Research Journal of Dairy Sciences 7 (2-6): 14-17, 2013
ISSN: 1993-5277

Accuracy of Genomic Selection Through Imputation of Sires in Dairy Cattle

C.I. Cho, M. Alam, T.J. Choi and K.H. Cho
National Institute of Animal Science, Rural Development Administration,
33-801 Cheonan, Republic of Korea

Abstract: The purpose of this study, was to determine the accuracy of imputation for non-genotyped sires using their progenies genotypes and to compare the accuracy of genomic breeding values from imputed sires with respect to their true genomic data. A total of 1,800 were simulated to construct phenotypic data and pedigree data. A genotype panel for all animals was prepared as well. Among them, 20 sires were selected randomly and imputed assuming that their genotypes were missing. The average accuracy of imputation for sires through 100 simulation repeats was 88.7% and the obtained range of accuracies was 87-90.4%. The accuracy of Genomic Breeding Values (GEBV) of whole population was slightly higher than for pedigree based Breeding Values (EBV) 0.639 and 0.611, respectively. The accuracy for GEBV from 20 selected sires using true or imputed genotypes were 0.673 and 0.669, respectively. These results indicated that imputation of missing sire genotypes can obtain an accuracy which is almost close to the accuracy of estimates from their true genotypes. Therefore, it can be concluded that there is a possibility in genomic selection for using imputed sires which have no genotypes but yet have their progenies genotypes available. These results deemed more applicable in Korean dairy genomic evaluation, i.e., for Holstein where most sire genotypes are difficult to obtain due to an extensive use of imported semen.

Key words: Genomic breeding values, genotype, pedigree, genomic, Korean

INTRODUCTION

Imputation is usually used to infer genotyping errors due to faulty data entry or by a misinterpretation of the pattern on a gel (Broman, 1999) or to impute from low density SNP chip to high density SNP chip. As soon as, the high density genomic SNP panels have emerged, i.e., Illumina Bovine 50K Bead Chip, the need for imputation has also become important in genetic studies related to QTL mapping, genome-wide association study or genomic selection (Johnston *et al*., 2011) due to the presence of missing genotypes or inaccuracies while genotyping. To date, both family and population based methods of imputation (Browning and Browning, 2009; Sargolzaei *et al*., 2011; Van Raden *et al*., 2011) are implemented in various computer programs, given some advantages and disadvantages of each programs regarding accuracy of imputation and speed of computation (Chen *et al*., 2011; Johnston *et al*., 2011).

Since 2010, South Korea has participated Multiple Across-Country Evaluation (MACE) program of interbull in order to evaluate a total of 23 dairy cattle traits related to production, body conformation and somatic cell score traits. Currently, genomic evaluations for Holstein bulls are regularly practiced in Canada, Germany and USA using Illumina Bovine 50K Chip (Chen *et al*., 2011). Throughout, the last few decades most of the advanced countries in dairy cattle research had collected DNA from many selected bulls and were able to make reference populations quickly as well. In contrast, through the existing Korean dairy cattle program only 2~3 bulls are selected in a year whereas the majority of genetic contributions (~100 heads) comes from imported bulls. Therefore, alongside the difficulties in collecting DNA samples from those limited imported bulls, it is also a great challenge to make a reference population for genomic selection by the few bulls from Korean dairy cattle program. This only leaves us one possibility in hand which is the tracing (imputing) the genotypes of unknown parents based on the current half-sib progenies available. This simulation report aimed to impute the unknown sires in a population based on progeny genotypes, as well as analyze the accuracy of imputation and the accuracy of genomic selection obtained through imputation.

MATERIALS AND METHODS

Simulation data: A historical population which increased gradually from 100-1000 heads by 100 generations was simulated using QMSim Software to create initial Linkage

Corresponding Author: Kwang-Hyeon Cho, Animal Breeding and Genetics Division, National Institute of Animal Science, 331-801 Cheonan, Shinbang, 1 Gil 114, Republic of Korea

Disequilibrium (LD), as well as establish mutation-drift equilibrium in the population (Sargolzaei and Schenkel, 2009). At the 101th generation, a total of 200 bulls and 800 dams were selected as recent population which in turn produced 800 progenies in the subsequent 102th generation, providing 4 progenies to each of the sires, thus raised the total to 1,800 heads in last two generations of the present study (Table 1). It was assumed to have phenotypic data for whole animals in this simulation. The Illumina Bovine 50K BeadChip V2.0 provided the genomic information in order to construct necessary SNP (Single Nucleotide Polymorphism) panel of all autosomes in the study (Fig. 1).

