Pakistan J. Zool., pp 1-8, 2022. **DOI**: https://dx.doi.org/10.17582/journal.pjz/20211221081250

In Silico Analysis of Genome Wide Non-Synonymous Single Nucleotide Polymorphisms in Indigenous Cattle Breeds of Pakistan

Umer Farooq[1,2], Nimra Murtaza[2], Abubakar Siddique[1], Bilal Saleem[1], Obaid Ur Rehman[1], Nageen Zahra[1], Muhammad Uzair[1], Muhammad Naeem Riaz[1,3]* and Muhammad Ramzan Khan[1]*

[1]National Institute for Genomics and Advanced Biotechnology (NIGAB), National Agricultural Research Centre, Park Road, Islamabad, Pakistan
[2]Department of Biosciences, COMSATS University Islamabad, Park Road, Islamabad, Pakistan
[3]Animal Biotechnology Program, Animal Sciences Institute, National Agricultural Research Centre, Park Road, Islamabad, Pakistan

Article Information
Received 21 December 2021
Revised 05 April 2022
Accepted 23 April 2022
Available online 17 June 2022
(early access)

Authors' Contribution
UF and MRK conceived the idea and designed the study. AB and UF retrieved and analyzed the data. MNR and UF interpreted the data and analyzed the results. NZ, MU, and NM completed the first draft of paper. BS and OR revised the first draft of paper. MRK and MNR finalized the final version of manuscript. All authors interpreted the data, critically revised the manuscript for important intellectual contents and approved the final version.

Key words
WGS, SNP, Single nucleotide polymorphisms, Pakistani cattle

ABSTRACT

Genomic selection programs for yield enhancement and disease resistance have become a reality with the availability of highly detailed genomic information. This information is critical to highlight genetic polymorphisms related with economically important traits including milk and meat yield, development, and resistance against diseases. In this present study, our main objective was to identify the deleterious SNPs and their associated genes which possibly disrupt protein's structure and function, as well as lead to genetic disease. We performed genome wide reference-based sequence alignment and functional annotation to identify deleterious non-synonymous SNPs (nsSNPs) in cattle breeds of Pakistan. For this purpose, genomic data of four different purpose cattle breeds including Bhagnari, Cholistani, Sahiwal and Red Sindhi was analyzed. Comparison with taurine reference genome ARS-UCD.1.2.99 discovered 29,032,662 genomic variations of which 25,469,157 were single nucleotide polymorphisms (SNPs) and 3,563,505 were Insertion/Deletions (InDels). Functional annotation identifies 122,943 missense SNPs that may possibly affect economically important traits. Using sequence and structure based computational tools SIFT, we identified 154 deleterious variants in 134 genes. Gene enrichment highlighted the presence of these genes in different biological processes including developmental, signaling, transport, metabolic and homeostasis. These findings are useful resource for further exploration into the molecular processes associated with these variances.

INTRODUCTION

Livestock have traditionally played important part in development of human societies for a long time. According to a rough estimate, livestock supports the livelihoods of approximately 0.6 billion farmers in underdeveloped nations (Thornton *et al*., 2006). Domestication of cattle has a major role in development of a country's agriculture economy. They contribute more than milk, meat, fiber skin and fuel productions. Based on their physical appearances nearly 800 cattle breeds are divided into two categories, *Bos taurus* (taurine) and *Bos indicus* (indicine) (Felius *et al*., 2011). Beside many other reported differences, the appearance of a hump over the shoulders is one of the major phenotypic characteristics used to distinguish indicine breeds form taurine breeds (Zeder *et al*., 2006). Indicine cattle grow in tropical and subtropical climates such as Africa, Southeast Asia, Brazil, northern Australia, southern China, and sections of the United States thanks to its physiological characteristics. Taurine, on the other hand, is typically found in more developed European countries, North America, and Australia, where it has a greater metabolic rate and dietary requirements. There are collectively 35 recognized cattle breeds in Pakistan and India (Felius *et al*., 2014) which make considerable contribution to the agriculture

* Corresponding author: mrkhan@PARC.gov.pk; naeemriaz@parc.gov.pk
0030-9923/2022/0001-0001 $ 9.00/0

economy. After goats, cows are the second most common animal species in Pakistan with an estimated 51.5 million animals (Finance Division, 2021), accounting for 3.2% of world cattle population (Singh *et al*., 2014). According to environmental conditions, most cattle breeds are created as drought breeds like Bhagnari, while Red Sindhi and Sahiwal are milk breeds and Tharparkar, Achai, Gabrali, and Cholistani are dual-purpose cattle breeds (Felius *et al*., 2014).

