Should the Markers on X Chromosome be Used for Genomic Prediction?

***G. Su*[*], *B. Guldbrandtsen*[*], *G. P. Aamand*[†], *I. Strandén*[§] *and M. S. Lund*[*]**

[*] *Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark*

[†] *Nordic Cattle Genetic Evaluation, DK-8200 Aarhus N, Denmark*

[§] *Biotechnology and Food Research, MTT Agrifood Research, 31600 Jokioinen, Finland*

Email: Guosheng.Su@agrsci.dk

Abstract

This study investigated the accuracy of imputation from LD (7K) to 54K panel and compared accuracy of genomic prediction with or without the X chromosome information, based on data of Nordic Holstein bulls. Beagle and Findhap were used for imputation. Averaged over two imputation datasets, the allele correct rates of imputation using Findhap were 98.2% for autosomal markers, 89.7% for markers on the pseudo autosomal region of the X chromosome, and 96.4% for X-specific markers. The allele correct rates were 98.9%, 91.2% and 96.8%, respectively, when using Beagle. Genomic predictions were carried out for 15 traits based on 54K marker data, imputed 54K for test animals, and imputed 54K for half of reference animals. GBLUP models with or without residual polygenic effect were used for genomic prediction. For all three data sets, genomic prediction using all markers gave slightly higher reliability than prediction excluding the X chromosome. Averaged over 15 traits, the gains in reliability from the X chromosome ranged from 0.3% to 0.5% points among the three data sets and models. Using a model with a G-matrix accounting for sex-linked relationship appropriately or a model which divided genomic breeding value into an autosomal component and an X chromosomal component did not led to better prediction based on the present data where all animals were bulls. A model including polygenic effect did not recover the loss of prediction accuracy due to exclusion of the X chromosome. It is recommended using markers on the X chromosome for routine genomic evaluation.

Keywords: genomic prediction, genotype imputation, the X chromosome

Introduction

Hereditary behaviour of the X chromosome is different from autosomes. For example, in cattle a sire passes its X chromosome to each daughter, but not to his sons. A male only inherits a copy of the X chromosome from his mother, while a female inherits a copy of the X chromosome from her father and one copy from her mother. Therefore, the relationships caused by the X chromosome are different for males and females. In addition, not all regions on X chromosome are specific only on X chromosome, but a small region is homologous with Y chromosome and is inherited like autosomes. This increases the complication of genetic relationship between individuals in terms of the X chromosome. Moreover, in genomic prediction of dairy cattle, de-regressed proof, DYD and EBV are usually used as response variables. These variables are derived from a model where a pedigree-based relationship matrix is constructed according to the inheritance of autosomes. In addition, the density of markers on the X chromosome was markedly lower than autosomes in current SNP chips. These features may reduce the efficiency of the X chromosomal markers for genomic prediction, and could be the reasons why the X chromosome is not used for genomic prediction in some countries and populations.

There are very few reports for imputation accuracy of the markers on the X chromosome (Johnston *et al.*, 2011) and contribution of the markers on X chromosome to accuracy of genomic predictions (VanRaden *et al.*, 2009). The objectives of this study are to investigate the accuracy of imputing missing genotypes on X chromosome, validate accuracy of genomic prediction with or without X chromosome markers, and compare genomic predictions

using genomic relationship matrices with or without differentiated calculation for sex-linked markers, based on the data of Nordic Holsteins.

Material and Methods

Marker data

The data used in this analysis comprised 5,643 progeny-tested bulls from Nordic Holstein population, born during the period from 1974 to 2010. The animals were genotyped with the Illumina Bovine SNP50 BeadChip (about 54K markers). The marker data were edited by deleting the markers with minor allele frequency (MAF) lower than 0.01 and the markers with average GenCall score lower than 0.60. After editing, 43,314 markers on 29 autosomes and 827 markers on the X chromosome remained.

The PAR was identified as the region where markers have a substantial level of heterozygous genotypes in the genotyped bulls. Among the markers on the X chromosome, 133 markers located in the PAR and the remaining 694 were X chromosome specific (X-specific) markers.

