Breed-of-origin specific genomic relationship matrix improves genomic prediction accuracy in crossbred Holstein Friesian cattle in India

S.G. Gajjar[1], B. Guldbrandtsen[2], G. Su[2], N.G. Nayee[1], G. Sahana[2], K.R. Trivedi[1] & M.S. Lund[2]

[1]National Dairy Development Board, Anand-388001, Gujarat, India
sggajjar@nddb.coop (Corresponding Author)
[2]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, 8830 Tjele, Denmark

Introduction

Populations arising from crossbreeding of distantly related breeds pose special challenges for genomic prediction. Markers may have quite different effects in the two parent breeds. India's National Dairy Development Board (NDDB) implements progeny testing programs in population of dairy cattle arising from cross of Holstein cattle with *B. indicus* cattle (CBHF). Here the size of the current reference population of bulls is limited for genomic selection. One way to aid genomic selection, would be by genomic prediction based on breed-of-origin of alleles as determined by chromosome painting. The current study was thus undertaken to explore the efficacy of using chromosome painting (Lawson *et al.*, 2012) for CBHF cattle to estimate breed-of-origin of alleles to construct breed-of-origin specific genomic relationship matrices for genomic prediction for milk yield using single-step genomic BLUP (ssGBLUP) in CBHF cattle.

Material and methods

Phenotypes were first lactation test-day milk yield records of 8749 CBHF daughters sired by 229 bulls under Sabarmati Ashram Gaushala (SAG) CBHF progeny testing program. Data were collected from the Information Network for Animal Productivity and Health (INAPH), managed by NDDB. Genotype data was available for 2,194 daughters and 109 sires, which were genotyped using a custom SNP panel (about 47,000 SNPs including Illumina's BovineLD Genotyping BeadChip, Illumina, San Diego, CA, USA) developed by the NDDB on Illumina platform for genotyping of *Bos indicus* and their taurine crosses. Genotype data comprised of 36,363 autosomal SNPs passing quality control criteria *viz*. average call rates above 95% per individual, 90% per locus, and MAF higher than 0.01. Pedigree records were verified by molecular data for any errors and were subsequently addressed by preparing and analysing a separate data set.

For estimation of allele-specific breed-of-origin by chromosome painting, genotype data was split by chromosome. Data were phased using BEAGLE 4.1 (Browning and Browning, 2007; Browning and Browning, 2016) using 10 burn-in and 5 phasing iterations with 100 marker sliding windows with an overlap of 48 markers and an effective population size of 2,000. Allele-specific breed-of-origin was inferred using ChromoPainterV2 (Lawson *et al*., 2012) separately for each chromosome. 47 Sahiwal, 96 Gir, 30 Kankrej and 20 HF animals were used as donor populations while the recipient populations contained the crossbred samples with 5 expectation maximization iterations; maximizing over copying proportions was used for painting, with an effective population size of 2,000. Genetic distances in Morgan were approximated by assuming that 1 Mb corresponded to 1 cM.

The output from chromosome painting was edited in R (R Core Team, 2017) for input data to calculate G-matrix. Though breed-of-origin probabilities were available for all donor populations, Kankrej, Sahiwal and Gir probabilities were combined as indicine origin for further analyses. Each allele at a locus was considered for two possible origins *viz*. *Bos taurus* and *Bos indicus;* and hence, an individual had four probabilities corresponding to the four breed-of-origin alleles at each locus (2 allele by two breed). Thus an individual had a value ranged between 0 to 2 for each breed-of-origin allele, and the sum of four breed-of-origin alleles was 2. In construction, each breed-of-origin allele was treated like a locus and alleles were considered identical-by-state if they had the same state and the same origin. In total, 36,363 markers resulted in 145,452 'genotypes' and G-matrix was calculated using this input in "invgmatrix" program.

