Indian J Dairy Sci 74(2): 138-144
https://doi.org/10.33785/IJDS.2021.v74i02.006

RESEARCH ARTICLE

Efficiency of imputing missing genotypes by INDUSCHIP v2 in HF Crossbred cattle

Sujit Saha[1], Nilesh Nayee[1], Heena A Shah[2], Swapnil Gajjar[1], A Sudhakar[1], Sandeep K Donthula[1] and Hardik V Poojara[1]

Received: 13 September 2020 / Accepted: 24 November 2020 / Published online: 07 June 2021

Abstract: INDUSCHIP- an Illumina platform based custom made genotyping chip was designed with 45K polymorphic markers for Indian cattle breeds and 8K base SNPs of Illumina BovineLD chip to genotype indigenous and crossbred cattle in India. Current study was undertaken to assess the genotype imputation efficiency of INDUSCHIP v2 microarray in HF crossbred cattle and compare its efficiency of imputation with that of GGP-35K microarray. HD genotyping data of total 869 cattle from 14 indicine breeds, 2 crossbred (HF and Jersey crossbred) and 2 exotic breeds (HF, Jersey) were used for this study. Post quality control, only 846 animals and 449955 SNPs remained for imputation study. Only 23.65% of 35339 SNPs in GGP-35K chip are found to be common with INDUSCHIP v2 SNP panel. Imputation was carried out with the help of Beagle 5.0 software using subset of both INDUSCHIP v2 and GGP-35K SNP panels. The study revealed higher average concordance rate (CR) and squared correlation (DR^2) for INDUSCHIP v2 as compared to GGP-35K in crossbred HF population.

Keywords: Genotype Imputation, HF Crossbred cattle, INDUSCHIP, Single Nucleotide Polymorphism

Introduction

Identification of polymorphic variants (SNPs) across the genome, development of high throughput genotyping and sequencing techniques has led to the generation of massive amount of genomic information on large number of individuals. In Livestock, these genomic information is mainly used for breeding purpose, known as Genomic selection (GS), where, superior individuals are selected for breeding at the very young age on the basis of Genomic enhanced Breeding values (GEBV), computed as a linear function of evenly spaced DNA markers (SNP)spread across the genome and their associated genotypes (Meuwissen et al. 2016). Genomic information from dense SNP chips provides an opportunity to increase rate of genetic progress in the breeding programs if a sufficient number of markers and animals with phenotypes are genotyped. More number of markers means greater linkage disequilibrium between SNPs and more chances of capturing genomic variation. However, several studies indicated that increase in SNP density, after a certain threshold, does not seem to improve the quality of realized genomic relationship in any significant way (Su et al. 2012, Chang et al. 2019).

Since, genotyping with HD SNP panels are expensive, it limits the number of animals to be genotyped. Hence, in practice, people preferred cost effective alternative called genotype imputation, which allows inference of the missing marker genotypes from individuals genotyped with low or medium density (LD) panels by using information from reference population genotyped with high-density panels (Carvalheiro et al. 2014).This not only makes it possible to increase the genomic information and predict missing genotypes (Marchini and Howie,2010) but to reduce genotyping costs and intensify genomic selection (Ventura et al. 2014) by genotyping more number of animals and combine data from different breeds (Larmer et al. 2014).

To implement genomic selection in India for indicus breeds and their taurine crosses, a medium-density customized chip i.e., INDUSCHIP v1 consisting of 45700 SNPs sampled from HD genotype of the mostly four indicus breeds (Gir, Sahiwal, Kankrej and Red Sindhi) and their taurine crosses (HF cross & Jersey cross) have been developed (Mrode et al. 2019). The genotyping

[1]National Dairy Development Board (NDDB), Anand, Gujarat-388 001, India

[2]Centre for DNA Fingerprinting and Diagnostics (CDFD), Uppal, Secunderabad, Telangana-500 039, India

Sujit Saha (✉)
NDDB, Anand, Gujarat-388001, India
Email: ssaha@nddb.coop; sujitssahaabc@gmail.com

chip contained around 41000 SNPs from HD data having high MAF (0.25), uniformly distributed across the genome for all the breeds under study with an average distance between two consecutive SNPs around 65 kbps. In addition to the above, 2000 ancestry informative SNPs for above mentioned six breeds, ISAG recommended parentage SNPs and some known open-source genetic markers were also included (Nayee et al. 2017). Subsequently, INDUSCHIP v1 was upgraded to INDUSCHIP v2 (52363 SNPs) incorporating additional 6663 highly polymorphic SNPs (Saha et al. 2020).

