

The impact of genotyping different groups of animals on accuracy when moving from traditional to genomic selection

M. Pszczola,*†‡¹ T. Strabel,‡ J. A. M. van Arendonk,† and M. P. L. Calus*

*Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 8200 AB Lelystad, the Netherlands †Animal Breeding and Genomics Centre, Wageningen University, 6700 AH Wageningen, the Netherlands ‡Department of Genetics and Animal Breeding, Poznan University of Life Sciences, Wolynska 33, 60-637 Poznan, Poland

ABSTRACT

Compared with traditional selection, the use of genomic information tends to increase the accuracy of estimated breeding values (EBV). The cause of this increase is, however, unknown. To explore this phenomenon, this study investigated whether the increase in accuracy when moving from traditional (AA) to genomic selection (GG) was mainly due to genotyping the reference population (GA) or the evaluated animals (AG). In it, a combined relationship matrix for simultaneous use of genotyped and ungenotyped animals was applied. A simulated data set reflected the dairy cattle population. Four differently designed (i.e., different average relationships within the reference population) small reference populations and 3 heritability levels were considered. The animals in the reference populations had high, moderate, low, and random (RND) relationships. The evaluated animals were juveniles. The small reference populations simulated difficult or expensive to measure traits (i.e., methane emission). The accuracy of selection was expressed as the reliability of (genomic) EBV and was predicted based on selection index theory using relationships. Connectedness between the reference populations and evaluated animals was calculated using the prediction error variance. Average (genomic) EBV reliabilities increased with heritability and with a decrease in the average relationship within the reference population. Reliabilities in AA and AG were lower than those in GG and were higher than those in GA (respectively, 0.039, 0.042, 0.052, and 0.048 for RND and a heritability of 0.01). Differences between AA and GA were small. Average connectedness with all animals in the reference population for all scenarios and reference populations ranged from 0.003 to 0.024; it was lowest when the animals were not genotyped (AA; e.g., 0.004 for RND) and

highest when all the animals were genotyped (GG; e.g., 0.024 for RND). Differences present across designs of the reference populations were very small. Genomic relationships among animals in the reference population might be less important than those for the evaluated animals with no phenotypic observations. Thus, the main origin of the gain in accuracy when using genomic selection is due to genotyping the evaluated animals. However, genotyping only one group of animals will always yield less accurate estimates.

Key words: genomic selection, reference population design, breeding value reliability, connectedness

INTRODUCTION

Genomic selection (**GS**) uses dense SNP marker arrays. These markers are assumed to be in linkage disequilibrium (**LD**) with QTL (Meuwissen et al., 2001), allowing their effects to be estimated. To estimate SNP marker effects, GS requires a set of genotyped and phenotyped animals, a so-called reference population used to evaluate genotyped animals without phenotypic information.

Another approach to evaluating animals without explicitly estimating SNP marker effects is to use SNP markers to estimate relationships between animals. Genomic relationships can capture Mendelian sampling and reveal links between animals that are seemingly unrelated through pedigree. Thus, genomic relationships are more precise, improving the connectedness between animals in a reference population and the evaluated animals. This higher connectedness reduces bias, and thus improves the genetic evaluation (Kennedy, 1981).

Based on such genomic relationships, a genomic relationship matrix (**G**) can be created using various methods (e.g., Nejati-Javaremi et al., 1997; VanRaden, 2008; Yang et al., 2010). This genomic relationship matrix can be used in the genomic BLUP procedure (G-BLUP), where **G** replaces the additive relationship matrix (**A**). Alternatively, **G** for genotyped animals can be merged with **A** for all animals, to enable simultaneous use of phenotypic information of genotyped

Received March 19, 2012. Accepted May 28, 2012.

¹Corresponding author: mbee@jay.up.poznan.pl

and ungenotyped animals in genetic evaluation (e.g., Legarra et al., 2009; Misztal et al., 2009; Aguilar et al., 2010; Christensen and Lund, 2010). This method, as shown by Christensen and Lund (2010) using the gene content prediction method (Gengler et al., 2007), provides an approximation to unseen genotypes and generates genomic EBV (GEBV) in a single step.

An important aspect of genetic improvement is the response to selection, which depends linearly on accuracy of selection. The accuracy of traditional selection depends on the availability of phenotypic information on relatives as well as the animal's own performance and on heritability of the considered trait. The accuracy of GS depends on heritability, but also on several other factors (Daetwyler et al., 2008; Goddard, 2009). First, the size of the reference population is relevant; this is equivalent to the availability of phenotypic information on relatives and the animal's own performance in traditional selection. The larger the reference population, the higher the accuracy of predicted breeding values. Second, the lower the effective population size (N_e) , the fewer independent segments in the genome, reducing the number of markers needed to tag all segments and the fewer records are needed to accurately estimate the effects of all independent segments (Goddard, 2009). Third, the effective number of loci, dependent on the mating structure and the recombination length of the genome, affects the number of markers required to tag all potential QTL (Goddard, 2009). Finally, the accuracy of GEBV is affected by both the relationship between the evaluated animals and the reference population (Habier et al., 2010; Wolc et al., 2011; Pszczola et al., 2012) and its design (the relationships within the reference population; Pszczola et al., 2012).

