Background: Indigenous cattle are well adapted to the hot tropical climate. Conservation and genetic improvement in the indigenous cattle breeds have become a necessity due to the changing climatic condition. Genome-based research in indigenous cattle for in-depth characterization of genetic variants will help in the improvement of economically important traits in these cattle breeds.

Methods: Blood samples were collected for five indigenous cattle breeds of Tamil Nadu from their breeding tracts. Pooled DNA samples were sequenced using the Illumina platform. Short sequences were aligned using BWA, variant calling and annotation were done using GATK and snpEff respectively.

Result: This study identified 29,366,340 variants, which include 25,944,935 SNPs and 3,421,405 indels with a variant rate of 91 bp across the genome. Around 215,820 SNPs were present in the coding region of 5,572 genes. This study will provide the framework for further genetic analysis of phenotypic variations in economically important traits in native cattle of Tamil Nadu.

Key words: BWA, GATK, Illumina, Indels, SnpEff, SNPs.

INTRODUCTION

In general, indigenous cattle are well-adapted to the tropical climate and resistant to most of the tropical diseases. They can survive well in a low input system by converting low-quality forage into high-quality milk and draught power Obeidat *et al.* (2002).

In India, the genetic improvement of native cattle has been implemented only in few breeds by selective breeding of proven bulls born to elite cows through artificial insemination. Selective breeding has contributed to significant genetic improvement in economically important traits, but the accuracy of selection has been low because of less availability of animals with records and the unavailability of deep pedigree information for selection. Genome-based research in indigenous animals has gained growing attention nowadays as evidenced by the release of the whole genome sequence of Nellore, Gir and Brahman cattle as well as the DNA chips such as Induschip, Indigau *etc.*

Over the last few years, a considerable number of genetic variants in the form of single nucleotide polymorphisms, indels and structural variations have been identified across the cattle genome as a result of a number of bovine whole-genome sequencing studies, HapMap and the 1000 bull genome projects, catalogued in dbSNP (https://www.ncbi.nlm.nih.gov/snp/). As of 2017, a total of 99.71 million SNPs have been deposited in dbSNP, among which, only 4.38 million SNPs are from indicine cattle. This indicates that more than 96 per cent of currently available SNPs are from taurine breeds. Consequently, the SNP arrays currently used for genotyping, genome-wide association studies and genomic selection are biased towards taurine cattle, impeding the accuracy of genomic selection in indicine cattle breeds (Iqbal *et al.* 2019). Hence, there is a need to discover a substantial number of SNPs across many indigenous breeds, belonging to various geographical regions, to build new high-density SNP arrays for unbiased genotyping and subsequent genomic selection for economically important traits in indigenous cattle.

In this study, we report for the first time the genetic characterization of five indigenous cattle breeds of Tamil Nadu, *viz.* Kangayam, Umblachery, Pulikulam, Bargur and Alambadi, which are known for their draught ability, through analyzing whole-genome sequence data. Moreover, this study will provide a valuable resource for genome-wide association studies and genomic selection, which will also help in further investigation of genetic mechanisms underlying the traits of interest in indigenous cattle.

MATERIALS AND METHODS

A population of 304 animals, in agreement with the breed characteristics from all five indigenous cattle breeds of Tamil Nadu, were selected from their breeding tracts for this study.

Department of Animal Genetics and Breeding, Madras Veterinary College, Tamil Nadu Veterinary and Animal Sciences University, Chennai-600 007, Tamil Nadu, India.

Corresponding Author: R. Thiagarajan, Department of Animal Genetics and Breeding, Madras Veterinary College, Tamil Nadu Veterinary and Animal Sciences University, Chennai-600 007, Tamil Nadu, India. Email: thiagarajan.r@tanuvas.ac.in

How to cite this article: Mahalakshmi, L., Thiagarajan, R., Karthickeyan, S.M.K. and Basha, S.H. (2022). Identification of Genome-wide Single Nucleotide Polymorphisms in Indigenous Cattle Breeds of Tamil Nadu. Indian Journal of Animal Research. DOI: 10.18805/IJAR.B-4934.

