Journal of Animal Science, 2020, Vol. 98, No. 4, 1–14

doi:10.1093/jas/skaa101
Advance Access publication April 8, 2020
Received: 29 January 2020 and **Accepted:** 7 April 2020
Board Invited Review

BOARD INVITED REVIEW

Current status of genomic evaluation

Ignacy Misztal,[†,1] Daniela Lourenco,[†] and Andres Legarra[‡]

[†]Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, [‡]Department of Animal Genetics, Institut National de la Recherche Agronomique, Castanet-Tolosan, France

[1]Corresponding author: ignacy@uga.edu

ORCiD number: 0000-0002-0382-1897 (I. Misztal).

Abstract

Early application of genomic selection relied on SNP estimation with phenotypes or de-regressed proofs (**DRP**). Chips of 50k SNP seemed sufficient for an accurate estimation of SNP effects. Genomic estimated breeding values (**GEBV**) were composed of an index with parent average, direct genomic value, and deduction of a parental index to eliminate double counting. Use of SNP selection or weighting increased accuracy with small data sets but had minimal to no impact with large data sets. Efforts to include potentially causative SNP derived from sequence data or high-density chips showed limited or no gain in accuracy. After the implementation of genomic selection, EBV by BLUP became biased because of genomic preselection and DRP computed based on EBV required adjustments, and the creation of DRP for females is hard and subject to double counting. Genomic selection was greatly simplified by single-step genomic BLUP (**ssGBLUP**). This method based on combining genomic and pedigree relationships automatically creates an index with all sources of information, can use any combination of male and female genotypes, and accounts for preselection. To avoid biases, especially under strong selection, ssGBLUP requires that pedigree and genomic relationships are compatible. Because the inversion of the genomic relationship matrix (**G**) becomes costly with more than 100k genotyped animals, large data computations in ssGBLUP were solved by exploiting limited dimensionality of genomic data due to limited effective population size. With such dimensionality ranging from 4k in chickens to about 15k in cattle, the inverse of **G** can be created directly (e.g., by the algorithm for proven and young) at a linear cost. Due to its simplicity and accuracy, ssGBLUP is routinely used for genomic selection by the major chicken, pig, and beef industries. Single step can be used to derive SNP effects for indirect prediction and for genome-wide association studies, including computations of the *P*-values. Alternative single-step formulations exist that use SNP effects for genotyped or for all animals. Although genomics is the new standard in breeding and genetics, there are still some problems that need to be solved. This involves new validation procedures that are unaffected by selection, parameter estimation that accounts for all the genomic data used in selection, and strategies to address reduction in genetic variances after genomic selection was implemented.

Key words: genomic evaluation, genomic selection, large data, single-step GBLUP

Introduction

Genomic selection is now widely practiced across the breeding and genetics industry. This is evident by large-scale genotyping using inexpensive SNP chips. As of November of 2019, genotypes were available for over 3 million U.S. Holsteins (https://queries.uscdcb.com/Genotype/cur_freq.html), over 700,000 for American Angus (S. P. Miller, American Angus Association, Saint Joseph, MO, personal communication), and over 100,000 animals per line for some pig and broiler breeding companies.

Generally, the beginning of genomic selection is attributed to a study by Meuwissen et al. (2001). They used simulated data to conduct analyses with a large number of equally spaced markers; no attempt was made to identify QTLs but

Abbreviations

AF	allele frequencies
APY	algorithm for proven and young
BOO	breed of origin
DGV	direct genomic value
DRP	deregressed proofs
DYD	daughter yield deviation
G	genomic relationship matrix
GBLUP	genomic BLUP
GEBV	genomic estimated breeding values
GPU	graphical processing units
GWAS	genome-wide association studies
ICS	independent chromosome segments
LR	linear regression
PCG	preconditioned conjugate gradient
PEV	prediction error variance
ssBR	single-step Bayesian regression
ssGBLUP	single-step genomic BLUP
SVD	singular value decomposition
UPG	unknown parent groups

some markers were by chance to be closely linked to QTLs. Computations included haplotypes, and analyses were done by methods called BayesA and BayesB that assumed different distribution of haplotype effects. With 2,200 genotyped animals, they obtained prediction accuracies of 0.85. The accuracies were >0.7 after five generations without phenotyping or with only 500 genotyped animals. Such high accuracies with small data created high hopes in the animal breeding community.