Choosing an appropriate design for the reference population, as suggested by Meuwissen (2009) and Calus (2010) and shown by Pszczola et al. (2012), may be a way to optimize current techniques. Such optimization is especially required when the number of animals in the reference population is limited. This can be a consequence of, for example, a small population size, a limited number of phenotypes available from research herds or (historic) experiments, a lack of routinely taken measurements for the considered trait, genotyping costs, or measuring a difficult or expensive trait. For instance, in the case of methane emission, measuring a single observation is very expensive because it requires sophisticated equipment.

A comparison of GS with traditional selection tools showed, especially in the case of unphenotyped juveniles, an increase in accuracy when GS was used (Meuwissen et al., 2001; Schaeffer, 2006; Pszczola et al., 2011). The origin of this increase in accuracy is unclear, given that it is unknown whether this increase is mainly due to

substituting pedigree with genomic information for the reference population or for the evaluated animals. This is also an important question from a practical point of view because genomic data are sometimes unavailable for animals with valuable phenotypes. Including these animals in the reference population is possible by using approaches that combine genomic with pedigree information.

The aim of this study was to investigate whether the increase in accuracy, when moving from traditional selection to GS, is mainly due to genotyping the reference population or the evaluated animals. More specifically, it asks whether the accuracy of GEBV increases 1) when the reference population is ungenotyped while the evaluated animals are genotyped, and 2) when the evaluated animals are ungenotyped and the reference population is genotyped, and the combined relationship matrix is used. We evaluated the amount of gain by comparing, across scenarios at different heritability levels, predicted reliabilities and connectedness levels.

MATERIALS AND METHODS

Data

The simulated data set used in this study, reflecting a dairy cattle population, has been described in detail by Pszczola et al. (2012). Briefly, the simulation scheme adopted from de Roos et al. (2009) mimics different sizes of N_e at different stages in the historic cattle population by using inflated values of recombination rates and frequency of mutation events. The use of inflated parameters heavily reduces the number of generations simulated at each of the stages of the historic cattle population and yields realistic LD levels (de Roos et al., 2008). To achieve LD levels for modern cattle breeds while simulating biallelic loci, this simulation scheme was further modified as described by Pszczola et al. (2012).

The data set retained for analysis comprised pedigrees of 10 generations of randomly mated animals. Each of the first 5 generations included 25 sires mated randomly with 400 dams, and each of the last 5 generations included 50 sires mated randomly with 200 dams. The last 2 generations were composed of animals with genotypic data. Data were available from 20 replications of the simulation process.

The simulated genome was 3 M long and consisted of 3 chromosomes of 1 M each. In the last generation, the average distance between segregating loci across the whole genome was approximately 0.03 cM, the average minor allele frequency was 0.25, and the average LD, measured as r^2 (Hill and Robertson, 1968) between adjacent loci, was 0.23.

Reference Populations and Evaluated Animals

Simulated animals were partitioned into an evaluation set and reference populations. The evaluated animals (n=1,000) were chosen from the second genotyped generation (i.e., 20 randomly chosen progeny each of 50 randomly chosen sires). The 4 reference populations were chosen from the animals in the first genotyped generation.

Because each reference population, consisting of 2,000 cows, had a different family structure with respect to the sizes of parental half-sib families, their average relationship varied. The average relationship of the first reference population (high relationship), which consisted of progeny of 5 sires, was 0.095. The average relationship for the second reference population (medium relationship), consisting of progeny of 20 sires, was 0.056. For the third reference population (low relationship), consisting of progeny of 40 randomly chosen sires, the average relationship was 0.050. The average relationship for the fourth reference population (RND), consisting of randomly selected animals, was the lowest: 0.049.

Calculation of (G)EBV Reliabilities

Reliabilities were calculated deterministically for all the evaluated animals for 4 situations: first, no animals were genotyped and the pedigree-based relationships were used (**AA**); second, animals in the reference population were genotyped and the evaluated animals were not (**GA**); third, the reference population was not genotyped whereas the evaluated animals were (**AG**); and fourth, all the animals were genotyped (**GG**).

The matrices were created as follows. Relationship matrix **H** combined pedigree and genomic data for GA and AG, following the method of Legarra et al. (2009):

$$\begin{bmatrix} \mathbf{A}_{11} + \mathbf{A}_{12} \mathbf{A}_{22}^{-1} (\mathbf{G} - \mathbf{A}_{22}) \mathbf{A}_{22}^{-1} \mathbf{A}_{21} & \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{G} \\ \mathbf{G} \mathbf{A}_{22}^{-1} \mathbf{A}_{21} & \mathbf{G} \end{bmatrix}, \qquad [1]$$

where the pedigree relationship matrix \mathbf{A} was partitioned for ungenotyped (1) and genotyped (2) animals. The genomic relationship matrix, \mathbf{G} , was created as

$$\frac{\mathbf{Z}\mathbf{Z}'}{^{2}\sum p_{i}\left(1-p_{i}\right)}$$

following the method of VanRaden (2008), where p_i is the frequency of the second allele at locus i for which the homozygote genotype is coded as 1, and \mathbf{Z} is derived from the genotypes of the animals in the reference