Breeding values for milk yield were estimated using test-day records by random regression analysis using ssGBLUP. Fixed effects under the model were herd (village), age at calving and year x season of calving interaction. Random effects under the model comprised herd x year of milk recording x month of milk recording interaction and herd x age at calving interaction. Days in milk was considered fixed regression and animal as well as permanent environment effects were considered random regressions. Legendre polynomial of order three was used for modeling fixed and random regressions. Model-I in the following will be referred to approach wherein above model using conventional G-matrix was used, while Model-II will be referred to the same model, but using G-matrix constructed by integrating allele-specific breed origins.

Variance components in both models were estimated separately using Average Information Restricted Maximum Likelihood (AIREML) algorithm in DMUV5.2 (http://dmu.agrsci.dk/DMU/). Heritability estimates for Model-I for data sets with and without corrections for pedigree, as detected by molecular data, were 0.4482 and 0.4569, respectively; while the same for Model-II for data sets with and without above mentioned pedigree corrections were 0.4307 and 0.4382, respectively. Breeding values were estimated using variance components estimated separately in each model. Validation was done by five-fold cross-validation. Sires were ranked on basis of number of daughter records and divided in to 5 sets randomly. Half-sib daughter records of sires in each set were then left out from analysis (while retaining them in pedigree) to prepare data for 5 validation sets (Data set - I). The pedigree links with errors detected by molecular data were deleted in Data set - II which was re-analysed with same validation scheme as described above. Data set - III was further prepared with different validation scheme, in which only sons (with 20 daughters each) of proven sires (also with 20 daughters each) were selected. Total 40 such sons of proven sires were selected and divided in to 5 random sets. Half-sib daughter records of sires in each of the set were left-out to prepare data for 5 validation sets (Data set - III). Thus, scenario in Data set - III assumes that the sons of progeny tested sires do not have daughter records and cross-validation was done to compare the results. Predicted breeding values of daughters in the cross validation test data were correlated with their corrected yield (Yc: breeding value + permanent environment effect + residual) estimated from the full data (no data left out) analysed without any genomic relationship matrix (i.e. only numerator relationship matrix used to model the pedigree). Predicted breeding values of sires whose daughters were in validation set were correlated with average Yc of their daughters estimated from full data (analysed without any genomic relationship matrix). Genomic relationship matrices were weighted by giving 20% weight to the numerator relationship matrix and 80% to the unweighted genomic relationship matrix.

Results and discussions

Table-I, II and III present pooled correlations (for all 5 data sets) between Yc of daughters with their breeding values (comparison criteria for daughters) and correlations between average Yc of daughters of sires and sire breeding values (comparison criteria for sires). The third column was the difference (in percentage) between the correlations of the models, in terms of gain/ loss in Model - II compared to Model - I, for three data sets. Correlations for all daughters in Data sets I and II did not show improvement, but the same was noted in Data set - III. Correlations for non-genotyped daughters improved in Data set - III but reduced in Data sets - I and II. Correlations across genotyped daughters improved consistently (around +5%) with all three data sets.

Correlations between average Yc of daughters of sires and sire breeding values did not improve when all sires were considered in Data sets - I and II and the same decreased for non-genotyped sires. Note that Data set - I & II included all data so sires with fewer daughters were also included. This may be a cause of no gain being observed due to large uncertainty. However, correlations improved for genotyped sires in all three data sets (there were no non-genotyped sires in Data set - III), biggest gain noted in Data set – III (15%). In addition, as shown in Table-4, the model accounting for breed-of-origin of alleles reduces bias for validation, while Table-3 confirms high gains in Model - II for Data set - III. Data set - III translates practical scenarios wherein genomic breeding values for young bulls from proven sires would be evaluated for use in breeding program.

Conclusions

Probability of breed-of-origin of alleles at each marker locus in CBHF cattle can be estimated by chromosome painting. This information can be used to construct breed-of-origin aware genomic relationship matrices. We observed that such G-matrix improved accuracy of predicted genomic breeding values for genotyped animals and reduced bias of predictions compared to conventional genomic breeding value predictions for CBHF cattle, which treats identical-by-state alleles the same, irrespective of breed origin. This new approach is expected to further enhance accuracies of selection of young sires and bull dams and thereby will improve realized genetic gains per unit of time in CBHF population of India.

