SUPPLEMENTARY MATERIAL AVAILABLE ONLINE

Indian Journal of Animal Sciences **94** (1): 44–49, January 2024/Article
https://doi.org/10.56093/ijans.v94.i1.135196

Assessment of linkage disequilibrium and haplotype block structure in indigenous cattle populations of Tamil Nadu, India using whole genome sequence data

MENDUGULLA TEJASWI SINDHU RATNAM[1], ALAGAPPAN GOPINATHAN[1✉], K THILAK PON JAWAHAR[2], P AZHAHIANAMBI[3], S VANI[1], K G TIRUMURUGAAN[1] and S M K KARTHICKEYAN[1]

Tamil Nadu Veterinary and Animal Sciences University, Tamil Nadu 600 007 India

Received: 8 April 2023; Accepted: 10 November 2023

ABSTRACT

Understanding the nature of linkage disequilibrium (LD) between molecular markers reflects the degree of non-random association between their alleles, which is essential for genome-wide association studies. The main objective of this study was to assess the chromosome-wise linkage disequilibrium and haplotype block structure in indigenous cattle breeds of Tamil Nadu, India. A total of 79 animals belonging to five indigenous cattle breeds were considered as single indigenous cattle population and three samples per breed were sequenced using Illumina NovaSeq™ 6000 and Illumina HiSeq 2500. The LD parameter (r^2) was estimated for a total of 62,095,778 pair-wise SNPs for all autosomes. The chromosome-wise number of SNP pairs, mean r^2±standard deviation and median r^2 values were 2,141,234, 0.484±0.246 and 0.493, respectively. Chromosome-wise mean r^2 of different distance bins were calculated and it showed the maximum r^2 value of 0.839 at 31-70 kb distance for Chromosome 11 and minimum r^2 value of zero for Chromosome 7, 11, 15, 20 and 29, respectively at a distance 71-100 kb. This study also revealed a total of 413, 277 haplotype blocks which covered 2.34% of the autosomal genome. There was a total of 1,589,118 SNPs distributed within the haplotype blocks, covering a total length of 53.62 mb. The results of the study suggest the need for breed-specific reference populations for indigenous cattle breeds, which have a greater density of molecular markers of economic significance and thereby to identify breed-specific haplotypes in future.

Keywords: Haplotype, Indigenous cattle breeds, Linkage disequilibrium, WGS data

Linkage disequilibrium (LD) is the non-random association of alleles at different loci among the individuals in a population and it is also known as gametic disequilibrium (Salem *et al*. 2018). Selection, linkage, admixture, mutation rate, migration, genetic drift, population structure, recombination rate, and effective population size were variables that might influence the pattern of LD within a population (Ardlie *et al*. 2002). LD studies in cattle were first based on the use of a low-density panel of microsatellite markers or a limited number of SNPs encompassing one or a few areas of chromosomes (Saravanan *et al*. 2020). Domestication and breed formation have resulted in high levels of LD in livestock species (Khatkar *et al*. 2006). LD decay is defined as the determination of the level of linkage disequilibrium between SNPs in relation to the physical distance. The measure of LD between SNPs (r^2) decreases as the inter-marker distance increases, a process known as LD decay (Qanbari *et al*. 2010). Therefore, the decline in the level of LD as a function of physical distance between markers, in turn was used to assess the recombination frequency.

Haplotype blocks are defined as long stretches of DNA, inherited together as discrete blocks, along a chromosome with a low recombination rate and high LD and they have been extensively studied in animal breeding and gained importance in genomic selection for predicting breeding values and genome-wide association studies for economically important traits in cattle (Mokry *et al*. 2014). Individual SNP markers were less useful than haploblocks because each haploblock may have a larger LD with causal mutations than any individual SNP, hence alleles within a haploblock can capture more variance in a trait (Cuyabano *et al*. 2015).

In this backdrop, the objective of the present study was to investigate the pattern of linkage disequilibrium, the extent of linkage disequilibrium level (r^2) and haplotype block structure analysis in indigenous cattle populations of Tamil Nadu using whole genome sequence data.

