Towards routine estimation of breeding values using one million genotyped animals

Jan ten Napel[1], Andrew Cromie[2], Ghyslaine Schopen[1], Jeremie Vandenplas[1] & Roel Veerkamp[1]

[1] Wageningen University & Research Animal Breeding and Genomics, P.O. Box 338, 6700 AH, Wageningen, the Netherlands
Jan.tenNapel@wur.nl (Corresponding Author)
[2] Irish Cattle Breeding Federation, Highfield House, Shinagh, Bandon, Co. Cork, Ireland

Summary

Breeding organisations generally estimate breeding values to identify the most appropriate individuals to become parents of the next generation. Inclusion of one million genotyped animals is realistic. Three methods to specify genetic similarity between individuals, as implemented in MiXBLUP, were compared in a routine genetic evaluation of age at first calving of beef cattle in Ireland using genotypes of 50,240 SNP on 613,984 animals. The methods were pedigree BLUP, single-step genomic BLUP using genomic recursions to approximate the inverse genomic relationship matrix (ssGBLUP-APY) and single-step ridge-regression BLUP (ssRRBLUP). Inclusion of genomic information dramatically increased memory requirement and number of IO operations. Convergence of ssRRBLUP is slower than pedigree BLUP and ssGBLUP-APY. Time per iteration was similar for ssRRBLUP and ssGBLUP-APY. Correlations between solutions of different methods varied from 0.35 to 0.80 for genotyped and from 0.70 to 0.94 for non-genotyped animals. Also slope and intercept between any two methods differed for genotyped and non-genotyped animals. Overall, utilising a large number of genotyped animals in routine genetic evaluations with ssRRBLUP is possible, but as yet very slow due to slower convergence and a large number of IO operations. BLUP with regression on SNP covariates seems to be the most suitable method when convergence and IO problems have been resolved.

Keywords: genomic selection, mixed model, computer program, relationship matrix, beef cattle

Introduction

Breeding organisations generally estimate breeding values to identify the most appropriate individuals to become parents of the next generation.

The availability of single nucleotide polymorphism (SNP) information on a large scale made it possible to utilise the observed, instead of expected genetic similarity between individuals. Genomic information combined with phenotypic data and pedigree can be used to predict more accurate breeding values (Pszczola et al., 2013). The number of genotyped animals is increasing rapidly to millions of animals in, for example, Irish cattle. This requires new approaches for specifying genetic similarity in breeding value estimation.

So-called single-step GBLUP (ssGBLUP) allows simultaneous use of phenotypic information of non-genotyped and genotyped animals, pedigree information and genomic

information thanks to combining the inverse genomic ($\mathbf{G}^{-1}$) and pedigree ($\mathbf{A}^{-1}$) relationship matrices into a blended inverse relationship matrix ($\mathbf{H}^{-1}$) (Aguilar et al., 2010, Christensen and Lund, 2010). The ssGBLUP is now widely used across species in private and national routine genetic evaluations due to its simplicity. This method, however, becomes computationally more demanding as the number of genotyped individuals increases. Especially time and memory required to invert the **G** matrix increase non-linearly with increasing number of genotyped individuals.

Two alternatives for ssGBLUP are so-called Algorithm for Proven and Young animals (APY) and single-step ridge regression BLUP (ssRRBLUP). The APY algorithm efficiently computes an approximation of the inverted **G**, based on a subset of the genotyped animals. The ssRRBLUP method fits the SNP as covariates in the BLUP model, instead of being used for the computation of a genomic relationship matrix (Fernando et al., 2014). This ssRRBLUP method explicitly imputes SNP covariates for non-genotyped individuals using pedigree relationships and fits a residual polygenic effect to account for imputation error (Fernando et al., 2014).

We implemented these methods in MIXBLUP (Ten Napel *et al.*, 2017), and the aim of this paper is to test and compare results and performance of pedigree BLUP, ssGBLUP and ssRRBLUP for routine genetic evaluation in the Irish suckler beef program with over 0.6 million genotyped individuals,12 million animals in the pedigree and 2 million phenotypes.

Material and methods

The routine genetic evaluation used to compare methods of specifying genetic similarity is the Irish fertility evaluation of beef cattle, provided by the Irish Cattle Breeding Federation.