Imputation of non-genotyped sires: To estimate the accuracy of imputation, the genotypes of randomly selected 20 sires having 4 progenies each in 101th generation were deleted for all autosomes (29 pairs) and respectively imputed using the genomic information of rest of the animals (1,780 heads). Imputation was done by Fimpute Software which has implemented a family based approach to impute animal's missing genotypes (Sargolzaei *et al*., 2011).

Table 1: The parameters considered in the simulation study

Items	Parameter/Condition
Historical population	
No. of generation	100
No. of animals (1st generation)	100
No. of animals (100th generation)	2,000
Recent population	
No. of generation	2
Total number of animals	1,800
Assumed heritability (h^2)	0.3
Assumed phenotypic variance	1
Genotype data panel	
No. of chromosomes pairs	29
Marker position	Evenly distributed
QTL position	Evenly distributed
Total number of QTLs	290
Distribution of additive allelic effects	Gamma (shape = 0.4)

Genetic evaluation: The Estimated Breeding Value (EBV) of animals using with phenotypic and pedigree data of related animals were analyzed by the following animal model:

$$y = 1\mu + Za + e \quad (1)$$

Where:

- y = The vector of phenotypes
- μ = The overall mean
- Z = An incidence matrix
- a = The vector of random additive genetic effects of the animal
- e = The vector of random residual errors

The additive genetic variance was assumed to be $A\sigma_a^2$ where A is the numerical relationship matrix using the pedigree information. Genomic Estimated Breeding Values (GEBV) of animals were calculated by the method proposed by Van Raden (2008), Eq. 2:

$$y = 1\mu + Zu + e \quad (2)$$

Where, y, μ, Z and e indicate to similar terms as mentioned earlier and u is the vector of random additive genetic effects that correspond to allele substitution effects for marker. The additive genetic variance was assumed to be $G\sigma_a^2$ where G is the genomic relationship matrix obtained by:

$$G = ZZ' / 2\Sigma p_i(1 - p_i)$$

The accuracy of GEBV (r_{GEBV}) and EBV (r_{EBV}) was calculated as the correlation of GEBV or EBV with respect to the True Breeding Value (TBV) of animals. Similar steps from data simulation to genetic evaluation were repeated 100 times to minimize estimation errors.

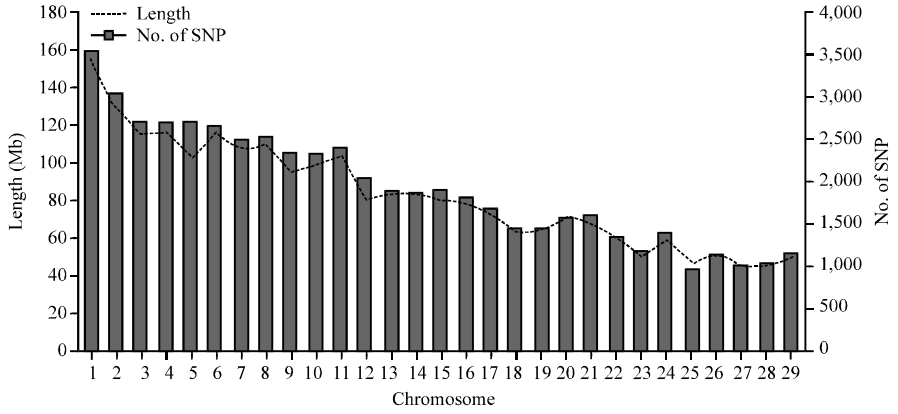

Fig. 1: Genome-wide distribution of SNP and total length in Illumina 50 K Bovine Chip panel

RESULTS AND DISCUSSION

Figure 1 presents a summary on the downloaded Illumina Bovine 50K SNP chip data after exclusion of markers for missing chromosome information or physical distances alongside the makers on sex chromosomes. The screening process retained a total of 52,886 markers in the chip data for later analysis. Total physical distance of markers observed across the genome was 2,508 Mb. The chromosome 1 and 28 harbored the highest and the lowest number of markers 3,430 and 981 SNPs, respectively. An average distance of 0.047 Mb was obtained between adjacent markers too.

In this study, the pre-determined genotypes of randomly selected 20 sires (each having four progenies) were considered as unknown and therefore imputed accordingly. The average accuracy of imputation for sires through 100 simulation repeats was 88.7% and the obtained range of accuracies was 87.0-90.4% (Table 2). A study by Van Raden *et al*. (2013) observed slightly higher imputation accuracies for non-genotyped dams having 4 or more genotyped progenies such as, 93.5% (Findhap Software) and 95.1% (FImpute Software) using relatively larger data sets (116,380 heads). These differences in accuracies for imputation between the studies could possibly be due to the unequal sample sizes in which the present study had a noticeably lesser genotyped animals. Additional factors, such as method of imputation and structure of a population could also have slightly biased the imputation accuracy in this study (Johnston *et al*., 2011; Van Raden *et al*., 2013; Daetwyler *et al*., 2011; Hayes *et al*., 2011).