Acknowledgements

The authors sincerely thank all officers implementing SAG CBHF progeny testing program and have arranged for collection of required samples. Genotyping for the samples was carried out at M/s. Sandor Life Sciences, Hyderabad. Present work was designed and executed as part of collaboration between Department of Molecular Biology and Genetics (MBG), Aarhus University, Denmark and National Dairy Development Board (NDDB), India. We sincerely thank NDDB management for financial support and QGG scientists of Aarhus University for technical support for this work.

List of References

Browning, S.R. & B.L. Browning, 2007. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81:1084-97. doi:10.1086/521987.

Browning, S.R. & B.L. Browning, 2016. Genotype imputation with millions of reference samples. Am. J. Hum. Genet. 98:116-126. doi:10.1086/j.ajhg.2015.11.020.

DMU Version 6, release 5.2. Madsen, P. & J. Jensen. A User's guide to DMU. A Package for Analysing Multivariate Mixed Models. Center for Quantitative Genetics & Genomics, Aarhus University, Research Centre Foulum, 8830 Tjele, Denmark.

Lawson, D., G. Hellenthal, S. Myers, & D. Falush, 2012. Inference of population structure using dense haplotype data. PLoS Genet 8(1):e1002453.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Table 1. Pooled correlations (for all 5 sub-data sets) between corrected yield of daughters and their EBV and correlations between average corrected yield of daughters of sires and sire EBV, in validation Data set - I

Category	Conventional ssGBLUP	Breed-of-origin ssGBLUP	Gain/loss
All daughters	0.2002	0.2022	1%
Non-genotyped daughters	0.1676	0.1629	-3%
Genotyped daughters	0.3022	0.3184	5%
All sires	0.2257	0.2289	1%
Non-genotyped sires	0.1529	0.1510	-1%
Genotyped sires	0.2758	0.2899	5%

Number of genotyped daughters, non-genotyped daughters and total daughters were 1848, 6901 and 8749, respectively.

Table 2. Pooled correlations (for all 5 sub-data sets) between corrected yield of daughters and their EBV and correlations between average corrected yield of daughters of sires and sire EBV, in validation Data set - II

Category	Conventional ssGBLUP	Breed-of-origin ssGBLUP	Gain/loss
All daughters	0.2002	0.2011	0%
Non-genotyped daughters	0.1693	0.1635	-3%
Genotyped daughters	0.3178	0.3332	5%
All sires	0.2331	0.2309	-1%
Non-genotyped sires	0.0940	0.0863	-8%
Genotyped sires	0.2941	0.3008	2%

Number of genotyped daughters, non-genotyped daughters and total daughters were 1473, 6901 and 8374, respectively.

Table 3. Pooled correlations (for all 5 sub-data sets) between corrected yield of daughters and their EBV and correlations between average corrected yield of daughters of sires and sire EBV, in validation Data set - III

Category	Conventional ssGBLUP	Breed-of-origin ssGBLUP	Gain/loss
All daughters	0.1247	0.1318	5%
Non-genotyped daughters	0.1097	0.1162	6%

Genotyped daughters	0.2662	0.2758	3%
Genotyped sires	0.1422	0.1670	15%

Number of genotyped daughters, non-genotyped daughters and total daughters were 1473, 6901 and 8374, respectively.

Table 4. Regression of corrected yield on EBV of daughters

Set	Conventional ssGBLUP	Breed-of-origin ssGBLUP	Gain/loss
Data set - I	0.9102	0.89148	−2%
Data set - II	0.92618	0.90682	−2%
Data set - III	0.62128	0.65984	6%

For Data set – I, Number of genotyped daughters, non-genotyped daughters and total daughters were 1848, 6901 and 8749, respectively; and for Data set – II and Data set – III, number of genotyped daughters, non-genotyped daughters and total daughters were 1473, 6901 and 8374, respectively.