In order to investigate accuracy of imputation for the markers on the X chromosome, low density marker data were created from the 54K marker data by masking the markers which did not exist in the Illumina BovineLD BeadChip (about 7K markers). After editing, the LD data had 6,699 markers among which there were 218 markers on the X chromosome (188 X-specific and 25 PAR markers).

The whole data set was divided into reference data and test data, such that 3,995 bulls born before January 1st, 2005 constituted the reference population and 1,648 bulls born after this date comprised the test population. However, the number of animals with phenotypic information differed among different traits due to different numbers of published EBV available. Three sets of data were used for validating the accuracy of genotype imputation and genomic prediction. 1) 54K: the marker data of the 54K chip with imputation of sporadic missing genotypes, 2) IMP_test: the animals in test population with imputed 54K data from the LD marker data, 3) IMP_0.5ref: half of reference animals (random sample) with imputed 54K marker data from the LD data.

Imputation methods

The LD marker data were imputed to 54K data using two programs. One was Beagle (Browning and Browning, 2009) which is a popular imputation program. The other was Findhap (VanRaden *et al.*, 2011) which is a fast imputation program and takes the hereditary behaviour of the X chromosome into account. Therefore, when using Findhap, the PAR was taken out as an independent part and treated as an autosome, and the rest markers of the X chromosome were treated as X-specific markers. The imputed genotypes were compared with the original genotypes. Accuracy of imputation was measured by allele correct rate (proportion of the number of correctly imputed alleles to total number of imputed alleles).

Phenotype data

The phenotypic data for genomic prediction were de-regressed proofs (DRP), which were derived from Nordic genetic evaluations in January 2013. The traits under analysis were 15 traits in the Nordic Total Merit index. These traits were: milk yield, fat yield, protein yield, growth, fertility, birth index, calving index, udder health, other diseases, feet and legs, longevity, body conformation, udder conformation, milking ability, and temperament.

Genomic prediction models

Genomic predictions based on marker data with or without the markers on the X chromosome were carried out using the following GBLUP models.

GBLUP_A: G-matrix was built using autosomal markers only.

GBLUP_All: G-matrix was built using all markers and treating X-specific markers as autosomal markers.

GBLUP_All$_X$: G-matrix was built using all markers and calculated by specifying X-specific markers.

GBLUP_A-X: breeding value was divided into an autosomal component and an X chromosomal component.

GBLUP_A-Pol: GBLUP-A including residual polygenic effect.

GBLUP_All$_X$-Pol: GBLUP_All$_X$ including residual polygenic effect.

G-matrix without specific calculation for X-specific markers was built as presented by VanRaden (VanRaden, 2008),

$$\mathbf{G} = \mathbf{MM}' / \sum 2p_j(1-p_j)$$

This G-matrix can actually reflect sex-linked relationship, for example the relationship between sire and son is zero in terms of X chromosome, but the relationship is scaled up, for example, the diagonal for a male is 2, instead of 1. To get a correct G-matrix, a specific calculation for X-specific markers is required. It can be done in the same way as the calculation for autosomal markers. The only difference is that for X-specific markers the elements of the genotype coefficient matrix (M matrix, after centering) are divided by $\sqrt{2}$ for males.

Genomic predictions using different marker data sets and different models were evaluated by comparing genomic estimated breeding values (GBV) with DRP for animals in the test data. GBV was calculated as the sum of genomic effect and residual polygenic effect when using GBLUP_A-Pol and GBLUP_Allx-Pol, and as the sum of autosomal effect and X chromosome effect when using GBLUP_A-X. Reliabilities of GBV were measured as squared correlation between GBV and DRP divided by the average reliability of DRP (Su *et al.*, 2012).

Genomic prediction was carried out using DMU package (Madsen *et al.*, 2010) and all variance components were estimated from the present data sets.

Results and Discussion

Genotype imputation

Imputation accuracy was lower for markers on the X chromosome than those on autosomes (Table 1). The main reason could be that the marker density on the X chromosome was lower than those on the autosomes. Based on the current marker data (after editing), the average interval between adjacent markers was about 0.175 Mbp for markers on the X chromosome, while about 0.058 for autosomes.