The process of natural and human driven selection has changed the cattle genotypes leading to diverse genetic and phenotypic profiles, and adaptation to temperature and tropical environments (MacHugh *et al*., 1997; Porto-Neto *et al*., 2013). Since the completion of the taurine cattle genome (Elsik *et al*., 2009), thousand bull genome project (Hayes and Daetwyler, 2019), the worldwide bovine genome sequencing, and the HapMap project (Gibbs *et al*., 2009), a considerable number of genetic variants including single nucleotide polymorphisms (SNPs) and InDels (Insertions/ Deletions) that have been recorded in a publicly available database (Sherry *et al*., 2001; Eck *et al*., 2009; Elsik *et al*., 2009; Gibbs *et al*., 2009; Iqbal *et al*., 2019).

Due to the advancements and cheap cost of next-generation sequencing (NGS), it is now feasible to simultaneously generate sequence data and estimate SNP allele frequencies in a range of reference populations (Kumar *et al*., 2012). Whole genome SNP genotyping analysis detected genomic regions targeted by selection, which, for example, contain immune and environmental adaptation related genes. This has opened new doors for studying the genetic variation and processes which help in adaptation to different biogeographic regions (Iso-Touru *et al*., 2016; Weldenegodguad *et al*., 2019). SNPs are the most prevalent type of genetic variation that are stable, have a low mutation rate, are inherited in a mendelian manner, and are relatively inexpensive to genotype (Snelling *et al*., 2005). As a result, they are routinely utilized as biomarkers in genetic research. Advances in bioinformatics and statistical tools have also improved our understanding of demographic evolution, the possible role of genomic structural variations, adaptation during domestication and selection, and the biological functions of these genomic variations in livestock breed (Gutenkunst *et al*., 2009; Li and Durbin, 2011; Jacob *et al*., 2020).

Regardless of these recent breakthroughs, there is a severe gap in whole-genome investigations of Asian cow breeds and European bovine breeds that are commonly employed for meat and dairy production (Kawahara-Miki *et al*., 2011; Choi *et al*., 2014). In this study, we analyzed the whole genome sequence data for four different purpose cattle breeds of Pakistan. Our main goal was to identify and analyze the missense variants resulting from nsSNP and predict their effect on resulting amino acid sequence, their structure, functionality, and stability. Missense variations not only modify the tertiary structure of proteins, but also lead to formation of deleterious phenotypes. Results of this study will be useful in furthering research into the genetic pathways underpinning features of interest in Indian cattle.

MATERIALS AND METHODS

Data collection

The whole genome sequences of Pakistani *Bos indicus* cattle samples included in this study were recently made available in public domain (Iqbal *et al*., 2019). Sequence data is available in China National GenBank Sequence Archive under project accession number CNP0000189. Total genomic DNA was extracted form whole blood (10 mL) samples and were sequenced using BGISEQ-500. For this study, we choose four different purpose breeds including Sahiwal (CNS0014619), Bhagnari (CNS0014606), Cholistani (CNS0014604), and Red Sindhi (CNS0014620).

Reads alignment and mapping

Quality check of sequence reads was performed using fast QC (Andrews, 2010) tool to identify low quality reads and adapter sequences. Raw sequences were first subjected to filtering process to identify and remove low quality reads and adapters using trimmomatic software (Bolger *et al*., 2014). Burrows-Wheeler Alignment (BWA) (Li and Durbin, 2009) algorithm was used with default parameters to create index files for reference genome ARS-UCD1.2 Btau5.0.1Y accessed from the 1000 Bull Genome project resources (Hayes and Daetwyler, 2019). Short reads of each selected cattle sample were mapped to reference genome using BWA with option "bwa-mem". SAM files generated as result of alignment were converted to their binary equivalent (BAM) files and sorted using SAM tools (Li *et al*., 2009). PCR duplicates were removed using Picard tools "Mark Duplicate" command line utility from aligned reads (Picard tools by broad institute https://broadinstitute.github.io/picard/). Base Quality Score Recalibration (BQSR) was performed on mapped reads to resolve high or low base quality scores estimation. Final BAM files were used for downstream variant calling.

