A Single Step SNP Model Applied to Test-Day Data of Dairy Cows

Z. Liu[*], **M. E. Goddard**[†], **F. Reinhardt**[*], **and R. Reents**[*]

[*]VIT, Heideweg 1, D-27283 Verden, Germany, [†]University of Melbourne, Australia.

ABSTRACT: A single step SNP model was applied to test-day data of German Holstein cows to evaluate genotyped and phenotyped animals jointly and account for the impact of genomic pre-selection. A total of 343 million test-day records from 19.2 million cows were analyzed with a multiple lactation random regression test-day model, and >100,000 genotyped animals were considered with a genomic model assuming all SNP markers explaining equal genetic variance. In addition, international phenotypes of c.a. 113,000 Holstein bulls were added in order to use a multiple-country genomic bull reference population. A residual polygenic effect was fitted to the single step SNP model to reduce the overestimation problem of genomic prediction. Computing strategies for estimating effects of the model, including SNP effects, were developed. Technical issues for a routine implementation were addressed using the data of German Holsteins.

Keywords: single step genomic model; marker effects; test-day data; dairy cows

Introduction

Genomic evaluation and selection based on the theory by Meuwissen et al. (2001) have been conducted on a routine basis in an increasing number of countries (VanRaden, 2008; Lund et al., 2011). Until now, most countries have applied a so-called multiple step model to genomic evaluation (Liu et al., 2011; VanRaden, 2008), in which a genomic model was fitted to deregressed estimated breed values (EBV) obtained from a conventional evaluation in a previous step. For German Holsteins, bull EBV from national and multiple-across country evaluation (MACE) were deregressed to generate a form of pseudo phenotype for genomic evaluation. The deregressed EBV were used as phenotypes for German Holsteins in SNP effect estimation with the EuroGenomics bull reference population (Lund et al., 2011), calculation of direct genomic values (DGV) and computing pedigree index for young genotyped animals. A selection index method was finally used to combine DGV and pedigree index for obtaining genomic estimated breeding values (GEBV).

The multi-step genomic model was relatively simple and easy to implement. However, the separate steps of genomic and conventional evaluations can not properly take the genomic pre-selection into account and thus conventional EBV and subsequently GEBV would be biased (Patry and Ducrocq, 2011). In addition, the multi-step genomic model can not optimally consider relationship among genotyped animals and between genotyped and non-genotyped animals. In contrast to the multi-step models, single step genomic models (Aguilar et al., 2010; Christensen and Lund, 2010) evaluated genotyped and non-genotyped animals jointly, thus can give unbiased predictions of genetic merits. However, the single step genomic models did not provide estimates of SNP effects which give meaningful biological interpretation of SNP genotypes on phenotypes.

Routine conventional genetic evaluation consider all cows with phenotypes typically born in last 20 years or more and their ancestors in pedigree. But only a small fraction of the evaluated animals have been genotyped. Therefore, the number of animals evaluated with a multi-step genomic model is far less than that in a conventional genetic evaluation. This means that the upgrading of the current multi-step genomic model to a single step genomic model must deal with a much higher number of animals. Efficient computing algorithms are thus needed for jointly estimate additive genetic effects of cows with phenotypes, genotyped animals and their relatives in pedigree. The objectives of our study were 1) to apply a single step SNP model to test-day data of dairy cows, 2) to develop efficient computational algorithms for estimating effects of the genomic model.

Materials and Methods

A single step SNP model. Test-day yields in first three lactations are analyzed with a random regression model (Liu et al. 2003):