The first large-scale genotyping was possible after the introduction of the SNP 50k bovine chip (Matukumalli et al., 2009), which provided an affordable and accurate technology for genotyping. Much subsequent work on the methodology of genomic selection focused on SNP effects and on the creation of a genomic relationship matrix (**G**) (VanRaden, 2008), a concept that allowed conceptual comparisons between pedigree-based and genome-based predictions. Methods using either SNP effects or genomic relationships led initially to field data analyses using a multistep methodology (VanRaden, 2008; VanRaden et al., 2009), where a regular genetic evaluation by pedigree BLUP (meant as non-genomic method throughout the paper) is followed by the extraction of pseudo-phenotypes of genotyped animals, a genomic analysis for genotyped animals, and the creation of an index combining results from BLUP and the genomic analysis (VanRaden et al., 2009). The multistep methodology is the natural choice when the genomic and pedigree/phenotypic data are owned by separate organizations.

When the genomic selection was introduced, the main focus was on testing models to increase accuracy, in particular increasing the accuracy of prediction by SNP selection (or differential weighting), assuming that it was possible to identify most pairs QTL-closest SNP from data. However, as the data grows bigger, gains with SNP selection become smaller or nonexistent (Karaman et al., 2016). Subsequently, most commercial evaluations do not use SNP selection.

The multistep method is relatively complicated and in its initial form relies on the existence of animals (bulls) with high accurate EBVs from pedigree information. It is also subject to double counting of the genomic information when both parents and progenies are genotyped. Because the genomic information can be expressed as genomic relationships (VanRaden, 2008), Misztal et al. (2009) proposed a single-step evaluation that enhanced the BLUP machinery with a relationship matrix that combines pedigree and genomic relationships. Subsequently, a pedigree-based BLUP with any model of analysis could support genomic models just by replacing the relationship matrix. A combined matrix was first shown by Legarra et al. (2009) and complete analysis using the so-called single-step genomic BLUP (**ssGBLUP**) was presented by Aguilar et al. (2010) and Christensen and Lund (2010). In the following studies, ssGBLUP was shown to be as accurate, or more, than multistep analyses.

Initially, the main focus of the single-step research was ensuring compatibility of genomic and pedigree information (Vitezica et al., 2011) because incompatibility creates biases, especially under strong selection. A later focus was extending ssGBLUP to larger numbers of genotyped animals (Legarra and Ducrocq, 2012; Fernando et al., 2014, Liu et al., 2014; Misztal et al., 2014a). Currently, ssGBLUP is the main tool for genomic evaluation in species other than dairy. If a population includes non-genotyped animals with phenotypes, the transition to some form of single step is unavoidable because BLUP, which is used to created pseudo-observations adopted in multistep, becomes biased by genomic preselection (Patry and Ducrocq, 2011b). The only alternative to ssGBLUP that has been explored is the use of segregation analysis to partially "infer" genotypes of the ancestors of genotyped animals, to later introduce this information in a refined ssGBLUP (Meuwissen et al., 2015). This strategy gave promising results but it is computationally complex and has not been pursued.

Advances in genotyping techniques are allowing sequence data to be generated at a lower cost; therefore, there is an interest to exploit these data (Georges et al., 2019). Sequence data can be used to identify recessive genes, targets for gene editing, and also potential causative SNP that can aid a genetic prediction across breeds or lines (Hayes and Daetwyler, 2019). However, gains with using the potential causative variants for genetic prediction appear to be limited (VanRaden et al., 2017; Fragomeni et al., 2019), with perhaps the exception of across-breed prediction (Moghaddar et al., 2019).