Variant calling

Single nucleotide polymorphisms (SNPs) along with insertion and deletions mutations (InDels) were identified using "Haplotype Caller" tool of genome analysis toolkit (GATK) which performs local de-novo assembly of haplotypes in genomic variation regions (McKenna *et al*.,

2010). GATK was used according to guidelines of 1000 Bull Genome Project available at http://www.1000bullgenomes.com/doco/1000bullsGATK3.8pipelineSpecifications_Run8_Revision 20191101.docx (Fernandes Júnior *et al*., 2020). GATK "Select Variants" mode was used to separate SNPs and InDels into separate files. All variants including SNPs and InDels were discovered as difference from the reference genome ARS-UCD1.2 sequence. To remove the false positive calls, variants were filtered with GATK "Variant FIltration" argument using hard-filters with the following exclusion criteria: (1) quality by depth - QD<2.0; (2) Fisher Strand test - FS>60.0; (3) root mean square of the mapping quality score - MQ < 40.0; (4) and SOR > 9.0. SNPs and INDELS of autosomal chromosomes + X were left after filtration process.

Annotation of variations

Genetic variants were annotated using SnpEff program with annotation database ARS-UCD1.2.99 (Cingolani *et al*., 2012). SnpEff software assigns each SNPs and InDels to a functional class and provide several fields of information describing the affected transcripts and proteins, if applicable. SNPs and InDels were assigned to different functional classes including exonic, intronic, intergenic, splice site acceptor, splice site donor, splice site region, downstream, upstream, UTR 3 prime and UTR 5 prime region. InDels were also allocated to the in-frame deletion and insertion, disruptive in-frame deletion, and insertion functional classes whereas functional classes stop lost, stop retained, initiator codon were assigned to SNPs only. Variants were categorized into missense, nonsense, and silent variants. Missense or non-synonymous single base pair substitution result in amino acid substitution which affects the proteins phenotypically and functionally.

Detection of deleterious SNPs and go enrichment

SIFT tool (Vaser *et al*., 2016) was used to predicts the genes with tolerated and deleterious SNPs from the missense variants. This program identifies the tolerated and deleterious SNPs to predict the impact of amino acid substitution on phenotypic and functional alterations in protein molecules. SIFT program to identifies tolerance score (TI) for each in range from 0.0 to 1.0. The SIFT score ≤ 0.05 labels the non-synonymous variant as deleterious to protein function (Ng and Henikoff, 2003; Sim *et al*., 2012). Gene enrichment analysis was performed to identify over-represented biological processes using genes with deleterious SNPs. This gene lists comprising was submitted to the ClueGO (Bindea *et al*., 2009), a Cytoscape (Shannon *et al*., 2003) plug-in that combines Gene Ontology and KEGG pathways to produce a well-organized GO/ pathway annotation network, allowing researchers to see if these genes are linked to key pathways.

RESULTS AND DISCUSSION

Sequence data and alignment

Genomic sequences data of 4 different cattle breeds of Pakistan was analyzed which included dairy breeds Sahiwal and Red Sindhi, dual purpose breed Cholistani and drought tolerant breed Bhagnari. Individual sample sequence reads ranged from 754,986,078 to 1,605,207,524. Mapping of each sample to cattle reference genome ARS-UCD1.2 using BWA software yielded 83% to 99% alignment or sample reads (Table I). SAM alignment files were sorted and then converted to their binary format (BAM) files. Flow diagram depicts the bioinformatics analysis performed on the new generation sequence data (Fig. 1).

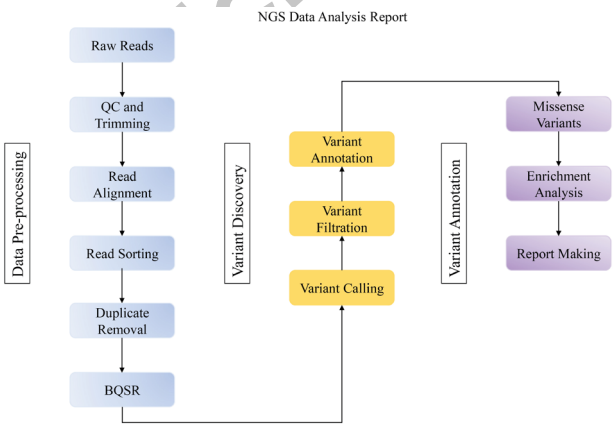

Fig. 1. Bioinformatics workflow for analyzing whole genome sequencing data.

