Received: 29 November 2017 | Revised: 24 July 2018 | Accepted: 25 July 2018

DOI: 10.1111/jbg.12356

ORIGINAL ARTICLE

WILEY Journal of Animal Breeding and Genetics

Suitability of existing commercial single nucleotide polymorphism chips for genomic studies in *Bos indicus* cattle breeds and their *Bos taurus* crosses

Nilesh Nayee[1] | **Goutam Sahana**[2] | **Swapnil Gajjar**[1] | **Ananthasayanam Sudhakar**[1] | **Kamlesh Trivedi**[1] | **Mogens Sandø Lund**[2] | **Bernt Guldbrandtsen**[2]

[1]National Dairy Development Board, Gujarat, India

[2]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark

Correspondence
Nilesh Nayee, National Dairy Development Board, Anand-388001, Gujarat, India.
Email: nileshn@nddb.coop
and
Goutam Sahana, Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, 8830 Tjele, Denmark.
Email: goutam.sahana@mbg.au.dk

Abstract

Bos indicus cattle breeds are genetically distinct from *Bos taurus* breeds. We examined the performance of three SNP arrays, the Illumina BovineHD BeadChip (777k; Illumina Inc.), the Illumina BovineSNP50 BeadChip (50k) and the GeneSeek 70k Indicus chip (75Ki; GeneSeek) in four *B. indicus* breeds (Gir, Kankrej, Sahiwal and Red Sindhi) and their *B. taurus* crosses, along with two *B. taurus* breeds, Holstein and Jersey. More SNPs on both Illumina SNP chips were monomorphic in *B. indicus* breeds (average 20.3%–29.3% on the 777k chip, 35.5%–45.5% on the 50k chip) than in Holstein (19.7% on the 777k chip, 17.1% on the 50k chip). The proportion of monomorphic SNPs on the 75Ki chip was much lower, 4% (2.8%–7%) in *B. indicus* breeds, while it was 33.5% in Holstein. With on average 164,357 heterozygous loci in *B. indicus* breeds, the 777k SNP chip has sufficient heterozygous loci to design a chip customized for *B. indicus* breeds. Principal component analysis clearly differentiated *B. indicus* from *B. taurus* breeds. Differentiation among *B. indicus* breeds was only achieved by plotting the third and fifth principal components using 777k genotype data. Admixture analysis showed that many *B. indicus* animals, previously believed to be of pure origin, are in fact had mixed ancestry. The extent of linkage disequilibrium showed comparatively higher effective population sizes in four *B. indicus* breeds compared to two *B. taurus* breeds. The results of admixture analyses show that it is important to assess the genomic composition of a bull before using it in a breeding programme.

1 | INTRODUCTION

India is rich in cattle genetic diversity. There are 41 well-defined breeds of *Bos indicus* origin (http://www.nbagr.res.in/regcat.html). With changing climatic conditions, indigenous breeds will become more important as they are well-adapted to topical climatic conditions (Hansen, 2004) and low input production systems. However, numbers of purebred animals in India have declined compared to numbers of crossbred cattle (Basic animal husbandry and fisheries statistics, 2015). This is mainly due to a lack of proper identification of purebred animals, intermixing due to uncontrolled interbreeding among *B. indicus* breeds, as well as large-scale organized crossbreeding with imported *Bos taurus* cattle to improve milk yield.

The average milk yield of *B. indicus* cows is 2.5 kg milk per day, whereas crossbred cows, with varying levels of *B. taurus* ancestry, on average yield 6.9 kg milk per day in India (Basic animal husbandry and fisheries statistics, 2015). Genetic improvement in indicine breeds would

improve economic output to farmers and thereby support conservation of *B. indicus* breeds. Currently in India, genetic improvement in dairy cattle is only implemented in a few breeds through pedigree-based selection of breeding bulls based on performance recording of elite cows. However, accuracy of selection remains low due to a limited numbers of animals under recording and a lack of deep pedigree records, causing low genetic response to selection. The long generation interval of *B. indicus* animals due to high age at maturity (Aroeria, da Silva, Fontes, & Sampaio, 1977; Kumar, 1969) further lowers genetic gain per year.

In addition to limited phenotype recording, lack of deep pedigree records, unknown breed of origin and long generation intervals act as constraints on effective breeding of dairy cattle in India. These constraints could partly be overcome by genomic selection. Genomic selection refers to selection decisions based on genomic breeding values predicted from the joint effects of genetic markers covering the whole genome (Hayes, Bowman, Chamberlain, & Goddard, 2009; Meuwissen, Hayes, & Goddard, 2001). Genomic selection has been successfully implemented for genetic improvement in various *B. taurus* breeds of dairy cattle (e.g., Lund et al., 2011; Su et al., 2012). Significantly increased response to selection has been reported using genomic selection compared to conventional progeny-based breeding value evaluations (Garcia-Ruiz et al., 2016). Genotyping with the Illumina BovineSNP50 BeadChip (50k; Illumina Inc., San Diego, CA) has been used extensively for genomic prediction in cattle. This 50k SNP chip was designed primarily based on polymorphisms observed in *B. taurus* cattle. Therefore, the 50k chip has a strong ascertainment bias towards *B. taurus* breeds (Utsunomiya et al., 2014). Besides the 50k chip, two other SNP chips differing in marker density, the Illumina BovineLD BeadChip (7k) and the Illumina BovineHD BeadChip (777k) are marketed by Illumina. These SNP chip are used extensively for genotyping cattle of *B. taurus* origin (https://www.uscdcb.com/Genotype/cur_freq.html). Commercial SNP chips for *B. indicus* cattle are also available, for example, the GeneSeek 70k Indicus chip (75Ki; GeneSeek, Lincoln, NE) designed for *B. indicus* cattle. However, the usefulness of these chips for genotyping various *B. indicus* cattle breeds from India for genomic studies has yet to be established.

Genotype information on *B. indicus* breeds will have several other major applications besides genomic prediction of breeding values. Many of these breeds are used for production of crossbred cows by crossing with bulls from Holstein and Jersey breeds. It is advocated to limit *B. taurus* ancestry to 50%–62.5% for crossbreds in India (National Breeding Policy, Government of India; http://dahd.nic.in/hi/node/86833) to maintain tolerance to adverse climate, poor feeding and management conditions. Lack of pedigree information makes it difficult to determine the proportion of *B. taurus* ancestry when selecting bull mothers to produce crossbred bulls. Genotyping data can accurately estimate the proportional ancestry of *B. indicus* animals in selection candidates for future pure breeding programs, as well as the breed composition in crossbreds.

