Short communication: The effect of genotyping cows to improve the reliability of genomic predictions for selection candidates

C. Edel,¹ E. C. G. Pimentel, L. Plieschke, R. Emmerling, and K.-U. Götz
Institute of Animal Breeding, Bavarian State Research Center for Agriculture, 85586 Grub, Germany

ABSTRACT

In this study we investigate the potential of enlarging the reference population for genomic prediction in dairy cattle by routinely genotyping a random sample of the first-crop daughters of every AI bull in the breeding program. We analyzed small nuclear pedigrees, each consisting of a genotyped selection candidate and 3 generations of genotyped male ancestors. Genotypes were taken from the genomic routine evaluation of Fleckvieh cattle in Germany and Austria. The phenotypic information of a daughter of any one male in each of these pedigrees was either considered to be part of the daughter yield deviation of the corresponding sire, or was assumed to be an individually observed genotyped daughter of this sire. Daughter genotypes in this case were simulated from phased haplotypes of their sires and random maternal gametes drawn from a haplotype library. We measured the gain from genotyping daughters as the increase in model-based theoretical reliability of the genomic prediction for a putative selection candidate. We expressed the improvements as a marginal increase, corresponding to an increase in reliability at a reliability baseline level of zero, to simplify comparisons. Results were encouraging with 2 to 40% of marginal reliability increase for selection candidates depending on the assumed heritability of the trait and the number of daughters modeled to be genotyped in the design.

Key words: genomic selection, reference population, genotyping of cows

Short Communication

For dairy cattle populations using genomic breeding value estimation, it can be assumed that all available bulls have already been genotyped and in many cases international collaborations have exploited all the in-

formation about males in other populations. A further increase in the reliability of genomic predictions can then only be achieved by the annual growth of the sire pool, which is often limited. Another option is to enlarge the reference population by genotyping cows with phenotypes (Calus et al., 2013; Thomassen et al., 2014). At present genotyping costs, a complete genotyping and integration of all cows under milk recording is not feasible. Furthermore, deliberately including potentially preselected cows, selected for genotyping by farmers, may introduce a bias to genomic breeding value estimation (Dassonneville et al., 2012).

In this pilot investigation, we analyze the potential of routinely genotyping samples of first crop daughters of AI bulls. We quantify the effects of such an approach in particular with respect to the reliability of genomic breeding values of current male selection candidates. Other possible advantages of genotyping cows (e.g., for the selection of bull dams) are not covered.

Our approach uses nuclear pedigree structures from the routine genomic evaluation of the Fleckvieh breed in Germany and Austria, each consisting of a genotyped selection candidate with genotyped male ancestors reaching back for 3 generations including sire, paternal, and maternal grandsires and all great-grandsires. From several thousands of these core-structures in our Fleckvieh genotype pool data (Illumina Bovine SNP50) we randomly selected 100 for the purposes of this paper. Phased haplotypes for chromosomes 1 to 29 as well as for the pseudo-autosomal part of chromosome X for all male individuals were taken from the results of the imputation procedure done in routine evaluation, which is done with findhap v2 (VanRaden et al., 2011). We combined these haplotypes to genotypes and constructed standard genomic relationship matrices **G** (VanRaden, 2008) for each nuclear pedigree separately using base-allele frequency estimates (Gengler et al., 2007) from the routine evaluation. In the next step, we calculated model-derived genomic reliabilities (R^2) based on a selection-index approach; that is, we calculated

$$R^2 = \frac{\text{diag}(\mathbf{C}\mathbf{P}^{-1}\mathbf{C}')}{\text{diag}(\mathbf{C})},$$

Received August 12, 2015.

Accepted November 4, 2015.