Current study was undertaken to assess the genotype imputation efficiency of INDUSCHIP v2 micro array to HD level in Holstein Friesian crossbred (HFCB) population and compare its performance with other commercially available medium density chip i.e., GGP indicus-35K microarray developed by Neogen Geneseek operation on Illumina platform specially designed for indicine cattle.

Materials and Methods

Source of data

Total 869 number of Cattle belong to 14 different Indicine breeds (Amritmahal, Deoni, Gir, Hariana, Hallikar, Kankrej, Khillar, Kangayam, Ongole, Red Sindhi, Rathi, Sahiwal, Siri and Tharparkar) and 2 crossbred (HF crossbred-HFCB and Jersey crossbred-JCB) breeds were genotyped with 777K Bovine HD BeadChip (Illumina, Inc.,San Diego, CA). The genotype data for 2 taurine breeds, Holstein Friesian (HF) and Jersey, were obtained from Aarhus University, Denmark. The genotype candidates were selected mainly from frozen semen stations in India and certain state run livestock farms maintaining purebred animals of those breeds.

Data editing

Quality control checks were applied to raw data. SNPs located on autosomes, with call rate >95% and genotyping rate>90% were kept. Further, SNPs with a minor allele frequency (MAF) less than 0.01 and Hardy Weinberg equilibrium having p value less than 10^{-4} were excluded.

After quality control, out of a total of 869 animals of 14 different breeds (multi-breed) and 777962 SNPs, only 846 animals and 449955 SNPs remained for imputation study.

Retrieval of INDUSCHIP and GGP indicus-35K SNP panels:

50K SNP panel data (52363 SNP) of INDUSCHIP was retrieved from customized INDUSCHIP v2 manifest file (NDDB_ Induschip2_ 15061153X355693_B1.bpm). Around 2949 SNPs, which were present in INDUSCHIP v2 manifest file but was not found to be matching with HD SNPs, thus were excluded from this study. After quality control, finally 49399 SNPs remained, whose HD genotyping data was extracted as a subset to study the imputation efficiency of INDUSCHIP. Similarly, The SNP panel list of GGP indicus-35K medium density chip was obtained from NAGRP community data repository.

Creation of test, reference and validation data sets

From this data, randomly 11 HFCB animals were selected at a time to form test groups animals. While remaining animals 835 animals of multiple breed were taken as reference group with HD data obtained after quality control. Five such test groups were created. Subsequently, genotyping information for the INDUSCHIP and GGP indicus 35K SNP panel were retrieved as a subset from HD data for all the five test groups of animals.

Further, in order to study the concordance of imputation for missing genotypes, five validation data sets with HD genotype data for each group of test animals were also created.

A schematic diagram of the experimental design of this imputation study is presented in Figure 1.

Imputation using INDUSCHIP and GGP indicus-35K SNP panels

Imputation was carried out for 5 test groups of animals using genotyping information of INDUSCHIP v2 SNP panel and GGP indicus-35K SNP panel, respectively. During the study, instead of taking all the 29 autosomes, imputation was carried out for 5 selected autosomes (i.e. Chromosome no.1, 5, 15, 20 and 25) to compare the imputation efficiency.

PLINK (Purcell et al. 2007) software was used for quality control of the data, creation of test, reference and validation data sets as well as for preparing inputs file for Beagle. Imputation was carried out using Beagle 5.0 software (Browning et al. 2018), a population-based imputation program (does not rely on pedigree information)that adopts a stochastic procedure based on a Hidden Markov Monte-Carlo process to infer the probabilities of each haplotype/genotype (Carvalheiro et al. 2014).Imputation accuracy was assessed in terms of concordance rate i.e. the proportion of alleles or genotypes that are correctly imputed (Weigel et al. 2010) and squared correlation between the estimated allele dose and the true allele dose i.e. dosage r^2 (DR^2).The animal wise concordance rate between imputed and actual genotype was estimated using R statistical software and DR^2 values between markers are obtained from Beagle software output.