Identification and annotation of variants

Alignment BAM files (including Bhagnari, Sahiwal, Cholistani and Red Sindhi) were pooled together to perform variant calling. Total ~25.46 M SNPs and ~3.56 M InDels were identified in the 29 autosomes and X chromosome against bovine reference genome ARS-UCD1.2 (Fig. 2 and Supplementary Table S1). Among the 3,563,505 InDels, 1,87,986 were deletions. Number of SNPs present per individual samples were 13.19, 18.96, 15.63, and 15.21 million in Bhagnari, Cholistani, Sahiwal, and Red Sindhi, respectively. The transition to transversion (Ts/Tv) ratio was pretty much similar across the samples, 2.27 in Bhagnari, 2.3 in Cholistani, 2.28 in Sahiwal and 2.3 in Red-Sindhi (Table II). This is indicative of high quality of our SNPs identified. Number of SNPs and Indels depend on the length of chromosomes. Out of total SNPs and InDels, only 104152 (0.41%) SNPs and 409 (0.01%) InDels matched to previously reported variations in dbSNP build 150.

Table I. Number of raw reads and aligned reads in each sample.

Breed	Sample ID	Read length (bp)	Total reads	Mapped reads	Total reads aligned %	Unmapped reads
Cholistani	CNS0014604	50	1,262,644,830	1,238,899,685	98.11	23,745,145
Bhagnari	CNS0014606	50	754,986,078	633,661,223	83.93	121,324,855
Sahiwal	CNS0014619	50	1,605,207,524	1,594,140,047	99.31	11,067,477
Red Sindhi	CNS0014620	50	1,579,398,030	1,568,961,018	99.33	10,437,012

Table II. Summary statistics of variants in each sample.

SNP Stats	Bhagnari	Cholistani	Sahiwal	Red Sindhi
Total SNPs	13,196,027	18,967,415	15,634,144	15,215,012
Total InDels	1563914	2263706	2028795	1850518
Singleton SNPs	1376894	3039282	2163090	1960479
Het/Hom	0.53	2.21	0.77	0.75
Ts/Tv Ratio	2.27	2.3	2.28	2.3

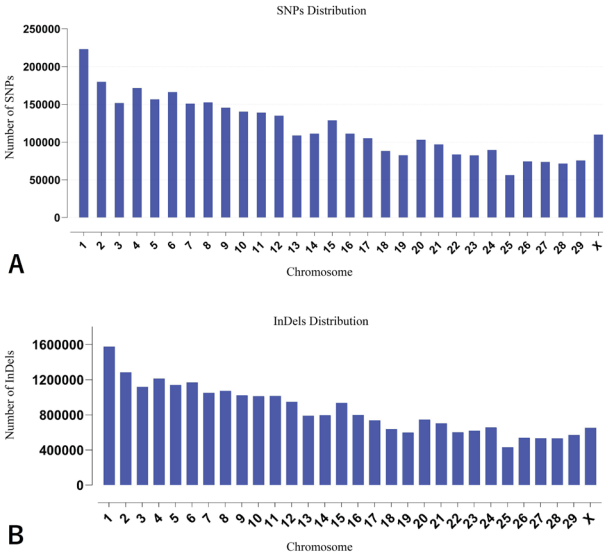

Fig. 2. Number of single nucleotide polymorphism (SNP) distribution by chromosome (A) and number of insertion/deletions (InDels) distribution by chromosome (B).

