RESEARCH ARTICLE

Effect of composition and size of the reference population in genotype imputation efficiency of INDUSCHIP in HF Crossbred cattle

Sujit Saha, Nilesh Nayee, Heena Shah, Swapnil Gajjar, G Kishore, R O Gupta and K R Trivedi

Received: 09 February 2020 / Accepted: 09 March 2020 / Published online: 12 July 2020
© Indian Dairy Association (India) 2020

Abstract: The objective of this study was to investigate the effect of composition and size of the reference population in imputation efficiency of INDUSCHIP v2 in Indian HF crossbred cattle. Data set consisted of a total of 869 cattle from 14 Indicine breeds, 2 crossbreds (HF and Jersey crossbreds) and 2 exotic breeds (HF, Jersey) genotyped with Illumina BovineHD (Illumina, San Diego, CA) panel. Post QC, 846 animals and 449955 SNPs remained for imputation study. 3 test groups each with randomly selected 25 HFCB animals with subset genotype of INDUSCHIP v2 were created, whereas with HD genotyping data of remaining animals, 3 different categories of reference groups were created namely reference 1 (HF, Jersey, all 14 Indicine breeds, HF and Jersey crossbreds), reference 2 (HF, HF crossbred, Sahiwal, Gir and Kankrej) and reference 3 (pure HF, Sahiwal, Gir and Kankrej). Imputation efficiency of INDUSCHIP v2 was expressed in terms of concordance rate and Dosage R² (DR²). Reference groups 1 and 2 were found to be better than Reference group 3. Further, the size of the reference population had an impact on imputation efficiency. The concordance rate and DR² decreased with decline about population size. However, a reference population with 280 animals was found to be sufficient to obtain a concordance rate of around 95% or more and DR² around 0.93. More number of HF, HF crossbred, Sahiwal, Gir and Kankrej animals need to be HD genotyped and incorporated in the reference population to improve the imputation efficiency of INDUSCHIP v2.

Keywords: Crossbred cattle, Genotype Imputation, HD chip, LD chip, Single Nucleotide Polymorphism, Reference population

Introduction

Under Genomic selection, evenly spaced DNA markers (Single Nucleotide Polymorphism-SNPs) spread across the genome are used to estimate breeding values (GEBV) for the target individuals (Meuwissen et al. 2016). Genomic information from dense SNPs chips provides the opportunity to increase the rate of genetic progress in the breeding programs if a sufficient number of markers and animals with phenotypes are genotyped (Carvalho et al. 2014). More number of markers means greater linkage disequilibrium between SNPs and more chances of capturing genomic variation. Since genotyping with HD SNP panels are expensive, it limits the number of animals to be genotyped. Hence, in practice, a cost-effective alternative called genotype imputation is preferred. Genotype imputation makes it possible to extrapolate genotypes from lower to higher density arrays based on a representative sample of individuals genotyped at higher density (Pausch, H, 2013). 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). The imputation efficiency of any chip depends upon several factors namely imputation method, software used for imputation, the MAF of the SNP to be imputed, linkage disequilibrium between SNPs, the chromosomal position of the SNP, the quality of SNP maps, size and composition of reference population, etc. (Schrooten 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, Red Sindhi) and their taurine crosses (HF cross & Jersey cross) have been developed. (Mrode et al. 2019). The genotyping 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

National Dairy Development Board, Anand-388 001, Gujarat, India

Sujit Saha (✉)
Animal Breeding, National Dairy Development Board, Anand-388
001, Gujarat, India
Email: ssaha@nddb.coop; sujitsahaabc@gmail.com

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.

Keeping this in mind the current investigation was undertaken to study the effect of the composition of the reference population and its size on the genotype imputation efficiency of INDUSCHIP v2, a custom made medium density genotyping chip designed on Illumina platform to genotype crossbred and indigenous cattle of India.