¹Corresponding author: Christian.Edel@LfL.bayern.de

with

$$\mathbf{C} = \mathbf{G}\sigma_A^2$$

and

$$\mathbf{P} = \mathbf{G}\sigma_A^2 + \mathbf{R}\sigma_E^2,$$

where $\mathbf{R}$ is a diagonal matrix with diagonal elements $1/W_i$ and

$$w_i = s^* N_{D_i}$$

with

$$s = \frac{1 - h^2}{4 - h^2},$$

where h^2 is the assumed heritability of the trait and N_{D_i} is the assumed number of daughter performances aggregated to the i th daughter yield deviation (**DYD**; we assumed each cow to have exactly one observation, so the i th DYD comprises N_{D_i} unique performances of N_{D_i} daughters). For each selection candidate, we calculated a theoretical model-based reliability of the genomic prediction given the nuclear $\mathbf{G}$ matrix and the assumed data structure. In the following text, we refer to this reliability as the base- $R^2(R_b^2)$ because here only males were assumed to be genotyped and all phenotypes of females were aggregated in the DYD of their sires, accordingly.

In the second step, some or all of the daughters constituting the DYD of the i th male were modeled explicitly as genotyped. Because no genotypes were available for females, they were simulated by simulating a meiosis based on the 2 haplotypes of each chromosome of its sire (ignoring crossing-over events). The resulting pseudo-gamete was then combined with another pseudo-gamete randomly chosen from a library of 10,000 female gametes built beforehand from a sample of the maternal gametes of all selection candidates from the routine evaluation to virtually produce a new daughter genotype. To avoid these generated daughter genotypes from eventually consisting of the same maternal gamete as a genotyped member of any of the nuclear pedigrees investigated, the respective maternal gametes were excluded when building the library. The new “synthetic” daughter genotypes were then incorporated into matrices $\mathbf{G}$ and $\mathbf{R}$, and the reliability for the selection candidate in this extended design (R_e^2) was calculated. To compensate for daughters modeled individually as

genotyped, a corresponding information equivalent was subtracted from the assumed reliability of her father's DYD. This approach was possible because we were only interested in reliabilities of genomic breeding values, which depend only on the structure of the information.

Simply reporting R_e^2 versus R_b^2 values for selection candidates can be misleading, especially when R_b^2 are relatively low as in the case of genomic predictions using the small nuclear pedigree structures as a reference population as it is done in this investigation. The effect of additionally genotyping daughters and including them into the reference population on reliabilities depends very much on the level of reliability achieved with the baseline genomic system. Therefore, we chose to express the increase in reliability as marginal reliability (R_m^2) with the aim of becoming independent of the base R^2 level. This R_m^2 is calculated as

$$R_m^2 = \frac{R_e^2 - R_b^2}{R_e^2 R_b^2 + 1 - 2R_b^2}$$

(Harris and Johnson, 1998). Marginal R^2 can be interpreted as the increase in reliability at a base R^2 level of zero. To facilitate the translation of results presented here to more reasonable levels of base R^2 , we included 2 graphs that illustrate the effective increase resulting from adding various R_m^2 values to increasing levels of base R^2 (Figures 1 and 2).

For each of the designs investigated and described in the following paragraphs, we modeled 3 scenarios. (1) In the sire (**S**) scenario, only the sire of candidates was assumed to have daughters phenotyped and genotyped, whereas all other males in the nuclear pedigree had only phenotyped daughters. This mimics a situation where only first-crop daughters of the most recently used sires are available for genotyping, and would be the typical situation encountered in the first years after starting a genotyping program for females. (2) In the sire, grandsire, great-grandsire (**S-GS-GGS**) scenario, every bull in the nuclear pedigree had the same number of genotyped daughters given by the respective design. This scenario was conceived to investigate a situation where the genotyping program has already been running for several generations. However, in such a situation, at least a fraction of sires of selection candidates will most likely be descendants of genomic young bulls without daughter information themselves. (3) Therefore, in the grandsire, great-grandsire (**GS-GGS**) scenario, we modeled genotyped and phenotyped daughters for grandsires and great-grandsires of selection candidates only, whereas the sires had no daughter information at all.

