Effects of Selective Genotyping and Selective Imputation in Single-Step GBLUP

C. Edel, E.C.G. Pimentel, L. Plieschke, R. Emmerling and K.-U. Götz

Bavarian State Research Center for Agriculture, 85586 Poing-Grub, Germany

Abstract

Single-Step genomic prediction has been advocated to be the next logical step in the development of large routine applications. Despite its intriguingly simple concept, there are still several problems to be solved from a theoretical standpoint. One problem is the inflation of genomic predictions frequently occurring in practical applications. Ad-hoc remedies have been proposed, like the use of scaling-factors when building the **H** matrix or pruning of data and pedigrees used for prediction. In this investigation we suppose that selective genotyping and selective imputation in Single-Step GBLUP are major components contributing to inflated predictions. Using the general reformulation of the Single-Step model given by R. Fernando as an illustration, single-step GBLUP can be conceptually divided into two separate steps of estimation: step one is the estimation of gene-contents for all animals in the pedigree from observed genotypes and step two is the estimation of SNP-effects using observed and imputed genotypes and the available phenotypic data. In recent studies we have examined the effect of selective genotyping and the role of genotypes without phenotypes in Single-Step GBLUP in simulation studies. Our conclusion is that selective genotyping and the selective quality of genotype imputation can lead to biased genomic estimates in Single-Step GBLUP. In order to support our argument we demonstrate the effect of the inclusion or exclusion of older birth years of genotyped bulls showing strong evidence for selective genotyping on two-step and single-step predictions using real Fleckvieh data and discuss the vital importance of the difference between exclusion of genotypes and exclusion of phenotypes and genotypes.

Key words: Single-step GBLUP, genomic breeding values, inflation of genomic predictions

Introduction

Selective genotyping describes a situation where, intentionally or not, animals are selected for genotyping based on a criterion that includes Mendelian Sampling information. In a recent publication we have investigated the effect of selective genotyping on the quality of genomic predictions. Using simulation we found that using reference animals affected by selective genotyping had a strong impact in reducing validation reliabilities of genomic predictions and resulted in highly inflated estimates in two-step genomic prediction (Plieschke *et al.,* 2016). The impact of selective genotyping in the framework of single-step GBLUP was, however, not covered by that investigation. Predicting genomic breeding-values with single-Step GBLUP can conceptually be divided into two estimation steps, where the first one is the imputation of unobserved genotypes using existing genotypes and the second one is the estimation of genomic breeding values based on a reference of animals with either imputed or observed genotypes (Fernando *et al.,* 2014). As a consequence genotyped animals without phenotypic information can influence genomic predictions by improving the imputation quality of their ungenotyped ancestors, an aspect that may not be obvious at first glance. Recently we have explored this mechanism by comparing genotypes implicitly imputed within single-step GBLUP with their known true state by the use of simulation (Shabalina *et al.,* 2017). We found that, depending on the number of genotyped offspring, the accuracy of imputation shows variation and can reach high values, virtually imputing the true genotype of an ungenotyped ancestor. In real world populations only successful sires which are highly selected, have many (genotyped) offspring and are thus imputed reliably. It is this aspect that we refer to as 'selective imputation' throughout this paper.

In this investigation we added further evidence from empirical data to our observations derived from the preceding simulations. The investigation intends to test a set of hypotheses: 1) Selective genotyping has a strong impact on two-step as well as on single-step genomic predictions, and 2) selective genotyping in single-step genomic prediction might occur in two ways, either directly, or indirectly, due to selective imputation. Presence and effects of selective imputation are not easily detected.

Materials and Methods

All results were derived within the framework of a standard forward-prediction validation scenario. Deregressed proofs, derived from the April 2013 routine evaluation for Fleckvieh (FV) production traits (milk-, fat- and protein-yield, MY, FY, PY) were used as phenotypes in either two- or single-step analyses. Resulting genomic proofs were compared to April 2017 deregressed proofs of validation bulls. Results presented are the slopes of the Interbull GEBV test (Mäntysaari *el al.*, 2012) and the so called realized reliabilities (VanRaden, 2009). The set of validation bulls was kept constant in all analyses. Several deregression methods were tested in advance to provide appropriate phenotype-aggregation (results not shown). We used deregressed proofs calculated beforehand with a full conventional model for either full- or pruned analyses. This was done to keep the quality of the phenotypes used constant and to be able to focus on the specific effects of the inclusion or exclusion of groups of reference animals in either single- or two-step analyses. This approach excludes potentially problematic effects when pruning raw phenotypes like for example changing definitions and estimates for fixed effects and/or genetic groups. A justification and strategy for absorbing the phenotypes of ungenotyped offspring into the proofs of genotyped parents and using these aggregated phenotypes within single-step evaluations was given by Meuwissen *et al.* (2011). In two-step analyses reference groups consisted entirely of genotyped bulls. In single-step analyses bull and cow genotypes were used and restrictions were only applied on birth years.