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_p\mathbf{p} + \mathbf{Wu} + \mathbf{e} \qquad [1]$$

where $\mathbf{y}$ is a vector of test-day yields, $\mathbf{b}$ is a vector of fixed effects of herd-test-date-parity-milking-frequency and lactation curves on regions x seasons x age x intervals of calving (Liu et al. 2003), $\mathbf{p}$ represents permanent environmental effects of cows, $\mathbf{u}$ is a vector of additive genetic effects, $\mathbf{e}$ represents residual effects. $\mathbf{X}, \mathbf{Z}_p, \mathbf{W}$ are incidence matrices for effects $\mathbf{b}$, $\mathbf{p}$, and $\mathbf{u}$, respectively. It is assumed that there are two groups of animals to be evaluated, group 1 without and group 2 with genotype data available. The genotyped animals may or may not have test-day records. We further assume that *m* SNP markers selected from a SNP chip for genomic evaluation cannot explain all additive genetic variance (σ_g^2), leaving $k\,\sigma_g^2$ residual polygenic variance, where *k* is the proportion of additive genetic variance not explained by all the *m* SNP markers. Thus, additive genetic effects of the genotyped animals in group 2 can be divided into:

$$\mathbf{u}_2 = \mathbf{Zg} + \mathbf{a}_2 \qquad [2]$$

where $\mathbf{u}_2$ represents additive genetic effects of the genotyped animals, $\mathbf{g}$ represents additive genetic effects of the *m* fitted SNP markers, $\mathbf{a}_2$ represents residual polygenic effects (RPG) of the genotyped animals, and $\mathbf{Z}$ is an incidence matrix for SNP marker effects of the genotyped animals. It is assumed further that the SNP marker effects have a (co)variance structure:

$$\mathrm{var}(\mathbf{g}) = \mathbf{B}\sigma_g^2 \qquad [3]$$

where $\mathbf{B} = \frac{1-k}{m}\mathbf{I}$ under a BLUP SNP model (Liu et al., 2011) assuming all the SNP explaining equal genetic variance.

For the genotyped animals, (co)variances of RPG and additive genetic effects are respectively:

$$\mathrm{var}(\mathbf{a}_2) = \mathbf{A}_{22}\, k\sigma_g^2\,, \qquad [4]$$

$$\mathrm{var}(\mathbf{u}_2) = (\mathbf{ZBZ}' + k\mathbf{A}_{22})\sigma_g^2 = \mathbf{G}_{22}\,\sigma_g^2\,. \qquad [5]$$

From Formula [5] we can see that our genomic relationship matrix $\mathbf{G}_{22}$ is a linear function of observed genomic relationship matrix $\mathbf{ZBZ}'$ with an implicit weight (1-k) and expected pedigree relationship matrix $\mathbf{A}_{22}$ with a weight k. (Co)variance matrix of additive genetic effects of non-genotyped animals in group 1 ($\mathbf{u}_1$) and the genotyped animals ($\mathbf{u}_2$) is:

$$\mathrm{var}(\mathbf{u}) = \mathrm{var}\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix} = \mathbf{G}\,, \qquad [6]$$

and its inverse matrix is:

$$\mathbf{G}^{-1} = \left[\mathbf{A}^{-1} \quad + \quad \begin{bmatrix}0 & 0\\ 0 & \mathbf{G}_{22}^{-1} - \mathbf{A}_{22}^{-1}\end{bmatrix}\right]\sigma_g^{-2}\,. \qquad [7]$$

Computing strategies. Appending the vector of SNP effects **g** to **u** gives a complete set of equations (Liu et al. 2014) which contain equations for the additive genetic effects of the genotyped animals and SNP effects:

$$\mathbf{W}_2{}'\mathbf{X}\hat{\mathbf{b}} + \mathbf{W}_2{}'\mathbf{Z}_p\hat{\mathbf{p}} + \lambda\mathbf{A}^{21}\hat{\mathbf{u}}_1 + (\mathbf{W}_2{}'\mathbf{W}_2 + \lambda\mathbf{A}^{22})\hat{\mathbf{u}}_2$$

$$= \mathbf{W}_2{}'\mathbf{y} + \lambda\mathbf{A}_{22}^{-1}(\hat{\mathbf{u}}_2 - \tfrac{1}{k}\hat{\mathbf{a}}_2) \qquad [8]$$

$$\left(\mathbf{B}^{-1} + \tfrac{1}{k}\mathbf{Z}'\mathbf{A}_{22}^{-1}\mathbf{Z}\right)\hat{\mathbf{g}} = \tfrac{1}{k}\mathbf{Z}'\mathbf{A}_{22}^{-1}\hat{\mathbf{u}}_2\ . \qquad [9]$$