MATERIALS AND METHODS

Sample collection and genomic DNA isolation: A

Present address: [1]Madras Veterinary College, Tamil Nadu Veterinary and Animal Sciences University, Tamil Nadu. [2]Livestock Farm Complex, Madhav Milk Colony, Chennai. [3]Translational Research Platform for Veterinary Biologicals, Madhavaram Milk Colony, Chennai. ✉Corresponding author email: gopinathan.a@tanuvas.ac.in

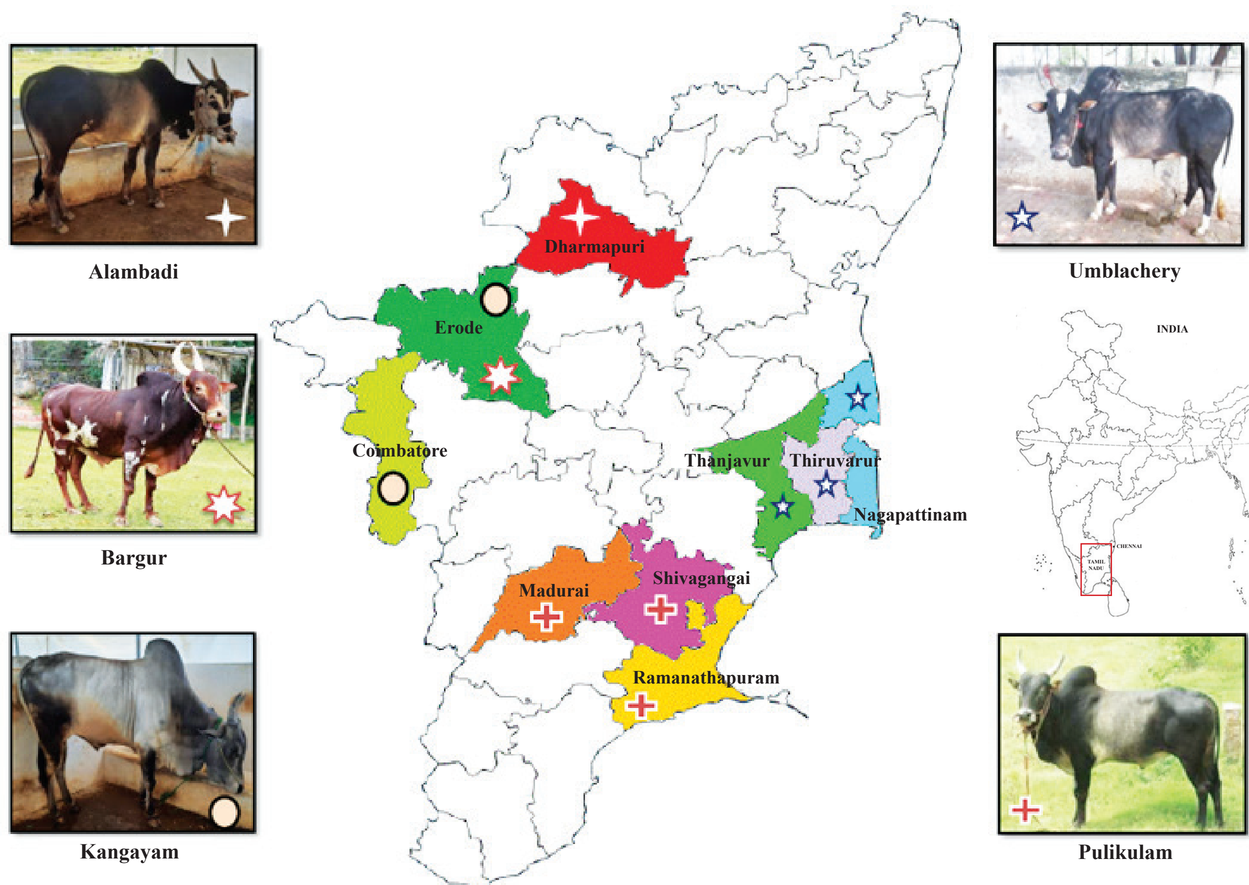

Fig. 1. Indigenous cattle breeds of Tamil Nadu and their breeding tract

total of 302 blood samples were collected from unrelated animals belonging to five indigenous cattle breeds of Tamil Nadu, viz. Alambadi (16), Bargur (106), Kangayam (63), Pulikulam (56) and Umblachery (61) for this study at respective breed conservation units (Fig.1).