population by subtracting 2 times the allele frequency expressed as a difference of 0.5 [i.e., $2(p_i - 0.5)$] from matrix \mathbf{M} , which specifies the marker genotypes for each individual as -1, 0, or 1. Because for building \mathbf{H} , \mathbf{A} and \mathbf{G} must be compatible, the \mathbf{G} matrix had to be adjusted. First, the inbreeding level in \mathbf{G} was scaled to the inbreeding level in \mathbf{A} . This was done by calculating the pedigree inbreeding coefficients averaged across all animals (\bar{f}_p) . Because the current population was used to calculate p_i , it was assumed that an average genomic inbreeding coefficient in \mathbf{G} is zero. \mathbf{G}^* was then calculated following the formula derived from Wright's F-statistics, as in Powell et al. (2010):

$$\mathbf{G}^* = \mathbf{G}(1 - \overline{f}_p) + 2\overline{f}_p \mathbf{J},$$

where G^* contains the relationships relative to the same base used in A, and J is a matrix of all ones. This adjustment is equivalent to what was proposed by Vitezica et al. (2011)

To create $\hat{\mathbf{G}}$, the second step, following Yang et al. (2010), accounts for the fact that \mathbf{G}^* is estimated with some error:

$$\hat{\mathbf{G}} = \mathbf{G} * + \mathbf{E} = \mathbf{A} + (\mathbf{G} * - \mathbf{A}) + \mathbf{E},$$

where **E** is the matrix containing estimation errors for $\hat{\mathbf{G}}$, assuming the following variances for those matrices $V(\hat{\mathbf{G}} - \mathbf{A}) = V(\mathbf{G}^* - \mathbf{A}) + V(\mathbf{E})$. $V(\mathbf{E})$ equals 1/N, where N is the SNP number used. To account for the sampling variance in $\hat{\mathbf{G}}$, $\hat{\mathbf{G}} - \mathbf{A}$ is regressed back toward \mathbf{A} , and $\tilde{\mathbf{G}}$ is

$$\begin{split} \tilde{\mathbf{G}} &= \mathbf{A} + \frac{(\hat{\mathbf{G}} - \mathbf{A}) \times V(\mathbf{G}^* - \mathbf{A})}{[V(\mathbf{G}^* - \mathbf{A}) + V(\mathbf{E})]} \\ &= \mathbf{A} + \frac{(\hat{\mathbf{G}} - \mathbf{A}) \times [V(\hat{\mathbf{G}} - \mathbf{A}) - \frac{1}{N}]}{V(\hat{\mathbf{G}} - \mathbf{A})}. \end{split}$$

Because the sampling error in relationships in $\hat{\mathbf{G}}$ depends on the value of the relationship, values for $V(\hat{\mathbf{G}} - \mathbf{A})$ were calculated separately for bins of relationships in \mathbf{A} . The bins of relationships were 0 to 0.10, >0.10 to 0.25, >0.25 to 0.50, and >0.50. The last bin excluded parent-offspring pairs because their relationship is expected to be 0.5. The diagonal elements of $\hat{\mathbf{G}}$ (i.e., self-relationships) were not regressed. The regression coefficients averaged across simulations for the 4 bins were, respectively, 0.963, 0.983, 0.991, and 0.989, which are similar to those reported in a different study based on real data (Veerkamp et al., 2011).

Deterministic predictions of GEBV reliabilities were based on formulas that can be derived either from selection index theory for a derivation see the Appendix in Pszczola et al. (2012)] or from the prediction error variances of the mixed model equations used to estimate the breeding values, as follows.

For AA, the reliability (r_{AA}^2) was

$$\mathbf{a} \left[\mathbf{A} + \mathbf{I} \left(\frac{\sigma_e^2}{\sigma_a^2} \right) \right]^{-1} \mathbf{a}',$$
 [2]

where a is a vector with pedigree-based relationships of an evaluated animal with the animals in the reference population; A is the additive relationship matrix for animals in the reference population; I is an identity matrix, σ_e^2 is the residual variance; and σ_a^2 is the genetic variance. Their ratio reflects heritability (h^2) . For GA and AG, $r_{\rm GA}^2$ and $r_{\rm AG}^2$ were calculated as

$$\mathbf{h} \left[\mathbf{H}_{\text{ref}} + \mathbf{I} \left(\frac{\sigma_e^2}{\sigma_a^2} \right) \right]^{-1} \mathbf{h}', \tag{3}$$