Success of genomic selection depends on the extent of linkage disequilibrium (LD) across the genome, which varies between populations. The extent of LD influences the number of markers required across the genome to associate genetic variation with economic traits effectively. A population with higher LD requires comparatively fewer markers compared to a population with lower LD (Daetwyler, Pong-Wong, Villanueva, & Woolliams, 2010; Goddard, 2009). LD patterns also provide insight into the demographic history of a population. Information on past effective population size (N_e) improves the understanding and modeling of the genetic architecture underlying complex traits (Garcia-Gamez, Sahana, Gutierrez-Gil, & Arranz, 2012). The range of LD in the genome can be used to estimate ancestral N_e (Weir & Hill, 1980). Estimates of LD at various map distances in *B. indicus* breeds allow the estimation of past N_e, which will indicate the SNPs density required to cover the genome sufficiently for genomic studies in major *B. indicus* breeds in India.

The Illumina BovineHD BeadChip contains 777,962 SNPs (777k). It was recommended as a reference panel for Gir (Gyr), a *B. indicus* breed (Boison et al., 2015). However, genotyping of large numbers of animals with this high-density chip would require a much larger investment compared to medium-/low-density chip. Given relatively large effective population sizes of *B. indicus* breeds (Bovine HapMap et al., 2009) and existing chips mostly having been designed for *B. taurus* breeds, available low-density chips may not be suitable for genotyping *B. indicus* breeds. Being able to observe a large amount of genomic information efficiently is crucial for the success of genomic selection and conservation of genetic resources. Thus, the objectives of the present study were: (a) to identify SNPs from the 777k panel of markers highly polymorphic in *B. indicus* breeds in India; (b) to examine the suitability of the Illumina BovineSNP50 BeadChip (50k) and the GeneSeek 75k Indicus chip (75Ki) for genomic studies in *B. indicus* breeds and their crosses with *B. taurus* cattle; and (c) to study how well breed composition can be estimated by the SNP genotypes for *B. indicus* breeds and their *B. taurus* crosses.

The usefulness of present genotyping chips on the Illumina platform was evaluated on four *B. indicus* breeds from India: Sahiwal (SHW), Gir (GIR), Kankrej (KNK), and Red Sindhi (RSN), and their crosses with two *B. taurus* breeds, Holstein (HOL) and Jersey (JER). These breeds/populations contribute 90% of the frozen semen produced in India (NDDB compilation, www.nddb.org).

2 | MATERIALS AND METHODS

2.1 | Animals and genotyping

Two hundred and forty-two animals, mainly bulls (219) used for frozen semen production and 23 females were selected based on their phenotypic appearance consistent with typical breed characteristics accepted by local breeders. Pedigree information was considered whenever available to avoid including close relatives in the study samples. The bulls selected each had sons used for semen collection and therefore are expected to have nontrivial contribution to the current gene pool of the breed. Due to a lack of sufficient number of unrelated bulls kept at semen stations for the Gir and Holstein crossbred (HCB) populations, 23 (13 Gir and 10 HCB) unrelated cows from the field were included in the study. In addition, 777k chip genotype data from 20 purebred Danish Jersey cows and 20 Nordic Holstein cows were provided by Aarhus University, Denmark. In total 110 *B. indicus*, 40 *B. taurus* and 132 *B. indicus* × *B. taurus* animals were genotyped. Numbers of animals from each breed used in this study are given in Table 1.

DNA was isolated from either semen or blood samples using standard protocols using commercially available kits (Quiagen). DNA samples were subjected to quality control using nanodrop (www.nanodrop.com) and Gel electrophoresis. Samples with at least 50 ng/μl DNA concentration, 260/280 OD ratios of 1.8–2.0, and a good quality gel picture were selected for genotyping with the 777k chip at M/s Sandor Life Sciences Ltd., Hyderabad, India. SNP genotyping data generated by Illumina Genome Studio were analyzed after performing standard quality control. Data were converted to PLINK input files (Purcell et al., 2007). Samples with genotyping rates less than 90% and SNPs with less than 90% individuals genotyped were discarded from further analysis. Genotyping data with the 777k chip from 20 Holstein and 20 Jersey purebred animals obtained from Aarhus University, Denmark (Su et al., 2012) were combined with the data of purebred animals of *B. indicus* breeds and their crosses for population genetics study.

2.2 | Comparing performance of various genotyping chips

Annotation files for SNPs in 75Ki genotyping chip and Illumina 50k chip were obtained from the SNPchimp database (http://bioinformatics.tecnoparco.org/SNPchimp/; (Nicolazzi et al., 2014)). Using the SNP list from the annotation files, the genotypes for SNPs in 75Ki and 50k chips were extracted from 777k genotypes for further analysis. The SNP positions were annotated based on UMD 3.1 reference genome (Zimin et al., 2009). Sample wise and SNP wise genotyping rates were calculated using PLINK software (Purcell et al., 2007).

Breed-wise minor allele frequencies (MAF) were calculated for the 777k panel SNPs (total 273 animals), and for subsets corresponding to the 75Ki (272 animals) and the

TABLE 1 Numbers of animals from *Bos indicus*, *Bos taurus* and their crosses genotyped with Illumina BovineHD BeadChip (777k) SNP array, average minor allele frequency (MAF) and expected heterozygosity in different breeds for three different SNP chips: Illumina BovineHD BeadChip (777k), Illumina BovineSNP50 BeadChip (50k) and GeneSeek 70k Indicus chip (75Ki)