Blood samples of 10 ml each were collected in EDTA vacutainers (0.5% EDTA; Beckton-Dickinson, India) under aseptic conditions through jugular vein puncture. Genomic DNA was isolated from the blood samples using the standard modified phenol-chloroform method (Sambrook *et al*. 1989). The concentration and quality assessment of DNA samples were done and followed byoptical densities (OD) were measured at 260 nm and 280 nm with TE buffer as a blank.

Pooling of DNA samples: A total of 79 samples Alambadi (14), Bargur (19), Kangayam (16), Pulikulam (16) and Umblachery (14)) were chosen from working samples (302 in no.) as per the group criteria such as breeding bulls; high yielders; and low yielders based on available production records at each breed conservation unit. To ensure a uniform representation of selected individuals in a group, the concentration of DNA from selected individuals was adjusted to 200 ng per µl with TE buffer. An equal volume of DNA (10 µl) from each selected individual was combined to obtain criteria-specific DNA pools, that gave enough genetic material (DNA) for subsequent whole genome sequencing. As a result, three groups per breed and a total of 15 groups for five indigenous cattle breeds were labelled as ACFG1, ACFG2 and ACFG3, respectively for Alambadi; BCG1, BCG2 and BCG3, respectively for Bargur; KCG1, KCG2 and KCG3, respectively for Kangayam; PCG1, PCG2 and PCG3, respectively for Pulikulam and UCG1, UCG2 and UCG3, respectively for Umblachery cattle and samples were submitted in frozen form to an outsourcing facility (M/s. Clevergene™, Bengaluru). After quality check, DNA samples were sequenced, and the data was generated using Illumina NovaSeq™ 6000 and Illumina HiSeq 2500 at 20X coverage and the results were obtained in PLINK format files using PLINK tools (.ped and .map files).

Quality control: Quality control was performed using PLINK v 1.9 software (Purcell *et al*. 2007). SNP markers with a call rate of more than 95% were adopted to screen SNPs. SNPs on X-chromosome, Y-chromosome, and mitochondrial DNA were removed for further analysis and therefore the final data set consists of SNPs present in 29 chromosomes (autosomes) only.

Estimation of linkage disequilibrium (LD): Two statistical techniques such as r^2 and D^l were widely used to measure the extent of LD and the most common measure is an estimation of r^2 (Hill and Robertson 1968), which is the squared correlation between alleles at two loci and is generally accepted as the most robust due to less sensitivity to sample size and allele frequency differences. r^2 is

expressed as follows:

$$r^2 = \frac{(f_{11}f_{22} - f_{12}f_{21})^2}{fA_1\, fA_2\, fB_1\, fB_2}$$

Where, A and B, two-loci with alleles denoted by A_1, A_2, B_1 and B_2, respectively; fA_1, fA_2, fB_1 and fB_2 were frequencies of A_1, A_2, B_1 and B_2 respectively; f_{11}, f_{22}, f_{12} and f_{21} were frequencies of the haplotypes A_1B_1, A_2B_2, A_1B_2 and A_2B_1, respectively.

The LD (r^2) measure was calculated for all marker pairs of each chromosome (Purcell *et al*. 2007). The first step is to determine how many markers are needed to quantify the extent of LD. Overall 75,000-3,00,000 informative SNPs would capture the majority of LD information in cattle breeds (Khatkar *et al*. 2008). Descriptive statistics were computed for LD using Linux commands.

To study the LD decay, SNP pairs were grouped as <30 kb, 31-70 kb, 71-100 kb, 101-200 kb, 201-300 kb, 301-400 kb, 401-500 kb and >500 kb physical distance (Salem *et al*. 2018). The mean r^2 was computed for each distance bin in whole autosomal chromosomes and the mean r^2 in each distance bin was plotted against the median of the distance bin range (Kiselyova *et al*. 2014).

Consider two loci with alleles 'A' and 's' at locus one, and 'B' and 'd' at locus two, with allele frequencies πA, πs, πB, and πd, respectively, the resulting haplotype frequencies are πAB, πAd, πsB, and πsd. The basic component of all LD statistics is the difference between the observed and expected haplotype frequencies:

$$Dsd = (\pi AB - \pi A \pi B)$$

The distinction between these statistics lies in the scaling of this difference. The first of the two measures (r^2) was calculated as:

$$r^2 = \frac{(Dsd)^2}{\pi A\ \pi s\ \pi B\ \pi d}$$

The correlation coefficient between the two loci is represented by r^2 (Hill and Robertson 1968). Fisher's exact test is applied to compare sites with two alleles at each locus (Fisher 1935) or multifactorial permutation analysis is used (Weir 1996).