Submitted: 18-05-2022 **Accepted:** 06-12-2022 **Online:** 12-12-2022

Among the total population, 225 animals were unrelated (38 Pulikulam, 63 Bargur, 48 Kangayam, 61 Umblachery and 15 Alambadi). Blood samples were collected from the jugular vein of the animals, under aseptic conditions, using vacutainers containing 0.5 per cent of Ethylene Diamine Tetra Acid (EDTA) anticoagulant. Collected samples were brought to the laboratory in a Thermo container and stored at -20°C until subjected to DNA extraction. Genomic DNA was extracted by using the standard Phenol-Chloroform method Sambrook and Russel. (2001). The quantity and quality of DNA were measured with Nanodrop™ 1000 spectrophotometer (Thermo Scientific, USA). Samples with an optical density ratio (260/280 nm) of 1.8 were considered for further processing. Agarose gel electrophoresis (0.8%) was used to check the integrity of genomic DNA.

DNA samples from each breed were classified into three groups based on sex and phenotype of the animals viz bulls, high milk yielders (more than two liters per day) and low milk yielders (less than one liter per day) for better SNP detection Liao *et al.* (2013). For unbiased and accurate estimation of allele frequencies, the concentration of DNA samples was standardized to 200 ng per microliter and samples were pooled by mixing equal amounts (10-20 µl) of DNA from each sample from a particular group Gautier *et al.* (2013).

Table 1: Mapping statistics of sequence data.

Sample	QC Passed reads	Mapped reads	unmapped reads	Mapped reads %	Depth
ACG1	34,41,33,466	34,05,19,828	36,13,638	98.95%	14.2
ACG2	30,56,34,511	30,27,82,502	28,52,009	99.07%	12.59
ACG3	41,32,62,383	41,02,74,180	29,88,203	99.28%	14.75
BCG1	27,72,03,681	26,73,21,089	98,82,592	96.43%	16.01
BCG2	39,58,29,790	36,94,36,604	2,63,93,186	93.33%	22.81
BCG3	40,78,33,164	39,13,78,528	1,64,54,636	95.97%	23.64
KCG1	34,26,03,699	33,02,93,118	1,23,10,581	96.41%	19.2
KCG2	53,38,50,846	51,26,95,893	2,11,54,953	96.04%	29.51
KCG3	39,62,21,095	37,94,44,509	1,67,76,586	95.77%	23.77
PCG1	40,39,13,588	38,92,64,812	1,46,48,776	96.37%	22.62
PCG2	47,87,67,728	45,96,53,138	1,91,14,590	96.01%	28.01
PUG3	45,89,92,665	45,48,46,650	41,46,015	99.10%	14.79
UCG1	29,45,99,920	29,07,81,625	38,18,295	98.70%	13.7
UCG2	38,31,47,058	37,94,04,405	37,42,653	99.02%	15.7
UCG3	42,93,22,400	42,38,91,695	54,30,705	98.74%	18.99
Mean	39,10,21,066	38,01,32,572	1,08,88,495	97.28%	19.35

ACG- Alambadi cattle group, BCG- Bargur cattle group, KCG- Kangayam cattle group, PCG- Pulikulam cattle group, UCG- Umblachery cattle group (Group 1- bulls, Group 2- High milk yielders (>2 kg/day), Group 3- low milk yielders (<1 kg/day).

Table 2: Per sample variant statistics.