Breed group	Breed abbreviations	No. of samples		Average MAF			Expected heterozygote loci		
		Bulls	Cows	777k	75Ki	50k	777k	75Ki	50k
Bos indicus	GIR	20	13	0.15	0.28	0.11	165,521	25,395	7,191
	KNK	22	0	0.15	0.27	0.11	162,290	25,655	7,015
	SHW	43	0	0.15	0.27	0.11	165,112	24,726	7,218
	RSN	12	0	0.15	0.26	0.11	164,503	24,055	7,233
Holstein crossbred	HGR	8	1	0.25	0.28	0.21	260,182	26,028	13,399
	HKN	21	0	0.26	0.32	0.21	272,446	29,185	13,577
	HSW	37	7	0.26	0.33	0.21	270,934	29,404	13,412
	HCB	9	2	0.26	0.3	0.21	265,876	27,556	13,445
Jersey crossbred	JSW	22	0	0.26	0.31	0.2	263,611	28,514	12,708
	JCB	24	0	0.26	0.31	0.2	261,988	28,053	12,921
Bos taurus	HOL[a]	0	20	0.22	0.18	0.22	215,650	16,770	13,320
	JER[a]	0	20	0.19	0.16	0.18	190,963	15,168	11,792
	Total	219	63						

[a]Genotype data obtained from Aarhus University, Denmark; GIR: Gir; HCB: HOL Crossbred; HGR: HOL-GIR crossbred; HKN: HOL-KNK crossbred; HOL: Holstein; HSW: HOL-SHW crossbred; JCB: JER crossbred; JER: Jersey; JSW: JER-SHW crossbred; KNK: Kankrej; RSN: Red Sindhi; SHW: Sahiwal.

50k chip (277 animals) to study heterozygosity of SNP markers using PLINK software. Small differences in numbers of animals for different SNP chips were due to removal of some individuals during quality control for the specific marker set in a chip. Average expected heterozygosity for each breed was calculated using formula $\frac{1}{n}\sum_i 2p_iq_i$ where p_i was the minor allele frequency, and q_i was the major allele frequency at SNPs i, n is the number of SNP. The genotype distribution for each SNP in the 777k panel was tested for deviations from Hardy-Weinberg proportions (HWP) using a χ^2 test as implemented in PLINK. SNPs with a test probability of $-\log_{10}(P) > 5$ were removed from analysis. MAF were compared for all the three chips for each breed.

2.3 | Population structure

The population differentiation and structure of four *B. indicus*, two *B. taurus* breeds and their crossbred populations were studied using Wright's F_{st} statistics, principal component analysis (PCA) and admixture analyses as described below. These were performed for all the three SNP chips (777k, 75Ki and 50k) to examine whether choice of chip might influence the results.

2.3.1 | Wright's F_{st} statistics

Weir and Cockerham's estimator of Wright's F_{st} (Weir & Cockerham, 1984) statistics was evaluated using PLINK software for determining genetic distance among all the breeds under study. Genomic relationships within breed among purebred individuals were calculated by VanRaden's method 1 (VanRaden, 2008) using *Gmatrix* (http://dmu.agrsci.dk/Gmatrix/). The genomic relationship matrix was normalized to keep diagonal elements close to 1 (Forni, Aguilar, & Misztal, 2011). From each pair of individuals with an estimated genomic relationship more than 0.2, one was removed prior to estimation of F_{st} statistics. The final dataset included 232 animals. 760,139 SNPs for 777k chip, 70,023 SNPs for 75Ki chip and 46,584 for 50k chip were used after removing SNPs with genotyping rates <0.90.

2.3.2 | Principle Component Analysis (PCA)

High-density SNP genotypes were filtered with the criteria of MAF of 0.01 and SNPs with pairwise LD (r^2) value of <0.1 with adjoining SNPs to obtain a pruned SNP panel for PCA analysis. This left 138,451 SNPs for 777k, 25,403 SNPs for 75Ki and 19,002 SNPs for 50k after filtering the datasets. PCA was performed using SMARTPCA script from EIGENSOFT (Price et al., 2006). Eigenvectors obtained were plotted to visualize breed differences using *ploteig* script from the EIGENSOFT package.

2.3.3 | ADMIXTURE analysis

Unsupervised clustering analysis using ADMIXTURE software (Alexander, Novembre, & Lange, 2009) was carried out to infer ancestry ratios for 120 purebred animals using pruned 777k data having 138,451 SNPs. To arrive at the number of K components, tenfold cross-validations was performed for $K = 1$ to $K = 9$ ancestral populations. The Q values (the ancestry coefficients, (Alexander et al., 2009)) obtained were plotted by breed to show proportion of ancestry from different breeds in each individual animal. A supervised clustering was carried out to infer ancestry ratio in crossbred animals keeping $K = 6$.

2.3.4 | Linkage disequilibrium patterns

The data set of selected HD genotyped individuals was used to estimate LD for SNPs on chromosome 1 using PLINK. Breed-wise LD (r^2) values were calculated for each pair of 1,000 consecutive SNPs on chromosome 1 at most 10 Mbp apart using a sliding window. The SNP pairs were binned according to their physical distance with a bin size of 10 kbp, and changes in LD with increasing distance were compared among *B. indicus* and *B. taurus* breeds. Average r^2 values were plotted against the physical distance. Red Sindhi breed was excluded for LD estimation due to small sample size.

Effective population sizes at different times points in the past were estimated for each breed. N_e was estimated by solving $r^2 = \frac{1}{1+4N_eC} + \frac{1}{n}$ (Sved & Feldman, 1973; Weir & Hill, 1980). Here r^2 was the squared correlation of alleles at a pair of loci, c was the distance in Morgan (here approximated by 1 Mb ≈ 1 cM), and n was the chromosome sample size (number of haplotypes = 2 × number of individuals). N_e estimated from LD at a genetic distance c (in Morgan) corresponds to N_e at $T = 1/(2c)$ generations in the past (Hayes, Visscher, McPartlan, & Goddard, 2003). N_e at 50 and 100 generations in the past were estimated.