Although the rate of increase in genetic gain delivered by genomic selection can be over 100 % in some cases (García-Ruiz et al., 2016), issues have recently emerged. One important issue is the fast reduction in additive genetic variance and more undesirable genetic correlations between important traits (Hidalgo et al., 2020). This reduction is even more noticeable when genomic information is not used for variance components estimation. The same phenomenon may be responsible for the reported reduction of 33% in heritability in computations of genomic predictions for production yield traits to avoid bias in genomic estimated breeding values (**GEBV**; VanRaden et al., 2014).

From the onset of genomic selection, many ideas were proposed and usually tested by simulation, and many of these ideas were later applied to real data sets, first small then large. Many of these studies led to questions about various aspects of genomic selection, for example: Is it better to use haplotypes instead of SNP? What is the optimal number of SNP? Why there are discrepancies between simulation and field studies? Why is SNP selection less useful with large data? How can we use unknown parent groups (**UPG**) in genomic models? Is there any limit to genomic selection?. The purpose of this paper is to present and evaluate proposed ideas on genomic selection considering most up-to-date experiences with field data.

Exploring Genomic Selection Developments

Initial developments

Genomic selection is generally attributed to a study by Meuwissen et al. (2001) where simulated data for up to 2,200 phenotyped animals with genomic information expressed

as 50k haplotypes. For prediction, haplotype effects were estimated by several methods including treating them as fixed effects, by haplotype-BLUP assuming haplotypes as normal random effects, by BayesA assuming a *t*-distribution of effects (allowing for large effects), and by BayesB assuming a mixture distribution where most haplotypes had null effects. The accuracy of predicting breeding values for the next generation was the highest using BayesB and reached 0.85. The persistence of accuracies over generations (without additional phenotypes) was excellent, decaying only to 0.72 after an additional five generations of data. Reducing the number of animals with phenotypes to 1,000 only slightly reduced the accuracy to 0.79. The study by Meuwissen et al. (2001) generated great excitement in the animal breeding community, showing the possibility of very high accuracy with small data. However, this turned out to be true because the simulation was unrealistic with a small genome and QTLs of large effect, and no selection. Muir (2007) showed low persistence of genomic predictions under selection and dependence of accuracy on population parameters.

Much of the work involving the methodology of genomic selection on a practical side was accomplished by VanRaden (2008) using SNP markers instead of haplotypes. Also, he showed the equivalence of BLUP with SNP effects to genomic BLUP (**GBLUP**) using **G**; where **G** = **ZZ**/*k*, with **Z** being the matrix of gene content, $k = 2\sum_{i=1}^{n\ SNP} p_i(1-p_i)$, and p_i the frequency of the *i*th SNP. While genomic and pedigree inbreeding were highly correlated (r = 0.68) using base allele frequencies (**AF**) but lowly using current AF (r = 0.12), any AF resulted in similar prediction accuracy. The SE for elements of **G** was inversely proportional to the square root of the number of markers. He stated that genomic relationships are due to shared alleles, and he related the distribution of such alleles to a study by Stam (1980). VanRaden (2008) findings on gene frequencies were validated by Strandén and Christensen (2011) who showed that in SNP, BLUP and GBLUP AF only affect the mean of predictions. The number of shared alleles, also known as independent chromosome segments (**ICS**), was used for approximating the accuracy of genomic selection based on the number of genotyped animals and heritability (Goddard, 2009); lower Ne means fewer segments to estimate and higher accuracy of genomic selection for the same population size.

Limited dimensionality of genomic information

Genomic prediction in farm animals is possible because of the small effective population size (Ne). Stretches of DNA from overrepresented ancestors (i.e., popular bulls) form relatively few segments called LD blocks (Muir, 2007), shared segments (VanRaden, 2008), or ICS (Goddard, 2009). While the segments are not easily identified and have fuzzy limits (i.e., they are broken at slightly different places across two sibs), they appear indirectly, for example, as singular **G** that needs to be blended to become full rank. The number of chromosome segments is usually quantified by the formula presented by Stam (1980) as 4 NeL, where L is genome length. In a simulated population, Pocrnic et al. (2016a) found that the accuracy of prediction using a recursion was maximized assuming 4 NeL segments. They also showed that for a large population the number of segments can be estimated as the number of the largest eigenvalues explaining 98% of the variation in **G**, with the remaining 2% interpreted as noise. Extension of their studies to farm animals (Pocrnic et al., 2016b) allowed to determine the number of segments, and indirectly Ne, and the optimal size of the SNP chip for several species (see Table 1).