Materials and Methods

Source of data

A total of 869 number of cattle belong to 14 different Indicene breeds (Amritmahal, Deoni, Gir, Haryana, Hallikar, Kankrej, Khillar, Kangayam, Kankrej, Ongole, RedSindhi, Rath, Sahiwal) and 2 crossbred (HFCB and Jersey crossbred) breeds were genotyped with 777K BovineHD BeadChip (Illumina, Inc., San Diego, CA). The genotype data for 2 Taurine breeds i.e. 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 HD genotype data. Only 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 less than 10^{-4} were excluded.

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

Creation of Test, Reference and Validation data sets

From this data, randomly 25 HFCB animals were selected at a time to form test groups of animals. Only subset genotypes of INDUSCHIP v2 were considered for test animals and the rest of the animals were taken as reference group with HD genotype data. Three such test groups were created. For each test group, 3 different categories of reference groups were created namely Reference 1 (821 animals comprising of HF, Jersey, all 14 Indigenous breeds, HF and Jersey crossbred animals), Reference 2 (404 animals comprising of only HF, HF crossbred, Sahiwal, Gir and Kankrej breed) and Reference 3 (266 animals with only pure HF, Sahiwal, Gir and Kankrej cattle). Imputation accuracy was measured as the concordance of actual HD genotype with imputed

genotypes and squared correlation (Dosage R²) between estimated allele dose and true allele doses for test animals. A schematic diagram of the experimental design of this imputation study is presented in Figure.1.

Imputation using INDUSCHIP v2 SNP panels

Thereafter, 50K SNP panel data (52363 SNP) 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 not found to be matching with HD SNPs, 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.

Imputation was carried out for all the 3 test groups of animals using genotyping information of INDUSCHIP v2 SNP panel considering all the 29 autosomes against 3 different reference populations.

Subsequently, to investigate the impact of the size of the reference population on imputation efficiency, around 70% (121), 50% (202), 40% (242) and 30% (283) animals were retained randomly in the reference population Group 2. Thereafter imputation was carried out for all the 3 test groups of animals.

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 input files for Beagle. Imputation was carried out using Beagle 5.0 software (Browning et al. 2018), which is 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 (Carvalho 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 expressed as Dosage R² (DR²) in Beagle 5.0 (Browning et al. 2018). The animal wise concordance rate between imputed and actual genotype was estimated using R statistical software and squared correlation values between markers were obtained after phasing and imputation using Beagle software.

Results and Discussion

Efficiency of INDUSCHIP v2 in imputing missing SNPs in HF crossbred cattle using a different reference population

Average concordance rates and DR² obtained for 3 test groups of animals using 3 different reference populations are presented in Figure 2. Results obtained from this study revealed (Table 1) that the highest average concordance rate was obtained for

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

Table 1 Average Concordance Rate using three different reference populations

Test Group	Reference group		
	Reference-1(821)	Reference-2(404)	Reference-3 (266)
Test -1	0.978	0.973	0.898
Test -2	0.970	0.963	0.887
Test -3	0.972	0.967	0.901
Average	0.974	0.968	0.895

Table 2 DR2 using three different reference populations

Test Group	Reference group		
	Reference-1(821)	Reference-2(404)	Reference-3 (266)
Test -1	0.952	0.946	0.861
Test -2	0.948	0.940	0.856
Test -3	0.946	0.941	0.865

reference group-1 (0.974) followed by reference group-2 (0.968), while the lowest concordance rate was found for reference group-3 (0.895). DR2 estimates as presented in Table 2 found to vary between 0.946-0.952 for reference group-1, 0.940-0.946 for reference group-2 and 0.856-0.865 reference group-3.

Efficiency of INDUSCHIP v2 in imputing missing SNPs in HF crossbred cattle for various size of the reference population

To assess the impact of varying size of reference population on imputation efficiency, average concordance rates and DR2 were estimated using reference group 2 for all three test groups of animals keeping only 70%, 50%, 40% and 30% of the animals, respectively in the said reference groups. The estimates are presented in Table 3 and Table 4, respectively. The results indicate an increasing trend in the average concordance rate as well as DR2 as the size of the reference population increases (Figure 3).

