360. Indirect genomic prediction reduces computational costs in large-scale single-step evaluations

I. Strandén[1*], J. ten Napel[2], R.F. Veerkamp[2], R. Evans[3], S. Naderi[3], E.A. Mäntysaari[1] and J. Vandenplas[2]

[1]Natural Resources Institute Finland (Luke), Jokioinen, Finland; [2]Wageningen University & Research, P.O. Box 338, 6700 AH Wageningen, the Netherlands; [3]Irish Cattle Breeding Federation, Highfield House, Newcestown Road, Bandon, Cork, Ireland; ismo.stranden@luke.fi

Abstract

Computing time and memory requirements increase with single-step methods to estimate genomic breeding values when the number of genotyped animals increases. Computational costs can be reduced by omitting genotypes of animals without phenotype and progeny, often the candidate animals for selection. Indirect prediction of a candidate animal GEBV can be based on the animal's genotype and SNP marker solutions (DGV). Alternatively, the sum of DGV and residual polygenic (RPG) effect can be computed, denoted GRV (Genomic and Residual polygenic Value). We applied indirect genomic prediction for a 6 trait calving difficulty evaluation. There were 1.50 million genotyped animals of which 36% were considered candidate animals. Based on our results, DGV showed high accuracy but also bias due to omitting the RPG effect. The GRV prediction had high accuracy and low bias. Computing time was reduced by 33%.

Introduction

Practical implementation of single-step genomic BLUP (ssGBLUP; Aguilar *et al.* 2010; Christensen and Lund 2010) has met many challenges (e.g. Mäntysaari *et al.* 2020; Misztal *et al.* 2020). Continual increases in volumes of genotyped animals also increase the computational costs. A practical aspect seldom considered is the efficient calculation of genomic estimated breeding values (GEBV) for genotyped animals without own and progeny information, so called candidate animals. Excluding these genotypes in the evaluation and estimating their GEBVs afterwards may reduce computational costs (Tsuruta *et al.* 2021). Including these genotypes in the routine evaluation may even be undesirable (Koivula *et al.* 2018). Liu *et al.* (2016) presented formulas for predicting GEBVs of genotyped candidate animals when the model has a residual polygenic (RPG) effect. The aim of this study was to compare the performance of approaches for indirect estimation of GEBV for genotyped candidate animals.

Materials & methods

The ssGTABLUP approach. Mäntysaari *et al.* (2017) presented an efficient computational approach for ssGBLUP named ssGTABLUP where the genomic relationship matrix has the form $\mathbf{G}_C = \mathbf{G}_0 + \mathbf{C}$. Here, $\mathbf{C} = w\mathbf{A}_{gg}$, w is the RPG proportion and $\mathbf{A}_{gg}$ is the pedigree-based relationship matrix among the genotyped animals. The genomic part is $\mathbf{G}_0 = \mathbf{ZBZ}'$ where Z is an $n \times m$ matrix of centered marker genotypes, B is an $m \times m$ diagonal scaling matrix, n is the number of genotyped animals and m is the number of SNP markers. In the VanRaden (2008) method 1, the centering uses base population allele frequencies $p_i, i = 1, \ldots, m$ and the scaling matrix is $\mathbf{B} = \mathbf{I}\frac{1-w}{k}$ with the scaling constant $k = 2\sum_{i=1}^{m} p_i(1-p_i)$.

The $\mathbf{G}_C^{-1}$ matrix in the model equations (MME) can be calculated as $\mathbf{G}_C^{-1} = \frac{1}{w}\mathbf{A}_{gg}^{-1} - \frac{1}{w^2}\mathbf{A}_{gg}^{-1}\mathbf{ZK}^{-1}\mathbf{Z}'\mathbf{A}_{gg}^{-1}$ where $\mathbf{K} = \frac{1}{w}\mathbf{Z}'\mathbf{A}_{gg}^{-1}\mathbf{Z} + \mathbf{B}^{-1}$ is a symmetric positive definite matrix. In practice, iterative methods can be used to solve the MME where the $\mathbf{G}_C^{-1}$ matrix is not explicitly computed and the needed product $\mathbf{G}_C^{-1}\mathbf{v}$ is computed efficiently (Mäntysaari *et al.* 2020).