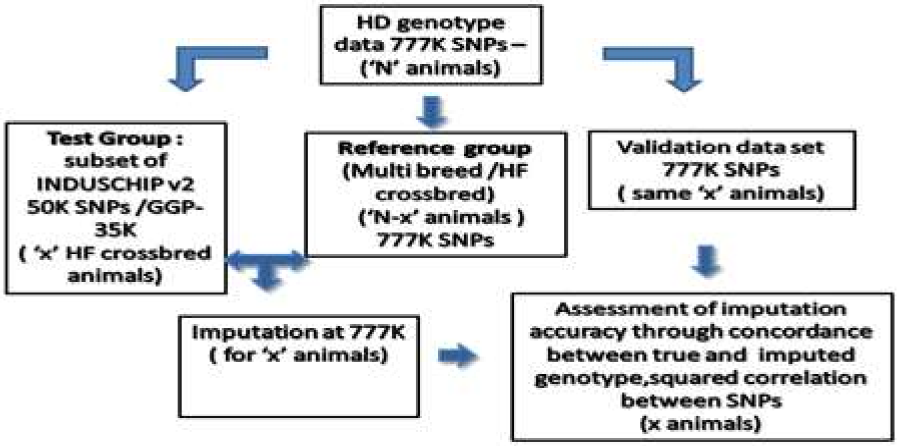
Results and Discussion

Characterization of INDUSCHIP v2 SNP chip

N*umber and Distribution of SNPs across autosomes:*

For an SNP array to be efficient in genotyping for a particular population, it is important to ensure that the selected SNPs are

Fig. 1 Schematic Diagram of the experimental design for imputation study


HD genotype data 777K SNPs – ('N' animals)
Test Group : subset of INDUSCHIP v2 50K SNPs /GGP-35K ('x' HF crossbred animals)
Reference group (Multi breed /HF crossbred) ('N-x' animals) 777K SNPs
Validation data set 777K SNPs (same 'x' animals)
Imputation at 777K (for 'x' animals)
Assessment of imputation accuracy through concordance between true and imputed genotype,squared correlation between SNPs (x animals)