The average concordance rates observed were 0.959, 0.946, 0.939 and 0.922 respectively for 70%, 50%, 40% and 30% of animals retained randomly in the reference groups. The DR2 values ranged between 0.931 to 0.934, 0.913 to 0.922, 0.905 to 0.911 and 0.884 to 0.887, respectively, for 70%, 50%, 40% and 30% of the animals retained in reference group.

The present study revealed that the composition of the reference population plays an important role in determining imputation efficiency at the HD level. For HF crossbred cattle, the reference population-1 comprising all the 14 Indigenous breeds, HF, Jersey, HF crossbred and Jersey crossbred resulted in higher concordance rate and DR2, while reference group 2, consisting of HF, HF crossbred and 3 indigenous breeds like Sahiwal, Gir and Kankrej cattle, despite of having nearly half of the HD genotyped animals than reference group-1 (404 against 821) resulted into very minor loss of imputation accuracy with

Fig. 2 Average Concordance rate (%) and DR2 (%) of HF crossbred test group of animals for different composition of the reference population

Fig. 3 Average concordance rate (%) and DR2(%) for different size of reference populations in HF crossbred cattle

Table 3 Average Concordance Rate for different size of the reference population

Test Group	Reference group size				
	With all available animals	% of animals retained			
	(404)	70%(283)	50%(202)	40%(162)	30%(121)
Test -1	0.973	0.963	0.953	0.944	0.925
Test -2	0.963	0.956	0.943	0.936	0.920
Test -3	0.967	0.959	0.943	0.938	0.929
Average	0.968	0.959	0.946	0.939	0.922

Figures in the parenthesis indicates no. of animals

concordance rate around 95% between imputed and actual genotype and average DR2 around 0.94. This may be due to the fact that as the HF crossbred cattle were mainly produced by interbreeding of exotic HF with indigenous Sahiwal, Gir and Kankrej cattle, the existing HF crossbreds are expected to be

more closely related to HF, Sahiwal, Gir and Kankrej breed and their crosses than other indigenous draft breeds, Jersey and their crosses. As a result, almost similar imputation efficiency was observed despite having almost half the size of the population in reference Group 2. Berry and Kearney (2011), Ma et al. (2013), Moghaddar et al. (2015), Bolormaa et al. (2015) and Ventura et al.

Table 4 DR2 for different size of the reference population

Test Group	Reference group size				
	With all available animals	% of animals retained			
	(404)	70%(283)	50%(202)	40%(162)	30%(121)
Test -1	0.952	0.934	0.922	0.911	0.887
Test -2	0.948	0.932	0.917	0.908	0.887
Test -3	0.946	0.931	0.913	0.905	0.884

(2016) in their studies also reported that genetically closer animals in the reference and imputation population produce higher imputation accuracies.

On the other hand, the present study also revealed that retaining only purebred animals i.e. HF, Sahiwal, Gir and Kankrej breed (excluding of HF crossbred) in reference group-3 resulted in relatively poor imputation efficiency. This indicates the possibility of the presence of crossbred specific haplotypes. The results are in agreement with the findings of Oliviera Junior et al. (2017), where the inclusion of crossbred Girolando in the reference population had a greater effect on the imputation accuracy than the purebred Gyr haplotypes.

The findings indicated that to improve the imputation efficiency of INDUSCHIP v2 by strengthening reference population, more number of HF, HF crossbred, Sahiwal, Gir and Kankrej cattle etc. from varied sources need to be HD genotyped and included in the reference population.

Subsequently, when imputation was carried out for all the three test groups of animals using the same reference (reference group 2) of varying sizes, the average concordance rates and DR2 were found to decrease as the size of the reference population decreases, which agrees with the findings reported by Schrooten, D.T. and De Roos, A.P.W.(2010) and Pausch, H. et al. (2013). However, the present study indicated a reference population size of around 280 was sufficient to obtain concordance rate around 95% or more and DR2 around 0.93. Ghoreishifar S.M et al. (2018) in Italian Mediterranean buffaloes also found that increasing the reference population size from small to intermediate (i.e., from 42 to 202) resulted in a greater improvement in imputation accuracy compared to increasing the Reference Population size from intermediate to large (i.e., from 202 to 736).