R.F. Veerkamp and Y. de Haas (eds)

DOI: 10.3920/978-90-8686-940-4_360, © I. Strandén *et al.* 2022

Let vector $\mathbf{u}_g$ ($\mathbf{u}_n$) have the single-step estimates for the genotyped (non-genotyped) animals. Assuming an RPG effect, the GEBVs for non-genotyped and genotyped animals can be decomposed as (Fernando *et al.* 2014; Fernando *et al.* 2016; Liu *et al.* 2016):

$$\mathbf{u} = \begin{bmatrix}\mathbf{u}_n \\ \mathbf{u}_g\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{Zg} + \boldsymbol{\epsilon} + \mathbf{d}_n \\ \mathbf{Zg} + \mathbf{d}_g\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{Zg} \\ \mathbf{Zg}\end{bmatrix} + \begin{bmatrix}\boldsymbol{\epsilon} \\ \mathbf{0}\end{bmatrix} + \begin{bmatrix}\mathbf{d}_n \\ \mathbf{d}_g\end{bmatrix} \quad (1)$$

where $\mathbf{A}_{ng}$ is the pedigree-based relationship matrix between the non-genotyped and genotyped animals, ε is the imputation residual, vector: $\mathbf{d} = \begin{bmatrix}\mathbf{d}_n \\ \mathbf{d}_g\end{bmatrix}$ corresponds to RPG, the part of the genetic effects not explained by the genomic data (Christensen and Lund 2012). Thus, genotyped animal GEBV has two components: (1) Zg, that is the direct genetic value (DGV) due to the marker effects; (2) d_g, that is the breeding value due to the RPG effect.

Indirect GEBVs of genotyped animals without own and progeny information. Computational costs in any ssGBLUP can be reduced using a two-step algorithm: (1) calculate GEBV without the candidate animals; (2) predict GEBV of the candidate animals. According to Equation 1, we need to compute $\hat{\mathbf{u}}_c = \mathbf{Z}_c\hat{\mathbf{g}} + \hat{\mathbf{d}}_c$ where subscript *c* refers to the candidate animals. The marker solutions can be easily calculated in ssGTABLUP after solving the MME using the formula $\hat{\mathbf{g}} = \frac{1}{w}\mathbf{K}^{-1}\mathbf{Z}'\mathbf{A}_{gg}^{-1}\hat{\mathbf{u}}_g$ (Liu *et al.* 2014).

Calculation of DGV for a candidate animal is straightforward but the RPG effects $\hat{\mathbf{d}}_c$ are unknown. According to Liu *et al.* (2016), $\hat{\mathbf{d}}_c = \mathbf{A}_{cg}\mathbf{A}_{gg}^{-1}\hat{\mathbf{d}}_g$ where $\hat{\mathbf{d}}_g = \hat{\mathbf{u}}_g - \mathbf{Z}\hat{\mathbf{g}}$ and $\mathbf{A}_{cg}$ is the pedigree-based relationship matrix between the genotyped candidate animals and the genotyped animals already included in the single-step evaluation. Calculation of $\hat{\mathbf{d}}_c$ can be done in two steps. First, $\mathbf{x} = \mathbf{A}_{gg}^{-1}\hat{\mathbf{d}}_g$ which can be done efficiently using sparse matrices (Strandén *et al.* 2017). Second, $\hat{\mathbf{d}}_c = \mathbf{A}_{cg}\mathbf{x}$ which can be computed efficiently (Colleau 2002) using the full pedigree-based relationship matrix times vector product:

$$\begin{bmatrix}\mathbf{b}_n \\ \mathbf{b}_g \\ \hat{\mathbf{d}}_c\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{nn} & \mathbf{A}_{ng} & \mathbf{A}_{nc} \\ \mathbf{A}_{gn} & \mathbf{A}_{gg} & \mathbf{A}_{gc} \\ \mathbf{A}_{cn} & \mathbf{A}_{cg} & \mathbf{A}_{cc}\end{bmatrix}\begin{bmatrix}\mathbf{0} \\ \mathbf{x} \\ \mathbf{0}\end{bmatrix}.$$

Data and models. The data and associated variance components were extracted from the 6 trait calving difficulty evaluation for Irish beef cattle performed by ICBF. There were 9.54 million animals in the pedigree file of which 5.76 million had a data record. There were 50,240 SNP markers available from 1.50 million genotyped animals. The animals had been genotyped using the Illumina Bovine SNP50 Bead Chip (Illumina, San Diego, CA). The MME had 123.32 million unknowns of which 114.50 million were additive and maternal genetic effects.