Table 1. Summary of the designs investigated¹

Design	Scenario	h^2	gs/ggs generation		Sire generation	
			nDau	nGeno	snDau	snGeno
D_P	S	0.35	1,000	0	50/100/200/500	50/100/200/500
	GS-GGS	0.35	1,000	50/100/200/500	0	0
	S-GS-GGS	0.35	1,000	50/100/200/500	50/100/200/500	50/100/200/500
D_F	Like D_P except $h^2 = 0.05$					
D_{CM}	S	0.25	200	0	50/100	50/100
	GS-GGS	0.25	200	50/100	0	0
	S-GS-GGS	0.25	200	50/100	50/100	50/100
D_{CL}	Like D_{CM} except $h^2 = 0.10$					

¹nDau is the number of daughters assumed to be phenotyped in the grandsire/great-grandsire (gs/ggs) generations. nGeno denotes the number thereof genotyped. snDau and snGeno are numbers of phenotyped and genotyped daughters assumed for the sires of selection candidates accordingly. S denotes a scenario where only the sires of candidates were assumed to have daughters phenotyped and genotyped, whereas in scenario GS-GGS genotyped and phenotyped daughters were modeled for grandsires and great-grandsires of selection candidates only. In scenario S-GS-GGS, every bull in the nuclear pedigree had the same number of genotyped daughters given by the respective design.

To assess the increase in R^2 resulting from the 3 scenarios described above, we investigated a set of trait designs, differing in heritabilities and the availability of phenotypes (additionally summarized in Table 1).

Figure 1. Example for the effect of adding a marginal reliability. The example is for design D_P with 100 daughters genotyped (see also Table 2). Marginal reliabilities of 0.17, 0.10, and 0.07 as found for scenarios S-GS-GGS (every bull in the nuclear pedigree with genotyped daughters), GS-GGS (genotyped daughters for grandsires and great-grandsires of selection candidates only), and S (only sires of candidates have genotyped daughters), respectively, will result in an increase of reliability of about 0.09, 0.05, and 0.04 at a base- R^2 level of 0.3, or an increase of about 0.03, 0.02, and 0.01 when the base R^2 level is already at 0.6.

In performance and fitness trait (D_P , D_F) designs, we tried to mimic a typical situation for traits with no limitation in the number of phenotypes. In the basic design each male of the grandsire and great-grandsire generation was assumed to have 1000 phenotyped daughters without genomic information. In the extended designs we generated genotypes for an increasing number of these daughters (50/100/200/500) and included them into the system as described above. Sires of candidates were either assumed to be genomic young sires with no

Figure 2. General illustration of the decreasing effect of adding a marginal reliability (R_m^2 , per definition at 0 base R^2 level) to an increasing base R^2 level.

Table 2. Marginal gain in reliability [R_m^2 ($\times 100$)] for the performance and the fitness trait design (D_P , D_F), respectively, for increasing numbers of genotyped daughters per sire (nGeno)¹

nGeno	D_P ($h^2 = 0.35$)			D_F ($h^2 = 0.05$)		
	Genotyped daughters for			Genotyped daughters for		
	S	GS-GGS	S-GS-GGS	S	GS-GGS	S-GS-GGS
50	4 (0/7)	6 (5/10)	10 (5/16)	1 (0/2)	1 (1/2)	2 (0/3)
100	7 (1/13)	10 (7/15)	17 (11/26)	1 (0/3)	2 (1/3)	3 (1/5)
200	11 (6/18)	17 (13/24)	26 (20/35)	2 (1/4)	4 (2/5)	5 (3/7)
500	19 (10/28)	27 (22/35)	39 (32/52)	5 (2/8)	7 (5/10)	11 (8/17)

¹Values shown are averages (minimum/maximum values in parentheses) across 100 randomly selected nuclear pedigrees. S = only sires of candidates have genotyped daughters. GS-GGS = genotyped daughters for grand-sires and great-grand-sires of selection candidates only. S-GS-GGS = every bull in the nuclear pedigree with genotyped daughters.

daughters or sires with first crop daughters (see description of scenarios below). Sires with first crop daughters had 50, 100, 200, or 500 daughters with genotypes, corresponding to the situation assumed for the grandparental generation. This approach was chosen to mimic an increasing number of first crop daughters becoming phenotyped and genotyped simultaneously.