relationships of an evaluated animal with the animals in the reference population ($\mathbf{H}_{\text{ref_eva}}$); \mathbf{H}_{ref} is the combined pedigree and genomic relationship matrix for animals in the reference population. For situation GA,
$$\begin{split} \mathbf{H}_{\mathrm{ref_eva}} &= \mathbf{A}_{\mathrm{ref_eva}} \mathbf{A}_{\mathrm{ref}}^{-1} \mathbf{G}_{\mathrm{ref}} \quad \mathrm{and} \quad \mathbf{H}_{\mathrm{ref_eva}} &= \mathbf{G}_{\mathrm{ref}}, \quad \mathrm{where} \\ \mathbf{A}_{\mathrm{ref_eva}} & \mathrm{is \ the \ part \ of \ } \mathbf{A} \ \mathrm{describing \ relationships \ between} \end{split}$$
the reference and evaluated animals, and \mathbf{A}_{ref} and \mathbf{G}_{ref} are created for the genotyped reference population. For $\begin{aligned} & \mathbf{AG}, & \mathbf{H}_{\mathrm{ref}_\mathrm{eva}} &= \mathbf{A}_{\mathrm{ref}_\mathrm{eva}} \mathbf{A}_{\mathrm{eva}}^{-1} \mathbf{G}_{\mathrm{eva}} \\ & \mathbf{H}_{\mathrm{ref}} &= \mathbf{A}_{\mathrm{ref}} + \mathbf{A}_{\mathrm{ref}_\mathrm{eva}} \mathbf{A}_{\mathrm{eva}}^{-1} \left(\mathbf{G}_{\mathrm{eva}} - \mathbf{A}_{\mathrm{eva}} \right) \mathbf{A}_{\mathrm{eva}}^{-1} \mathbf{A}_{\mathrm{eva}_\mathrm{ref}}, \end{aligned}$ where $\mathbf{A}_{\text{ref_eva}}$ and $\mathbf{A}_{\text{eva_ref}}$ are parts of \mathbf{A} describing the relationships between the reference and evaluated animals; A_{ref} is part of A for the ungenotyped reference

where h is a vector based on part of H that contains

For GG, $r_{\rm GG}^2$ was as in Van Raden (2008), Goddard et al. (2011), or Pszczola et al. (2012):

population; \mathbf{A}_{eva} and \mathbf{G}_{eva} are created for the genotyped

evaluated population.

$$\mathbf{c} \left[\mathbf{G} + \mathbf{I} \left(\frac{\sigma_e^2}{\sigma_a^2} \right) \right]^{-1} \mathbf{c}', \tag{4}$$

where \mathbf{c} is a vector with the genomic relationships of an evaluated animal with the animals in the reference population. This vector is a column of the C matrix for a particular evaluated animal. The C matrix itself is created as

$$\frac{\mathbf{Z}_{2}\mathbf{Z}'}{^{2}\displaystyle\sum p_{i}\left(1-p_{i}\right)}$$

and \mathbb{Z}_2 is constructed from the genotypes of the evaluated and reference animals. In the GG scenario, G for all animals (i.e., the reference and evaluated animals) was created as $\tilde{\mathbf{G}}$. The \mathbf{C} matrix, therefore, is an offdiagonal part of $\tilde{\mathbf{G}}$ that describes relationships between the reference population and the evaluated animals.

Connectedness

Connectedness was calculated as in Lewis et al. (1999) between the evaluated animals and animals in the reference population. Because negative relationship coefficients are present when genomic information is incorporated, the absolute values were taken; thus, the connectedness level was calculated as:

$$con = \left| \frac{\text{PEC}(\hat{a}_i, \hat{a}_j)}{\sqrt{\text{PEV}(\hat{a}_i)\text{PEV}(\hat{a}_j)}} \right|,$$
 [5]

where $\operatorname{PEV}\left(\hat{a}_{i}\right)$ is the prediction error variance of an EBV of an evaluated animal i, $PEV(\hat{a}_i)$ is the prediction error variance of an EBV of a reference animal j, $PEC(\hat{a}_i, \hat{a}_i)$ is the prediction error covariance between the breeding values of the animals i and j. Note that con = 0 means that the animals are not connected. The PEV and PEC were obtained from inverted coefficient matrices of mixed model equations. For each of the evaluated animals, its average connectedness with all the animals in the reference population was calculated.

RESULTS

Reliability

The average reliabilities of (G)EBV predictions are shown in Figures 1, 2, and 3 for the different heritability levels and across all the considered scenarios and reference populations. The significance of the differences between the scenarios was tested within each reference population. Almost all of these differences were highly significant (P < 0.01). The exceptions were differences between scenarios AA and GA for the reference population medium relationship and low relationship.

Overall, the reliabilities of the scenarios, as can be seen in Figures 1 to 3, increased with increasing heritability and with a decreasing average relationship within the reference population. In GG, when all the animals

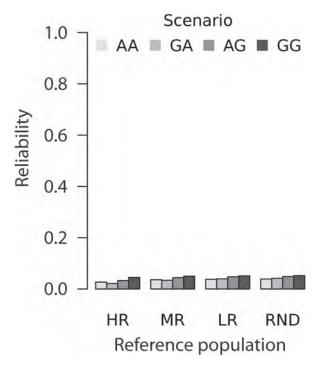


Figure 1. Average reliabilities of breeding values for the evaluated set of animals calculated based on scenarios in which pedigree information only is used (AA); the reference population is genotyped and the evaluated animals are ungenotyped (GA); the reference population is ungenotyped and the evaluated animals are genotyped (AG); and all animals are genotyped (GG) across 4 reference populations with different family structures (high relationship = HR; medium relationship = MR; low relationship = LR; random selection of animals = RND) for the heritability of 0.01, averaged over all replicates.