Haplotype block analysis: The haplotype block analyses were carried out using gPLINK v 2.05. The total number of SNPs covered by all the blocks, chromosome length (mb), total number of haploblocks, total block length (Mb), chromosome block coverage (percent) and SNPs in blocks (per cent) for all the autosomes in the population were estimated (Saravanan *et al*. 2020).

RESULTS AND DISCUSSION

Data description: A total of 26,406,037 SNPs were sequenced with a variant rate of one in every seven bases across the five indigenous cattle breeds investigated. The locations of genome-wide variants were restricted to 30,866,203 sites across all the samples; out of which 25,944,935 were found to be biallelic single nucleotide polymorphism sites; 3,421,405 InDel sites, 1,042,672 multi-allelic sites and 4,57,191 multi-allelic SNP sites.

Linkage disequilibrium (LD): The LD parameter (r^2) was estimated for a total of 62,095,778 pairwise SNPs for all autosomes. The chromosome-wise number of SNP pairs, mean r^2 ± standard deviation and median r^2 values are summarized in Table 1. The maximum number of SNP pairs (3,836,001) and the minimum number of SNP pairs (1,233,641) were found on chromosomes 1 and 25, respectively. The chromosome-wise r^2 values ranged from 0.475±0.245 (on chromosome 23) to 0.492±0.248 (on chromosome 25) with an overall mean of 0.484±0.246. The median r^2 values ranged from 0.205 (chromosome 2) to 1.000 (chromosomes 5, 25 and 29) with an overall median r^2 of 0.493.

Table 1. Number of SNP pairs, mean r^2±S.D. and median r^2 for different chromosomes

Chromosome Number	No. of SNP pairs	Mean r^2±S.D.	Median r^2
1	3,836,001	0.482 ± 0.244	0.318
2	3,266,793	0.481 ± 0.244	0.205
3	2,716,952	0. 481 ± 0.243	0.427
4	3,247,836	0.485 ± 0.246	0.615
5	2,951,472	0.483 ± 0.245	1.000
6	3,302,616	0.492 ± 0.246	0.727
7	2,678,285	0.484 ± 0.246	0.460
8	2,655,465	0.482 ± 0.245	0.444
9	2,752,755	0. 482 ± 0.246	0.259
10	2,570,809	0.480 ± 0.246	0.356
11	2,355,664	0.480 ± 0.244	0.719
12	2,328,483	0.481 ± 0.245	0.308
13	1,928,509	0.486 ± 0.247	0.333
14	1,935,049	0.485 ± 0.246	0.238
15	2,129,265	0.487 ± 0.249	0.360
16	1,809,994	0.481 ± 0.245	0.680
17	1,877,918	0.482 ± 0.246	0.365
18	1,516,725	0.487 ± 0.248	0.349
19	1,427,828	0.484 ± 0.246	0.351
20	1,902,849	0.483 ± 0.245	0.467
21	1,680,681	0.484 ± 0.247	0.284
22	1,481,116	0.488 ± 0.247	0.881
23	1,482,017	0.475 ± 0.245	0.467
24	1,731,767	0.484 ± 0.246	0.464
25	1,233,641	0.492 ± 0.248	1.000
26	1,272,592	0.485 ± 0.247	0.231
27	1,315,427	0.482 ± 0.245	0.615
28	1,301,170	0.486 ± 0.247	0.376
29	1,406,099	0.487 ± 0.248	1.000
Total	62,095,778	--	--
Mean	2,141,234	0.484 ± 0.246	0.493

Chromosome-wise mean r^2 values for <30 kb, 31-70 kb, 71-100 kb, 101-200 kb, 201-300 kb, 301-400 kb, 401-500 kb and >500 kb distance bin sets were 0.595, 0.536, 0.516, 0.502, 0.489, 0.477, 0.464 and 0.461, respectively (Supplementary Table 1). Average r^2 values of more than 0.40 to 0.49 was observed at 201-300 kb, 301-400

