EXOME WIDE VARIANT DISCOVERY BY NEXT GENERATION DNA SEQUENCING IN VECHUR CATTLE OF KERALA

R. S. Reshma[1], T. V. Aravindakshan[2], G. Radhika[3], T. Naicy[4] and K. Raji[5]
School of Applied Animal Production and Biotechnology, College of Veterinary and Animal Sciences, Mannuthy-680 651, Thrissur, Kerala.

Received : 02.06.2020
Accepted : 28.06.2020

Abstract

Vechur, the native cattle breed of Kerala, India is well-known for economically valuable phenotypic traits like disease resistance, adaptability to hot and humid tropical climatic conditions, low feed requirements and high quality milk. However, genomic information underlying these traits is rare. In the present study, the whole exome sequencing of a Vechur cow using Illumina HiSeq 2500 platform is reported. Comparison of sequences with Bos taurus reference genome assembly (UMD 3.1) identified 1,716,847 variants including 1,578,749 Single Nucleotide Polymorphisms and 138,098 Insertion/Deletions, of which 359,034 variants (20.91%) were novel. Detailed annotation of the identified variants showed that majority were situated in the intergenic region. Out of the 724,808 variants found inside the gene region, 107,880 were exonic variants. In the exonic variant, substantial proportion of non-synonymous (34.83%), frameshift (12.47%), nonsense (0.3%), start loss (0.06%) and stop loss (0.0009%) variants were identified. This information will provide a better understanding of genetic differences responsible for the peculiar phenotypic traits inherent to Vechur cattle.

Keywords: *Whole-exome sequencing, Vechur cattle, disease resistance, adaptability*

The cattle breeds in the world occur as two major subspecies – *Bos taurus* and *Bos indicus* which diverged from each other, more than 10,000 years ago from a common ancestor Auroch (*Bos primigenius*) (McTavish *et al.,* 2013). When compared to *Bos taurus* (hump less – European, Asian and African), *Bos indicus* (humped – South Asian and East African) is highly adapted to survive well in tropical and sub-tropical environmental conditions. Of the 50 registered cattle breeds of India surviving in various geographical and agro climatic regions (http://www.nbagr.res.in/regcat.html), Vechur (*Bos indicus*) is the cattle breed indigenous to Kerala, the southernmost state of India. This small

** Forms part of the MVSc thesis of the first author submitted to the Kerala Veterinary Animal Sciences University, Pookode, Wayanad, Kerala*

1. *MVSc scholar, Animal Biotechnology, CVAS, Mannuthy. Ph.7902875712*
2. *Director, Centre for Advanced studies in Animal Genetics and Breeding*
3. *Associate Professor, Centre for Advanced studies in Animal Genetics and Breeding*
4. *Assistant Professor, Department of Animal Genetics and Breeding and corresponding author Ph.No. 9446119307, naicy@kvasu.ac.in*
5. *Associate Professor, Department of Veterinary Physiology*

cattle breed is noted for certain physiological traits like high disease resistance, low feed requirements, heat tolerance and adaptability to hot climatic conditions (Raghunandanan, 2006). In order to reveal the essence of these features, the genome wide analysis of Vechur cattle is needed.

After the completion of cattle genome sequencing, a number of genome sequencing studies were performed in different cattle breeds – Black Angus and Holstein (Stothard *et al.,* 2011), Fleckvieh and Braunvieh (Schwarzenbacher *et al.,* 2016), Hanwoo, Yanbian (Choi *et al.,* 2014 and 2015) Jeju Heugu and Korean Holstein cattle (Choi *et al.,* 2014) resulting in the identification of large number of different genomic variants. Apart from these, there are still a number of cattle breeds whose genome has not been well scanned. Understanding these genetic variations will be an important initiative to reveal the genomic information underlying trait variations.