were genotyped, as was especially evident in case of the highest heritability (see Figure 3), the reliabilities were considerably higher than in the other cases. When the evaluated animals were genotyped and the reference population was not (AG), reliabilities were noticeably lower than those in GG. When the reference population was genotyped and the evaluated animals were not (GA), reliabilities were somewhat higher than those in AG. For example, in the case of RND for the heritability of 0.01, the average reliabilities for AA, GA, AG, and GG were, respectively, 0.039, 0.042, 0.048, and 0.052 (Figure 1). Differences between these scenarios increased together with an increase in the heritability level (Figures 1 to 3). Differences between the traditional breeding scheme using pedigree information (AA) and GA were always small, with ranking changing across heritability levels. For example, at the heritability of 0.9 (Figure 3) and with the RND reference population, AA (0.207) was slightly higher than GA (0.204), whereas at the heritability of 0.3 (Figure 2), AA (0.165) was slightly lower than GA (0.173). For the heritability of 0.01, the GA scenario tended to be

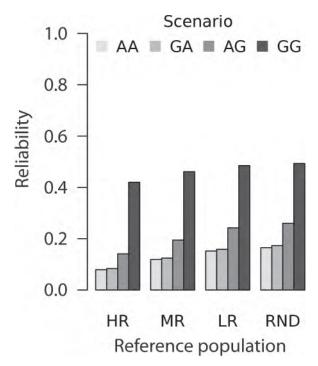


Figure 2. Average reliabilities of breeding values for the evaluated set of animals calculated based on scenarios in which pedigree information only is used (AA); the reference population is genotyped and the evaluated animals are ungenotyped (GA); the reference population is ungenotyped and the evaluated animals are genotyped (AG); and all animals are genotyped (GG) across 4 reference populations with different family structures (high relationship = HR; medium relationship = MR; low relationship = LR; random selection of animals = RND) for the heritability of 0.3, averaged over all replicates.

more accurate than the AA scenario (Figure 1; as for example in case of RND, where GA was 0.042 and AA was 0.039) only when the average relationship within the reference was low.

Connectedness

The average connectedness level of evaluated animals with all the animals in the reference population (con) for different scenarios and reference populations and for a heritability of 0.3 is shown in Figure 4. The con ranged from 0.003 to 0.024. Differences present across designs of the reference populations were very small. The con level was lowest when the animals were not genotyped (AA; e.g., 0.004 for RND) and highest when all the animals were genotyped (GG; e.g., 0.024 for RND). Genotyping only the reference population (GA) slightly improved con when compared with AA (e.g., from 0.004 for AA to 0.005 for AG in the RND scenario). Genotyping only the evaluated animals (AG) caused a higher increase (to 0.008 for AG in the RND scenario). The

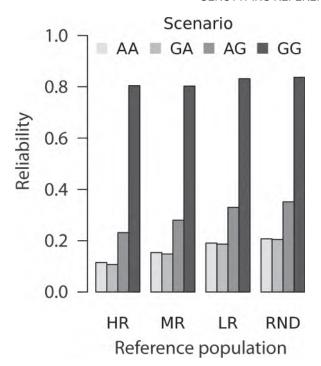


Figure 3. Average reliabilities of breeding values for the evaluated set of animals calculated based on scenarios in which pedigree information only is used (AA); the reference population is genotyped and the evaluated animals are ungenotyped (GA); the reference population is ungenotyped and the evaluated animals are genotyped (AG); all animals are genotyped (GG) across 4 reference populations with different family structures (high relationship = HR; medium relationship = MR; low relationship = LR; random selection of animals = RND) for the heritability of 0.9, averaged over all replicates.

con level increased when the heritability level increased (not shown; i.e., the heritability level acted as a scaling factor).

DISCUSSION

The aim of this study was to investigate whether the increase in accuracy, when moving from traditional selection to GS, is mainly due to genotyping the reference population or the evaluated animals. In it, accuracy of selection is expressed as the reliability of (G)EBV. These reliabilities were calculated using deterministic prediction that can be derived from selection index theory or the prediction error variances of the mixed model equations.

Combined Relationship Matrix

In the GA and AG scenarios, the genomic-based relationships were combined with the pedigree-based relationships into the **H** matrix (see Legarra et al., 2009; Misztal et al., 2009). Combining the pedigree-based

and the genomic-based relationships may yield different breeding value reliabilities across differently created **G** matrices (Forni et al., 2011). Matrices **A** and **G** have to be on the same scale before they are combined into **H**. To ensure this, in the present study, the inbreeding level in **G** was scaled to the inbreeding level in **A** as in Powell et al. (2010), and then, using the method of Yang et al. (2010), **G** was regressed toward **A**. Further improvement of the compatibility of the 2 matrices could possibly be achieved using LD-linkage analysis methodology (Meuwissen et al., 2011); however, this option was not explored here.

Although the applied adjustments did result in a similar scale for both matrices, **G** and **A** still have different properties. This is because genomic relationships include Mendelian sampling and therefore have higher variance than do pedigree-based relationships. This difference in the matrix properties may be a reason why, in the case of GA, it was unexpected that the use of the combined relationship matrix, was only somewhat (or not) beneficial when compared with traditional BLUP (AA).