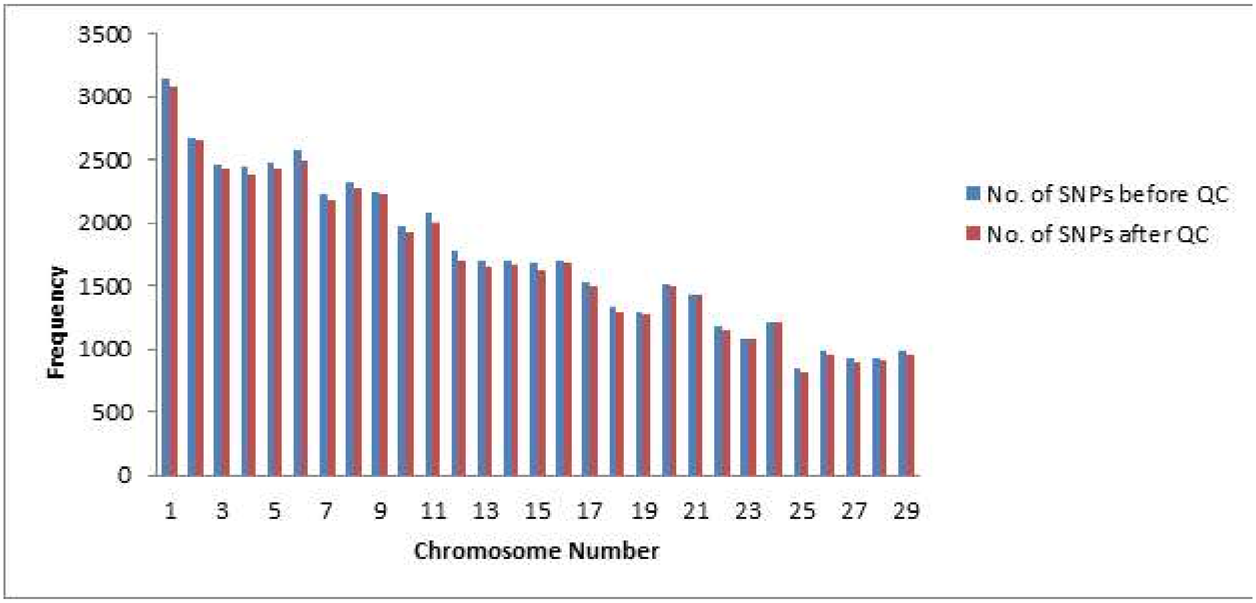


Fig. 2 Chromosome-wise distribution SNPs in INDUSCHIP v2 before and after quality control

distributed evenly covering the entire genome. INDUSCHIP was designed by selecting a subset of SNPs from Illumina BovineHD genotyping array. INDUSCHIP v2 manifest file revealed that there were altogether 52363 SNPs located in all chromosomes. Out of which only 50436 SNPs are located in 29 autosomes (96.3%). Distribution of SNPs across the autosomes in INDUSCHIP v2 vis-à-vis Illumina Bovine HD chip is presented in Table No.1. The data revealed that on an average 6.8% of the HD SNPs located per autosomes were selected in customized INDUSCHIP v2 microarray.

The average distance between the SNPs was found to be around 49.7 Kb across the autosomes. The maximum distance between SNPs was found in chromosome number 10 (52.52 Kb), while minimum distance (46.79 Kb) was observed in chromosome number 9.

Post quality control (QC), out of a total of 50436 SNPs located in autosomes, only 49399 SNPs remained for imputation study. The autosome wise distribution of SNPs before and after quality control (QC) is presented in Figure 2.

Minor allele Frequency

Autosome-wise distribution of minor allele frequencies (MAF) in HFCB population was estimated using PLINK and presented in Table no.2. MAF was classified into three different categories viz. Rare SNPs (MAF > 0 – <0.05), Intermediate SNPs (MAF >= 0.05 – 0.25), and Highly polymorphic SNPs (MAF > 0.25).The distribution of SNPs based on MAF in HFCB population for INDUSCHIP v2 SNP panel indicated that the majority of SNPs (around 73.27%) existing in INDUSCHIP v2 SNP panels are polymorphic having MAF >0.25 (Figure 3).

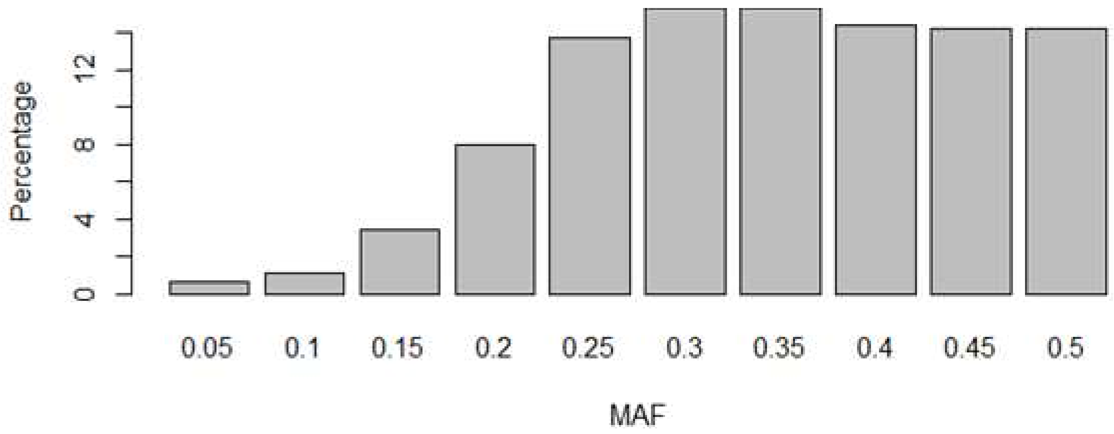


Fig. 3 MAF-wise distribution of SNPs (%) in INDUSCHIP v2

Table 1 Chromosome-wise distribution of SNPs in Illumina Bovine HD chip and INDUSCHIP v2 microarray