The accuracies for breeding value estimation using genomic information (GEBV) and pedigree information (EBV) are listed in Table 2. The accuracies of breeding values using true genotypes of whole population were 0.639 (r_{GEBV}) and 0.611 (r_{EBV}), respectively. Therefore, it indicates that the gain in accuracy by using GEBV would be of 0.028 higher while compared to the accuracy of respective pedigree based EBV. Besides that these, mentioned earliar accuracies from true genotypes of selected sires were slightly higher than whole population. The higher accuracies with selected sires group could be accounted to the greater relationships of sire and progenies in the analyzed model (4 progenies per sire) than in the overall population. However, the accuracy for GEBV using imputed sire genotype, either for whole population or selected sires were found to be very identical to their respective accuracies with true genotypes (Table 2). These outcomes indicated that imputation of missing sire genotypes can obtain an accuracy which is almost close to the accuracy of the estimates for genomic EBV using true genotypes.

Table 2: The accuracy of imputation and genetic evaluation in whole population with randomly selected sires[1]

Items	Mean	SD	Min.	Max.
Accuracy of imputation	0.887	0.007	0.870	0.904
	Whole population (N = 1,800)		Randomly selected sires (N = 20)	
Type of genotype data	r_{GEBV}	r_{EBV}	r_{GEBV}	r_{EBV}
All true	0.639	0.611	0.673	0.643
With imputed sires	0.639		0.669	

[1]SD = Standard Deviation; Min. = Minimum value; Max. = Maximum value; r_{GEBV} = Accuracy (correlation) of genomic breeding value to its true breeding; r_{EBV} = Accuracy (correlation) of estimated breeding value based on pedigree to its true breeding value

CONCLUSION

Therefore, it can be concluded that there is a good possibility of using imputed genomic data for sires with missing genotypes which at least have 4 progenies each together with similar given conditions (i.e., SNP densities, reference population size, etc.) in genomic selection of animals. These results can also favor the genomic selection in Holstein cattle in Korea where sires genotypes are usually difficult to obtain due to the extensive use of imported semen all over the country.

ACKNOWLEDGEMENTS

This research was carried out with the supported of development of selection program in dairy bull using genomic information (Project No. PJ009260) project and Postdoctoral Fellowship Program of the National Institute of Animal Science, RDA, Korea.

REFERENCES

Broman, K.W., 1999. Cleaning genotype data. Gen. Epidemiol., 17: 79-83.

Browning, B.L. and S.R. Browning, 2009. A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. Am. J. Human Gen., 84: 210-223.

Chen, J., Z. Liu, F. Reinhardt and R. Reents, 2011. Reliability of genomic prediction using imputed genotypes for german holsteins: Illumina 3K to 54K Bovine chip. Interbull Bulletin No. 44. Stavanger, Norway, August 26-29, 2011, pp: 51-54. https://journal.interbull.org/index.php/ib/article/view/1191/1259.

Daetwyler, H.D., G.R. Wiggans, B.J. Hayes, J.A. Woolliams and M.E. Goddard, 2011. Imputation of missing genotypes from sparse to high-density using long-range phasing. Genetics, 189: 317-327.

Hayes, B.J., P.J. Bowman, H.D. Daetwyler, J.W. Kijas and J.H.J. van der Werf, 2011. Accuracy of genotype imputation in sheep breeds. Anim. Gen., 43: 72-80.

Johnston, J., G. Kistemaker and P.G. Sullivan, 2011. Comparison of different imputation methods. Interbull Bulletin No. 44. Stavanger, Norway, August 26-29, 2011, pp: 25-33. https://journal.interbull.org/index.php/ib/article/view/1186/1254.

Sargolzaei, M. and F.S. Schenkel, 2009. QMSim: A large-scale genome simulator for livestock. Bioinformatics, 25: 680-681.

Sargolzaei, M., J.P. Chesnais and F.S. Schenkel, 2011. FImpute: An efficient imputation algorithm for dairy cattle populations. J. Anim. Sci. Abst., 89: 421-421.

Van Raden, P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci., 91: 4414-4423.

Van Raden, P.M., J.R. O'Connell, G.R. Wiggans and K.A. Weigel, 2011. Genomic evaluations with many more genotypes. Genet. Sel. Evol., Vol. 43. 10.1186/1297-9686-43-10

Van Raden, P.M., D.J. Null, M. Sargolzaei, G.R. Wiggans and M.E. Tooker *et al*., 2013. Genomic imputation and evaluation using high-density Holstein genotypes. J. Dairy Sci., 96: 668-678.