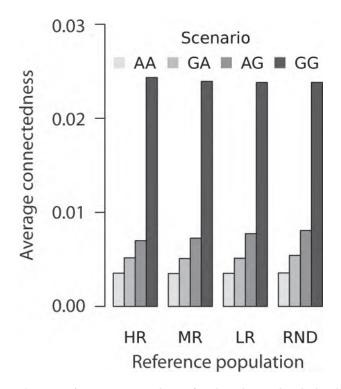


Figure 4. Average connectedness of evaluated animals calculated based on scenarios in which pedigree information only is used (AA); the reference population is genotyped and the evaluated animals are ungenotyped (GA); the reference population is ungenotyped and the evaluated animals are genotyped (AG); all animals are genotyped (GG) across 4 reference populations with different family structures (high relationship = HR; medium relationship = MR; low relationship = LR; random selection of animals = RND) for the heritability of 0.3, averaged over all replicates.

Scenario AG was always better than GA, which can be explained by the availability of information on Mendelian sampling. When genotypes are to be predicted from offspring to parents, this information is available, which may lead to high accuracy of the genotype prediction. For instance, if the number of genotyped offspring per ungenotyped parent is large (i.e., 30 or more), the accuracy of genotype prediction can even reach 100%, as shown by Boettcher et al. (2004). In the opposite situation (i.e., when missing genotypes are to be predicted from parents to offspring), prediction accuracy will be poorer because no information on Mendelian sampling is available; therefore, the distinction between, for example, half sibs is not possible. When both parents are genotyped, in fact, the maximum genotype prediction accuracy, measured as the correlation between the true and predicted genotypes, assuming a heritability of 1, is equal to $r = \sqrt{\mathbf{bG}} / \sigma_a = \sqrt{\frac{1}{2}}$, where both **b** and **G** contain values of 0.5 and $\sigma_a = 1$ (Mrode, 2005). It is therefore important to genotype both parents if the genotypes of their offspring are to be predicted (Calus et al., 2011; Pszczola et al., 2011). If both parents (or more ancestors) and offspring of an animal are known, the accuracy of its genotype prediction would be substantially higher (Gengler et al., 2008). Therefore, it is expected that genotyping evaluated animals, which are usually younger than the reference animals, is more beneficial than the opposite. This is especially evident when the evaluated animals are offspring of the reference animals, as discussed above. The closest relationships available between reference and evaluated animals in our design were due to sharing the same ungenotyped sire (i.e., being half-sib family members), maternal grandsires, or other more distant relatives. Closer relationships between genotyped and ungenotyped animals in the AG and GA scenarios would increase the accuracy of predicting genomic relationships for ungenotyped animals, in a similar way as the accuracy of genotype imputation depends on relationships between genotyped and imputed animals (Mulder et al., 2012). In general, the level of accuracy in the simulated scenarios is expected to be lower than when closer relationships between reference and evaluated animals would be present. In most practical situations, however, the evaluated animals are genotyped as well as at least part of the reference animals, and stronger family links between these 2 groups may exist. Such a setup leads to considerable differences between a pedigree-based scenario and a single-step approach favoring the latter method (e.g., Aguilar et al., 2010; Christensen and Lund, 2010; Chen et al., 2011; Forni et al., 2011; Vitezica et al., 2011).

The genome simulated in this study was about 10 times smaller than the cattle genome. As indicated in Pszczola et al. (2012), the simulated LD was close to values observed in real data (e.g., those reported by de Roos et al., 2008). To further check similarities to expectations found in real data analyses, we estimated the number of effective chromosome segments (Me)calculated based on the formula of Daetwyler et al. (2008).Rearranging this equation vields $Me = Nh^2/(r^2 - Nh^2)$, where N is the reference population size, h^2 is the heritability level, and r^2 is the reliability obtained from our analysis. The calculated value for Me ranged from 352 to 437 for an h^2 of 0.9, from 617 to 829 for an h² of 0.3, and from 365 to 424 for an h² of 0.01. The lowest values were observed for reference populations with a weaker family structure. Our Me values agreed reasonably well with the results of other studies for small cow reference populations. For example, Verbyla et al. (2010) showed that Me calculated for a small population of Holstein-Friesian heifers for energy balance with an h^2 of 0.325 was 472.

As our results have shown, compared with traditional BLUP, genotyping only the evaluated animals achieved substantially higher reliabilities. Although the realized reliabilities will be substantially lower than if all or most of the reference and the evaluated animals are genotyped, genotyping only the evaluated animals is more beneficial than genotyping only animals with phenotypic records.

Impact of Relationships

The average relationship within the reference population, as reported previously by Pérez-Cabal et al. (2010) and Pszczola et al. (2012), affected the average reliability regardless of which animals were genotyped. When the animals in the reference population were loosely related to each other, as shown in Figures 1–3, the reliabilities were higher. This means that the design of the reference population is also important when genotyped and ungenotyped animals are analyzed jointly using the H matrix.

The availability of more precise relationship data for the reference population (GA) resulted in small differences in reliabilities compared with a situation in which none of the animals were genotyped (AA), whereas, as shown in Figures 1 to 3, genotyping the evaluated animals (AG) increased reliabilities noticeably. This may be because, from regular BLUP models (Henderson, 1985), it is known that when an animal has phenotypic records itself, the emphasis on information about relatives is reduced. Translated to the scenarios studied here, this reduced emphasis on information concerning relatives implies that very precise genomic relationships among animals in the reference population (i.e., animals with phenotypes) might be less important than those for the evaluated animals with no phenotypic observations on themselves or on their descendants.