Sample	Non-Ref. Hom	Het	Transitions	Transver	depth	Hom	Hom/Het	Ts/Tv
ACFG1	1419164	10882819	8399571	3902412	15.5	18483521	1.70	2.15
ACFG2	866771	13551304	9870022	4548053	13.7	15815036	1.17	2.17
ACFG3	874115	13390047	9752926	4511236	17.2	15976293	1.19	2.16
BCG1	870356	13246130	9681651	4434835	12.9	16120210	1.22	2.18
BCG2	722496	14094264	10161677	4655083	18	15272076	1.08	2.18
BCG3	727220	14001629	10109058	4619791	19	15364711	1.10	2.19
KCG1	926807	12823197	9411399	4338605	15.7	16543143	1.29	2.17
KCG2	1051092	12458516	9260126	4249482	25	16907824	1.36	2.18
KCG3	1256481	11359620	8638350	3977751	18.4	18006720	1.59	2.17
PCG1	1393040	11040745	8512917	3920868	18.9	18325595	1.66	2.17
PCG2	784417	13773055	9983178	4574294	22.6	15593285	1.13	2.18
PUG3	818185	13728842	9950141	4596886	17.9	15637498	1.14	2.16
UCG1	1105921	12307496	9173602	4239815	13.4	17058844	1.39	2.16
UCG2	1147220	12156236	9090141	4213315	17.3	17210104	1.42	2.16
UCG3	1259061	11659056	8822390	4095727	19.5	17707284	1.52	2.15
Mean	1014823.07	12698197.1	9387809.93	4325210.2	17.67	16668142.9	1.33	2.17

Non-Ref. Hom- Non-reference homozygous, Het- Heterozygous, Hom- Homozygous, Transver- Transvertions, Hom/Het- Homozygous/ Heterozygous, Ts/Tv- Transitions/transvertions.

The whole-genome sequencing library was prepared using QIAseq FX DNA Library Kit for Illumina (QIAGEN). The sequence data was generated using Illumina Hiseq 2500 and NovoSeq 6000 and quality control of sequenced data was carried out using FastQC (Andrews, 2017) and MultiQC (Ewels *et al.,* 2016) programmes. After quality control, the reads were aligned to reference genome (Bos-indicus-1.0) using BWA algorithm Li and Durbin (2009) and duplicate reads were removed using PICARD. Variant calling was done using GATK (McKenna *et al.* 2010) and snpEff (Cingolani *et al.* 2012) was used for annotation of variants.

RESULTS AND DISCUSSION

Whole-genome sequencing and mapping of short reads

Sequencing was carried out using Illumina Hiseq 2500 and Novoseq 6000, which produced an average of 72.65±4.79 GBs of data ranging from 49.46 GB to 92.21 GB in each sample. Out of 424,750,051 raw paired-end reads produced for each sample, 391,021,066 reads (ranging from 267,321,089 to 512,695,893 reads) with a length of 151 bp, an average GC content of 44±0.25 per cent per sequence and Phred score of more than 30 were mapped to the reference genome (*Bos-indicus*-1.0) successfully. Variants identified using the *Bos indicus* reference genome will be more specific to indicine breeds Devadasan *et al.* (2020). Around 38,01,32,572±17,371,788.63 reads were uniquely mapped with an average depth of 19.35-fold ranging from 12.59 to 29.51 and coverage of 97.28 per cent (Table 1). Coverage was affected by GC content, sequencing technology, read length, library preparation, structural variants and novel sequences (Liao *et al.* 2013). An average of 15x depth coverage was sufficient to identify almost 75 per cent of heterozygous variants and the accuracy increases with the increase in depth (Bentley *et al.* 2008).

Identification of SNPs and indels

The total length of the genome was 2,673,965,444 bp. The average depth achieved was 17.67-fold (ranging from 12.9

Table 3: Variant rate across all chromosomes.