The genomic prediction does not act on individual segments but on their clusters, where the four largest clusters could account for 10% of the genetic variation (Pocrnic et al., 2019a). Small data only allow to estimate only the largest eigenvalues (or clusters), but they explain a large portion of the genomic variation in **G**. Subsequently, moderate accuracy of genomic selection can be achieved with small data sets, and large data sets are needed for additional improvements. The same study explains why SNP selection improves accuracy in small but not in large data sets. Genomic selection works by estimating the effects of chromosome segments, and once nearly all are well estimated, the accuracy is high without SNP selection (Karaman et al., 2016) or weighting (Lourenco et al., 2017).

Estimation of haplotypes or SNP effects

If the DNA information is inherited as chromosome segments, it would be natural to base the estimation on haplotypes rather than on SNP effects. Using haplotypes would potentially account for epistasis within each block, as for instance, a segment of 5 SNP can be estimated as having 25 different SNP combinations, as opposed to only 5 SNP effects. In practice, the difference in accuracy in models with haplotypes and SNP effects is negligible (Cuyabano et al., 2015; Jónás et al., 2016). Problems using haplotypes are the need for a complex data preparation and arbitrary choices in their definition, poor estimates for rare haplotypes and the existence of spurious haplotypes due to genotyping errors.

Multistep genomic evaluation

A study by VanRaden et al. (2009) using field data sets established a mature multistep methodology for genomic selection in dairy cattle. The steps included running pedigree-based BLUP with the national database, creating pseudo-observations for genotyped animals (bulls) such as daughter yield deviation (**DYD**). These pseudo-observations are fit into a model estimating on SNP effects assuming normal (linear) or non-normal (nonlinear) distributions. Finally, genomic predictions for a genotyped animal or candidate to selection are obtained combining pedigree and genomic-based predictions into an index:

$$\mathrm{GEBV} = w_1\mathrm{PA} + w_2\mathrm{DGV} - w_3\mathrm{PI},$$

Table 1. Estimated number of chromosome segments, effective population size, and the optimal size of SNP chip following Pocrnic et al. (2016)

Species	Estimated number of segments	Estimated optimal size of SNP chip	Estimated effective population size
Broiler chicken	4.2k	50k	44
Pig	4.1k	49k	48
Angus cattle	10.6k	127k	113
Jersey cattle	11.5k	138k	101
Holstein cattle	14.0k	168k	149

where GEBV is the genomic estimated breeding value, PA is parent average, **DGV** is direct genomic value, and **PI** is parental index created based on pedigree relationships for genotyped animals. Essentially, PI removes double counting of relationship information because part of PA is included in DGV, and weights w_1 to w_3 could be approximated from reliabilities of each component. Genomic prediction in the aforementioned paper was validated by forward prediction, where prediction for young bulls using truncated data were compared with pseudo-phenotypes of those bulls obtained by BLUP with complete data. That study also showed that increasing the reference size of the genotyped population has a higher impact on prediction accuracy than the number of SNP markers and that assuming non-normal SNP distribution has a positive effect only on traits with large effect QTL. Instead of creating DYD, which is time-consuming, pseudo-observation could be calculated as deregressed proofs (**DRP**) (VanRaden and Wiggans, 1991; Garrick et al., 2009; Wiggans et al., 2011).

Another version of a multistep method depended on using genomic predictions called molecular breeding value as a correlated trait (Kachman, 2008; MacNeil et al., 2010). In this version, GEBVs for genotyped animals based on previously estimated SNP effects were added as an extra trait with genetic correlation computed separately. Because genomic predictions indirectly include parent average, it was hard to account for double counting, and genetic trend was abnormally high even for early time periods when the genomic selection was not practiced (Lourenco et al., 2018).