3 | RESULTS

3.1 | Comparison of variation captured by various genotyping chips

Between 20.3% and 29.3% of the SNPs on the 777k chip and between 35.5% and 45.5% SNPs on the 50k chip were monomorphic in *B. indicus* breeds. In Holstein, 19.7% (777k) and 17.1% (50k) SNPs were monomorphic. The proportion of monomorphic SNPs in the 75Ki chip was, as expected, much lower, 4% (2.8%–7%) in *B. indicus* breeds, while it was 33.5% in Holstein. The proportion of monomorphic SNPs for Jersey breed were 26.3%, 38.4% and 26.1% in 777K, 75Ki and 50K chips, respectively. The

75Ki chip had higher average MAF in *B. indicus* breeds (0.26–0.28), but had lower average MAF in the Holstein (0.18) and Jersey (0.16). The 777K chip had an average MAF of 0.15 for *B. indicus* breeds, 0.22 for Holstein and 0.19 for Jersey. The 50k chip had the lowest average MAF (0.11) for *B. indicus* breeds and had higher average MAF in *B. taurus* breeds (0.22 for Holstein and 0.18 for Jersey).

Average MAF in each breed and expected heterozygosity for different SNP chips are presented in Table 1. The number of SNPs monomorphic in each breed and the number of SNPs simultaneously monomorphic in two breeds for 777k chip are presented in Table 2. The proportions of SNPs with MAF < 10% and MAF > 25% are given in Table 3. Supporting information Figure S1A-L for the distributions of MAF for various breeds for 777k chip, Supporting information Figure S2A-L for 75Ki and Supporting information Figure S3A-L for 50k chip in various breeds.

The proportion of SNPs deviating from HWP for 777k, 50k and 75Ki chips are presented in Table 4. Overall, less than 0.15% SNPs deviated significantly from HWP. Supporting information Figure S4A-G for distribution of $-\log_{10}(p)$ values for deviations from HWP in various breeds studied.

3.2 | Wright's F_{st} estimates

Mean F_{st} value estimated with 777k chip for all *B. indicus* breeds considered together was 0.058 indicating some genetic differentiation among these breeds. Pairwise Wright's F_{st} values calculated for all pairs of four *B. indicus* breeds, two *B. taurus* breeds are given in Table 5. *B. indicus* and *B. taurus* breeds showed great divergence. Between 6.6% and 9.3% of the total genetic variation could be attributed to subdivision into *B. indicus* breeds. These analyses were repeated with the 75Ki and 50k chips for four *B. indicus* breeds (Supporting information Table S1). The F_{st} estimates from 75Ki to 50k chips were in the similar range as observed with the 777k chip.

3.3 | Principal component analyses for population structure

After pruning the SNPs based on LD and removal of individuals to keep sample sizes reasonably balanced, genotyping data of 138,451 SNPs and 232 individuals were used for PCA using EIGENSOFT (Price et al., 2006). The software, by default, computes the first 10 principal components (PC). The first PC explained around 17% of the variation, whereas the first five PCs together explained around 30% variation from data using 777k chip. The loadings onto the first and second PCs for all purebred animals are plotted in Figure 1a. To assess breed structure within *B. indicus* breeds, PC from only *B. indicus* breeds were plotted (Figures 1b,c). Loadings onto first and second PCs for all 232 animals based on PCA using 75Ki and 50k SNP panels are presented in Supporting information Figures S5 and S6. Holstein (HOL) and Jersey (JER) animals formed two distinct clusters using PC1 and PC2. All *B. indicus* breeds fell into one cluster when PC1 and PC2 were plotted. Crossbred cattle are spread between these three clusters. Using the 75Ki and 50k data, the difference between *B. taurus* and *B. indicus* was represented by the first PC (Supporting information Figures S5 and S6). The first PC explained 13.4% and 12.3% of variation in genotypes for 75Ki and 50k chips, while top five PC explained 26.9% and 24.9% of the variance, respectively.

3.4 | ADMIXTURE analysis

Unsupervised clustering analysis using ADMIXTURE software (Alexander et al., 2009) showed that the majority of the *B. indicus* animals were assigned to their respective breeds (Figure 2). A supervised model-based ancestry estimation using 777K chip and $K = 6$ for crossbred animals of four *B. indicus* and two *B. taurus* breeds is presented in Figure 3. Some individuals from *B. indicus* breeds showed evidence of mixed ancestry (Figure 2). The degree of admixture varied from individual to individual. The cross

TABLE 2 Number of SNPs in Illumina BovineHD BeadChip (777k) monomorphic in different breeds

Breed	Gir	Sahiwal	Kankrej	Red Sindhi	Holstein	Jersey
Gir	157,800					
Sahiwal	97,727	157,618				
Kankrej	127,506	120,342	227,691			
Red Sindhi	118,991	112,184	148,015	222,015		
Holstein	32,884	31,352	40,784	41,996	150,159	
Jersey	52,119	45,356	61,367	65,667	123,144	200,405

Diagonal in the table shows number of monomorphic SNPs in a breed, whereas off diagonal shows number of SNPs monomorphic and common to both the breeds.

TABLE 3 Distribution of SNPs based on minor allele frequency (MAF) ranges for three different SNP chips: Illumina BovineHD BeadChip (777k), Illumina BovineSNP50 BeadChip (50k) and GeneSeek 70k Indicus chip (75Ki) for purebred and crossbreds (CB)

		% Monomorphic SNPs			%SNP MAF ≤ 0.1			%SNP MAF > 0.25		
Breed group	**Breed**	**777k**	**75Ki**	**50k**	**777k**	**75Ki**	**50k**	**777k**	**75Ki**	**50k**
Bos indicus	Gir	20.3	2.8	36.1	50.8	17.1	64.9	28.1	58.8	18.0
	Kankrej	29.3	5.4	45.5	49.0	16.3	64.1	27.9	57.8	17.9
	Sahiwal	20.3	3.0	35.5	49.5	18.0	65.3	28.1	54.8	18.6
	Red Sindhi	28.5	7.0	44.3	51.6	20.9	65.5	25.6	49.6	16.9
Holstein crossbred	Holstein-Gir CB	9.7	3.4	21.8	18.5	10.4	31.1	46.7	51.3	37.9
	Holstein-Kankrej CB	4.6	0.4	15.3	17.8	4.3	32.3	56.2	72.1	41.5
	Holstein-Sahiwal CB	3.7	0.3	13.2	18.0	4.3	32.5	55.8	72.6	40.4
	Holstein-indicus CB	7.7	1.8	19.5	22.4	11.0	35.5	49.9	59.6	38.5
Jersey crossbred	Jersey-Sahiwal CB	6.8	0.8	19.4	19.9	6.0	36.0	54.8	69.9	39.2
	Jersey-indicus CB	6.7	1.0	19.2	18.4	5.4	34.0	53.0	66.7	40.0
Bos taurus	Holstein	19.7	33.5	17.1	34.1	44.4	33.1	46.4	39.9	45.9
	Jersey	26.3	38.4	26.1	41.6	50.5	43.1	39.9	34.4	38.5