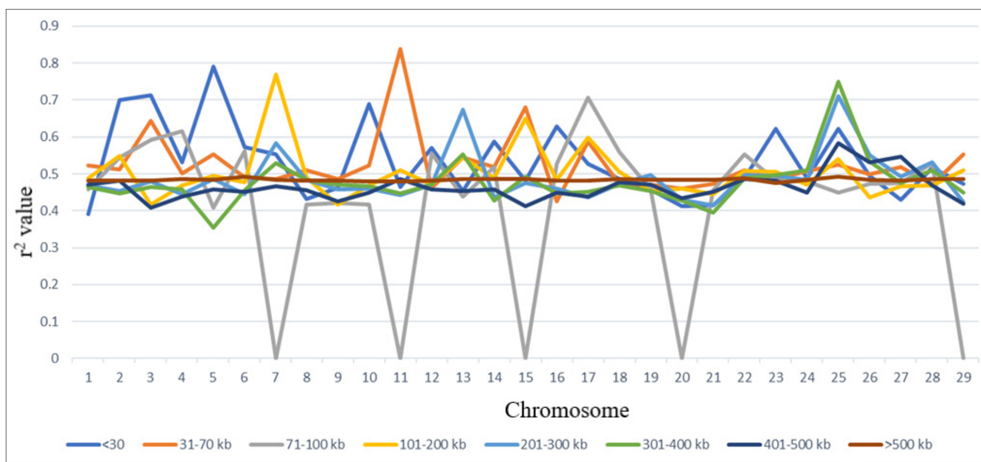

Fig. 2. Chromosome-wise mean r^2 values at different distance bin sets (kb).

kb, 401-500 kb and >500 kb distance intervals and r^2 values of more than 0.5 was noticed in <30 kb, 31-70 kb, 71-100 kb and 101-200 kb distances. The maximum r^2 value of 0.839 (Chromosome 11) was found at a 31-70 kb distance and the minimum r^2 value of zero (Chromosomes 7, 11, 15, 20, and 29) at 71-100 kb (Fig. 2). The pattern of r^2 values at different physical distances pooled over all the autosomes clearly showed that the average r^2 declines with increasing physical distance between markers in the population (Fig. 3).

An overall mean r^2 of 0.484±0.246 was observed in this study, which indicates good prediction accuracy for genomic selection in future, which is in agreement with an earlier study Meuwissen *et al*. (2001). A similar mean r^2 of 0.45 was also observed in Angus cattle (Porto-Neto *et al.* 2014). On the contrary, the lowest LD values of 0.24, 0.21 and 0.21 were reported in Swiss Eringer (Flury *et al*. 2010), Gyr (Silva *et al*. 2010) and Pinzgau cattle (Sidlova *et al*. 2016) breeds, respectively. The wide variation in r^2 values will be useful for the customization of SNP chips for future breeding programs, genome-wide association studies and genomic selection indices (Mokry *et al*. 2014).

The LD decay was also analyzed for various distance bins and these results showed that the range of r^2 appeared as 0.527 to 0.791 for less than 30 kb distance bins and reached 0.422-0.485 range for more than 500 kb distance bins for indigenous cattle population. The LD levels were considerably different and the differences in LD extent were more detectable. The decline in mean r^2 value occurred slowly with an increase in physical distances between markers. Further, more slow LD decay was observed at larger distances (401-500 kb to more than 500 kb), which showed that population decline has been more intense in recent generations.

Compared to the other studies, it appeared that r^2 values were overestimated in this study and this might be due to relatively small sample sizes being subject to bias and this bias may vary with inter-marker distance (Zhu *et al*. 2013). Despite the sample size bias, it would be reasonable to expect extensive LD in indigenous cattle populations, because these small populations were exposed to serious extinction risk for more than a decade due to a reduction in population size, increased inbreeding, and uncontrolled crossbreeding in their respective breeding tract. This *Bos*