Following the idea of Legarra and Ducrocq (2012), the genomic contribution term:

$$\mathbf{A}_{22}^{-1}(\hat{\mathbf{u}}_2 - \tfrac{1}{k}\hat{\mathbf{a}}_2) = \mathbf{A}_{22}^{-1}\hat{\mathbf{u}}_2^* = \hat{\varphi} \qquad [10]$$

can be computed equivalently by solving the following equations:

$$\mathbf{A}_{22}\hat{\varphi} = \hat{\mathbf{u}}_2^*\,. \qquad [11]$$

From Equation [8] we can see that right-hand-sides (RHS) of conventional mixed model equations need to be corrected for the genomic contribution $\lambda\hat{\varphi}$ for the genotyped animals. Re-arranging the terms of SNP effects in Equation [9] gives:

$$\hat{\mathbf{g}} = \tfrac{1}{k}\mathbf{BZ}'\mathbf{A}_{22}^{-1}(\hat{\mathbf{u}}_2 - \mathbf{Z}\hat{\mathbf{g}}) = \tfrac{1}{k}\mathbf{BZ}'\mathbf{A}_{22}^{-1}\hat{\mathbf{a}}_2 = \tfrac{1}{k}\mathbf{BZ}'\hat{\boldsymbol{\gamma}} \qquad [12]$$

where $\hat{\boldsymbol{\gamma}} = \mathbf{A}_{22}^{-1}\hat{\mathbf{a}}_2$. [13]

Formulas [10] and [13] represent two extra major calculations for genomic contribution of the single step SNP model [1]. Numerical iterative procedures, such as preconditioned conjugate gradients (Strandén and Lidauer, 1999), may be used to obtain $\hat{\varphi}$ and $\hat{\boldsymbol{\gamma}}$.

Estimating SNP effects with a special algorithm. In contrast to single step genomic BLUP models (Aguilar et al., 2010; Legarra and Ducrocq, 2012), our genomic model [1] has an extra step of estimating SNP effects. This extra step enables us to actively control the information flow from genomic reference population to candidates. In order to decide which genotyped animals are allowed to make contribution to SNP effect estimation, we introduce a filter matrix **F** to all genotyped animals:

$$\mathbf{F} = diag\{1,0,0,1,\cdots,1,1,0\} \qquad [14]$$

where diagonal element 1 or 0 means a genotyped animal defined as a reference animal or not, respectively. The **F** matrix can be inserted into the SNP effect estimation equation [12]:

$$\hat{\mathbf{g}} = \tfrac{1}{k}\mathbf{BZ}'\mathbf{FA}_{22}^{-1}\hat{\mathbf{a}}_2 \qquad [15]$$

to allow additive genetic effects of only selected genotyped animals to be used in the conversion to SNP effects.

Estimating SNP effects represents a so-called 'large p and small n' computational issue, where more SNP effects are to be estimated using fewer reference animals. Residual polygenic effects, $\hat{\mathbf{a}}_2$, of reference animals need to be separated from $\hat{\mathbf{u}}_2$ in order to obtain SNP effect estimates, $\hat{\mathbf{g}}$. For estimating both components of $\hat{\mathbf{u}}_2$, an iterative procedure has been developed. For a given set of $\hat{\mathbf{u}}_2$ estimates of the reference animals, following relations exist between two consecutive rounds of iteration:

$$\mathbf{Z}(\hat{\mathbf{g}}^{[r+1]} - \hat{\mathbf{g}}^{[r]}) = (\hat{\mathbf{u}}_2 - \hat{\mathbf{a}}_2^{[r+1]}) - (\hat{\mathbf{u}}_2 - \hat{\mathbf{a}}_2^{[r]}) = \hat{\mathbf{a}}_2^{[r]} - \hat{\mathbf{a}}_2^{[r+1]} \qquad [16]$$

$$\hat{\mathbf{a}}_2^{[r+1]} = \hat{\mathbf{a}}_2^{[r]} - \mathbf{Z}(\hat{\mathbf{g}}^{[r+1]} - \hat{\mathbf{g}}^{[r]})\ . \qquad [17]$$