In confirmation trait designs (D_{CM} and D_{CL}), capacity for phenotyping is limited for some traits. We investigated 2 scenarios with low and moderate heritability (0.10 and 0.25). In the basic design, we assumed 200 daughters phenotyped per sire in the grandsire and great-grand-sire generation of which 50 or 100 were assumed to be genotyped in the extended designs. Sires of selection candidates were either assumed to be genomic young sires with no daughters or first-crop sires with as many genotyped and phenotyped daughters as grand-sires and great-grand-sires have genotyped daughters.

Obviously, adding phenotyped daughters to a given design would increase reliability of the selection candidate's breeding value per se. We want to emphasize here that all comparisons were done between equal numbers of "phenotyped only" or "phenotyped and genotyped" daughters to describe the marginal contribution of genotyping some of the daughters that were contained in the DYD of their sire in the basic design.

Genotyping a random sample of daughters of AI bulls and including them into the reference population can have a positive effect on the reliability of genomic predictions of candidates (see Tables 2 and 3). The size of the effect is a function of the number of meioses between the candidate and the group of females providing the information, the size of this group, and the heritability of their phenotypic information as pointed out for example by Hayes et al. (2009). For traits with medium to high heritability ($h^2 = 0.35$), 100 genotyped daughters provide a considerable amount of additional information (R_m^2 from 7 to 17% depending on the scenario), whereas in situations where heritability is very low ($h^2 = 0.05$) and the immediate sire of the selection candidate is a genomic young sire without daughter information even 500 genotyped daughters in the grandsire and great grandsire generation result in a very limited increase in marginal reliability ($R_m^2 = 8\%$).

The improvements show a relatively wide range as indicated by the minimum/maximum values additionally given in parentheses in Tables 2 and 3. This variation is presumably due to the fact that we used real genotypes for nuclear pedigrees from the population of dual-purpose Fleckvieh and is a consequence of the realized pattern of inheritance. This means that if a candidate received relatively more grandpaternal segments that

Table 3. Marginal gain in reliability R_m^2 ($\times 100$) for the conformation trait designs (D_{CM} , D_{CL}) with increasing numbers of genotyped daughters per sire (nGeno)¹

nGeno	D_{CM} ($h^2 = 0.25$)			D_{CL} ($h^2 = 0.10$)		
	S	GS-GGS	S-GS-GGS	S	GS-GGS	S-GS-GGS
50	3 (0/7)	5 (3/8)	7 (3/12)	2 (0/4)	3 (1/4)	3 (0/6)
100	5 (2/9)	8 (5/15)	12 (7/19)	3 (0/5)	4 (2/7)	5 (1/9)

¹Values shown are averages (minimum/maximum values in parentheses) across 100 randomly selected nuclear pedigrees. S = only sires of candidates have genotyped daughters. GS-GGS = genotyped daughters for grand-sires and great-grand-sires of selection candidates only. S-GS-GGS = every bull in the nuclear pedigree with genotyped daughters.

are well characterized by genotyped daughters, this will lead on average to a higher reliability of its genomic breeding value. This additional information about the range of benefits is a byproduct of the simple approach followed here.