Single-step genomic evaluations

The multistep methodology was well suited for scenarios where the phenotype and genomic data belong to separate organizations, and especially when most information can be condensed in a small number of animals to genotype. This includes dairy populations with a large number of average information bulls with many daughters. When a population includes both males and females, creating DRP free of double counting is hard, especially when genotyping includes parents and their progeny (Legarra et al., 2014). As the genomic information can be used to capture relationships, Misztal et al. (2009) proposed combining pedigree and genomic relationships into a combined relationship matrix. Subsequently, pedigree-only analyses could be converted to genomic analyses only by replacing the pedigree relationship matrix by the combined matrix, and the steps of construction pseudo-phenotypes and the index would no longer be needed. This combined matrix (**H**) was first presented by Legarra et al. (2009) who proposed to extend the genomic information to non-genotyped animals based on the joint distribution of breeding values of non-genotyped (u_1) and genotyped (u_2) animals:

$$\mathbf{H} = \begin{bmatrix} var(u_1) & cov(u_1, u_2) \\ cov(u_2, u_1) & var(u_2) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{bmatrix}$$

where subscripts 1 and 2 refer to non-genotyped and genotyped animals, respectively; **A** is the pedigree relationship matrix and **G** is the genomic relationship matrix or **G**. Christensen and Lund (2010) arrived to the same results, using the notion of predicting the genotype at non-genotyped individuals using pedigree information.

The inverse of **H** was presented by Aguilar et al. (2010) and Christensen and Lund (2010) as:

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$$

In the analyses done by Aguilar et al. (2010), the reliability of the new method called ssGBLUP was as high or higher than in multistep.

Further research involving ssGBLUP was split into several directions. The first was compatibility between pedigree and genomic relationships, as incompatibility can generate biases or losses of accuracy under selection (Vitezica et al., 2011). The second was an extension to a very large number of genotyped animals as initial implementation was based on dense matrices (Aguilar et al., 2011), which restricted the number of genotyped animals to about 100k animals. Finally, there was an interest in accommodating SNP weighting via a weighted **G**, especially with potential causative SNP obtained from sequence data. The interest in single-step methods increased as the genomic selection was underway, because pedigree BLUP and, therefore, multistep methods were becoming biased due to genomic preselection (Patry and Ducrocq, 2011b), whereas ssGBLUP accounts for preselection.

Different single-step formulations

Several alternative single-step formulas were proposed. These included equations where **G** is not inverted (Legarra and Ducrocq, 2012), where SNP effects are estimated for the genotyped animals and a polygenic effect is fit for non-genotyped animals (Legarra and Ducrocq, 2012; Liu et al., 2014), and where SNP effects were fit for all animals using imputed genotypes (Fernando et al., 2014; Taskinen et al., 2017). The purpose of these formulas was to reduce computations with many genotyped animals. As opposed to a regular ssGBLUP, which can be applied to an existing BLUP software just by replacing the relationship matrix, SNP-based models require new programming. Meuwissen et al. (2014) proposed an alternative single-step approach by combining identical by descent and identical by state approaches.

Compatibility between genomic and pedigree relationships

An important issue in single-step methodology is the compatibility of genomic and pedigree relationships. While the genomic relationships indirectly account for all the ancestors but have an arbitrary scale depending on gene frequencies, the pedigree relationships have a well-defined scale but are limited by the depth and completeness of the pedigree. When pedigrees were complete up to a base population, scaling **G** for compatibility (same means for diagonals and off-diagonals) with the pedigree relationship matrix for genotyped animals ($\mathbf{A}_{22}$) improved accuracy and eliminated bias for a population under strong selection (Chen et al., 2011a; Vitezica et al., 2011). With no selection, the impact of scaling was minimal. Similar scaling could be accomplished automatically by using base population gene frequencies (Strandén and Christensen, 2011; Christensen, 2012), although finding those frequencies when the base population is not genotyped, for example, using the method of Gengler et al. (2008), can be time-consuming, and it suffers sometimes from a clear definition of base population as described below.