TABLE 4 Number of SNPs deviating from HWP ($p < 0.00001$) in various breeds/crossbred populations[a] using Illumina BovineHD BeadChip (777k), Illumina BovineSNP50 BeadChip (50k) and GeneSeek 70k Indicus chip (75Ki)

		Deviation from HWE					
Breed group	**Breed/crossbred**	**No. of SNPs (777k)**	**% of 777k SNPs**	**No. of SNPs (50k)**	**% of 50k SNPs**	**No. of SNPs (75Ki)**	**% of 75Ki SNPs**
Bos indicus	GIR	2,105	0.3	65	0.1	17	0.03
	KNK	10,637	1.4	121	0.2	7	0.01
	SHW	16,510	2.1	275	0.6	56	0.08
Holstein crossbreds	HKN	26,449	3.4	347	0.7	21	0.03
	HSW	30,632	3.9	551	1.1	19	0.03
Jersey crossbreds	JSW	29,272	3.8	462	0.9	13	0.02
	JCB	16,604	2.1	435	0.9	7	0.01

[a]Some breed/crossbred had small sample size (11 for HCB, 12 for RSN and 9 for HGR) and therefore, not included in this table.

validation errors for the number of K components in the Admixture analysis are presented in Supporting information Figure S7A. As cross-validation errors differed little between $K = 4$–8 with 777k chip; $K = 6$ was chosen considering there were 6 breeds in the study. Plots of cross-validation errors for 50k and 75Ki chips are presented in Supporting information Figure S7B-C. The admixture plots using 50k data were similar to admixture plots using the 777k data (Supporting information Figure S8), whereas 75Ki data clearly separated *B. indicus* breeds into more components compared with 50k (Supporting information Figure S8).

3.5 | Linkage disequilibrium pattern

The distribution of LD (r^2) estimates as a function of physical distance among the SNPs located on chromosome 1 was calculated for all three SNP chips (Supporting information Table S2A-C). Mean r^2 values for various distance bins are presented Figure 4 for 777k chip and Supporting information Figure S9A-B for 75Ki and 50k SNP chips, respectively. LD over long distances followed similar pattern in three *B. indicus* breeds as in *B. taurus* breeds; however, mean r^2 was around 0.15 at a distance above 100 kbp which was 0.2 for Holstein and still higher, 0.27 for Jersey. This indicates that historic effective population sizes of *B. indicus* breed must have been somewhat larger than that of the *B. taurus* breeds studied. LD calculations for Red Sindhi breed were not carried out due to less number of samples. Past effective population size estimated based on the LD measures using 777k chip are presented in Table 6. The effective population sizes in two *B. taurus* breeds had declined at a faster rates compared with four *B. indicus* breeds. N_e at 50 and 100 generations in the past

TABLE 5 Pairwise Wright's F_{st} values among four *Bos indicus* breeds: Gir (GIR), Kankrej (KNK), Red Sindhi (RSN) and Sahiwal (SHW), two *Bos taurus* breeds: Holstein (HOL) and Jersey (JER), using Illumina BovineHD BeadChip (777k)

Breeds	RSN	SHW	KNK	GIR	HOL	JER
RSN	0	0.093 ± 0.0014	0.081 ± 0.0015	0.092 ± 0.0014	0.404 ± 0.0038	0.441 ± 0.0038
SHW	0.093 ± 0.0014	0	0.076 ± 0.0012	0.09 ± 0.0012	0.428 ± 0.0040	0.467 ± 0.0039
KNK	0.081 ± 0.0015	0.076 ± 0.0012	0	0.066 ± 0.0012	0.423 ± 0.0041	0.462 ± 0.0040
GIR	0.092 ± 0.0014	0.09 ± 0.0012	0.066 ± 0.0012	0	0.43 ± 0.0040	0.469 ± 0.0039
HOL	0.404 ± 0.0038	0.428 ± 0.0040	0.423 ± 0.0041	0.43 ± 0.0040	0	0.159 ± 0.0022
JER	0.441 ± 0.0038	0.467 ± 0.0039	0.462 ± 0.0040	0.469 ± 0.0039	0.159 ± 0.0022	0