Methods to specify genetic similarity

The general BLUP model that was used to estimate breeding values can be represented as follows:

(1)

where **y** is the vector of phenotypes, **X** and **Z** are incidence matrices relating observations with fixed and direct additive genetic effects, **β** is the vector of fixed effects and **u** is the vector of direct additive genetic effects.
The mixed-model equations are

(2)

where var(**u**) = **V** with **V** a relationship matrix between individuals and var(**e**) = **I** with **I** an identity matrix of order equal to the number of observations.
For pedigree BLUP, the pedigree relationship matrix **A** was used for **V**. For ssGBLUP, the blended pedigree and genomic relationship matrix **H** was used for **V**. The inverse of **H** was calculated as

(3)

where is an approximation of the inverse of the genomic relationship matrix **G** using genomic recursions with 40,000 randomly chosen core animals (Misztal, 2016). **G** was calculated according to the first method of VanRaden (2008) as implemented in calc_grm (Calus and Vandenplas, 2015).
The model ssRRBLUP that was used to calculate the GEBV can be represented as follows:

(4)

where **M** is a SNP covariate matrix and relates imputed (*non*) or observed (*gen*) SNP to individuals, contains the SNP effects and contains the residual polygenic effects for non-genotyped individuals and is zero, otherwise.

With and , the mixed model equations are

(5)

The variance was assumed to be the same for all SNP and was calculated as

(6)

where is the number of fully informative SNP and is the allele frequency of the major allele of SNP *i*.

Statistical model

The statistical model to evaluate age at first calving was obtained from routine evaluations and included herd-year-season of first calving (N=959,426 classes) as a fixed effect, heterosis and recombination as fixed covariates, contribution of each genetic group to the total genetic value of an animal as random covariates (Westell et al., 1988) and the random additive genetic effect. Definition of genetic group is based on breed of founder animals.

Data

A total of 2,117,526 records of age at first calving from 70,410 herds were available. The Irish beef cattle population is a mixture of purebred and crossbred animals of at least 50 breeds. The most commonly used breeds are Charolais, Limousin, Angus, Simmental and Hereford. A high proportion of cows is a crossbred of a dairy breed and a beef breed. The pedigree consisted of 11,987,772 animals and up to 16 generations of ancestors.

Animals were genotyped using three different custom international dairy & beef SNP chips (IDB SNP chip; Mullen *et al*., 2013). All missing SNP genotypes across the three SNP chips were imputed. After screening, a total of 50,240 SNP were available on 613,984 animals.

Software

Breeding values were estimated using MiXBLUP (Ten Napel et al., 2017). Genomic relationship matrices were calculated with calc_grm (Calus and Vandenplas, 2015), as integrated in MiXBLUP.

Results and discussion

Computational demands

Utilizing genomic information for large-scale routine evaluations dramatically increased the computational demands compared to a pedigree BLUP evaluation (Table 1). The genomic relationship matrix **G** and the SNP covariate matrix **M** are dense and large matrices when the number of genotyped individuals is large. Processing this vast amount of information either requires a substantial memory allocation or dramatically increases the number of Input/Output (IO) operations. The size of the binary APY inverse of **G** to be read every iteration was 278 Gb. The size of the binary SNP covariate matrix in this study was 413 Gb. Time and memory requirements can be balanced in the software.

Table 1. Computational demands of the three methods to specify genetic similarity between individuals.

	Pedigree BLUP	ssGBLUP-APY	ssRRBLUP
Time per iteration, h:m:s	0:00:04	0:15:01	0:12:31
Iterations to convergence, N	185	430	2421
Memory allocation solving[1]	8 Gb	64 Gb	64 Gb

[1] Memory that was actually used was not recorded

For the three models the preconditioned conjugate gradient (PCG) algorithm was used with block-Jacobi preconditioner matrix. Number of iterations to achieve convergence was substantially higher for the ssRRBLUP model. The condition number of the coefficient matrix of ssRRBLUP models is probably high (Vandenplas et al., 2018), which adversely affects convergence. A solution may be to use a different preconditioner matrix that reduces the condition number of the coefficient matrix even more. A different algorithm, such as deflated Preconditioned Conjugate Gradient (Vandenplas et al., 2018), may be another approach for efficient solving.