Study design. GEBVs were estimated with two data sets. The full data had all genomic information. In the reduced data, genomic information from the candidate animals were removed. There were 533,116 candidate animals, i.e. the reduced data had 965,868 genotyped animals. For both evaluations, the genomic relationship matrix was based on VanRaden method 1 with a 20% RPG proportion. The Z matrix was centered using base population allele frequencies estimated by the Bpop program (Strandén and Mäntysaari 2020).

Candidate animal GEBV was estimated by three approaches: (1) PA: mean of parent GEBVs; (2) DGV: $\mathbf{Z}_c\hat{\mathbf{g}}$ using estimated SNP marker effects $\hat{\mathbf{g}}$ from the reduced data analysis; and (3) GRV: formula $\mathbf{Z}_c\hat{\mathbf{g}} + \mathbf{A}_{cg}\mathbf{A}_{gg}^{-1}\hat{\mathbf{d}}_g$. Comparison of these candidate animal GEBV with the full data GEBV for the direct and maternal effects was based on a Pearson correlation, linear regression, and level bias for each trait separately. Regression was from the full data GEBV on the reduced data GEBV. Level bias for trait *j* was defined as average of $(\hat{\mathbf{u}}_c - \hat{\mathbf{u}}_{c,full})/\boldsymbol{\sigma}_j$ where $\hat{\mathbf{u}}_c$ and $\hat{\mathbf{u}}_{c,full}$ are GEBV from the reduced and full data analysis, respectively, and σ_j is genetic standard deviation of the trait.

Solver. The ssGTABLUP MME were solved using preconditioned conjugate gradient (PCG) iteration by MiX99 software with 10 CPU threads (Strandén *et al.* 2018). The PCG method was assumed to be converged when $\sqrt{(\mathbf{Cs}^{[k]}-\mathbf{r})'(\mathbf{Cs}^{[k]}-\mathbf{r})/\mathbf{r}'\mathbf{r}} < 10^{-7}$ where C is the coefficient matrix of MME, $s^{[k]}$ is vector of solutions at round *k*, and r is the MME right-hand side vector.

Results

Preprocessing and solver computing times were reduced effectively in the same proportion as the number of genotyped animals used in the full vs reduced data (Table 1). For the non-candidate animals and both the direct and maternal effects, correlations in GEBV between the full and the reduced evaluations were between 98.9 and 99.8% for the genotyped animals and between 99.0 and 99.6% for the non-genotyped animals across traits. The indirect GEBVs approximated by GRV gave the highest correlations, the lowest level biases, and no over- or under-dispersion (i.e. regression coefficient close to: (1), across all traits and genetic effects (Table 2 and 3). In comparison, DGV had correlations close to 1 (on average 0.97), but the regression coefficients indicated under-dispersion and level biases were higher.

Discussion

Computing time was reduced effectively by omitting genotypes of candidate animals because their number was large. In the indirect GEBV prediction, DGV showed high accuracy but consistent bias due to omitting the RPG contribution. In contrast, GRV gave high correlations and low bias. When the RPG weight is smaller than in our study (e.g. 5%), DGV and GRV are likely to give similar values. The GRV approach can also be used in intermediate GEBV prediction of newly genotyped animals when the full data single-step evaluations are performed at longer intervals such as once every 4 months.

Table 1. Statistics of solving ssGTABLUP using the full and reduced data.

Data	Pre time[1]	RAM[2]	N[3]	Solver time	Indirect time	Total time
Full	9.6h	592	665	12.2h	-	21.8h
Reduced (GRV)	5.5h	390	664	7.5h	0.7h	13.7h

[1] Preprocessing computing time.
[2] Peak RAM in GB.
[3] Number of PCG iterations.

Table 2. Correlations, regression coefficients, and level bias for direct GEBV of genotyped candidate animals computed from the full data vs reduced data.