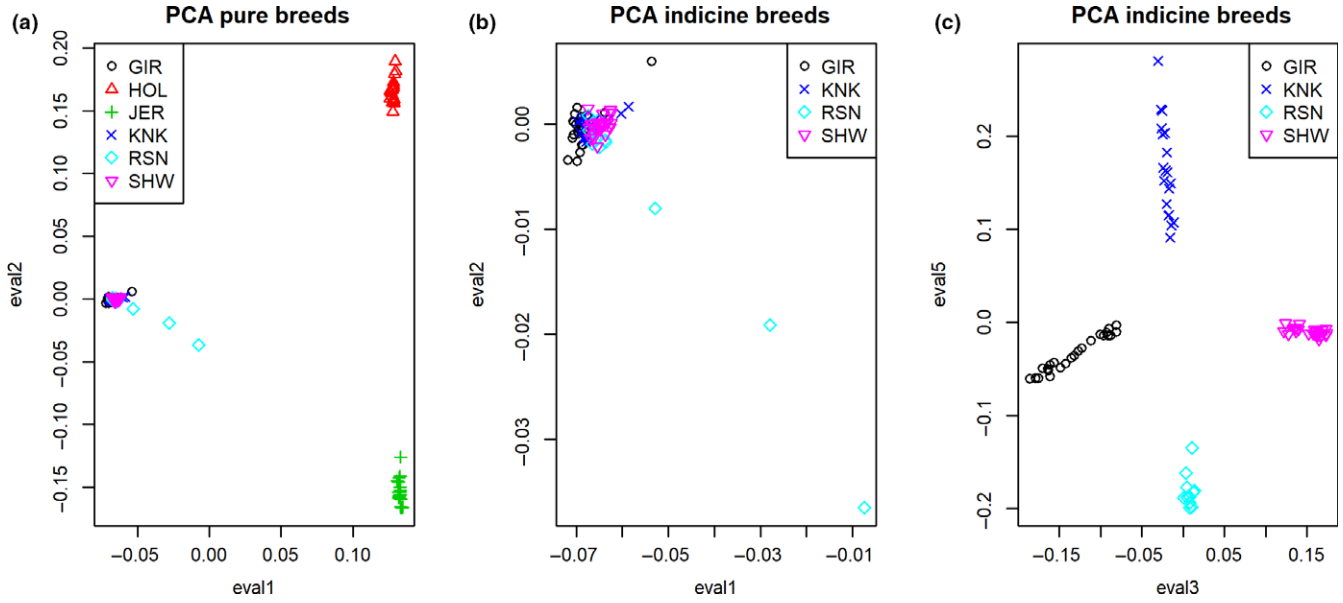

FIGURE 1 Plots of the principal components (PCs) constructed using Illumina BovineHD SNP (777k) chip. (a) First and second PCs for PCA analysis keeping both *Bos indicus* and *Bos taurus* breeds; (b) first and second PCs for PCA analysis, only indicus breeds; (c) third and fifth PCs for PCA analysis, only *B. indicus* breeds. GIR: Gir; KNK: Kankrej; SHW: Sahiwal; RSN: Red Sindhi; HOL, Holstein; JER, Jersey

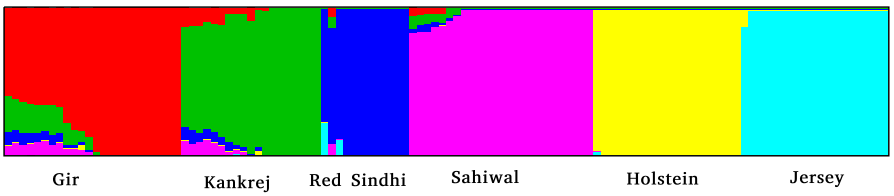

FIGURE 2 A Individual unsupervised model-based ancestry estimation using Illumina BovineHD SNP (777k) chip and $K = 6$ for purebred animals of four *Bos indicus* and two *Bos taurus* breeds. Vertical bars represent individuals and the breeds are indicated on the *X*-axis

estimates based on 75Ki and 50k SNP chips are presented in Supporting information Table S3A-B.

4 | DISCUSSION

The design history of various SNP chips is clearly reflected in the results we obtained. For the 777k and 50k chips, the highest average MAF was observed in Holstein crossbreds followed by the Holstein breed. Around 25% and 40% of the SNPs in the 777k and 50k SNP chips were monomorphic in the four *B. indicus* breeds studied, much higher than the numbers for the *B. taurus* breeds. There was limited overlap between the SNPs monomorphic in *B. indicus* breeds and those monomorphic in *B. taurus* breeds (Table 2). This reflects that SNPs included in these two SNP arrays were primarily selected based on being polymorphic in *B. taurus* breeds (https://www.illumina.com/Documents/products/datasheets/datasheet_bovineHD.pdf).

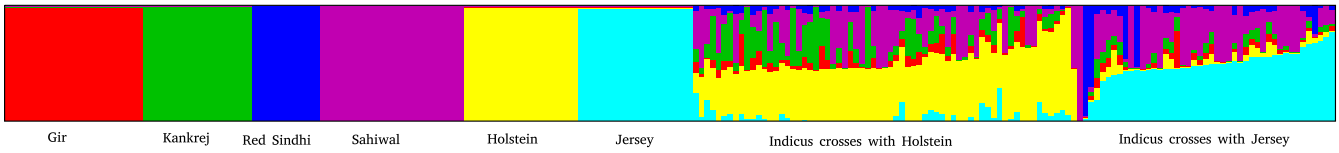

FIGURE 3 Supervised model-based ancestry estimation using Illumina BovineHD SNP (777k) chip and $K = 6$ for crossbred animals of four *Bos indicus* and two *Bos taurus* breeds. Vertical bars represent individuals and the breeds are indicated on the *X*-axis

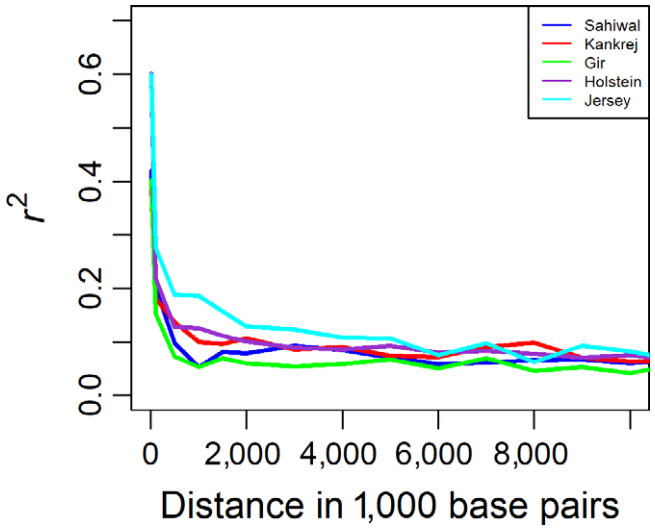

FIGURE 4 linkage disequilibrium (r^2) over long range distances between pair of SNPs using Illumina BovineHD SNP (777k) chip. Mean r^2 values are plotted against distance (kbp) between pair of SNP over long range

TABLE 6 Estimated effective population sizes in 50 and 100 generations back for *Bos indicus* and *Bos taurus* breeds based on Illumina BovineHD BeadChip (777k) SNP chip

Generations	Sahiwal	Kankrej	Gir	Holstein	Jersey
50	665.58	328.59	673.60	224.47	130.21
100	566.36	416.60	859.16	437.33	256.58

75Ki SNP panel has higher number of polymorphic SNPs than 50k SNP panel for *B. indicus* breeds. Around 50% of SNPs had a MAF > 0.25 for *B. indicus* breeds. Still, loci with MAF < 0.1 constituted more than 16% (around 20% for Sahiwal). On the other hand, considering lower expected heterozygosity of the 75Ki SNP panel in pure HOL and JER breeds, this chip is less suited to genotype pure Holstein and Jersey animals and their crossbreds and especially those with a high proportion of *B. taurus* origin. Given the high number of SNP on the 777k chip polymorphic in both *B. taurus* and *B. indicus* breeds, it should be feasible to design a SNP chip with high expected heterozygosity in both subspecies. That would be very useful for studies that simultaneously incorporate *B. taurus* and *B. indicus* breeds and their crosses.