Two-step analyses were done using standard weighted GBLUP. Before calculating the matrix **G**, the matrix of gene-contents was centered by subtracting two times the base-allele frequency (Gengler *et al.*, 2008). After solving for direct genomic breeding values, these were blended with conventional estimates to produce GEBV (for details see Edel *et al.*, 2011). Single-step analyses were performed by an application of Fernando's single-step SNP-BLUP model (Fernando *et al.*, 2014). This approach uses an explicit step of genotype imputation followed by a solution step for SNP-effects. It has been shown that this approach provides an equivalent model to standard single-step BLUP using the **H** matrix (Fernando *et al.*, 2014; Taskinen *et al.*, 2017). However, application of the 'Fernando-model' additionally allows for analyzing in detail the quality of the genotype imputation that is an implicit part of single-step prediction. Both in single- and two step analyses the same amounts of residual polygenic variance from the FV routine application were used (MY: 20%, FY: 15%, PY: 25%).

We additionally present two-step results that are scaled according to an approach that has been developed for the use in the FV routine application. This scaling aims at correcting for the too large dispersion of genomic predictions that is assumed to arise from selective genotyping in reference animals.

As mentioned in the introduction, a fundamental assumption to investigate was that in single-step applications selective genotyping might also arise as an effect of selective genotype imputation. To test whether this assumption holds, we investigated two approaches of pruning data influenced by selection: 1.) pruning of all data including phenotypes and genotypes, and 2.) pruning genotypes only. If our assumptions hold, we would expect the pruning of genotypes being relatively inefficient in controlling negative effects of selective genotyping.

Results

Selective Genotyping

When focusing on reference bulls, selective genotyping is predominantly occurring in older birth years. Figure 1 summarizes the most relevant aspects. In the pre-genomic era 500-600 bulls per year were progeny tested in the FV population. All progeny-tested bulls were genotyped back to birth year 1998 to establish the FV reference population. Prior to 1998 only samples of second-crop bulls with high impact on the population were available.

Two-step analyses

Table 1 summarizes the results of two-step analyses. Omitting selected genotypes in two-step analysis ('pruned' vs. 'raw') has a beneficial effect on the slope of the GEBV test and reduces the inflation of estimates. In this respect it has an effect very much comparable to the scaled estimates ('scaled') that are produced in the Fleckvieh routine evaluation. Scaling does not influence realized reliabilities since it does not change the ranking of animals, whereas with pruning some information is lost and reliabilities are slightly reduced.

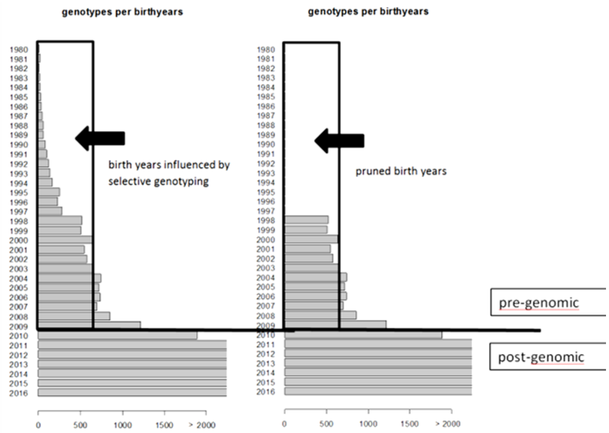

Figure 1. Fleckvieh bull reference population: Birth year before 1998 are influenced by selective genotyping.