Breeding values

Breeding values of ssGBLUP-APY varied less than pedigree BLUP (Figure 1) and ssRRBLUP (Figure 3), but there was a linear relationship in both cases. The correlations were 0.69 and 0.80, respectively. There was only a weak relationship between the breeding values of ssRRBLUP and pedigree BLUP (Figure 2). The correlation was 0.35. In fact, each figure contains two scatter plots, one for genotyped animals and one for non-genotyped animals, and each with a different regression coefficient. For example, the correlation between ssGBLUP-APY and ssRRBLUP was 0.33 for genotyped and 0.69 for non-genotyped animals. Corresponding regression coefficients of ssRRBLUP on ssGBLUP-APY were 1.38 and 0.82.

The reduced variation in breeding values of ssGBLUP-APY and the low correlation between ssGBLUP-APY and ssRRBLUP may be due to the core animals not being representative for the many breeds and crosses in the Irish beef cattle population. Stratified random sampling may be more appropriate in this case.

The correlations between the solutions of the three methods were lower than we

observed in other, smaller data sets. For non-genotyped animals, the correlation between solutions of pedigree BLUP and ssRRBLUP is generally close to 1.00 and for genotyped animals around 0.90. Further study revealed that some solutions of ssRRBLUP still had not converged. The same convergence criterion was used for all three methods, but ssRRBLUP probably needs an additional convergence criterion as SNP effects and residual polygenic effects are modelled separately.

Overall, utilising a very large number of genotyped animals in routine genetic evaluations is possible, but very slow due to slower convergence and an increased number of IO operations. In analyses that allowed the full inversion of G, the solutions of ssRRBLUP and ssGBLUP were very similar (Manzanilla-Pech et al., 2017). So ssRRBLUP can be used as a gold standard when a full inverse of G for ssGBLUP is not feasible.

The number of genotyped animals in the Irish suckler beef genetic programme exceeded one million in 2017 and every year 300,000 genotyped animals are added. Method ssGBLUP using an APY inverse of G will become more demanding on memory and IO operations as the size of the APY inverse increases. With increasing number of genotyped animals, ssRRBLUP benefits as fewer genotypes need to be imputed and the number of (SNP) effects does not increase. Hence, time per iteration is not affected as an imputed or observed SNP covariate record is fitted for any animal with data. Also, convergence may benefit from the smaller number of residual polygenic effects that need to be estimated. Therefore, BLUP with regression on SNP covariates seems to be the most suitable method when convergence and IO problems have been resolved.

List of References

Aguilar, I., I. Misztal, D.L. Johnson, A. Legarra, S. Tsuruta & T.J. Lawlor, 2010. Hot Topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93(2):743-752.

Calus, M.P.L. & J. Vandenplas, 2017. Calc_grm: a program to compute pedigree, genomic, and combined relationship matrices. Wageningen University & Research Animal Breeding and Genomics, Wageningen, the Netherlands.

Christensen, O.F. & M.S. Lund, 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:2.

Fernando, R.L., J.C.M. Dekkers, & D.J. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet. Sel. Evol. 46:50.

Manzanilla-Pech, C.I.V., R.F. Veerkamp, Y. de Haas, M.P.L. Calus & J. ten Napel, 2017. Accuracies of breeding values for dry matter intake using non-genotyped animals and predictor traits in different lactations. J. Dairy Sci. (in press) https://doi.org/10.3168/jds.2017-12741.

Misztal, I., 2016. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. Genetics 202: 401-409.

M.P. Mullen, M.C. McClure, J.F. Kearney, S.M. Waters, R. Weld, P. Flynn, C.J. Creevey, A.R. Cromie & D.P. Berry, 2013. Development of a custom SNP chip for dairy and beef cattle

breeding, parentage and research. Interbull Bulletin No. 47.

Pszczola, M., R.F. Veerkamp, Y. de Haas, E. Wall, T. Strabel & M.P.L. Calus, 2013. Effect of predictor traits on accuracy of genomic breeding values for feed intake based on a limited cow reference population. Animal 7:1759–1768.

Ten Napel, J., M.P.L. Calus, M. Lidauer, I. Stranden, E. Mäntysaari, H.A. Mulder & R.F. Veerkamp, 2017. MiXBLUP, user-friendly software for large genetic evaluation systems. Manual v2.1. Wageningen University & Research Animal Breeding and Genomics, Wageningen, the Netherlands.

Vandenplas, J., H. Eding, M.P.L. Calus & C. Vuik, 2018. Deflated preconditioned conjugate gradient method for solving single-step single nucleotide polymorphism BLUP. Paper 11.25. Proceedings, 11th World Congress of Genetics Applied to Livestock Production, Auckland, New Zealand.