When the base populations are heterogeneous with missing pedigrees across generations, as is typical in ruminants, ssGBLUP may diverge or become biased, and the standard way to ensure convergence was by including a parameter ω as in Tsuruta et al. (2011):

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - \omega A_{22}^{-1} \end{bmatrix}$$

The parameter ω compensates for incomplete pedigree and incomplete accounting of inbreeding (Misztal et al., 2017). Incompleteness of pedigree in $\mathbf{A}_{22}$ can be minimized by truncating old data and pedigree (Misztal et al., 2013b) or by assigning nonzero inbreeding to unknown parents (VanRaden, 1992); old missing pedigree becomes irrelevant with truncation of data. Truncation to two generations of phenotypes and three generations of pedigree reduced bias without lowering accuracy (Lourenco et al., 2014; Howard et al., 2018).

In a single-step SNP-based model known as single-step Bayesian Regression (**ssBR**) developed by Fernando et al. (2014), the compatibility between genomic and pedigree information is provided partially by the use of fixed effects in a model for genotyped animals (Hsu et al., 2017), similar to Vitezica et al. (2011) where this effect is implicitly fit as random. This arises from the findings of Strandén and Christensen (2011) that solutions from SNP BLUP and GBLUP are independent of gene frequencies if the model includes a mean. However, the missing pedigree problem is present in all single-step formulations and it becomes more complex with several base populations as described below.

Missing pedigree and UPG

In several species, there is a need to define several populations. This is the case in ruminants with missing parents (whereas unknown parents of animals born in 2000 are better than unknown parents of animals born in 2016), and the case in pigs and birds (with several lines collapsing into one, and with 2-, 3-, and 4-way crosses). These base populations have different means due to selection and not considering them leads to strong biases.

In BLUP, the genetic merit of these different base populations is often modeled by genetic or UPG (Quaas and Pollak, 1981; Quaas, 1988; Westell et al., 1988). In ssGBLUP, when UPG are applied only to pedigree relationships (**A**) as follows:

$$\mathbf{H}^* = \mathbf{A}^* + \begin{bmatrix} 00 & 0 \\ 0\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & 0 \\ 00 & 0 \end{bmatrix}$$

The convergence rate can be slow or no convergence may be reached (Tsuruta et al., 2014; Matilainen et al., 2016), partly because UPG were ignored in pedigree relationships for genotyped animals ($\mathbf{A}_{22}$). Indeed, construction of $\mathbf{A}_{22}$ implicitly assumes complete pedigrees. Misztal et al. (2013b) revised UPG equations to include groups also in the genomic portion of **H** based on Quaas–Pollak (QP) transformation:

$$\mathbf{H}^* = \mathbf{A}^* + \begin{bmatrix} 00 & 0 \\ 0\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & -\left(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}\right)\mathbf{Q}_2 \\ 0 & -\mathbf{Q}_2'\left(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}\right)\mathbf{Q}_2' \left(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}\right)\mathbf{Q}_2 \end{bmatrix}$$

When UPGs were applied to all components of **H** as above, convergence dramatically improved for a multitrait model in the Nordic dairy cattle population (Matilainen et al., 2016). Revised UPGs also worked well for the U.S. Holstein data up to 2014 (Misztal et al., 2017). However, using data updated to 2015, Masuda et al. (2018a), based on cross-validation, reported lower reliabilities using revised UPG than not using UPG at all. While most animals genotyped earlier were potentially elite, with complete pedigree, most genotyped animals after 2014 were commercial cows, often with incomplete pedigree and high pedigree error rate (Bradford et al., 2019a).

It is not clear whether the equations for UPG in ssGBLUP should be considered for **G** for a single breed as genomic relationships are not affected by missing pedigree, and, therefore, UPG are automatically accounted for. In other words, if all animals were genotyped, terms involving UPG should disappear from $\mathbf{H}^*$ above. Tsuruta et al. (2019a) found that removing **G** from the equation above improved accuracy and reduced bias. In GBLUP, using UPG for **G** did not increase accuracy for multi-breed populations (Plieschke et al., 2015).