Moreover, PAR has much lower imputation accuracy than X-specific region, though PAR markers were much denser (about double) than X-specific markers in both the LD and the 54K data. This might be explained by the fact that PAR was a small segment (about 11 Mbp according to our detection) which could reduce imputation efficiency. On the other hand, X-specific markers could have lower recombination rate than PAR makers, since crossovers can happen only in females, which could increase imputation accuracy. Poor imputation accuracy for PAR markers was also reported by Johnston *et al.* (2011) in the imputation from 3K to 54K panel.

With regard to imputation programs, Beagle led to slightly higher accuracy than Findhap in all scenarios. However, Beagle took much more time (about 8 hours for chromosome 1) than Findhap (about 2 minutes for chromosome 1) in IMP_test data.

3.1 Genomic prediction

Table 2 presents reliability of genomic prediction for 15 traits, using marker data with or without X chromosomal markers. For all three data sets, genomic prediction using all markers gave slightly higher reliability than predictions excluding the X chromosome, regardless whether the model included residual polygenic effect. Averaged over 15 traits inclusion of X chromosome markers increased the reliability of genomic predictions by 0.3% to 0.5% points.

GBLUP_Allx and GBLUP_All gave the same reliability of genomic predictions, indicating a G-matrix with specific calculation for sex-linked relationship of X-specific markers did not lead to better genomic predictions based on the current data. This could be because animals in the current data were all males. A G-matrix correctly reflecting sex-linked relationship for X-specific markers is expected to improve genomic prediction when data include both males and females. Moreover, GBLUP_A-X did not led to better prediction either, indicating that it is reasonable to assume that the effects of the X chromosomal markers and autosomal markers have the same distribution.

Including residual polygenic effect in the model improved the reliability of predicted breeding values on average by 0.9% point. The largest improvement was for longevity (3.6%) and other diseases (3.7%). For other traits the average improvement was 0.3%. Gao *et al.* (2012) reported an average increase of reliability by 0.3% across 16 traits (including longevity and other diseases). However, they used a constant weight of 0.20 on polygenic effect for all traits, while in the present study variances of polygenic effects were estimated for each trait. The estimated variance components showed that proportions of residual polygenic variance to total additive genetic variance ranged from zero to 53.4% with an average of 17.2%. In addition, the model including residual polygenic effect reduced bias, which was in line with findings reported by Liu *et al.* (2009) and Gao *et al.* (2012). In practical genetic evaluation, genomic estimated breeding values are usually blended with the EBV from conventional pedigree-based BLUP model. It is necessary to investigate if genomic breeding values including residual polygenic effect have any problem with double counting in the step of blending. This could happen because the residual polygenic effect is already included in genomic breeding value, and the blending procedure uses the residual polygenic effect once again.

Genomic predictions based on data sets of IMP_test and IMP_0.5ref had reliabilities close to the predictions based on real 54K data. The results were inconsistent with previous studies on genomic predictions using imputed 54K marker data from 3K markers (Dassonneville *et al.*, 2011). However Ma *et al.* (2013) reported that an improvement (2%) of imputation from 54K to HD by using a joint HD reference data did not result in a corresponding improvement of genomic prediction.

The contribution of markers on X chromosome to the reliability of genomic predictions differed among traits (Table 3). An increase in reliability close to 2% points was observed for fertility and other diseases. Correspondingly, the variances explained by the X chromosome for these two traits were much higher than those for the other traits. On average, markers on the X chromosome accounted for 1.7% of the additive genetic variance. Similarly, VanRaden *et al.* (VanRaden *et al.*, 2009) reported that the X chromosome accounted about 1% genetic variance in the USA Holstein population.

4. Conclusions

Accuracy of genotype imputation for the markers on the X chromosome was lower than the imputation for autosomal markers. Even though, the accuracy of imputation from 7K to 54K panel for the markers on X chromosome was still high in the Holstein population. Genomic prediction using all markers gave slightly higher reliability than predictions excluding markers on the X chromosome. It is recommended to use the markers on the X chromosome for genomic evaluation.

Acknowledgments

This work was performed in the project “Genomic Selection—From function to efficient utilization in cattle breeding (grant no. 3405-10-0137)”, funded under Green Development and Demonstration Programme by the Danish Directorate for Food, Fisheries and Agri Business, the Milk Levy Fund, VikingGenetics, Nordic Cattle Genetic Evaluation, and Aarhus University.