Chromosome No.	No. of SNPs in Bovine HD chip	No. of SNPs in INDUSCHIP v2	% SNP in INDUSCHIP v2 compared to Bovine HD chip	Average distance (in KB) between SNPs in INDUSCHIP v2
1	46495	3155	6.8	50.14
2	40056	2677	6.7	51.00
3	35579	2468	6.9	49.15
4	34980	2442	7.0	49.26
5	34842	2483	7.1	48.70
6	35519	2572	7.2	47.11
7	33168	2227	6.7	50.46
8	33529	2320	6.9	48.70
9	31060	2250	7.2	46.79
10	30449	1975	6.5	52.52
11	32015	2078	6.5	51.50
12	26127	1782	6.8	51.00
13	23594	1700	7.2	49.32
14	24780	1697	6.8	49.00
15	24755	1680	6.8	50.53
16	24178	1695	7.0	47.92
17	22266	1522	6.8	49.28
18	19386	1342	6.9	49.05
19	18908	1284	6.8	49.72
20	21490	1508	7.0	47.39
21	21175	1440	6.8	49.67
22	18034	1178	6.5	51.76
23	15215	1091	7.2	47.66
24	18620	1217	6.5	50.95
25	12931	838	6.5	50.92
26	15242	988	6.5	52.21
27	13152	922	7.0	49.18
28	13038	921	7.1	50.13
29	14710	984	6.7	51.31

Comparison of the efficiency of INDUSCHIP v2 and GGP indicus-35K microarray in imputing missing SNPs in HF crossbred cattle

Investigation on SNP markers available in GGP indicus-35K chip, respectively, revealed that out of total 35339 SNPs present in GGP indicus-35K chip, only 8361 SNPs are found (23.65 %)to be

Table 2 Chromosome-wise distribution of MAF in INDUSCHIP v2 microarray

Chromosome No.	Categories of Minor Allele Frequency (MAF)			Grand Total
	>0-0.05	>0.05-0.25	>-0.25	
1	14	814	2259	3087
2	34	768	1849	2651
3	21	588	1819	2428
4	14	586	1791	2391
5	21	651	1755	2427
6	13	660	1817	2490
7	14	586	1574	2174
8	20	856	1396	2272
9	5	603	1621	2229
10	5	453	1470	1928
11	5	468	1539	2012
12	10	455	1245	1710
13	9	595	1055	1659
14	18	499	1150	1667
15	9	401	1213	1623
16	8	471	1202	1681
17	12	351	1130	1493
18	6	302	989	1297
19	13	274	988	1275
20	13	375	1104	1492
21	13	394	1026	1433
22	7	236	910	1153
23	2	224	856	1082
24	3	326	877	1206
25	1	165	646	812
26	1	209	754	964
27	3	187	709	899
28	2	190	711	903
29	2	219	740	961
Total	298	12906	36195	49399
%	0.60	26.13	73.27	

common with INDUSCHIP v2 SNP panel. In GGP indicus-35K chip, around 81% of SNPs were found to be polymorphic with MAF > 0.25.

Imputation was carried out using genotype information at INDUSCHIP v2 SNP panel and GGP indicus-35K SNP panel for 5 chromosomes (i.e. Chromosome no. 1, 10, 15, 20 and 25, respectively) for all the five test group of animals.

The Concordance rate obtained from this study found to vary between 0.971 (Chromosome no.10) to 0.980 (Chromosome No.15) while imputing INDUSCHIP v2 SNP panel to HD level, while the same was varying from 0.961 (Chromosome no.10) to 0.974 (Chromosome No.15) for GGP indicus-35K (Table 3).

Caravalheiro et al. (2014), while imputing GGP20Ki and GGP75Ki panel to HD panel in Nellore animals, observed concordance rate of 97 and 99%, respectively.

The average DR^2 found to vary between 0.892-0.922 in INDUSCHIP v2, while it was 0.888-0.913 in GGP indicus-35K (Table 4).

The present study revealed that selected SNPs in customized INDUSCHIP v2, which was specifically designed for genotyping of indicine breeds and their crosses, were distributed uniformly covering the entire genome. Distribution of SNPs in INDUSCHIP v2 is found to be similar to the distribution of SNPs in other Bovine SNP chips like Illumina 50K and GeneSeek 75K (Mutukumalli et al. 2009).