Missing relationships also cause underestimation of inbreeding as animals with missing parents are automatically treated as not inbred. One solution is assigning nonzero inbreeding to missing parents (VanRaden, 1992; Lutaaya et al., 1999; Aguilar and Misztal, 2008). Such assignment improved convergence rate and bias in ssGBLUP in Holsteins (Misztal et al., 2017; Tsuruta et al., 2019a) although it only slightly affected the accuracy.

The concept of metafounders

Legarra et al. (2015) proposed to account for UPG while providing proper scaling by generalizing UPG to metafounders. In their approach, **G** would be derived using 0.5 AF as an "absolute reference" (Christensen, 2012), and **A** would be scaled for compatibility with **G** using relationships among and within metafounders, which are seen as pseudo-individuals. These relationships represent sizes and overlaps of the different base populations (Legarra et al., 2014). They can be estimated in such a way so that they account for scaling, unaccounted inbreeding, different genetic level (e.g., when using multi-breed animals or selected populations), and multiple breeds and crosses. Several methods were proposed to estimate the relationships, and, in practice, they imply estimating gene frequencies in the different base populations (Garcia-Baccino et al., 2017). In simulations and real data, the concept of metafounders delivered the least biased predictions (Garcia-Baccino et al., 2017; Meyer et al., 2018; Bradford et al., 2019b). When applied to dairy cattle, the relationships across metafounders could be well estimated only for metafounders associated with sufficient number of genotypes (S. Tsuruta, University of Georgia, Athens GA, personal communication). In dairy sheep, the use of metafounders reduces biases in predictions and instability of UPG estimates for small data sizes (F Macedo, INRAE, Toulouse, France, personal communication).

Evaluations of crossbred populations

Genomic evaluation of crossbred populations may be separated into two types, for specific crosses or for complex crosses. In pigs and chicken, there is an interest in using F1 and possibly three- to four-way crosses for the evaluation of purebreds on the commercial scale. In beef and dairy, the interest is to have a joint analysis of many breeds with complex crosses (e.g., "Kiwi" Jersey–Holstein crosses in New Zealand, 10+ breed crosses by International Genetic Solutions, and 50+ beef crosses by The Irish Cattle Breeding Federation). More recently, across-breed prediction with genomic data is not successful (Erbe et al., 2012; Kachman et al., 2013) because the breeds do not share the same chromosome segments. Also, the crossbreeds generate limited information if the amount of crossbred data is small and if they are progeny of very few parents (Pocrnic et al., 2019b). Genetic by environment interaction and purebred-crossbred correlations can be considered using multiple-trait models (Xiang et al., 2016; Vandenplas et al., 2017). With purebreds and defined crosses (F1), the genomic relationships can be adjusted

separately for each breed combination using gene frequencies or other methods (Makgahlela et al., 2014; Lourenco et al., 2016) although the impact of such adjustment is small if the selection pressure is low. With many crosses, a simple approach is to ignore gene frequencies and have one set of SNP effects (Golden et al., 2018) or one **G** (Mäntysaari et al., 2017) for all breeds and breed combinations. Steyn et al. (2019) simulated five breeds using either shared or separate relationships. In the second case, the accuracy was compromised if the number of SNPs was reduced from 45k to 9k, and despite all breeds having identical QTLs, interbreed predictions had low accuracy. In U.S. dairy, SNP effects are estimated separately for each breed as otherwise the predictions would be based on the dominating breed—Holsteins (VanRaden et al., 2020); phenotypes of crossbreds are not used in the regular genetic evaluation of purebreds because of concerns of compromising the evaluation of purebreds. The most refined method for the F1 crossbreds is by phasing haplotypes in crossbreds originating from two parental lines and building a model with two H matrices, one per breed (sometimes called the breed of origin [**BOO**] model) (Christensen et al., 2014). Xiang et al. (2016) observed an increase in accuracy compared with fitting a single H matrix in an analysis of Landrace, Yorkshire, and crosses. The method becomes complex for more complex crosses as the origin of alleles in each crossbred is more difficult to establish.