Genome wide annotation of variants

Variant sets including SNPs and InDels were functionally annotated them to their attribute and genes using SnpEff tool. In our dataset of total SNPs identified, 16,750,695 (41.53%) were discovered in between genes. 1,785,969 (4.43%) SNPs were in thousand base pair (bp) upstream region of genes and 1,826,857 (4.53%) were in thousand base-pair downstream region of genes (Table II). Due to distinct isoforms or overlapping genes, multiple functional effects were identified when compared the total number of SNPs and InDels used for variant. 58,970,854 transitions and 25,313,817 transversions were detected with Ts/Tv ratio of 2.3296 in SNPs. SnpEff identified 2,909 high impact (disruptive), 122,494 moderate and 239,647 low impact SNPs in all sequenced samples (Supplementary Table SII). Functional Annotation of InDels predicted the presence of 2,331,013 (40.78%) and 2,779,680 (48.63%) InDels in intergenic and intronic regions, respectively. Moreover, 285,210 (5.00%) and 273, 913 (4.80%) indels were located within thousand base pair upstream and downstream genic regions (Table III). 5,155 high impact (disruptive), 2,359 moderate and 5,730 low impact InDels were identified in all sequenced samples (Supplementary Table SII). SnpEff categorized the identified SNPs as missense, nonsense and silent according to their functional class (Table V). Missense variations (nsSNPs) can possibly cause amino acid substitutions which phenotypically and functionally effect important traits of cattle. For further downstream analysis, missense variants were carried forward.

Identification and enrichment of functional SNPs in coding region

Phenotypic effects of the missense variants were identified by SIFT tool. SIFT uses sequence homology-based approach to characterize the effect of amino acid substitution on protein function. SIFT analysis predicted 154 deleterious missense variants overlapping 134 genes in coding sequence regions with a score between 0.00-0.04. Moreover, 325 missense variants in coding sequence regions were predicted as tolerated (Table VI and Supplementary Table S3). Even though a small number of deleterious

SNPs were identified, all these variants overlapped protein coding genes. Gene list comprising deleterious SNPs was submitted in ClueGO to identify the functional gene ontology (GO) biological processes (BP). Gene ontology (GO) network is represented in Figure 3. Enrichment analysis revealed presence of these genes in several biological processes. Highlighted biological processes in gene interaction network included lipid transport (GO:0010876), transmembrane transport (GO:0055085), response to organic cyclic compounds (GO:0014070), response to endoplasmic reticulum stress (GO:0034976), negative regulation of transport (GO:0051051), detection of chemical stimulus (GO:0009593), skin development (GO:0043588), endomembrane system organization (GO:0010256), cytoskeleton organization (GO:0007010), response to growth factors (GO:0070848) and DNA conformation change (GO:0071103). List of all biological pathways and genes related is present in Supplementary Table S3.

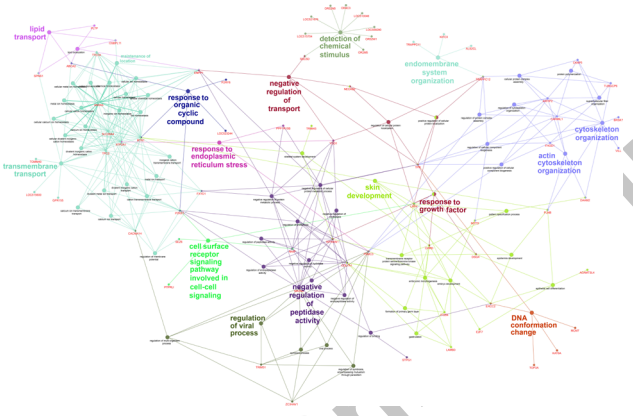

Fig. 3. ClueGO gene ontology analysis of 134 genes with deleterious nsSNP identified in all samples. ClueGO identifies the enriched go terms and visualizes them in grouped annotation network. This network shows the relationship between the terms based on the similarity of their associated genes. Each node represents a gene ontology term and the associated genes.

CONCLUSION

In silico approach to characterize genetic variations can assist us in predicting the consequences of mutations and explain their affecting role in biological mechanisms. In this study, we investigated whole genome sequence data of four different purpose cattle breeds in Pakistan (Cholistani, Sahiwal, Bhagnari and Red Sindhi) to identify genetic variations which lead toward deleterious effect on protein structure and functions. This analysis led to the discovery of 25,469,157 SNPs and 3,563,505 InDels in these breeds.

Table III. Annotation of SNPs.