The majority of the SNPs with high MAF (>0.25) across the autosomes, indicated existence of considerable heterozygosity in crossbred population and INDUSCHIP v2 appeared to be effective in capturing variability in the crossbred population. Malik et al. 2018 in his study using high throughput genotyping-by-sequencing (GBS) markers found that the MAF within the

Table 3 Average concordance rate of INDUSCHIP v2 and GGP indicus-35K in HFCB cattle

Group	INDUSCHIP v2					GGP indicus-35K				
	Chr1	Chr5	Chr10	Chr15	Chr25	Chr1	Chr5	Chr10	Chr15	Chr25
Test -1	0.986	0.984	0.970	0.987	0.978	0.983	0.977	0.958	0.984	0.969
Test -2	0.975	0.984	0.977	0.986	0.982	0.968	0.977	0.970	0.982	0.974
Test -3	0.982	0.980	0.968	0.975	0.980	0.977	0.972	0.958	0.966	0.976
Test -4	0.968	0.968	0.970	0.972	0.965	0.953	0.957	0.960	0.970	0.956
Test -5	0.971	0.968	0.971	0.975	0.969	0.963	0.956	0.960	0.968	0.956
Average	0.977	0.977	0.971	0.98	0.975	0.969	0.968	0.961	0.974	0.966

Table 4 Average DR^2 of INDUSCHIP v2 and GGP indicus-35K in HFCB cattle

Group	INDUSCHIP v2					GGP indicus-35K				
	Chr1	Chr5	Chr10	Chr15	Chr25	Chr1	Chr5	Chr10	Chr15	Chr25
Test -1	0.920	0.926	0.893	0.933	0.890	0.916	0.919	0.887	0.931	0.877
Test -2	0.900	0.923	0.918	0.930	0.908	0.897	0.916	0.915	0.922	0.898
Test -3	0.924	0.921	0.907	0.910	0.909	0.918	0.913	0.897	0.905	0.901
Test -4	0.892	0.906	0.902	0.918	0.877	0.883	0.886	0.897	0.895	0.896
Test -5	0.892	0.906	0.902	0.918	0.877	0.881	0.890	0.891	0.914	0.868
Average	0.906	0.916	0.904	0.922	0.892	0.899	0.905	0.897	0.913	0.888

Indian cattle varied from 0.103 (in Ongole cattle) to 0.177 (in Siri cattle), whereas the Holstein cattle had the lowest value of 0.089. Chagunda et al. 2018 reported average minor allele frequency of 0.29, 0.23, 0.18 and 0.13 for Holstein, Jersey, N'Dama and Gir cattle, respectively.

Comparing imputation efficiency between INDUSCHIP v2 and GGP indicus-35K expressed in terms of average concordance rate as well as squared correlation estimate (DR^2) between Imputed and actual genotypes revealed marginally better performance of INDUSCHIP v2 over GGP indicus-35K chip in Indian HF crossbred population. It may be attributed due to the fact that design of INDUSCHIP v2 chip was based on Indigenous breeds and its crosses (Nayee et al. 2017), while the SNP panels for GGP indicus-35K chip were selected from Australian Brahman, Droughtmaster, Guzerath, Gyr, Nellore, Santa Gertrudis, and tropical composite (Ferraz et al. 2018).

Conclusions

From the present study it can be concluded that the current version of customized INDUSCHIP micro array i.e. INDUSCHIP v2 was quite efficient in imputation at HD level, hence can be effectively used for genotyping and subsequent analysis. However, with the passage of time, as more and more number animals of different breeds spread across the country are genotyped and incorporated in reference population, it would be possible to improve its imputation efficiency further through expanding reference population and incorporating more informative SNPs for the Indian cattle population in future versions of INDUSCHIP micro array. Further, it may also lead to development of low density (LD) microarray with around 10000 informative SNPs and make genotyping facility available to the common dairy farmers at affordable cost.

Acknowledgements

Under Indo-Danish collaboration, technical guidance provided by the Scientists of Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Research Center Foulum, Aarhus University, Denmark, in selecting INDUSCHIP panel is duly acknowledged. The authors humbly thank NDDB management for funding and necessary support for this project. The cooperation extended by the Field staff and officials from various PT, PS projects, Livestock Farms for sample collection is duly acknowledged.

