Genomic predictions by single-step genomic BLUP with heterogeneous SNP variance for Japanese Holsteins

T. Baba[1], Y. Gotoh[1], T. Osawa[2], H. Abe[3], S. Nakagawa[3], S. Yamaguchi[3], Y. Masuda[4] & T. Kawahara[1]

[1] Holstein Cattle Association of Japan, Hokkaido Branch, Sapporo 001-8555, Japan. baba@holstein.jp (Corresponding Author)
[2] Department of Animal Breeding, National Livestock Breeding Center,Nishigo, Fukushima 961-8511, Japan
[3] Hokkaido Dairy Milk Recording and Testing Association, Sapporo 060-0004, Japan.
[4] Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA.

Summary

The objective of this study was to evaluate genomic predictions by the single-step genomic BLUP (ssGBLUP) considering heterogeneous variance of single nucleotide polymorphisms (SNP) in Japanese Holsteins. We used the record for 6 production traits (milk yield, fat yield, protein yield, fat percentage, protein percentage and somatic cell score) at the first three lactations, 3 type traits (feet & legs, mammary system and final score) at the first lactation and 1 reproduction trait (days open) at the first lactation. These data were truncated by the year of 2012 to validate the reliability of genomic enhanced breeding value (GEBV) with or without heterogeneous SNP variance. Genomic data included 4,849 bulls that were born before 2012. To estimate GEBV with heterogeneous SNP variance (wGEBV), the SNP variances were calculated with two methods; square of single SNP effect (M1) and mean of 20 adjacent SNP effects (M2). Validation reliability was assessed by the coefficient of determination (R^2) from the linear regression of deregressed-EBV from the full data on GEBV and wGEBV for 580 to 658 validation bulls. The R^2 for fat percentage from wGEBV with both M1 and M2 were 0.07 point higher than R^2 for GEBV. For the other traits, R^2 for wGEBV with M1 was similar to or slightly lower than one for GEBV and R^2 with M2 had the similar to or slightly higher than one for M1, except for somatic cell score. Our results suggested that the traits influenced by major SNP should be evaluated with the heterogeneity of SNP variances and the ssGBLUP with an approach of SNP window can provide more reliable results.

Keywords: heterogeneous SNP variance, Holstein, single-step genomic BLUP

Introduction

Genomic evaluation by the single–step genomic BLUP (ssGBLUP) with phenotype, pedigree and genotype information has an advantage over multi-step genomic evaluation in unnecessity of pseudo phenotypes and accounting for preselection bias. Although ssGBLUP had had a problem of the computing cost in the inverse of genomic relationship matrix with large number of genotyped animals, this problem has been resolved by the implementation of genomic recursion algorithm presented by Misztal *et al.* (2014).

In the genomic relationship matrix included in the ssGBLUP, it was typically assumed to be equal variance of all single nucleotide polymorphisms (SNP). Assuming heterogeneous

variance of SNPs would be preferred under the existence of causative QTL or major SNP. This consideration in the ssGBLUP was conducted by constructing a genomic relationship matrix weighted by SNP variances and this method had a better prediction compared with Bayesian method (Wang *et al.,* 2012). When used large number of genotyped animals, gain in reliability with heterogeneous SNP variance may be small. However, the number of reference animals in Japanese Holstein is limited (Baba *et al.,* 2017) and genomic evaluation with this approach can be effective to increase the reliability. Out objective was to evaluate genomic predictions from the ssGBLUP with heterogeneous SNP variance in Japanese Holsteins.

Materials and methods

Data

We used the record for 6 production traits (milk yield, fat yield, protein yield, fat percentage, protein percentage and somatic cell score) at the first three lactations from 1990 to 2017, 3 type traits (feet & legs, mammary system and final score) at the first lactation from 1984 to 2017 (feet & legs: collected from 1994 to 2017) and 1 reproduction trait (days open) at the first lactation from 1990 to 2017 in Japanese Holstein cows. These data were referred as full data. To implement validation analysis, the full data were truncated by 2012, and genomic enhanced breeding values (GEBV) with or without heterogeneous SNP variance were estimated from the truncated data, respectively. Table 1 shows the number of records in full and truncated data. Genomic data included 4,849 bulls that were born before 2012, genotyped by Illumina Bovine SNP50 BeadChip (Illumina Inc., San Diego, CA, USA).

Table 1. Number of records in full and truncated data.

Trait	Full data	Truncated data
Production traits[1,2]	5,185,927	4,413,986
Somatic cell score[3]	3,280,790	2,508,849
Feet & legs	591,258	483,506
Mammary system, Final score	659,734	551,982
Days open	2,040,144	1,680,675

[1] Milk yield, fat yield, protein yield, fat percentage and protein percentage
[2] The number of animals with records was 2,391,364 in full data and 2,037,634 in truncated data
[3] The number of animals with records was 1,474,967 in full data and 1,121,237 in truncated data

Model

The single-trait repeatability animal model for production traits and single-trait animal model for type traits and days open were used in this study. In the mixed model equations, the inverse of a relationship matrix (**H**) is shown as follows:

(1)

where $\mathbf{A}^{-1}$ is the inverse of the numerator relationship matrix, $\mathbf{G}^{-1}$ is the inverse of the genomic relationship matrix, $\mathbf{A}_{22}^{-1}$ is the inverse of the numerator relationship matrix for genotyped animals and ω is a scaling parameter. The value of ω was set to 0.30 for production traits and 0.50 for type traits and days open, respectively. The genomic relationship matrix was calculated as **G=** where **Z** is a centered matrix for SNP, **D** is a diagonal matrix for weights of SNP variances and p_i is the allele frequency of the i-th SNP

(VanRaden 2008). The initial GEBV were calculated with equal weights for all SNP (i.e. **D** = **I**) to estimate the SNP variances. The GEBV with heterogeneous SNP variance (wGEBV) were estimated with the updated SNP weights (Wang *et al.* 2012). Although this process can be iterative, between the update of SNP variances and the recalculation of wGEBV, we obtained wGEBV with no iterations. The SNP variances were calculated with two methods as following equations; (a) (refer as M1) or (b) (refer as M2), where is the *i*-th SNP effect and *n* is the number of size for the non-overlapping windows (Zhang *et al.,* 2016). In the method of M2, the SNP variances were calculated from a group of 20 adjacent SNPs. The SNP effects () were obtained by solving where is a vector of direct genomic values in reference animals.

Validation

We implemented the linear regression analysis of deregressed-EBV from the full data on GEBV and wGEBV for the 580 (production), 641(days open) or 658 (type) validation bulls which had no daughters in truncated data and more than 20 daughters in the full data. The coefficients of determination (R^2) and slope (*b*) from the regression analysis were used as indicators of the reliability and inflation, respectively.

Results and discussion

Table 2 and 3 show the $R^2$and *b* obtained from the regression analysis. The R^2 for fat percentage from wGEBV with both M1 and M2 were higher than R^2 for GEBV (gained 0.07). For the other traits, R^2 for wGEBV with M1 was similar to or slightly lower than one for GEBV and R^2 with M2 had the similar to or slightly higher than one for M1, except for somatic cell score. The *b* from the method M1 tends to be more variable than the values from M2 when compared the results from GEBV.

The better genomic prediction of fat percentage was obtained from ssGBLUP with the heterogeneous variance of SNPs as reported by VanRaden *et al.* (2009). The presence of major QTL like DGAT1 for fat percentage have been known and this result should be because wGEBV accounts for the major genes through SNP markers. In the other traits, the effect explained by heterogeneous SNP variance was small. The use of the higher density marker information may improve the reliability of genomic predictions.

The genomic predictions from ssGBLUP with M1 reduced the reliability in some traits. On the other hand, M2 tended to provide better results. It was agreed with Zhang *et al.* (2016) who used simulated data. Our result suggests that the assumption of heterogeneity of SNP variances in the single-step evaluation can also be more appropriate in real data.

Table 2. The coefficients of determination (R^2) from the regression analysis of GEBV and wGEBV on deregressed-EBV.

		wGEBV	
Trait	GEBV	M1	M2
Milk yield	0.23	0.22	0.23
Fat yield	0.21	0.21	0.22
Protein yield	0.21	0.19	0.22
Fat percentage	0.48	0.55	0.55
Protein percentage	0.46	0.46	0.47

Somatic cell score	0.15	0.14	0.14
Feet & legs	0.18	0.17	0.18
Mammary system	0.23	0.22	0.23
Final score	0.29	0.29	0.29
Days open	0.18	0.16	0.18

Table 3. The slope (b) from the regression analysis of GEBV and wGEBV on deregressed-EBV.

		wGEBV	
Trait	GEBV	M1	M2
Milk yield	0.92	0.87	0.92
Fat yield	1.00	0.97	1.03
Protein yield	0.94	0.87	0.94
Fat percentage	1.12	1.14	1.16
Protein percentage	1.05	1.03	1.06
Somatic cell score	1.02	1.06	1.02
Feet & legs	0.82	0.81	0.82
Mammary system	1.00	0.96	0.99
Final score	0.92	0.92	0.93
Days open	0.96	0.92	0.96

Conclusion

Genomic prediction with heterogeneous SNP variance considerably improved the reliability in fat percentage because of accountability for major SNP related to causative QTL. The traits influenced by major SNP should be evaluated by such method. The single-step genomic BLUP used heterogeneous SNP variance by an approach of SNP window can provide more reliable results.

List of References

Baba, T., Y. Gotoh, S. Yamaguchi, S. Nakagawa, H. Abe, Y. Masuda & T. Kawahara, 2017. Application of single-step genomic best linear unbiased predictions with a multiple-lactation random regression test-day model for Japanese Holsteins. Animal. Sci. 88(8):1226-1231.

Misztal. I., A. Legarra & I. Aguilar, 2014. Using recursion to compute the inverse of the genomic relationship matrix. J. Dairy. Sci. 97: 3943-3952.

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

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

Wang, H., I. Misztal, I. Auilar, A. Legarra & W.M. Muir, 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. 94: 73-83.

Zhang, X., D. Lourenco, I. Aguilar, A. Legarra & I. Misztal, 2016. Weighting strategies for single-step genomic BLUP: An iterative approach for accurate calculation of GEBV and

GWAS. Front. Genet. 7:151.