With the development of massively parallel sequencing, now it is not difficult to obtain the genomic information of ecologically important organisms. Recent advances in sequencing technology established a basis for the understanding of evolutionary biology, ancestry and economically important phenotypic traits in animals. Now the emergence of Next generation DNA sequencing (NGS) technology have reduced the cost and increased the speed of sequencing by several orders of magnitude (Lee *et al.,* 2013). Regardless of the declining cost of sequencing, it is still expensive to perform whole genome sequencing of an eukaryotic organism like cattle having around 3 Gb genome size. So the cost effective approach is to sequence only the meaningful regions of the genome. Exome comprises 1 to 2 per cent of a typical eukaryotic genome depending on species (Warr *et al.,* 2015), representing the major portion for searching the variants with immense effect on phenotypic traits. With the advent of methods for separating exome DNA (Parla *et al.,* 2011), it is now feasible to sequence the exome genome wide. With this background, the present study was undertaken to sequence the exome of Vechur cow.

Materials and Methods

Library preparation and sequencing

Genomic DNA used for sequencing was extracted from the whole blood of Vechur cow. Since DNA quality was extremely important for obtaining quality sequence data, careful quality checking was applied for extracted DNA and DNA with A260/280 ratio of >1.8 was used for library preparation. DNA was sheared and exonic regions were captured using Agilent Sure Select- Bovine (All Exome - 54 Mb) oligonucleotide probes. It targeted 54 mega bases of bovine genomic DNA of interest. Captured fragments were adaptor ligated and produced libraries. Illumina HiSeq 2500 was used to generate paired end 2 × 100 bp sequences resulting in the generation of 3.22 Gb of raw sequence data. Of the total reads obtained, 92.7per cent had a phred quality score of 30 or more. Generated sequence data was subjected to variant calling.

Before alignment, quality distribution, base distribution and GC distribution of forward and reverse reads in paired end sequencing was checked from the fastq files. Quality assessment involved removal of low quality reads, contamination control and trimming of adaptor sequences (Nayarisseri *et al.,* 2013). Based on quality report, sequence reads were trimmed to retain only high quality sequence. Reads were aligned to the reference *Bos taurus* genome and gene model downloaded from Ensembl (ftp://ftp.ensembl.org/pub/release-81/gtf/bos_taurus/Bos_taurus.UMD3.1.81.gtf.gz) after trimming the raw reads based on base quality, base composition and adapter sequences with Burrows – Wheeler Aligner (BWA) (O'Rawe *et al.,* 2013) Version - 0.7.5. Aligned reads were sorted with Picard tool Version – 1.100 Sort Sam command. Picard Mark Duplicates command was used to remove the read duplicates. SNPs and short indels were identified using SAMtools (Li *et al.,* 2009) Version – 0.1.18. Identified variants were further subjected to detailed annotation.

Results and Discussion

Sequencing and Aligning

In the present study, only the coding region of Vechur genome was targeted for sequencing. On an average 3.22 Gb of raw sequence data was generated as paired end 100 bp read length. The raw data quality summary is provided in Table 1. Low quality and contaminating reads were common artifacts in raw sequence data which would adversely affect the downstream analysis (Zhou *et al.,* 2013). Hence, quality control (QC) was critical for this NGS raw data. Out of the total 64,507,890 reads obtained, 64,506,674 (99.99%) reads passed QC.

When mapped to the *Bos taurus* reference genome (UMD 3.1), 98.44 per cent (63,500,350 reads) of the reads from sample aligned accurately (Table 1). Presence of duplicates might produce bias in variant allele identification and should be removed before variant calling, thereby reducing false calls and improving variant detection accuracy (Gao *et al.,* 2015). Duplicate reads were identified and removed. Chromosome wise coverage and dept is given in Table 2. It has been estimated that a minimum of 20 – 30 fold coverage is essential for detection of almost all variants (Li *et al.,* 2008). The coverage obtained was sufficient for detecting the variants.