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

Westell, R.A.; R.L. Quaas & L.D. Van Vleck, 1988. Genetic Groups in an Animal Model. Faculty Papers and Publications in Animal Science. Paper 309. http://digitalcommons.unl.edu/animalscifacpub/309.

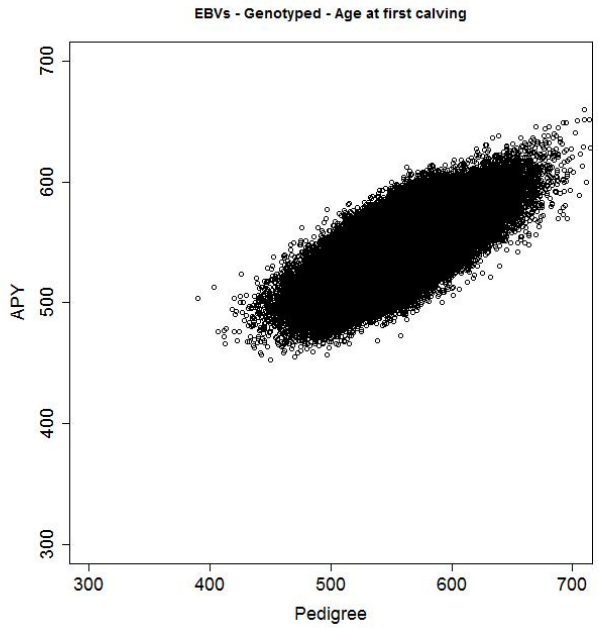

Figure 1. Scatter plot of breeding values of ssGBLUP-APY on pedigree BLUP.

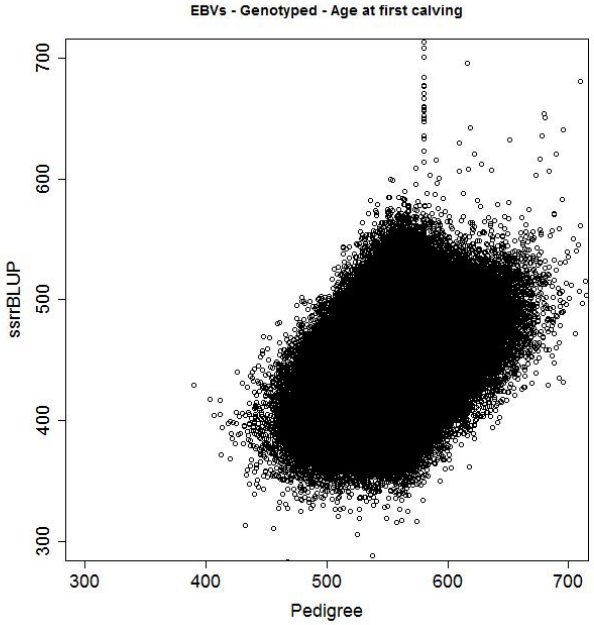

Figure 2. Scatter plot of breeding values of ssRRBLUP on pedigree BLUP.

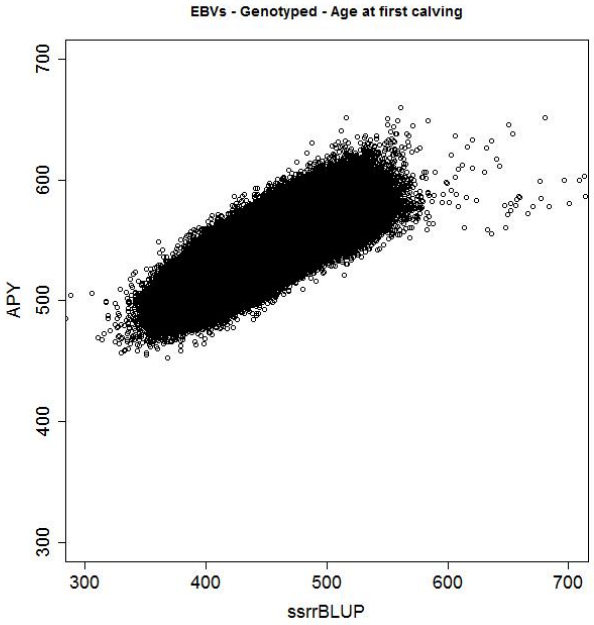

Figure 3. Scatter plot of breeding values of ssGBLUP-APY on ssRRBLUP.