References

Browning BL, Zhou Y, Browning SR (2018) A one penny imputed genome from next generation reference panels. American J Hum Genet 103: 338-348. doi: 10.1016/j.ajhg.2018.07.015.

Carvalheiro R, Boison SA, Neves HHR, Sargolzaei M, Schenkel FS, Utsunomiya YT, O'Brien AMP, SolknerJ, McEwan JC, Van Tassell CP, Sonstegard TS, Garica, JF (2014) Accuracy of genotype imputation in Nelore Cattle. Genet Sel Evol.46: 69. doi:10.1186/s12711-014-0069-1.

Chang LY, Toghiani S, Aggrey SE and Rekaya R (2019) Increasing accuracy of genomic selection in presence of high density marker panels through prioritization of relevant polymorphisms. BMC Genet.20: 21. doi.org/10.1186/s12863-019-0720-5.

Chagunda MGG, Mujibi FDN, Dusingizimana T,Kamana O, Cheruiyot E, Mwai OA (2018) Use of high density single nucleotide polymorphism (SNP) arrays to assess genetic diversity and population structure of dairy cattle in smallholder dairy systems: The Case of Girinka programme in Rwanda. Front Genet.9: 438. doi:10.3389/fgene.2018.00438

Ferraz JBS, Wu X, Li H, Xu J, FerrettiR, Simpson B, Walker J, Silva LR, Garica JF, TaitRG Jr, Bauck S (2018) Design of a low density SNP chip for Bos indicus:GGP indicus technical characterization and imputation accuracy of higher density SNP genotypes. In: Proceedings of the 11th World Congress of Genetics Applied to Livestock Production. Auckland, New Zealand,6-11 February,2018.

Larmer SG, Sargolzaei M, Schenkel FS (2014) Extent of linkage disequilibrium, consistency of gametic phase, and imputation accuracy within and across Canadian dairy breeds. J Dairy Sci.97: 1-14. doi:10.3168/jds.2013-6826.

Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet.11: 499-511. doi:10.1038/nrg 2796.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MJ,O'Connell J, Moore SS, Smith TPL, Sonstegard TS, Van Tassell CP (2009) Development and Characterization of a High Density SNP Genotyping Assay for Cattle. PloS One 4: e5350-5063

Meuwissen T, Hayes BJ, Goddard ME (2016) Genomic Selection: A paradigm shift in animal breeding. Anim. Front. 6(1): 6-14. https://doi.org/10.2527/af.2016-0002

Mrode R, Ojango JMK, Okeyo AM, Mwacharo M (2019) Genomic Selection and use of Molecular tools in Breeding Programs for Indigenous and Crossbred Cattle in Developing Countries: Current Status and Future Prospects. Front Genet. 9: 694. doi:10.3389/fgene.2018.00694.

Nayee N, Saha S, Gajjar S, Sudhakar A, Trivedi KR (2017) Compendium of International Workshop on Genomic Selection for Genetic Improvement in Indian Dairy Animals. November 28-29, 2017, BAIF, Pune, India: 15-16

Purcell S, Neale B, Todd-Brown K, ThomasL, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole –genome association and population based linkage analysis. American J of Hum Genet.81: 559-75.doi: 10.1086/519795.

Saha S, Nayee N, Shah H, Gajjar S, Kishore G, Gupta RO, Trivedi KR (2020) Effect of composition and size of the reference population in genotype imputation efficiency of INDUSCHIP in HF Crossbred cattle. Indian J Dairy Sci.73: 250-255

Su G, Brondum BF, Ma P, Guldbrandtsen B, Aamanand GP, Lund MS (2012) Comparison of genomic predictions using medium-density(~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and red dairy cattle populations. J Dairy Sci 95: 4657-4665

Ventura RV, Lu D, Schenkel FS, Wang Z, Li C, Miller SP (2014) Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbred beef cattle. J Anim Sci.92: 1433-1444. doi:10.2527/jas.2013-6638.

Weigel KA, Van Tassell CP, O'Connell JR, VanRaden PM, Wiggans GR (2010) Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. J Dairy Sci.93: 2229-2238. doi: 10.3168/jds.2009-2849