Single Nucleotide Polymorphisms (SNPs) detection

Stringent QC measures were applied for reducing false positive variant detection. Variant calling using SAM tools identified 1,578,749 SNPs (91.95%). SNP is the abundant variant, controlling variations in traits of interest in cattle (Choi *et al.,* 2013). Of the SNPs identified 70.36 per cent were homozygous and 29.63 per cent were heterozygous. Low proportion of heterozygous SNPs may be due to strict variant calling requirements (Eck *et al.,* 2009). The variants were compared with dbSNP after downloading the dbSNP files from NCBI (ftp://ftp.ncbi.nih.gov./snp/organisms/cow_9913/VCF/). As expected there was apparent difference between the *Bos taurus* and Vechur coding sequences. The

Table 2: Chromosome wise coverage and dept

Chromosome	Coverage	Depth
1	22.57	2.05
2	24.21	2.36
3	26.69	2.7
4	23.53	2.09
5	26.33	2.56
6	22.38	2.06
7	26.12	2.42
8	24.53	2.06
9	22.13	1.91
10	25.98	2.66
11	27.83	2.48
12	22.06	1.77
13	28.81	2.3
14	24.55	1.96
15	25.86	2.43
16	27.25	2.32
17	25.87	2.13
18	33.23	3.07
19	35.3	3.55
20	23.58	1.94
21	27.3	2.11
22	28.17	2.42
23	28.39	2.47
24	24.94	1.85
25	36.02	3.08
26	27.15	2.43
27	24.78	1.98
28	26.17	2.22
29	29.45	2.47
X	22.16	1.62

Table 1: Raw read and alignment summary

Sample	Read orientation	Raw reads (paired end)	Bases (Gb)	GC%	Total reads	QC passed reads	Aligned read count	Properly paired
Vechur	R1	32,253,945	3.22	43.72	64,507,890	64,506,674 (99.99%)	63,500,350 (98.44%)	96.56%
	R2	32,253,945	3.22	43.72				

Table 3: Statistics of dbSNP filtered variants

Sample	Vechur
With dbSNP	1,357,813 (79.09%)
Without dbSNP	359,034 (20.91%)
Total	1,716,847

Table 4: Variant calling summary

Sample Name	Vechur
Total variants	1,716,847
Total SNPs	1,578,749 (91.95%)
Total indels	138,098 (8.04%)
Total homozygous SNPs	1,110,818 (70.36%)
Total heterozygous SNPs	467,931 (29.63%)
Total transition type SNPs	1,141,587 (72.30%)
Total transversion type SNPs	437,162 (27.69%)
T_s/T_v	2.61

comparison identified 359,034 novel variants and is summarized in Table 3. SNPs will occur either as transitions (purine to purine or pyramidine to pyramidine) or transversions (purine to pyramidine or vice versa). The total number of transition changes (T_S) observed was more than twice as compared to transversion changes (T_V). The whole exome transition to transversion ratio was calculated to be 2.61. During SNP detection studies using massively parallel sequencing technology, limitations in sensitivity and specificity can be expected. T_s/T_v ratio is helpful in assessing the errors in sequencing (Kraus *et al.,* 2011). A higher value is indicative of higher quality SNP calls (Liu *et al.,* 2012). The obtained T_s/T_v was close to the expected ratio for exome sequencing. Variants identified were summarized and provided in Table 4.

Insertion/deletions (Indels) detection

In the present study we identified 138,098 Indels (8.04%), out of which majority were insertion or deletion of a single base pair.

Variants having a depth of ≥5 and Q score of ≥20 were high quality variants (Bodi *et al.,* 2013). From total 1,716,849 variants (1,578,749 SNPs and 138,098 indels) identified, 825,916 SNPs (52.31%) and 93,559 indels (67.75%) were having a Q score of >20. When depth of identified variants were analyzed 606,886 SNPs (38.44%) and 57,147 indels (41.38) were with a depth of >5.