Chromosome	Accession number	No. of variants	Length of chromosome	Variant rate
1	NC_32650.1	1772676	161108492	90.88
2	NC_32651.1	1517120	140680885	92.73
3	NC_32652.1	1296060	127865068	98.66
4	NC_32653.1	1449029	124429915	85.87
5	NC_32654.1	1353119	125642736	92.85
6	NC_32655.1	1382335	122646612	88.72
7	NC_32656.1	1226134	111948857	91.30
8	NC_32657.1	1244964	116941312	93.93
9	NC_32658.1	1259065	108100205	85.86
10	NC_32659.1	1197280	106310653	88.79
11	NC_32660.1	1161168	110256634	94.95
12	NC_32661.1	1076147	85442668	79.40
13	NC_32662.1	894906	84433108	94.35
14	NC_32663.1	884507	81409030	92.04
15	NC_32664.1	1017217	84800079	83.36
16	NC_32665.1	863140	77906047	90.26
17	NC_32666.1	887589	76519027	86.21
18	NC_32667.1	704798	65948792	93.57
19	NC_32668.1	670005	65317831	97.49
20	NC_32669.1	871711	75862604	87.03
21	NC_32670.1	789252	69307409	87.81
22	NC_32671.1	674818	61892534	91.72
23	NC_32672.1	723227	53331160	73.74
24	NC_32673.1	779077	65017658	83.45
25	NC_32674.1	503973	44044299	87.39
26	NC_32675.1	595431	51861174	87.10
27	NC_32676.1	609191	48749331	80.02
28	NC_32677.1	588154	46105673	78.39
29	NC_32678.1	642838	52131593	81.10
x	NC_32679.1	692592	88516652	127.80
y	NC_32680.1	38544	39421065	1022.75
Mitochondrial	NC_005971.1	273	16341	59.86
	Total	29366340	2673965444	91.06

to 22.6). 29,366,340 variants per sample were identified across the genome, which included 25,944,935 SNPs and 3,421,405 indels. Multiallelic SNP sites were 1,042,672. Out of the total variants identified, 57 per cent (ranging from 52 to 62 per cent) were homozygous and 37 per cent (ranging from 37 to 47 per cent) were heterozygous. The mean ratio of transitions and transversions was 2.17±0.003 (2.14 to 2.19) in each sample (Table 2), which is in agreement with previous studies (Choi *et al.* 2015, Das *et al.* 2015 and Stafuzza *et al.* 2017), indicating accuracy of SNP identification. The variant rate of the genome was around 91bp (Table 3). A high variant rate was observed in chromosome 23 with one variant for every 74 bp, Y chromosome had the lowest variant rate with one variant per 1022 bp. The lower variant rate in Y chromosome may be due to its haploid state in males, which results in the reduction of sequencing depth and lower rate of variant identification. Also, selection process reduces the retention of mutation due to the effect of recessive allele in hemizygous condition Stafuzza *et al.* (2017). The number of variants is proportionate to the length of chromosome (Fig 1), which is in agreement with the previous studies by Kawahara-Miki *et al.* (2011) and Stafuzza *et al.* (2017).