SNPs	Count	Percent
3 prime UTR variant	121,002	0.30%
5 prime UTR premature start codon gain variant	6,728	0.02%
5 prime UTR variant	43,109	0.11%
Downstream gene variant	1,826,857	4.53%
Initiator codon variant	18	0%
Intergenic region	16,750,695	41.53%
Intron variant	19,403,911	48.11%
Missense variant	122,494	0.30%
Noncoding transcript exon variant	27,818	0.07%
Splice acceptor variant	388	0.00%
Splice donor variant	684	0.00%
Splice region variant	36,706	0.09%
Start lost	239	0.00%
Stop gained	1,410	0.00%
Stop lost	192	0%
Stop retained variant	136	0%
Synonymous variant	202,961	0.50%
Upstream gene variant	1,785,969	4.43%

Table IV. Annotation of InDels.

InDels	Count	Percent
3 prime UTR variant	22,164	0.39%
5 prime UTR variant	6,639	0.12%
Bidirectional gene fusion	15	0%
Conservative in frame deletion	497	0.01%
Conservative in frame insertion	509	0.01%
Disruptive in frame deletion	867	0.02%
Disruptive in frame insertion	546	0.01%
Downstream gene variant	285,213	4.99%
Frameshift variant	4,395	0.08%
Gene fusion	15	0%
Intergenic region	2,331,013	40.78%
Intragenic variant	3	0%
Intron variant	2,779,680	48.63%
Noncoding transcript exon variant	3,020	0.05%
Noncoding transcript variant	86	0.00%
Splice acceptor variant	439	0.01%
Splice donor variant	432	0.01%
Splice region variant	6,277	0.11%
Start lost	36	0.00%
Start retained variant	5	0%
Stop gained	55	0.00%
Stop lost	54	0.00%
Stop retained variant	7	0%
Transcript ablation	14	0%
Upstream gene variant	273,913	4.79%

Table V. Number of effects by functional class.

SNPs	Count	Percent
MISSENSE	122,943	37.55%
NONSENSE	1,410	0.43%
SILENT	203,097	62.02%

Table VI. Result of nsSNP analysis by SIFT.

SIFT Prediction	Count
Total	648
CDS Region	516
Non-Coding	132
Nonsynonymous	344
Synonymous	155
Stop-Gain	13
Deleterious	154
Tolerated	325

Numerous SNPs, InDels and genes were found that have not been annotated yet. We annotated each variant's possible functional role, allowing us to find numerous functionally significant candidate variants. Functional annotation revealed 122,943 missense (non-synonymous) SNPs in all samples 2,909 and 5,155 high impact (disruptive) SNPs and InDels in were also predicted by annotation. Analysis of missense variations revealed several genes with deleterious variation, involved in different biological and cellular functions. Go enrichment analysis revealed that genes harboring deleterious variations are significantly enriched in important biological processes, such as metabolic processes, development, transport, and homeostasis processes. Because this whole-genome sequencing study used just one animal of each breed, more research is needed to understand the exact dynamics of each gene–trait combination. The findings of this study can further be used as significant resource for further research on genomic characteristics to find variations in economically important traits and to develop precise genomic tools for cow breeding.

ACKNOWLEDGMENTS

The authors are highly grateful to Director NIGAB and NARC for providing all the necessary facilities.

Supplementary material

There is supplementary material associated with this article. Access the material online at: https://dx.doi.org/10.17582/journal.pjz/20211221081250

Statement of conflict of interest

The authors have declared no conflict of interest.

REFERENCES

Andrews, S., 2010. *Fastqc: A quality control tool for high throughput sequence data*. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.H., Pagès, F., Trajanoski, Z. and Galon, J., 2009. Cluego: A cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, **25**: 1091-1093. https://doi.org/10.1093/bioinformatics/btp101

Bolger, A.M., Lohse, M. and Usadel, B., 2014. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*, **30**: 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Choi, J.W., Liao, X., Stothard, P., Chung, W.H., Jeon, H.J., Miller, S.P., Choi, S.Y., Lee, J.K., Yang, B. and Lee, K.T., 2014. Whole-genome analyses of korean native and holstein cattle breeds by massively parallel sequencing. *PLoS One*, **9**: e101127. https://doi.org/10.1371/journal.pone.0101127

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 w1118; iso-2; iso-3. *Fly*, **6**: 80-92. https://doi.org/10.4161/fly.19695

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 Biol.*, **10**: 1-8. https://doi.org/10.1186/gb-2009-10-8-r82

Elsik, C.G., Tellam, R.L., Worley, K.C., Gibbs, R.A., Muzny, D.M., Weinstock, G.M., Adelson, D.L., Eichler, E.E. and Elnitski, L., 2009. The genome sequence of taurine cattle: A window to ruminant biology and evolution. *Science*, **324**: 522-528.