Variant annotation

Out of the total SNPs (1,578,749) identified, 724,808 SNPs (45.91 %) were found inside the genes, of which 107,880 SNPs (14.88%) were in exonic region and 616,928 (85.12%) were in intronic region. Exonic SNPs consist of SNPs in the coding region, non-coding gene, non-coding RNA (ncRNA) and untranslated region (UTR). There were 91,442 coding region SNPs (84.76%), 2,399 non-coding gene SNPs (2.22%), 454 SNPs (0.42%) in ncRNA and 13,585 SNPs (12.59%) in UTR. Out of 13,584 UTR SNPs, 3,053 (22.47%) were in 5'UTR and 10,532 (77.53%) were in 3'UTR. In 5-splice site 289 SNPs and in 3-splice site 417 SNPs were identified.

Out of the total indels (138,098) identified, 66,005 indels (47.8%) were identified to be inside the genes, of which 3,620 (5.48%) were in exonic region and 62,385 (94.51%) were in intronic region. Indels in the exonic regions were composed of coding region indels, non-coding gene indels, indels in ncRNA and UTR indels, including 1,852 coding region indels (51.16%), 159 non-coding gene indels (4.39%), 41 indels (1.13%) in ncRNA and 1,568 indels (43.31%) in UTR. UTR indels (1,568 indels) consisted of 319 (20.34%) 5'UTR and 1,249 (79.66%) 3'UTR indels. There were 137 indels identified in 5-splice site and 158 indels in 3-splice site.

Functional annotation of variants

The variant functional annotation identified 62,693 synonymous mutation, which may not significantly affect the phenotype of an organism, but can affect protein folding and function (Parmley and Hurst, 2007). Also 37,564 missense mutations were detected that leads to non-synonymous amino acid substitutions. There were 1,345 frame-shift mutations leading to addition or deletion of a base pair or base pairs resulting in an alteration of reading frame from the position of mutation. Nonsense mutation which cause change of a sense codon to a chain-terminating codon was identified to be 322 numbers. Annotation also detected 67 start loss and a single stop loss variant.

Conclusion

Vechur cattle is renowned for its immunity, adaptability to local environmental conditions and high quality milk. During late twentieth century, cross breeding program of indigenous cattle breeds of Kerala with exotic high producing breeds with the objective to increase productivity, drastically reduced the population size of Vechur cattle. Now it is being realized that these cross bred cattle although have high productivity, lacks important traits required for better adaptability to hot humid conditions of Kerala. There is a growing demand for conserving this declining local genetic resource. Researches on complex traits have been benefited with the rising of NGS technologies. To reveal the genetic basis of phenotypes inherent to Vechur cattle, exome of Vechur cow was sequenced and mapped to *Bos taurus* reference genome assembly which generated more than one million variants. Exons although consists of only very small fraction of the complete genome, it had got a tremendous impact on the phenotype of an individual. Detailed analysis of this result identified 1,346 frameshift, 37,564 nonsynonymous, 323 nonsense, 66 startloss and a single stoploss variations which will lead to amino acid changes in the final protein product and can result in significant impacts on phenotype. From the total variants, 359,034 were novel including a number of nonsynonymous, frameshift and nonsense variants which might be responsible for the peculiar and unique phenotypic traits of Vechur cattle.

Acknowledgments

We are thankful to the Kerala Veterinary and Animal Sciences University for providing the facilities for the conduct of research. We thank SciGenom Labs Pvt Ltd, Kakkanad, Kerala for the sequencing and bioinformatics services.

References

Bodi, K., Perera, A.G., Adams, P.S., Bintzler, D., Dewar, K., Grove, D.S., Kieleczawa, J., Lyons, R.H., Neubert, T.A., Noll, A.C., Singh, S., Steen, R. and Zianni, M. 2013. Comparison of Commercially Available Target Enrichment Methods for Next-Generation Sequencing. *J. Biomol. Tech.* **24**:73–86.

Choi, J.W., Chung, W.H., Lee, K.T., Choi, J.W., Jung, K.S., Cho, Y., Kim, N. and Kim, T.H. 2013. Whole Genome Resequencing of Heugu (Korean Black Cattle) for the Genome-Wide SNP Discovery. *Korean J. Food Sci. An.* **33**:715-722.