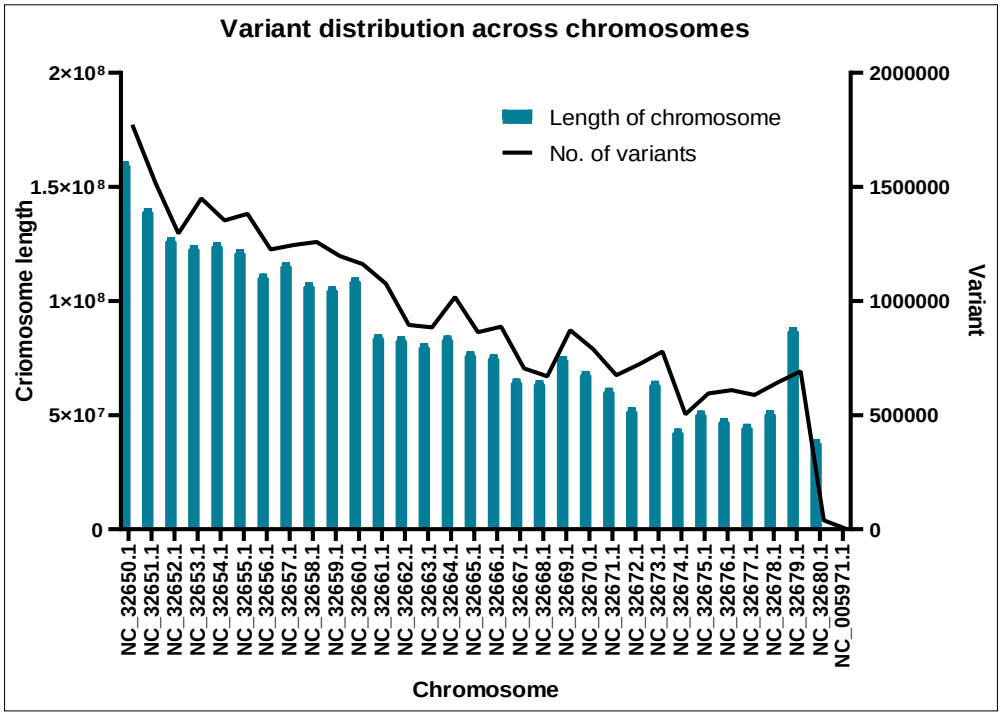

Fig 1: Variant distribution according to the length of chromosomes.

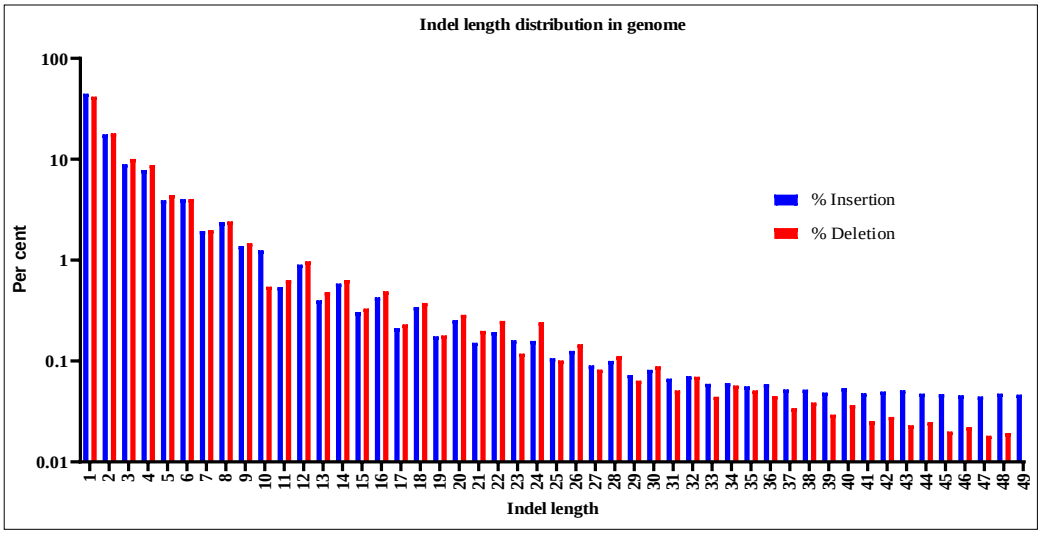

Fig 2: Indel length distribution.

Among the total indels identified, 1,287,879 were insertions and 2,133,526 were deletions. The length of the indel varied from 1 to 49. Single nucleotide insertions and deletions (948,585 and 1,046,590 respectively) were more commonly present. These results are similar to the findings of Eck *et al.* (2009), Liao *et al.* (2013) and Iqbal *et al.* (2019). Number of indels reduced when the length increased (Fig 2).

Around 25,237,418 variants out of the total were common for all the five breeds of cattle. Whereas, 1225844, 568,634, 528,093, 838,145 and 968,206 variants were specific to Alambadi, Bargur, Kangayam, Pulikulam and Umblachery cattle respectively.

Functional annotation of SNPs and indels in genome

Among the total SNPs, 61 per cent (17071364) of SNPs were present in intergenic region and 28.6 per cent (799258) were present in the intron region. SNPs found within 5kb upstream and downstream of a gene were 952518 (3.41 per cent) and 904544 (3.24 per cent) respectively. Around 191294 (0.7 per cent) SNPs were present in 3 prime and 5 prime UTR regions. Splice site SNPs included 761 splice acceptors, 911 splice donor and 20068 splice region SNPs (Table 4). These results are in accordance with the previous studies carried out by Kawahara-Miki *et al.* (2011), Liao *et al.* (2013), Das *et al.* (2015), Choi *et al.* (2015), Rosse *et al.* (2017) and Iqbal *et al.* (2019) in Kuchinoshima-Ushi cattle, Gir cattle, Danish Holstein cattle, Hanwoo, yanbian cattle, Guzera cattle and Native cattle breeds of Pakistan.