Formula [17] shows that, for a given set of $\hat{\mathbf{u}}_2$, residual polygenic effect $\hat{\mathbf{a}}_2$ can be updated using the changes in SNP effect estimates between two rounds r+1 and r. Iterating Equation [15] for SNP effects and Formula [17] for residual polygenic effects, both effects can be estimated efficiently for a given set $\hat{\mathbf{u}}_2$ of the reference animals, because the updating of residual polygenic effects involves reading genotypes of reference animals only once. As Equation [15] estimates the effects of all SNP markers simultaneously, the order of SNP markers no longer plays a role in convergence behavior as observed by Liu et al. (2011) with an estimation algorithm on a SNP-by-SNP basis.

Data. A total of 342,839,500 test-day records of 19,207,226 cows were taken from December 2013 routine conventional evaluation for German Holsteins. The number of genotyped animals reached 100,865, including 7999 cows with test-day data. For the evaluated trait milk yield, 113,259 Holstein bulls in the December 2013 MACE evaluation had only or more MACE than German domestic phenotype information, including 102,307 non-genotyped Holstein bulls. In total, 26,660 genotyped Holstein bulls were selected for the genomic reference population. For the genotyped population, 199,991 ancestors or non-genotyped relatives were found, resulting in a total of 300,856 animals in the pedigree for the genotyped animals. For the single step SNP model the total number of real animals in the pedigree and phantom parent groups was 24,819,062.

Because a Legendre polynomial with 3 parameters was used to model the additive genetic and permanent environmental effects of the first three lactations (Liu et al. 2003), the total number of equations of the single step SNP model amounted to 420,358,154 plus 45,163 SNP effects and 300,856 RPG effects.

Results and Discussion

Computing $\hat{\varphi}$ and $\hat{\boldsymbol{\gamma}}$. The single step model [1] needs to repeatedly compute $\hat{\varphi} = \mathbf{A}_{22}^{-1}\hat{\mathbf{u}}_2^*$ and $\hat{\boldsymbol{\gamma}} = \mathbf{A}_{22}^{-1}\hat{\mathbf{a}}_2$ by solving $\mathbf{A}_{22}\hat{\varphi} = \hat{\mathbf{u}}_2^*$ and $\mathbf{A}_{22}\hat{\boldsymbol{\gamma}} = \hat{\mathbf{a}}_2$ with an iterative numerical procedure. The PCG algorithm (Strandén and Strandén, 1999) was applied to the genotype data. Every 100 rounds a full PCG residual update was performed. For investigating accuracy of the numerical approach, true values of $\hat{\boldsymbol{\gamma}}$ were set to $\boldsymbol{\gamma}_{true} = \mathbf{1}$, and $\mathbf{a}_2^{true} = \mathbf{A}_{22}\boldsymbol{\gamma}_{true} = \mathbf{A}_{22}\mathbf{1}$ were generated which contained sum of pedigree relationship coefficients with all the genotyped animals. Solutions of the equations $\mathbf{A}_{22}\hat{\boldsymbol{\gamma}} = \mathbf{a}_2^{true}$ were compared to their true values $\boldsymbol{\gamma}_{true} = \mathbf{1}$. Figure 1 shows the convergence criteria of the PCG algorithm for solving the equations. The pre-defined criterion -15 was reached at round 1766. Difference between the estimates and true values of $\boldsymbol{\gamma}$ was 0.00000003, on average. For the 100,865 genotyped animals and 300,856 animals in the pedigree, the total computing time was 13 seconds for all 1766 rounds of iteration on a Linux server. Memory usage was very small as a result of no explicit setup of $\mathbf{A}_{22}$ matrix.