It should be kept in mind that even remarkable marginal improvements have little effect if the overall reliability of the genomic system is already high (see Figures 1 and 2 for an explanation). However, several additional aspects have to be considered. First, by investigating small nuclear pedigrees with the approach followed here, no conclusions can be drawn about cumulative effect on the accuracy of selection and considering all possible paths of selection in the breeding program. Second, with the strategy proposed, the information from genotyped females arguably contributes predominantly to the reliability of the Mendelian sampling part of the breeding value estimate of genomic candidates. This can be concluded because the reliability of parent average part is already well established by the according base designs using DYD. However, because haplotypes might not be family specific and might segregate elsewhere in the population, it is not possible to definitely quantify benefits on accuracy in large reference populations by analyzing small nuclear pedigrees due to compensating and cumulative effects. Nevertheless, we believe the approach presented here is valid. If the availability of first crop daughter-phenotypes is not severely limited, the heritability is reasonably high, and genotyping is cheap, this approach should provide useful information on the within family (Mendelian sampling) variation. This should be especially valuable for traits with high selection pressure, where the between-family variation is already limited as a consequence of the Bulmer effect. The strategy should also help to equalize the quality and quantity of genomic information between families. Obviously an additional complication arises from the fact that in many genomic breeding programs half-sibs of selection candidates have no phenotypes at the time of selection. In this case, the level providing the information slips back for one generation and the information on the within-family variance is reduced. This reduction might not be too severe as the results from scenario GS-GGS indicate. Third, the limited number of meioses between the genotyped females and the selection candidate will result in large haplotype blocks that are transmitted without recombination. The inference drawn from these samples should therefore be rather robust with respect to the reliability of the selection of candidates. Last but not least, we want to emphasize that with the approach followed here, no general conclusion can be drawn

about the cumulative effect on genetic gain considering all possible paths of selection in the breeding program.

Other benefits might be obtained by including genotyped females into the estimation. For example, we would expect that adding phenotypes and genotypes of unselected animals to the reference population has a stabilizing effect with respect to inflation and bias of the predictions. All the above aspects have to be verified, however, by additional analyses and were beyond the possibilities of the approach presented here.

In this short investigation, we have limited the assumed number of genotyped daughters to 500 for traits with an abundance of phenotypes and to 100 for traits where the capacities for phenotyping are limited. From an economic perspective, the increase in reliability presented here, given the current prices for genotyping, might not justify the additional costs for the breeding program. However, with decreasing costs and in view of other applications of genotyping for cows (e.g., genomic-based mating schemes, phenotype prediction) the additional benefits for the reliability of selection candidates may come along at no cost.

Reorganizing the current system for performance testing should be considered to establish phenotyping genotyped daughters in all traits with limited phenotyping capacity. Additionally, performance tests for health traits can be envisioned. Here, genotyping cows, phenotyped for the new traits, might be the only way to establish a reference population for genomic predictions, because DYD of historic bulls simply are not available. This would also provide a valuable pool of data for the detection of causal variants and for genetic improvement in these traits that would be hard to achieve otherwise. However, the technical question of how to handle genomic systems with potentially hundreds of thousands genotyped animals still has to be resolved.

ACKNOWLEDGMENTS

We gratefully acknowledge the Arbeitsgemeinschaft Süddeutscher Rinderzucht- und Besamungsorganisationen e.V. for their financial support within the research cooperation "Zukunftswege" and the contributors of the genotype pool Germany-Austria for providing the genotypes used in this investigation.

REFERENCES

- Calus, M. P. L., Y. de Haas, and R. F. Veerkamp. 2013. Combining cow and bull reference populations to increase accuracy of genomic prediction and genome-wide association studies. *J. Dairy Sci.* 96:6703–6715.

- Dassonneville, R., A. Baur, S. Fritz, D. Boichard, and V. Ducrocq. 2012. Inclusion of cow records in genomic evaluations and impact on bias due to preferential treatment. *Genet. Sel. Evol.* 44:40.
- Gengler, N., P. Mayeres, and M. Szydlowski. 2007. 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.
- Harris, B., and D. Johnson. 1998. Approximate reliability of genetic evaluations under an animal model. *J. Dairy Sci.* 81:2723–2728.
- Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res. (Camb.)* 91:47–60. <http://dx.doi.org/10.1017/S0016672308009981>.
- Thomasen, J. R., A. C. Sorensen, M. S. Lund, and B. Guldbrandtsen. 2014. Adding cows to the reference population makes a small dairy population competitive. *J. Dairy Sci.* 97:5822–5832.
- VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414–4423. <http://dx.doi.org/10.3168/jds.2007-0980>.
- VanRaden, P. M., J. R. O'Connell, G. R. Wiggans, and K. A. Weigel. 2011. Genomic evaluations with many more genotypes. *Genet. Sel. Evol.* 43:10.