Table 4: Functional annotation of SNPs.

Type	Count	Percentage
3_prime_UTR_variant	148457	0.53
5_prime_UTR_premature_start _codon_gain_variant	7058	0.03
5_prime_UTR_truncation	0	0.00
5_prime_UTR_variant	42837	0.15
conservative_inframe_deletion	0	0.00
conservative_inframe_insertion	0	0.00
disruptive_inframe_deletion	0	0.00
disruptive_inframe_insertion	0	0.00
downstream_gene_variant	904544	3.24
exon_loss_variant	0	0.00
frameshift_variant	0	0.00
initiator_codon_variant	36	0.00
intergenic_region	17071364	61.07
intragenic_variant	582613	2.08
intron_variant	7999258	28.62
missense_variant	108845	0.39
non_coding_transcript_exon_variant	4447	0.02
non_coding_transcript_variant	67	0.00
splice_acceptor_variant	761	0.00
splice_donor_variant	911	0.00
splice_region_variant	20068	0.07
start_lost	220	0.00
start_retained_variant	0	0.00
stop_gained	3915	0.01
stop_lost	1736	0.01
stop_retained_variant	563	0.00
synonymous_variant	102855	0.37
upstream_gene_variant	952518	3.41

Table 5: Functional annotation of indels.

Type	Count	Percentage
3_prime_UTR_variant	25569	0.69
5_prime_UTR_premature_start _codon_gain_variant	0	0.00
5_prime_UTR_truncation	2	0.00
5_prime_UTR_variant	5219	0.14
conservative_inframe_deletion	711	0.02
conservative_inframe_insertion	711	0.02
disruptive_inframe_deletion	1140	0.03
disruptive_inframe_insertion	696	0.02
downstream_gene_variant	136975	3.68
exon_loss_variant	2	0.00
frameshift_variant	9235	0.25
initiator_codon_variant	0	0.00
intergenic_region	2202674	59.19
intragenic_variant	79604	2.14
intron_variant	1112682	29.90
missense_variant	0	0.00
non_coding_transcript_exon_variant	613	0.02
non_coding_transcript_variant	120	0.00
splice_acceptor_variant	200	0.01
splice_donor_variant	237	0.01
splice_region_variant	3187	0.09
start_lost	27	0.00
start_retained_variant	8	0.00
stop_gained	205	0.01
stop_lost	392	0.01
stop_retained_variant	8	0.00
synonymous_variant	0	0.00
upstream_gene_variant	141200	3.79

Functional annotation of indels showed that around 59.18 per cent (2,202,674) of indels were observed in the intergenic region and 29.9 per cent (1,112,682) were in the intronic region. Also, two indels lead to loss of exon, two cause 5 prime UTR truncation, 9235 leads to frameshift variation and 3624 affects splice sites (Table 5).

CONCLUSION

This is the first study to perform whole-genome sequencing to identify variants in Tamil Nadu native cattle breeds and identified 25,944,935 SNPs and 3,421,405 indels against the Nellore cattle genome. Alambadi cattle showed more specific variants than other breeds. Functional annotation revealed more of intergenic variants and intronic variants. Non-synonymous mutations are more in the coding region of 5572 genes. The variants identified in this study will serve as a useful genetic tool and as candidates in genomic

selection and genome-wide association studies to improve economically important traits in indigenous cattle breeds of Tamil Nadu.

Conflict of interest: None.

REFERENCES

Andrews, S. (2017). FastQC: A Quality Control Tool for High throughput Sequence Data. 2010.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P. *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 456(7218): 53-59.