The study showed lower F_{st} values using 760,139 SNPs from 777k chip among *B. indicus* breeds (0.066–0.093) compared to F_{st} among *B. taurus* breeds (0.159) indicating comparatively less divergence among *B. indicus* breeds. A higher F_{st} value (13.3%) considering a larger diverse panel of *B. indicus* breeds genotyped with microsatellite markers has previously been reported (Sharma et al., 2015). Results of PCA analysis for genotype data of 120 purebred animals using pruned 777k data having 138,451 SNPs showed that the first and second PC clearly separated the *B. taurus* from the *B. indicus* breeds (Figure 1a), while all four *B. indicus* breeds clustered together expect a few Red Sindhi animals (Figure 1b). PCs 3 and 5 separated all four *B. indicus* breeds (Figure 1c). Sahiwal and Gir breeds were separated by principal component 3, whereas Red Sindhi and Kankrej were separated by principal component 5 (Figure 1c). The results are in agreement with previous studies of molecular differentiation of *B. indicus* breeds by Kale et al. (2010) and Sharma et al. (2015). Differences in ordering of PCs were observed depending on which chip was used. This indicates that PCA results are sensitive to the ascertainment happening during chip design.

Several of the *B. indicus* animals showed evidence of recent admixture. There was evidence of some admixture of Kankrej in Gir bulls and *vice versa* indicating mixing or crossing among the *B. indicus* breeds (Figure 2). The variability in the degree of admixture suggests that the admixture happened relatively recently. Some of the Sahiwal bulls showed the presence of Kankrej or Gir components. Surprisingly, some of the Kankrej and Gir animals had up to 5% of Red Sindhi as well as Sahiwal ancestry. The shared geographical area of these breeds and general lack of deep pedigree information might have contributed to this admixture.

As expected, crossbreds had various proportions of *B. indicus* and either Holstein or Jersey ancestry (Figure 3). Lack of detailed pedigrees for bull-mothers explains admixture of various indigenous breeds in the crossbred bulls. We observed that genotyped crossbred bulls had high degrees of *B. taurus* ancestry. The Indian national cattle breeding policy recommends keeping *B. taurus* genome levels to around 50%–62.5%. Some Jersey crossbred bulls had admixture from Holstein and *vice versa* (Figure 3). This may have happened when procuring bulls from

villages where the bull dams' pedigree information was incomplete or missing.

Linkage disequilibrium measured by r^2 decreased with increasing distance between SNPs (Figure 4). Average r^2 in *B. indicus* breeds decreased from around 0.40 at 10 kbp to 0.19 at 100 kbp and to 0.09–0.1 at 500 kbp. In HOL animals, the corresponding values for average r^2 were 0.60, 0.20 and 0.12. LD in JER breed were at still higher levels at 0.66, 0.27 and 0.17 for the same distances. Considering total bovine genome length of 2.7 GB, 50,000 SNPs (present in the widely used Illumina BovineSNP50 chip) corresponds to a distance of on average 60 kbp between neighbouring SNPs. The average LD (r^2) at 60 kb distance in HOL cattle was 0.26 in the present study. LD and N_e for Red Sindhi were not calculated because of small sample size in the present study. To achieve the same average LD among SNPs for *B. indicus* breeds included in the present study, average marker distance needs be kept at 40 kbp or less between SNPs. This corresponds to around 80,000 SNPs in the genotyping chip for *B. indicus* breeds of cattle from India. Average LD (r^2) values observed in current study were comparable to those reported by McKay et al., 2008 who used 2,641 SNP genotype data from 383 *B. taurus* and 137 *B. indicus.* For Holstein, Prasad et al. (2008) reported average r^2 of 0.60, 0.26 and 0.10 at distances of 5, 100 and 500 kbp, respectively (McKay et al., 2008; Prasad et al., 2008).

Lower levels of LD at larger distances translate in to higher past effective population size for *B. indicus* breeds than for the *B. taurus* breeds. Bulls of pure *B. indicus* breeds in the study were generally sourced from few elite herds and hence LD pattern observed might underestimate the historical effective population size. The estimated effective population sizes in the past (Table 6) in comparison with *B. taurus* breeds still indicate higher effective population sizes in *B. indicus* breeds. The effective population sizes in Holstein and Jersey were nearly halved from 100 to 50 generations in past, while the rate of decrease in *B. indicus* breeds were much lower, between 4% (Sahiwal) and 30% (Gir).

5 | CONCLUSION

The present study indicates the Illumina bovineSNP50 chip is not optimal for genotyping cattle of *B. indicus* origin. GeneSeek 75Ki chip represents a better choice for the *B. indicus* breeds, but not for *B. taurus* breeds. A customized SNP panel with more SNPs polymorphic in a majority of Indian *B. indicus* breeds while also moderately polymorphic in Holstein and Jersey cattle would be desirable and feasible. The LD patterns suggest including approximately 80,000 SNPs for such a custom chip. The performance of the designed custom chip should be evaluated for imputation accuracy to higher density marker sets. The results of admixture analyses show that it is important to assess the genomic composition of a bull before using it in pure breeding programme. Crossbred bulls also need to be screened to keep *B. taurus* genetic level around 50%–62.5% as recommended by the Indian national cattle breeding policy.

ACKNOWLEDGEMENTS

Under Indo-Danish collaboration, Department of Molecular Biology and Genetics, Aarhus University, Denmark and National Dairy Development Board, Anand, Gujarat, India have collaborated to explore and implement genomic selection in Indian cattle. The present work is the first study undertaken by this collaboration.

ORCID

Goutam Sahana http://orcid.org/0000-0001-7608-7577

REFERENCES

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Aroeria, J. A. D., da Silva, H. M., Fontes, L. R., & Sampaio, I. B. M. (1977). Age at first calving, length of reproductive life and life expectancy in Zebu cows. *Animal Breeding Abstracts*, *46*, 41.