Table 1. Slopes of the Interbull GEBV-test (b_1) and realized reliabilities (Rel_{real}) for the two-step analyses.

	b_1		Rel_{real}	
	MY	PY	MY	PY
raw	.87	.89	63	63
scaled	.93	.96	63	63
pruned	.92	.94	62	61

Single-step analyses

In the unpruned and unscaled version ('raw') single-step analysis shows an inflation of estimates well above the level observed with two-step analyses (table 2). This degree of inflation is substantially reduced when pruning phenotypic data before birth years 1998 (this was achieved by using only phenotypes of daughters of bulls born after 1997 to be comparable to two-step analyses). At least in one trait pruning has also a beneficial effect on realized reliabilities. In contrast to that finding, an approach only omitting the genotypes of animals born before 1998 (pruned:G in table 2) has no noticeable effect on reducing inflation.

Table 2. Slopes of the Interbull GEBV-test (b_1) and realized reliabilities (Rel_{real}) for the single-step analyses.

	b_1		Rel_{real}	
	MY	PY	MY	PY
raw	.81	.80	64	61
pruned: P	.91	.89	65	61
pruned: G	.82	.80	64	61

Discussion

The results support our two general hypotheses. Selective genotyping has a strong influence on genomic predictions by generating inflated estimates. This negative effect can also be observed in single-step analyses. Moreover, it seems to be amplified by a mechanism we called 'selective imputation' here. This indicates that single-step systems selectively restore information in the block of ungenotyped ancestors via imputation. Since only genotypes of selected parents can be restored efficiently by imputation, the negative effects on estimates of selective traits seem to be more or less inevitable in selected populations. We presented here a very strict pruning of data as one possible solution. This pruning actually implies to omit all data (phenotypes and consequently genotypes also) from a time before the start of a regular genotyping of unselected birth years. From our experience there seems to be no other approach similarly effective in reducing the effects of selective genotyping, although there had been some rather general proposals (Vitezica *et al.*, 2011) in the past.

It might be argued that the influence of older birth years of selectively genotyped animals might decrease over time. In fact, empirical observations from the FV routine evaluation support this hypothesis (not shown). However, the genotyping of potentially preselected females as possible candidates for selection in the bull-dam path has begun only recently in our population and females are currently not used in our routine reference (two-step). Even in our single-step test-application this group of females still has no strong impact on our forward-prediction validations. The answer to the question of whether opening the reference population or not (either directly or by introducing single-step technology) however critically depends on the impact of this group of animals on our estimates. We hopefully will be able to continue and extend our investigations with more genotyped female in the near future.

References

Edel, C., Schwarzenbacher, H. , Hamann, H., Neuner, S., Emmerling, R. & Götz, K.-U. 2011. The German-Austrian genomic evaluation system for Fleckvieh (Simmental) cattle. *Interbull Bulletin 44*, 152-156.

Fernando, R.L., Dekkers, J.C.M. & Garrick, D.J. 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.*

Gengler, N., Mayeres, P. & Szydlowski, M. 2008. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal 1,* 21–28.

Mäntysaari, E., Liu, Z. & VanRaden, P. 2010. Interbull Validation Test for Genomic Evaluations. *Interbull Bulletin 41*, 17-22.

Meuwissen, T.H.E., Luan, T. & Wooliams, J.A. 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. *J. Anim. Breed. Genet. 128,* 429–439.

Plieschke, L., Edel, C., Pimentel, E.C.G., Emmerling, R., Bennewitz, J. & Götz, K.-U. 2016. Systematic genotyping of groups of cows to improve genomic estimated breeding values of selection candidates. *Genet Sel Evol. 48:73.*

Shabalina, T., Pimentel, E.C.G, Edel, C., Plieschke, L., Emmerling, R. & Götz, K.-U. 2017. Short communication: The role of genotypes from animals without phenotypes in single-step genomic evaluations. *J. Dairy Sci. 100,* 8277–8281.

Taskinen, M., Mäntysaari, E. & Stranden, I. 2017. Single‑step SNP‑BLUP with on‑the‑fly imputed genotypes and residual polygenic effects. *Genet Sel Evol 49:36.*

VanRaden, P.M., Van Tassell, C.P.,Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F. & Schenkel, F.S. 2008. Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci. 92.* 16–24.

Vitezica, Z.G., Aguilar, I., Misztal, I. & Legarra, A. 2011. Bias in genomic predictions for populations under selection. *Genet. Res. 93:5,* 357-366.