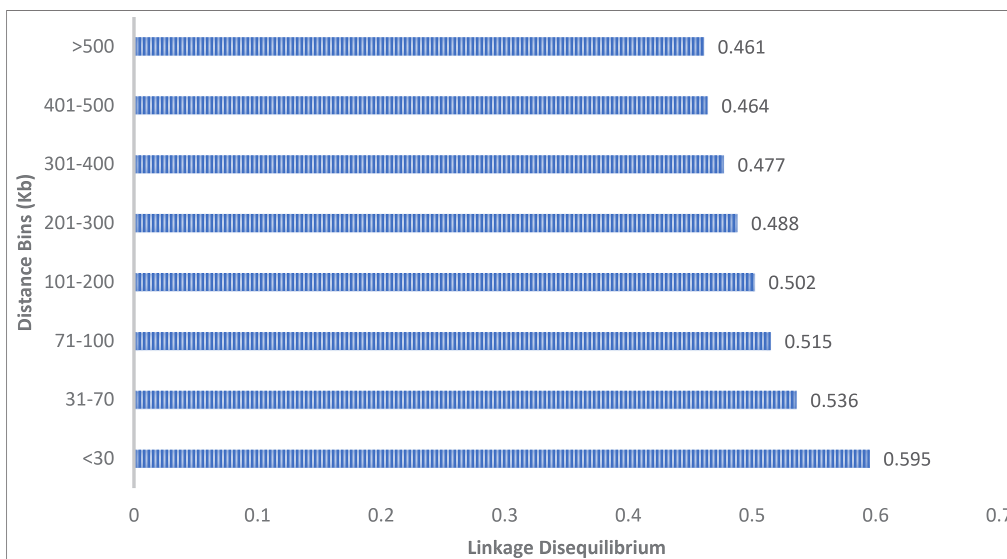

Fig. 3. Mean r^2 values at different physical distances (kb) pooled over all the autosomes.

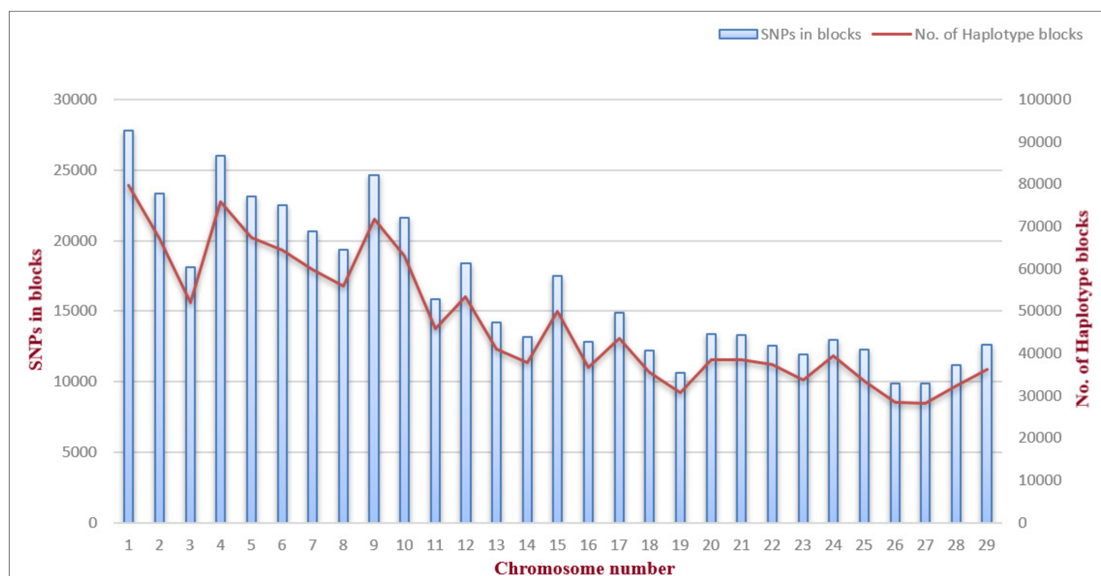

Fig. 4. Chromosome-wise number of SNPs and number of haplotype blocks.

indicus or *indicine* population had higher LD levels and this could be attributed to a smaller effective population size (Villa-Angulo *et al*. 2009) or a higher ancestral relatedness (Purfield *et al*. 2012). Hence, necessary steps should be taken to conserve these indigenous genetic resources in their respective breeding tract. In addition, LD levels were influenced by other factors such as minor allele frequency threshold, density of SNP panels, and physical distance between the markers (Zhu *et al*. 2013), which would warrant further research on this genome.