References

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

Dassonneville, R., Brondum, R.F., Druet, T., Fritz, S., Guillaume, F., Guldbrandtsen, B., Lund, M.S., Ducrocq, V. & Su, G. 2011. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. *J Dairy Sci 94*, 3679-3686.

Gao, H.D., Christensen, O.F., Madsen, P., Nielsen, U.S., Zhang, Y., Lund, M.S. & Su, G. 2012. Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. *Genetics Selection Evolution 44:8.*

Johnston, J., Kistemaker, G. & Sullivan, P.G. 2011. Comparison of Different Imputation Methods. *Interbull Bulleten 44*, 25-33.

Liu, Z., Seefried, F., Reinhardt, F. & Reents, R. 2009. A simple method for correcting the bias caused by pre-selection in conventional genetic evaluation. . *Interbull Bulletin 40*, 184-188.

Ma, P., Lund, M.S., Ding, X., Zhang, Q. & Su, G. 2013. Increasing imputation and prediction accuracy for Chinese Holsteins using joint Chinese-Nordic reference population. (Submitted to *J. Dairy Sci.*).

Madsen, P., Su, G., Labouriau, R. & Christensen, O.F. 2010. DMU - A Package for analyzing multivariate mixed models. I CD communication – Proceeding, paper 732, *Book of Abstracts, p. 137, the 9th WCGALP,* Leipzig, Germany, August 1-6, 2010.

Su, G., Madsen, P., Nielsen, U.S., Mäntysaari, E.A., Aamand, G.P., Christensen, O.F. & Lund, M.S. 2012. Genomic prediction for Nordic Red Cattle using one-step and selection index blending. *J Dairy Sci 95*, 909-917.

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. *J Dairy Sci 91*, 4414-4423.

VanRaden, P.M., O'Connell, J.R., Wiggans, G.R. & Weigel, K.A. 2011. Genomic evaluations with many more genotypes. *Genetics Selection Evolution 43.*

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F. & Schenkel, F.S. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci 92*, 16-24.

Table 1. Allele correct rate (%) for makers on autosomes, PAR and X-specific region of the X chromosome.

Dataset	Method	Autosomes	PAR	X-specific
IMP_test	Findhap	98.3	89.6	96.7
	Beagle	98.9	91.2	97.0
IMP_0.5ref	Findhap	98.0	89.9	96.2
	Beagle	98.8	91.1	96.5

Table 2. Reliability (%) of genomic predictions based on three marker data sets with or without markers on the X chromosome and using different models, averaged over 15 traits.

Data set	GBLUP A	GBLUP All	GBLUP Allx	GBLUP A-X	GBLUP A-Pol	GBLUP Allx-Pol
54K	38.0	38.5	38.5	38.5	38.9	39.3
IMP_test	37.9	38.3	38.3	38.4	38.9	39.2
IMP_0.5ref2	37.8	38.3	38.3	38.3	38.8	39.1

Table 3. Reliability (%) of genomic predictions with or without markers on the X chromosome for each trait (GBLUP_A vs. GBLUP_A-X, 54K data) and percentage of additive genetic variance explained by markers on the X chromosome (Var-Xchr, %).

Trait	N	GBLUP A	GBLUP A-X	Difference	Var-Xchr
Milk	1159	48.7	48.9	0.2	0.9
Fat	1159	47.1	47.6	0.5	1.3
Protein	1159	45.9	46.2	0.3	1.5
Fertility	1158	40.7	42.6	1.9	3.6
Birth index	1642	32.5	32.7	0.2	0.8
Calving index	1239	30.3	30.5	0.2	0.7
Udder health	1204	39.5	40.1	0.6	2.7
Other diseases	1050	36.3	38.2	1.9	4.1
Body conform.	1156	27.6	27.4	-0.3	2.2
Feet & legs	1150	33.2	33.7	0.6	1.5
Udder conform.	1156	44.0	44.5	0.5	1.8
Growth	1351	47.2	47.2	0.0	0.0
Milking ability	1155	47.1	47.4	0.3	1.2
Temperament	1142	18.3	18.3	0.0	2.5
Longevity	817	31.1	31.8	0.6	0.8
Average	**1180**	**38.0**	**38.5**	**0.5**	**1.7**