Basic animal husbrandry and fisheries statistics (2015). *Department of animal husbandry, dairying & fisheries*, ed. New Delhi, India: Government of India.

Boison, S. A., Santos, D. J., Utsunomiya, A. H., Carvalheiro, R., Neves, H. H., O'Brien, A. M., … da Silva, M. V. (2015). Strategies for single nucleotide polymorphism (SNP) genotyping to enhance genotype imputation in Gyr (*Bos indicus*) dairy cattle: Comparison of commercially available SNP chips. *Journal of dairy science*, *98*(7), 4969–4989. https://doi.org/10.3168/jds.2014-9213

Bovine HapMap, C., Gibbs, R. A., Taylor, J. F., Van Tassell, C. P., Barendse, W., Eversole, K. A., … Dodds, K. G. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science*, *324*(5926), 528–532. https://doi.org/10.1126/science.1167936

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, *185*(3), 1021–1031. https://doi.org/10.1534/genetics.110.116855

Forni, S., Aguilar, I., & Misztal, I. (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genetics, Selection, Evolution: GSE*, *43*, 1. https://doi.org/10.1186/1297-9686-43-1

Garcia-Gamez, E., Sahana, G., Gutierrez-Gil, B., & Arranz, J. J. (2012). Linkage disequilibrium and inbreeding estimation in

Spanish Churra sheep. *BMC Genetics*, *13*, 43. https://doi.org/10.1186/1471-2156-13-43

Garcia-Ruiz, A., Cole, J. B., VanRaden, P. M., Wiggans, G. R., Ruiz-Lopez, F. J., & Van Tassell, C. P. (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(28), E3995–E4004. https://doi.org/10.1073/pnas.1519061113

Goddard, M. (2009). Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica*, *136*(2), 245–257. https://doi.org/10.1007/s10709-008-9308-0

Hansen, P. J. (2004). Physiological and cellular adaptations of zebu cattle to thermal stress. *Animal Reproduction Science*, *82–83*, 349–360. https://doi.org/10.1016/j.anireprosci.2004.04.011

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges (vol 92, pg 433, 2009). *Journal of Dairy Science*, *92*(2), 433–443.

Hayes, B. J., Visscher, P. M., McPartlan, H. C., & Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research*, *13*(4), 635–643. https://doi.org/10.1101/gr.387103

Kale, D. S., Rank, D. N., Joshi, C. G., Yadav, B. R., Koringa, P. G., Thakkar, K. M., … Solanki, J. V. (2010). Genetic diversity among Indian Gir, Deoni and Kankrej cattle breeds based on microsatellite markers. *Indian Journal of Biotechnology*, *9*(2), 126–130.

Kumar, S. S. R. (1969). A report on some important economic traits of Red Sindhi and Jersey grades. *Indian Veterinary Journal*, *46*, 5.

Lund, M. S., Roos, A. P., Vries, A. G., Druet, T., Ducrocq, V., Fritz, S., … Su, G. (2011). A common reference population from four European Holstein populations increases reliability of genomic predictions. *Genetics, Selection, Evolution : GSE*, *43*, 43. https://doi.org/10.1186/1297-9686-43-43

McKay, S. D., Schnabel, R. D., Murdoch, B. M., Matukumalli, L. K., Aerts, J., Coppieters, W., … Moore, S. S. (2008). An assessment of population structure in eight breeds of cattle using a whole genome SNP panel. *BMC Genetics*, *9*(1), 37. https://doi.org/10.1186/1471-2156-9-37

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

Nicolazzi, E. L., Picciolini, M., Strozzi, F., Schnabel, R. D., Lawley, C., Pirani, A., … Stella, A. (2014). SNPchiMp: A database to disentangle the SNPchip jungle in bovine livestock. *BMC Genomics*, *15*, 123. https://doi.org/10.1186/1471-2164-15-123

Prasad, A., Schnabel, R. D., McKay, S. D., Murdoch, B., Stothard, P., Kolbehdari, D., … Moore, S. S. (2008). Linkage disequilibrium and signatures of selection on chromosomes 19 and 29 in beef and dairy cattle. *Animal Genetics*, *39*(6), 597–605. https://doi.org/10.1111/j.1365-2052.2008.01772.x

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Sharma, R., Kishore, A., Mukesh, M., Ahlawat, S., Maitra, A., Pandey, A. K., & Tantia, M. S. (2015). Genetic diversity and relationship of Indian cattle inferred from microsatellite and mitochondrial DNA markers. *BMC Genetics*, *16*, 73. https://doi.org/10.1186/s12863-015-0221-0

Su, G., Brøndum, R. F., Ma, P., Guldbrandtsen, B., Aamand, G. P., & Lund, M. S. (2012). Comparison of genomic predictions using medium-density (approximately 54,000) and high-density (approximately 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. *Journal of Dairy Science*, *95*(8), 4657–4665. https://doi.org/10.3168/jds.2012-5379

Sved, J. A., & Feldman, M. W. (1973). Correlation and Probability Methods for One and 2 Loci. *Theoretical Population Biology*, *4*(1), 129–132. https://doi.org/10.1016/0040-5809(73)90008-7

Utsunomiya, Y. T., Bomba, L., Lucente, G., Colli, L., Negrini, R., Lenstra, J. A., … European Cattle Genetic Diversity Consortium. (2014). Revisiting AFLP fingerprinting for an unbiased assessment of genetic structure and differentiation of *B. taurus* and zebu cattle. *BMC Genetics*, *15*, 47. https://doi.org/10.1186/1471-2156-15-47

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*(6), 1358–1370.

Weir, B. S., & Hill, W. G. (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics*, *95*(2), 477–488.

Zimin, A. V., Delcher, A. L., Florea, L., Kelley, D. R., Schatz, M. C., Puiu, D., … Salzberg, S. L. (2009). A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biology*, *10*(4), R42. https://doi.org/10.1186/gb-2009-10-4-r42

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Nayee N, Sahana G, Gajjar S, et al. Suitability of existing commercial single nucleotide polymorphism chips for genomic studies in *Bos indicus* cattle breeds and their *Bos taurus* crosses. *J Anim Breed Genet*. 2018;00:1–10. https://doi.org/10.1111/jbg.12356