In addition, different LD patterns on individual chromosomes could be created through uneven selection pressures on QTLs distributed throughout the genome. So, higher LD values can be expected for chromosomes harbouring quantitative trait loci (QTL) undergoing natural selection (Oyelami *et al*. 2020). However, patterns of LD within a population can be affected by several factors, such as selection, mutation rate, migration, genetic drift, population structure, recombination rate and effective population size (Hayes *et al*. 2009). Hence, a more detailed study on selected regions of the *indicine* genome and assessment of positive selection signatures are carried out in the future.

Haplotype block structure: The summary of haplotype blocks in different chromosome are given in Supplementary Table 2. This result, identified a total of 413,277 haplotype blocks which covered 2.34% of the autosomal genome. A total of 1,589,118 SNPs was distributed within the haplotype blocks, covering a total length of 53.62 Mb. Chromosome 1 had the largest number of haploblocks (23,939 numbers), while chromosome 27 had the least (8,475 numbers). Regarding the length of chromosome coverage (Mb), chromosomes 1 to 11 were largely covered, while chromosome 25 was less covered. In terms of chromosomal block coverage, chromosome 1 had the highest percentage of coverage (3.32) while chromosomes 16 and 19 had the least coverage (1.95). With respect to per cent of SNPs in blocks, chromosome 1 was found to have the highest (5.83) and chromosomes 26 and 27 had the lowest (2.06 each) values. Chromosome-wise number of SNPs and the number of haplotype blocks are depicted in Fig. 4.

In this study, a total of 413,277 haplotype blocks were identified, which comprised a total block length of 53.62 Mb which was higher when compared to German Holstein-Friesian cattle (Qanbari *et al*. 2010) and in Australian Holstein Friesian (Khatkar *et al*. 2008). In addition, another lower estimate of 76,673 haplotype blocks was reported in the Canchim breed of cattle (Mokry *et al*. 2014). The wide differences in haplotype blocks irrespective of *taurine* or *indicine* origin might be due to differences among breeds, marker types, density of markers and variation in chromosome regions (Marques *et al*. 2008).

The highest number of haplotype blocks stated in this study might be accredited to the method of genotyping used (whole genome sequencing), contrary to other studies where SNP bead chips were utilized for genotyping. Further, the higher haplotype diversity could be due to the samples being taken from the highly variant population kept at conservation farms and the diversity of haplotypes can also be influenced by its population size. This highest number of haplotypes could be useful for fine mapping of QTLs, genomic association studies and to guide the breeding programs to ensure the effective conservation and optimum utilisation of this valuable genetic resource.

Furthermore, block analysis in present study revealed that this population had more SNPs that were clustered into haplotype blocks, an indication that the population contains a lot of variants that are inherited from their ancestors.

ACKNOWLEDGEMENTS

Authors extend their gratitude to Molecular Genetics Laboratory, Department of Animal Genetics and Breeding, Madras Veterinary College, TANUVAS for their kind

support to carry out the work. In addition, they would also like to acknowledge the Government of India-NPBB-Tamil Nadu Livestock Development Agency for their financial assistance to carry out the work.

REFERENCES

Ardlie K G, Kruglyak L and Seielstad M. 2002. Patterns of linkage disequilibrium in the human genome. *Nature Review Genetics* **3**(4): 299–309.

Cuyabano B C D, Su G, Rosa G J M, Lund M S and Gianola D. 2015. Bootstrap study of genome-enabled prediction reliabilities using haplotype blocks across Nordic Red cattle breeds. *Journal of Dairy Science* **98**(10): 7351–363.

Fisher R A. 1935. The logic of inductive interference. *Journal of Royal Statistical Society* **98**(1): 39–82.

Flury C, Tapio M, Sonstegard T, Drogemuller C, Leeb T, Simianer H, Hanotte O and Rieder S. 2010. Effective population size of an indigenous Swiss cattle breed estimated from linkage disequilibrium. *Journal of Animal Breeding Genetics* **127**(5): 339–47.

Hayes B J, Bowman P J, Chamberlain A J and Goddard M E. 2009. Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science* **92**(2): 433–33.

Hill W G and Robertson A. 1968. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**(6): 226–31.

Khatkar M S, Collins A, Cavanagh J A, Hawken R J, Hobbs M, Zenger K R, Barris W, McClintock A E, Thomson P C, Nicholas F W and Raadsma H W. 2006. A first-generation metric linkage disequilibrium map of bovine chromosome 6. *Genetics* **174**(1): 79–85.