Conclusions

Imputation in HF crossbred population in India using custom made microarray INDUSCHIP v2 was found to be affected by both the composition of the reference population and its size. For further improvement in the efficiency of imputation by INDUSCHIP v2, more number of HF, HF crossbred, Sahiwal, Gir and Kankrej animals may need to be HD genotyped and incorporated in the reference population.

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 supporting liberally for this project. The cooperation extended by the Field staff and officials from various PT, PS projects, Livestock Farms for sample collection and services extended by the staff of Genomics lab, NDDB is duly acknowledged.

References

- Berry DP, Kearney JF (2011) Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. *Animal* 5: 1162-1169. doi:10.1017/S1751731111000309.
- Bolormaa S, Gore K, Van Der Werf JHJ, Hayes BJ and Daetwyler HD (2015) Design of a low-density SNP chip for main Australian sheep and its effect on imputation and genomic prediction accuracy. *Anim Genet* 46: 544-556. doi.org/10.1111/age.12340
- 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.
- Carvalho R, Boison SA, Neves HHR, Sargolzaei M, Schenkel FS, Utsunomiya YT, O'Brien AMP, Solkner J, 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.
- Ghoreishifar SM, Moradi-Shahrabak H, Moradi-Shahrabak M, Nicolazzi EL, Williams JL, lamartino D, Nejati-Javaremi A (2018) Accuracy of imputation of single-nucleotide polymorphism marker genotypes for water buffaloes (*Bubalus bubalis*) using different reference population sizes and imputation tools. *Livest Sci* 216: 174-182. doi.org/10.1016/j.livsci.2018.08.009.
- 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.
- Ma P, Brøndum RF, Zhang Q., Lund MS, Su G (2013) Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. *J Dairy Sci* 96: 4666-4677. doi:10.3168/jds.2012-6316.

- Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11: 499-511. doi:10.1038/nrg2796.
- Meuwissen T, Hayes BJ, Goddard ME (2016) Genomic Selection: A paradigm shift in animal breeding. *Anim Front* 6: 6-14. <https://doi.org/10.2527/af.2016-0002>
- Moghaddar N, Gore KP, Daetwyler HD, Hayes BJ, van der Werf JHJ (2015) Accuracy of genotype imputation based on random and selected references sets in purebred and crossbred sheep populations and its effect on accuracy of genomic prediction. *Genet Sel Evol* 47: 97. doi.org/10.1186/s12711-015-0175-8.
- 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
- Oliveira Junior GA, Tatiane CS, Chud CS, Ventura RV, Garrick DJ, Cole JB, Munari DP, Ferraz JBS, Mullart E, DeNise S, Smith S, da Silva MVGB (2017) Genotype imputation in a tropical crossbred dairy cattle population. *J Dairy Sci*. 100: 9623-9634. doi.org/10.3168/jds.2017-12732.
- Pausch H, Aigner B, Emmerling R, Edel C, Gotz KU, Fries R (2013) Imputation of high-density genotypes in the Fleckvieh cattle population. *Genet Sel Evol* 45: 3. doi:10.1186/1297-9686-45-3.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker P I, Daly MJ, Sham PC (2007) PLINK: a tool set for whole –genome association and population based linkage analysis. *American J Hum Genet* 81: 559-75. doi: 10.1086/519795
- Schrooten C, Dassonnneville R, Ducrocq V, Brøndum RF, Lund MS, Chen J, Liu Z, Gonzalez-Recio O, Pena J, Druet T (2014) Error rate for imputation from Illumina BovineSNP50 chip to Illumina BovineHD chip. *Genet Sel Evol* 46: 10
- Schrooten DT, De Roos APW (2010) Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. *J Dairy Sci* 93: 5443-5454
- 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: 1433-44. doi:10.2527/jas.2013-6638.
- Ventura RV, Miller SP, Dodds KG, Auvray B, Lee M, Bixley M, Clarke SM, McEwan JC (2016) Assessing accuracy of imputation using different SNP panel densities in a multibreed sheep population. *Genet Sel Evol*. 48: 71. doi.org/10.1186/s12711-016-0244-7.
- 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