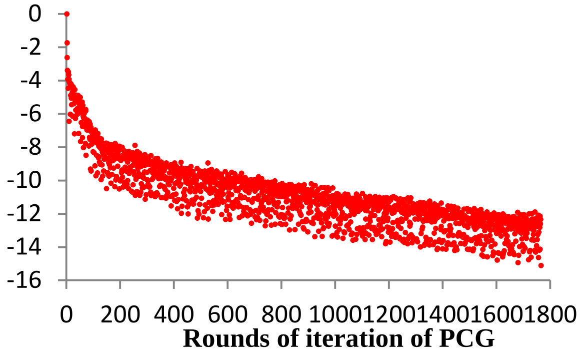

Figure 1. Convergence criteria for solving $\mathbf{A}_{22}\hat{\boldsymbol{\gamma}} = \hat{\mathbf{a}}_2$

Genomic and genetic evaluation. A fast algorithm, a lactation-based iteration procedure for the random regression test-day model (Liu et al. 2004), was retained for the single step SNP model. Foreign bulls, in particular those included in the EuroGenomics reference population, had deregressed MACE EBV expressed on a single trait basis. The deregressed EBV were extended to first three lactations by assuming the foreign bulls have equal additive genetic effects in the three lactations on a standardized scale. Moreover, we assumed that the second and third parameters of the Legendre polynomials were 0 for those bulls. Under those assumptions, a single value of deregressed MACE EBV was transformed into three intercepts for the three-lactation random regression test-day model. The whole set of equations was solved using Gauss-Seidel and PCG algorithms for comparing computing speed and resource usages.

There were four components implemented in the iteration program: conventional parts with national and international phenotypes, adjusting RHS for the genomic contributions, SNP effect estimation, and calculating GEBV of candidates without phenotypes. Firstly, the conventional parts of the single step SNP model were processed for the regular random regression model based on the test-day data of domestic cows and for including the deregressed MACE EBV of foreign bulls. Secondly, genomic contributions were calculated for adjusting RHS for the genotyped animals with phenotypes. Thirdly, SNP effects were estimated by converting GEBV of the selected reference bulls. And lastly, genotyped candidates without phenotypes were included for calculating their GEBV, because they did not provide any phenotype information. These four components/steps were executed repeatedly to guarantee convergence of the whole system.

Conclusion

A single step genomic model allowing direct estimation of SNP effects was described and applied to test-day data of German Holstein cows. Besides the test-day records of domestic cows, deregressed MACE EBV of foreign bulls were added as additional source of phenotypes. A multiple-lactation random regression test-day model for analyzing the phenotypic data was combined with a BLUP SNP model for evaluating genotype data. Computing strategies were developed for solving the equations, including estimating the SNP marker effects. Some intermediate results were presented for the implementation of the single step SNP model.

Literature Cited

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., and Lawlor, T. J. (2010). J. Dairy Sci. 93:743–752.

Christensen, O. F., and Lund, M.S. (2010). Genet. Sel. Evol. 42:2.

Legarra, A., and Ducrocq V. (2012). J. Dairy Sci. 95:1-17.

Liu, Z., Reinhardt, F., Bünger, A., and Reents, R. (2004). J. Dairy Sci. 87:1896-1907.

Liu, Z., Seefried, F. R., Reinhardt, F., Rensing, S., Thaller, G., and Reents, R. (2011). Genet. Sel. Evol. 43:19.

Liu, Z., Goddard, M.E., Reinhardt, F., and Reents, R. (2014) submitted.

Lund, M. S., A. P. W. De Roos, A. G. De Vries, et al. Genet. Sel. Evol. 43:43.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Genetics 157:1819-1829.

Patry, C., and Ducrocq, V. (2011). J Dairy Sci 94:1011-1020.

Strandén, I., and Lidauer, M. (1999). J. Dairy Sci. 82:2779-2787.

VanRaden, P. M. (2008). J Dairy Sci 91:4414-4423.