Choi, J.W., Choi, B.H., Lee, S.H., Lee, S.S., Kim, H.C., Yu, D., Chung, W.H., Lee, K.T., Chai, H.H., Cho, Y.M. and Lim, D. (2015). Whole-genome resequencing analysis of hanwoo and yanbian cattle to identify genome-wide snps and signatures of selection. Molecules and Cells. 38(5): 466-473.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w ^{1118}/ ; iso-2; iso-3. Fly. 6(2): 80-92.

Das, A., Panitz, F., Gregersen, V.R., Bendixen, C. and Holm, L.E. (2015). Deep sequencing of Danish Holstein dairy cattle for variant detection and insight into potential loss-of-function variants in protein coding genes. BMC Genomics. 16(1): 1043: (2015). https://doi.org/10.1186/s12864-015-2249-y.

Devadasan, M.J., Kumar, D.R., Vineeth, M.R., Choudhary, A., Surya, T., Niranjan, S.K., Verma, A. and Sivalingam, J. (2020). Reduced representation approach for identification of genome-wide SNPs and their annotation for economically important traits in Indian Tharparkar cattle. 3 Biotech. 10(7): 309. doi: 10.1007/s13205-020-02297-z. Epub 2020 Jun 16.

Eck, S.H., Benet-Pagès, A., Flisikowski, K., Meitinger, T., Fries, R. and Strom, T.M. (2009). Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. Genome Biology. 10(8): R82. doi: 10.1186/gb-2009-10-8-r82. Epub 2009 Aug 6.

Ewels, P., Magnusson, M., Lundin, S. and Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 32(19): 3047-3048.

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C. and Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: Pool-versus individual-based genotyping. Molecular Ecology. 22(14): 3766-3779.

Iqbal, N., Liu, X., Yang, T., Huang, Z., Hanif, Q., Asif, M., Khan, Q.M. and Mansoor, S. (2019). Genomic variants identified from whole-genome resequencing of indicine cattle breeds from Pakistan. Plos One. 14(4): e0215065.

Kawahara-Miki, R., Tsuda, K., Shiwa, Y., Arai-Kichise, Y., Matsumoto, T., Kanesaki, Y., Oda, S., Ebihara, S., Yajima, S., Yoshikawa, H. and Kono, T. (2011). Whole-genome resequencing shows numerous genes with nonsynonymous SNPs in the Japanese native cattle Kuchinoshima-Ushi. BMC Genomics. 12(1): 103. doi: 10.1186/1471-2164-12-103.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 25(14): 1754-1760.

Liao, X., Peng, F., Forni, S., McLaren, D., Plastow, G. and Stothard, P. (2013). Whole genome sequencing of Gir cattle for identifying polymorphisms and loci under selection. Genome. 56(10): 592-598.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M.A. (2010). The genome analysis toolkit: A map reduce framework for analyzing next-generation DNA sequencing data. Genome Research. 20(9): 1297-1303.

Obeidat, B.S., Thomas, M.G., Hallford, D.M., Keisler, D.H., Petersen, M.K., Bryant, W.D., Garcia, M.D., Narro, L. and Lopez, R. (n.d.). (2002). Metabolic characteristics of multiparous Angus and Brahman cows grazing in the Chihuahuan Desert. 13. DOI: 10.2527/2002.8092223x.

Rosse, I.C., Assis, J.G., Oliveira, F.S., Leite, L.R., Araujo, F., Zerlotini, A., Volpini, A. *et al.* (2017). Whole genome sequencing of Guzerá cattle reveals genetic variants in candidate genes for production, disease resistance and heat tolerance. Mammalian Genome. 28(1-2): 66-80.

Russell, D.W. and Sambrook, J. (2001). Molecular Cloning: A Laboratory Manual (Vol. 1, p. 112). Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Stafuzza, N.B., Zerlotini, A., Lobo, F.P., Yamagishi, M.E.B., Chud, T.C.S., Caetano, A.R. *et al.* (2017). Single nucleotide variants and InDels identified from whole-genome re-sequencing of Guzerat, Gyr, Girolando and Holstein cattle breeds. Plos One. 12(3): e0173954.