		Trait 1	2	3	4	5	6
Correlation	PA	0.737	0.775	0.687	0.656	0.671	0.713
	DGV	0.978	0.981	0.971	0.966	0.969	0.982
	GRV	0.995	0.995	0.996	0.996	0.997	0.997
Regression coefficient	PA	0.968	0.964	0.942	0.938	0.972	0.970
	DGV	1.036	1.045	1.042	1.033	1.038	1.043
	GRV	0.994	0.993	1.002	1.003	1.004	1.000
Level bias	PA	0.021	0.022	0.027	0.026	0.022	0.014
	DGV	0.006	-0.004	0.026	0.038	0.038	0.001
	GRV	0.004	0.003	-0.002	-0.004	0.001	-0.005

Table 3. Correlations, regression coefficients, and level bias for maternal GEBV of genotyped candidate animals from the full data vs reduced data.

		Trait 1	2	3	4	5	6
Correlation	PA	0.675	0.662	0.659	0.647	0.725	0.720
	DGV	0.976	0.976	0.979	0.978	0.983	0.983
	GRV	0.988	0.989	0.996	0.995	0.997	0.994
Regression coefficient	PA	0.977	0.961	0.959	0.952	0.973	0.985
	DGV	1.043	1.033	1.037	1.028	1.036	1.050
	GRV	1.004	0.997	1.003	1.002	0.997	0.999
Level bias	PA	0.011	0.005	0.016	0.015	0.005	0.008
	DGV	0.026	0.028	0.017	0.021	0.052	0.035
	GRV	0.006	-0.001	0.004	0.002	0.000	0.005

References

Aguilar I., Misztal I., Johnson D.-L., Legarra A., Tsuruta S, and Lawlor T.J. (2010) J. Dairy Sci. 93:743-752. https://doi.org/10.3168/jds.2009-2730

Christensen O.F., and Lund M.S. (2010) Genet. Sel. Evol. 42:2. https://doi.org/10.1186/1297-9686-42-2

Colleau J.J. (2002) Genet. Sel. Evol. 34:409-421. https://doi.org/10.1186/1297-9686-34-4-409

Koivula M., Strandén I., Aamand G.P., and Mäntysaari E.A. (2018) J. Animal Breed. Genet. 135:107–115. https://doi.org/10.1111/jbg.12318

Liu Z., Goddard M.E., Reinhardt F., and Reents R. (2014) J. Dairy Sci. 97:5833-5850. https://doi.org/10.3168/jds.2014-7924.

Liu Z., Goddard M.E., Hayes B.J., Reinhardt F., and Reents R. (2016) J Dairy Sci 99:2016-2025. https://doi.org/10.3168/jds.2015-10394

Misztal I., Lourenco D., and Legarra A. (2020) J. Animal Sci. 98: 1-14. https://doi.org/10.1093/jas/skaa101

Misztal I., Lourenco D., Legarra A. (2020) J. Animal Sci. 98: 1-14. https://doi.org/10.1093/jas/skaa101

Mäntysaari E.A., Evans R.D., and Strandén I. (2017) J. Animal Sci. 95:4728-4737. https://doi.org/10.2527/jas2017.1912.

Mäntysaari E.A., Koivula M., and Strandén I. (2020) J. Dairy Sci. 103:5314-5326. https://doi.org/10.3168/jds.2019-17754

Strandén I., Matilainen K., Aamand G., and Mäntysaari E.A. (2017) J. Animal Breed. Genet., 134:264-274. https://doi.org/10.1111/jbg.12257

Strandén I., Taskinen M., Matilainen K., Lidauer M., and Mäntysaari E.A. (2018) Proc. of the 11th WCGALP, Auckland, New Zealand.

Strandén I., Mäntysaari E.A. (2020) Agr. Food Sci. 29:166-176. https://doi.org/10.23986/afsci.90955

Tsuruta S., Lourenco D.A.L., Masuda Y., Lawlor T.J., and Misztal I. (2021) J. Dairy Sci. Comm. 2:356-360. https://doi.org/10.3168/jdsc.2021-0097

VanRaden P.M. (2008) J. Dairy Sci. 91: 4414-4423. https://doi.org/10.3168/jds.2007-0980.