We also considered a situation in which the evaluated animals originated from the same generation as the reference population. In this situation, the average reliabilities were higher than in the presented results, but the tendencies were similar (results not shown). For animals originating from different generations, differences in the reliability of (G)EBV were in line with those in the literature (Habier et al., 2010; Pszczola et al., 2011, 2012; Wolc et al., 2011).

Deterministic approaches can be used to approximate the reliability, given known and simple family structures of the considered population (Hayes et al., 2009b) or randomly mated population (Daetwyler et al., 2008; Goddard, 2009). In practice, however, more complex family structures are present and animals are not mated at random, and for such a situation, deriving a general deterministic prediction formula is not trivial. On the basis of our results, however, in general one may expect to achieve higher reliabilities when the reference population is composed of many small half-sib families as opposed to a few large ones; that is, the use of an RND reference population always yielded higher average reliabilities than using a high relationship (Figures 1 to 3).

Bias of Prediction

The reliability of (G)EBV can be assessed by inverting the left-hand side of the mixed model equations or, in an empirical way, by cross-validation (e.g., Calus et al., 2010). The former method, as applied in this study, tends to give somewhat overestimated results (Calus et al., 2009; Hayes et al., 2009a; Lund et al., 2009; Su et al., 2010). One reason for this overestimation is the underlying assumption that the markers explain all the genetic variance. The genetic variance, however, is at least partially also explained by loci between the markers (VanRaden, 2008). Thus, the reliabilities presented here may be somewhat overestimated. Despite this expected slight overestimation of the reliabilities, the method used here, unlike the cross-validation method, makes an assessment of individual reliabilities possible, which is important from the practical point of view when comparing animals' breeding values. Yet even if some overestimation would be present in this study, it is expected not to affect the conclusions presented here. This is because reliabilities for the breeding values based on pedigree information are expected to be unbiased, and replacing the pedigree with genomic data, to some extent, only adds more information.

Connectedness

The reliability predictions formulas presented here are equivalent to those obtained with use of the PEV of breeding values. The PEV of differences between animals from different management units can be also used to determine connectedness (Kennedy and Trus, 1993). Genetically unconnected herds or other management units with different genetic means cannot be distinguished in the genetic evaluation; thus, the comparison of breeding values among them is biased (Kennedy, 1981). This bias would be reduced if a positive genetic covariance existed between the units (i.e., they would be connected). In other words, 2 animals are connected when the PEC between them is nonzero; the difference between their breeding values is then expected to have smaller bias than would a pair of animals with zero PEC. Using the pedigree-based relationships, PEC is nonzero only when the animals are linked through the pedigree. Seemingly unlinked animals, in fact, may be linked through distant unrecorded ancestors, as can be revealed with genomic relationships. Using genomic relationships thus improves comparisons across the management units. This was shown in Figure 4, where the average connectedness level in the AA scenario was considerably lower than that in GG. Small differences across differently designed reference populations may be attributed to the fact that differences in the average relationship within the reference population based on genomic data were smaller than calculated based on the pedigree data.

CONCLUSIONS

This study aimed to investigate the contribution of using genomic information on a reference population or the evaluated animals to the increase in selection accuracy. Compared with traditional selection, genotyping only the evaluated group of animals significantly increased the accuracy of the estimates, whereas genotyping only the reference population yielded minor, and sometimes unfavorable, changes in accuracy. This was attributed to the fact that the emphasis on information concerning relatives is reduced when an animal has phenotypic records. The reduced emphasis on the information of relatives implies that very precise genomic relationships among animals in the reference population are less important than those for the evaluated animals with no phenotypic observations on themselves or on their descendants. Nevertheless, although the main origin of the gain in accuracy from using GS is

genotyping the evaluated animals, genotyping only one group of animals will always yield estimates that are substantially less accurate than when all the animals are genotyped. An additional benefit of using a genomic relationship matrix is reducing bias across herd, region, or country evaluations, which is demonstrated by the improved connectedness between the reference population and the evaluated animals.

ACKNOWLEDGMENTS

Roel Veerkamp (Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Lelystad, the Netherlands) and Han Mulder (Animal Breeding and Genomics Centre, Wageningen University, Wageningen, the Netherlands) are acknowledged for their useful comments and discussion. Linda McPhee (Linda McPhee Consulting Ltd., Cambridge, UK) is thanked for helping edit the manuscript. Marcin Pszczola gratefully acknowledges the financial support of the Koepon Stichting (Leusden, the Netherlands) and GreenHouse-Milk. The GreenHouseMilk project is financially supported by the European Commission under the Seventh Research Framework Programme (Grant Agreement KBBE-238562). This publication represents the views of the authors, not the European Commission, and the Commission is not liable for any use that may be made of the information. Financial support of the National Science Centre (Krakow, Poland) is acknowledged (decision number DEC-2011/01/N/NZ9/00614).