Felius, M., Beerling, M.L., Buchanan, D.S., Theunissen, B., Koolmees, P.A. and Lenstra, J.A., 2014. On the history of cattle genetic resources. *Diversity*, **6**: 705-750. https://doi.org/10.3390/d6040705

Felius, M., Koolmees, P.A., Theunissen, B., Consortium, E.C.G.D. and Lenstra, J.A., 2011. On the breeds of cattle historic and current classifications. *Diversity*, **3**: 660-692. https://doi.org/10.3390/d3040660

Fernandes Júnior, G.A., de Oliveira, H.N., Carvalheiro,

R., Cardoso, D.F., Fonseca, L.F.S., Ventura, R.V., and de Albuquerque, L.G., 2020. Whole-genome sequencing provides new insights into genetic mechanisms of tropical adaptation in nellore (*Bos primigenius indicus*). *Sci. Rep*., **10**: 1-7. https://doi.org/10.1038/s41598-020-66272-7

Finance Division, 2021. *Pakistan ecnomic survey*. Accessed on 15 November, 2021.

Gibbs, R.A., Taylor, J.F., Van Tassell, C.P., Barendse, W., Eversole, K.A., Gill, C.A., Green, R.D., Hamernik, D.L., and Kappes, S.M., 2009. Genome-wide survey of snp variation uncovers the genetic structure of cattle breeds. *Science*, **324**: 528-532. https://doi.org/10.1126/science.1167936

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. and Bustamante, C.D., 2009. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genet*., **5**: e1000695. https://doi.org/10.1371/journal.pgen.1000695

Hayes, B.J. and Daetwyler, H.D., 2019. 1000 bull genomes project to map simple and complex genetic traits in cattle: Applications and outcomes. *Annu. Rev. Anim. Biosci*., **7**: 89-102. https://doi.org/10.1146/annurev-animal-020518-115024

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**: e0215065. https://doi.org/10.1371/journal.pone.0215065

Iso-Touru, T., Tapio, M., Vilkki, J., Kiseleva, T., Ammosov, I., Ivanova, Z., Popov, R., Ozerov, M. and Kantanen, J., 2016. Genetic diversity and genomic signatures of selection among cattle breeds from siberia, eastern and northern europe. *Anim. Genet*., **47**: 647-657. https://doi.org/10.1111/age.12473

Jacob, K.K., Radhika, G. and Aravindakshan, T., 2020. An in silico evaluation of non-synonymous single nucleotide polymorphisms of mastitis resistance genes in cattle. *Anim. Biotechnol*., **31**: 25-31. https://doi.org/10.1080/10495398.2018.1524770

Kawahara-Miki, R., Tsuda, K., Shiwa, Y., Arai-Kichise, Y., Matsumoto, T., Kanesaki, Y., Oda, S.I., Ebihara, S., Yajima, S. and Yoshikawa, H., 2011. Whole-genome resequencing shows numerous genes with nonsynonymous snps in the japanese native cattle kuchinoshima-ushi. *BMC Genom*., **12**: 1-8. https://doi.org/10.1186/1471-2164-12-103

Kumar, S., Banks, T.W. and Cloutier, S., 2012. Snp discovery through next-generation sequencing and its applications. *Int. J. Pl. Genom*., **2012**. https://doi.org/10.1155/2012/831460

Li, H. and Durbin, R., 2009. Fast and accurate short read alignment with burrows wheeler transform. *Bioinformatics*, **25**: 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., and Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. *Nature*, **475**: 493-496. https://doi.org/10.1038/nature10231

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The sequence alignment/ map format and samtools. *Bioinformatics*, **25**: 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

MacHugh, D.E., Shriver, M.D., Loftus, R.T., Cunningham, P. and Bradley, D.G., 1997. Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*). *Genetics*, **146**: 1071-1086. https://doi.org/10.1093/genetics/146.3.1071

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., and Daly, M., 2010. The genome analysis toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res*., **20**: 1297-1303. https://doi.org/10.1101/gr.107524.110

Ng, P.C. and Henikoff, S., 2003. Sift: Predicting amino acid changes that affect protein function. *Nucl. Acids Res*., **31**: 3812-3814. https://doi.org/10.1093/nar/gkg509