Khatkar M S, Nicholas F W, Collins A R, Zenger K R, Cavanagh J A, Barris W, Schnabel R D, Taylor J F and Raadsma H W. 2008. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. *BMC Genomics* **9**: 187–95.

Kiselyova T Y, Kantanen J, Vorobyov N I, Podoba B E and Terletsky V P. 2014. Linkage disequilibrium analysis for microsatellite loci in six cattle breeds. *Russian Journal of Genetics* **50**(4): 406–14.

Marques E, Schnabel R D, Stothard P, Kolbehdari D, Wang Z, Taylor J F and Moore S S. 2008. High density linkage disequilibrium maps of chromosome 14 in Holstein and Angus cattle. *BMC Genetics* **9**(1): 1–12.

Meuwissen T H, Hayes B J and Goddard M E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**(4): 1819–829.

Mokry F B, Buzanskas M E, Mudadu M D A, Grossi D D A, Higa R H, Ventura R V, De Lima A O, Sargolzaei M, Meirelles S L C, Schenkel F S and Da Silva M V G B. 2014. Linkage disequilibrium and haplotype block structure in a composite beef cattle breed. *BMC Genomics* **15**(7): 1–9.

Oyelami F O, Zhao Q, Xu Z, Zhang Z, Sun H, Zhang Z, Ma P, Wang Q and Pan Y. 2020. Haplotype block analysis reveals candidate genes and QTLs for meat quality and disease resistance in chinese Jiangquhai pig breed. *Frontiers in Genetics* **11**: 752–59.

Porto-Neto L R, Kijas J W and Reverter A. 2014. The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. *Genetics Selection Evolution* **46**(1): 1–5.

Purcell S, Neale B, Todd Brown K, Thomas L, Ferreira M A, Bender D and Sham P C. 2007. PLINK: A tool set for whole genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**(3): 559–75.

Purfield D, Berry D, McParland S and Bradley D. 2012. Runs of homozygosity and population history in cattle. *BMC Genetics* **13**:70–73.

Qanbari S, Pimentel E C G, Tetens J, Thaller G, Lichtner P, Sharifi A R and Simianer H. 2010. The pattern of linkage disequilibrium in German Holstein cattle. *Animal Genetics* **41**(4): 346–56.

Salem M M I, Thompson G and Chen S. 2018. Linkage disequilibrium and haplotype block structure in Portuguese Holstein cattle. *Czech Journal of Animal Sciences* **63**(2): 61–69.

Sambrook J, Fritsol E F and Maniatis T. 1989. *Molecular Cloning- A Laboratory Manual*, 2nd ed. Cold Spring Harbor, New York.

Saravanan K A, Panigrahi M, Kumar H, Parida S, Bhushan B, Gaur G K, Kumar P, Dutt T, Mishra B P and Singh R K. 2020. Genome-wide assessment of genetic diversity, linkage disequilibrium and haplotype block structure in Tharparkar cattle breed of India. *Animal Biotechnology* **33**(2): 297–311.

Sidlova V, Kasarda R, Moravcíkova N, Trakovicka A, Curik I and Ferencakovic M. 2016. Extent of genome-wide linkage disequilibrium in Pinzgau cattle. *Journal of Central European Agriculture* **17**(2): 294–303.

Silva C R, Neves H H R, Queiroz S A, Sena J A D and Pimentel E C G. 2010. Extent of linkage disequilibrium in Brazilian Gyr dairy cattle based on genotypes of AI sires for dense SNP markers. *Leipzig* **4**: 1–29.

Villa-Angulo R, Matukumalli L K, Gill C A, Choi J, Van Tassell C P and Grefenstette J J. 2009. High-resolution haplotype block structure in the cattle genome. *BMC Genetics* **10**(1):19–28.

Weir B S. 1996. *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. 2nd Ed., Oxford University Press.

Zhu M, Zhu B, Wang Y H, Wu Y, Xu L, Guo L P, Yuan Z R, Zhang L P, Gao X, Gao H J and Xu S Z. 2013. Linkage disequilibrium estimation of chinese beef Simmental cattle using high-density SNP panels. *Asian-Australasian Journal of Animal Sciences* **26**(6): 772–79.