REFERENCES

- Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score1. J. Dairy Sci. 93:743–752.
- Boettcher, P. J., G. Pagnacco, and A. Stella. 2004. A Monte Carlo approach for estimation of haplotype probabilities in half-sib families. J. Dairy Sci. 87:4303–4310.
- Calus, M., H. Mulder, and R. Veerkamp. 2010. Comparison of reliabilities of direct genomic values. Interbull Bull. 41:25–28.
- Calus, M., H. Mulder, K. Verbyla, and R. Veerkamp. 2009. Estimating reliabilities of genomic breeding values. Interbull Bull. 40:198– 201.
- Calus, M. P. L. 2010. Genomic breeding value prediction: Methods and procedures. Animal 4:157–164.
- Calus, M. P. L., R. F. Veerkamp, and H. A. Mulder. 2011. Imputation of missing single nucleotide polymorphism genotypes using a multivariate mixed model framework. J. Anim. Sci. 89:2042–2049.
- Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. J. Anim. Sci. 89:2673–2679.
- Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:2.
- Daetwyler, H. D., B. Villanueva, and J. A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3:e3395.
- de Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of genomic predictions across multiple populations. Genetics 183:1545–1553.

- de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkgae disequilibrium and persistence of phase on Holstein-Friesian, Jersey, and Angus cattle. Genetics 179:1503-1512.
- Forni, S., I. Aguilar, and I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet. Sel. Evol. 43:1.
- Gengler, N., S. Abras, C. Verkenne, S. Vanderick, M. Szydlowski, and R. Renaville. 2008. Accuracy of prediction of gene content in large animal populations and its use for candidate gene detection and genetic evaluation. J. Dairy Sci. 91:1652–1659.
- 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.
- Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica 136:245–257.
- Goddard, M. E., B. J. Hayes, and T. H. E. Meuwissen. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128:409–421.
- Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42:5.
- Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009a. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. 41:51.
- Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009b. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. (Camb.) 91:47–60.
- Henderson, C. R. 1985. Best linear unbiased prediction using relationship matrices derived from selected base populations. J. Dairy Sci. 68:443–448.
- Hill, W., and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38:226-231.
- Kennedy, B. W. 1981. Bias and mean square error from ignoring genetic groups in mixed model sire evaluation. J. Dairy Sci. 64:689–697
- Kennedy, B. W., and D. Trus. 1993. Considerations on genetic connectedness between management units under an animal model. J. Anim. Sci. 71:2341–2352.
- Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656–4663.
- Lewis, R. M., R. E. Crump, G. Simm, and R. Thompson. 1999. Assessing connectedness in across-flock genetic evaluations. Page 121 in Proc. Br. Soc. Anim. Sci. Br. Soc. Anim. Sci., Penicuik, UK.
- Lund, M., G. Su, U. Nielsen, and G. Aamand. 2009. Relation between accuracies of genomic predictions and ancestral links to the training data. Interbull Bull. 40:162–166.
- Meuwissen, T. H. E. 2009. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet. Sel. Evol. 41:35.
- Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.
- Meuwissen, T. H. E., T. Luan, and J. A. Woolliams. 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. J. Anim. Breed. Genet. 128:429–439.
- Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information. J. Dairy Sci. 92:4648–4655.
- Mrode, R. 2005. Linear Models for the Prediction of Animal Breeding Values. CAB Int., Wallingford, UK.
- Mulder, H. A., M. P. L. Calus, T. Druet, and C. Schrooten. 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. J. Dairy Sci. 95:876–889.
- Nejati-Javaremi, A., C. Smith, and J. P. Gibson. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. J. Anim. Sci. 75:1738–1745.
- Pérez-Cabal, M. A., A. I. Vazquez, D. Gianola, G. J. M. Rosa, and K. A. Weigiel. 2010. Accuracy of genomic predictions in USA Holstein

- cattle from different training-testing designs. Page 150 in Book of Abstracts: 9th World Congr. Genet. Appl. Livest. Prod. (WC-GALP), Leipzig, Germany. EAAP, Rome, Italy.
- Powell, J. E., P. M. Visscher, and M. E. Goddard. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Genet. 11:800–805.
- Pszczola, M., H. A. Mulder, and M. P. L. Calus. 2011. Effect of enlarging the reference population with (un)genotyped animals on the accuracy of genomic selection in dairy cattle. J. Dairy Sci. 94:431–441.
- Pszczola, M., T. Strabel, H. A. Mulder, and M. P. L. Calus. 2012. Reliability of genomic selection for animals with different relationships within and to the reference population. J. Dairy Sci. 95:389–400.
- Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218–223.
- Su, G., B. Guldbrandtsen, V. R. Gregersen, and M. S. Lund. 2010. Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. J. Dairy Sci. 93:1175–1183.
- VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.

- Veerkamp, R. F., H. A. Mulder, R. Thompson, and M. P. L. Calus. 2011. Genomic and pedigree-based genetic parameters for scarcely recorded traits when some animals are genotyped. J. Dairy Sci. 94:4189–4197.
- Verbyla, K. L., M. P. L. Calus, H. A. Mulder, Y. de Haas, and R. F. Veerkamp. 2010. Predicting energy balance for dairy cows using high-density single nucleotide polymorphism information. J. Dairy Sci. 93:2757–2764.
- Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genet. Res. (Camb.) 93:357–366.
- Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D. J. Garrick, and J. C. M. Dekkers. 2011. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. Genet. Sel. Evol. 43:23.
- Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42:565–569.