Porto-Neto, L.R., Sonstegard, T.S., Liu, G.E., Bickhart, D.M., Da Silva, M.V., Machado, M.A., Utsunomiya, Y.T., Garcia, J.F., Gondro, C. and Van Tassell, C.P., 2013. Genomic divergence of zebu and taurine cattle identified through high-density snp genotyping. *BMC Genom*., **14**: 1-12. https://doi.org/10.1186/1471-2164-14-876

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T., 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**: 2498-2504. https://doi.org/10.1101/gr.1239303

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K., 2001. Dbsnp: The NCBI database of genetic variation. *Nucl. Acids Res*., **29**: 308-311. https://doi.org/10.1093/nar/29.1.308

Sim, N.L., Kumar, P., Hu, J., Henikoff, S., Schneider,

G. and Ng, P.C., 2012. Sift web server: Predicting effects of amino acid substitutions on proteins. *Nucl. Acids Res*., **40**: W452-W457. https://doi.org/10.1093/nar/gks539

Singh, U., Deb, R., Alyethodi, R.R., Alex, R., Kumar, S., Chakraborty, S., Dhama, K. and Sharma, A., 2014. Molecular markers and their applications in cattle genetic research: A review. *Biomark. Genom. Med*., **6**: 49-58. https://doi.org/10.1016/j.bgm.2014.03.001

Snelling, W.M., Casas, E., Stone, R.T., Keele, J.W., Harhay, G.P., Bennett, G.L. and Smith, T.P., 2005. Linkage mapping bovine est-based snp. *BMC Genom*., **6**: 1-10. https://doi.org/10.1186/1471-2164-6-74

Thornton, P.K., Jones, P.G., Owiyo, T., Kruska, R.L., Herrero, M.T., Kristjanson, P.M., Notenbaert, A.M.O., Bekele, N. and Omolo, A., 2006. *Mapping climate vulnerability and poverty in Africa*.

Vaser, R., Adusumalli, S., Leng, S.N., Sikic, M. and Ng, P.C., 2016. Sift missense predictions for genomes. *Nat. Protoc*., **11**: 1-9. https://doi.org/10.1038/nprot.2015.123

Weldenegodguad, M., Popov, R., Pokharel, K., Ammosov, I., Ming, Y., Ivanova, Z. and Kantanen, J., 2019. Whole-genome sequencing of three native cattle breeds originating from the northernmost cattle farming regions. *Front. Genet*., **9**: 728. https://doi.org/10.3389/fgene.2018.00728

Zeder, M.A., Bradley, D.G., Smith, B.D. and Emshwiller, E., 2006. *Documenting domestication: New genetic and archaeological paradigms*. Univ of California Press. https://doi.org/10.1016/j.tig.2006.01.007

Pakistan J. Zool., pp 1-8, 2022. **DOI**: https://dx.doi.org/10.17582/journal.pjz/20211221081250

Supplementary Material

In Silico Analysis of Genome Wide Non-Synonymous Single Nucleotide Polymorphisms in Indigenous Cattle Breeds of Pakistan

Umer Farooq[1,2], Nimra Murtaza[2], Abubakar Siddique[1], Bilal Saleem[1], Obaid Ur Rehman[1], Nageen Zahra[1], Muhammad Uzair[1], Muhammad Naeem Riaz[1,3]* and Muhammad Ramzan Khan[1]*

[1]National Institute for Genomics and Advanced Biotechnology (NIGAB), National Agricultural Research Centre, Park Road, Islamabad, Pakistan
[2]Department of Biosciences, COMSATS University Islamabad, Park Road, Islamabad, Pakistan
[3]Animal Biotechnology Program, Animal Sciences Institute, National Agricultural Research Centre, Park Road, Islamabad, Pakistan

Supplementary Table S1. Distribution of SNPs and InDels across chromosomes.

Supplementary Table S2. Number of variants by their impact.

Supplementary Table S3. Results of the deleterious nsSNP analysis by SIFT tool.

Supplementary Table S4. Gene Ontology biological processes (bp) enrichment by ClueGo.

* Corresponding author: mrkhan@PARC.gov.pk; naeemriaz@parc.gov.pk
0030-9923/2022/0001-0001 $ 9.00/0