Choi, J.W., Liao, X., Stothard, P., Chung, W.H., Jeon, H.J., Miller, S.P., Choi, S.Y., Lee, J.K., Yang, B., Lee, K.T., Han, K.J., Kim, H.C., Jeong, D., Oh, J.D., Kim, N., Kim, T.H., Lee, H.K. and Lee, S.J. 2014. Whole-Genome Analyses of Korean Native and Holstein Cattle Breeds by Massively Parallel Sequencing. *PLoS ONE* doi: 10.1371/journal.pone.0101127

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. *Mol. Cells*. **38** : 466-473.

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 :** R82.1- R82.8.

Gao, X., Xu, J. and Starmer, J. 2015. Notes. Fastq2vcf: a concise and transparent pipeline for whole-exome sequencing data analyses. *BMC Res. Notes*. **8**:72-76.

Kraus, R.H.S., Kerstens, H.H.D., Hooft, P.V., Crooijmans, R.P.M.A., Poel, J.J.V.D., Elmberg, J., Vignal, A., Huang, Y., Li, N., Prins, H.H.T. and Groenen, M.A.M. 2011. Genome wide SNP discovery, analysis and evaluation in mallard (*Anas platyrhynchos*). *BMC Genomics*. **12**:150

Lee, C.Y., Chiu, Y.C., Wang, L.B., Kuo, Y.L., Chuang, E.Y., Lai, L.C. and Tsai, M.H. 2013. Common applications of next-generation sequencing technologies in genomic research. *Transl. Cancer Res.* **2** : 33- 45.

Li, H., Ruan, J. and Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*. **18**:1851-1858.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G.,and Durbin, R. 2009. 1000 Genome Project Data. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**: 2078–2079.

Liu, Q., Guo, Y., Li, J., Long, J., Zhang, B. and Shyr, Y. 2012. Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. *BMC Genomics.* **13** : 1-8.

McTavish, E.J., Decker, J.E., Schnabel, R.D., Taylor, J.F. and Hillis, D.M. 2013. New World cattle show ancestry from multiple independent domestication events. *Proc. Natl. Acad. Sci*. **110**: E1398-E1406.

Nayarisseri, A., Yadav, M., Bhatia, M., Pandey, A., Elkunchwar, A., Paul, N., Sharma, D. and Kumar, G. 2013. Impact of Next-Generation Whole-Exome sequencing in molecular diagnostics. *Drug Invent. Today*. **5**: 327-334.

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W.E., Wei, Z., Wang, K. and Lyon, G.J. 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med*. **5**: 1-18.

Parla, J.S., Iossifov, I., Grabill, I., Spector, M.S., Kramer, M. and McCombie, W.R. 2011. A comparative analysis of exome capture. *Genome Biol*. **12**:1-17.

Parmley, J.L. and Hurst, L.D. 2007. How do synonymous mutations affect fitness? *BioEssays*. **29**:515–519.

Raghunandanan, K.V. 2006. Vechur cattle of Kerala. *The Indian Cow*. **7**:48-49.

Stothard, P., Choi, J., Basu, U., Sumner-Thomson, J.M., Meng, Y., Liao, X. and Moore, S.S. 2011. Whole genome resequencing of Black Angus and Holstein cattle for SNP and CNV discovery. *BMC Genomics*. **12**:559

Schwarzenbacher, H., Burgstaller, J., Seefried, F.R., Wurmser, C., Hilbe, M., Jung, S., Fuerst, C., Dinhop, N., Weissenböck, H., Fuerst-Walt, B., Doleza, M., Winkler, R., Grueter, O., Bleu, U., Wittek, T., Fries R. and Pausch, H. 2016. A missense mutation in TUBD1 is associated with high juvenile mortality in Braunvieh and Fleckvieh cattle. *BMC Genomics*. **17**:400. doi 10.1186/s12864-016-2742-y.

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N. and Watson, M. 2015. Exome Sequencing: Current and Future Perspectives. *Genes Genomes Genet.* **5**:1543-1549.

Zhou, Q., Su, X., Wang, A., Xu, J. and Ning, K. 2013. QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data. *PLoS ONE.* **8** :1 ■