The concept of metafounders provides a convenient solution to ssGBLUP applied to purebreds and crossbreds (Christensen et al., 2015; Xiang et al., 2017). In such a case, the relationship across breeds represent a distance from a common genetic origin (usually a small relationship, but potentially different across pairs of breeds), and the variances within breed reflect correct scaling separately for each breed and for all breeds simultaneously (Legarra et al., 2015). Xiang et al. (2017) fit this model treating each breed combination as a different trait to account for G × E and observed the same accuracy as in the BOO model of Xiang et al. (2016).

Modifying single step for large data sets

Single-step GBLUP requires explicit or implicit computations of $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$. When created using dense matrix techniques (Aguilar et al., 2011), the practical limit is about 100k animals. This is because computations increase cubically and storage quadratically with the number of genotyped animals. Several strategies were proposed to overcome size limitations.

Indirect computations of A_{22}^{-1}

Matrix $\mathbf{A}_{22}^{-1}$ is dense and, therefore, cannot be created efficiently for a large number of genotyped animals. Henderson (1976) showed that the inverse of a submatrix of **A** could be obtained based on the rules for inversion of a partitioned matrix:

$$\mathbf{A}_{22}^{-1} = \mathbf{A}^{22} - (\mathbf{A}^{12})'\left(\mathbf{A}^{11}\right)^{-1}\mathbf{A}^{12}$$

When only a product of $\mathbf{A}_{22}^{-1}$ and a vector is required in the iteration process as in the preconditioned conjugate gradient (**PCG**) algorithm, that product can be calculated sequentially every round as follows (Masuda et al., 2017; Strandén et al., 2017):

$$\mathbf{A}_{22}^{-1}\mathbf{q} = [\mathbf{A}^{22} - (\mathbf{A}^{12})'\left(\mathbf{A}^{11}\right)^{-1}\mathbf{A}^{12}]\mathbf{q}$$

where the product:

$$s = \left(\mathbf{A}^{11}\right)^{-1}\mathbf{A}^{12}\mathbf{q}$$

is computed as a solution to:

$$\mathbf{A}^{11}s = \mathbf{A}^{12}\mathbf{q}$$

using sparse matrix techniques, in particular, because $\mathbf{A}^{11}$ is sparse and small. Masuda et al. (2017) found that, for a U.S. Holstein population, this algorithm required 2 min to set up and less than 1 s per round of multiplication.

Algorithm for proven and young

Because of small effective population size in farm animals, **G** has a rank of about 5k for pigs and chicken to about 15k for beef and dairy (Pocrnic et al., 2016b), indicating the existence of that many LD blocks or chromosome segments. Subsequently, the inverse of **G** can be obtained by recursion on a number of "core" animals equal to the rank of **G**, indirectly assuming that breeding values of N animals contain the same information as the effects of N chromosome segments. When animals are designated as core (c) or noncore (n), the inverse of **G** can be directly obtained as (Misztal, 2016):

$$\mathbf{G}_{\text{APY}}^{-1} = \begin{bmatrix}\mathbf{G}_{cc}^{-1}0\\ 0\quad 0\end{bmatrix} + \begin{bmatrix}-\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn}\\ \mathbf{I}\end{bmatrix}\mathbf{M}^{-1}\left[-\mathbf{G}_{nc}\mathbf{G}_{cc}^{-1}\mathbf{I}\right]$$

where **M** is a diagonal matrix with elements:

$$m_i = g_{ii-}\mathbf{g}_{ic}\mathbf{G}_{cc}^{-1}\mathbf{g}_{ci}$$

where i refers to the *i*th genotyped, noncore animal. This method has almost a linear cost (computations and memory) with the number of animals (Fragomeni et al., 2015) and has been successfully applied to 2.3 M genotyped animals (Tsuruta et al., 2019b). The choice of core animals for recursion is not critical for accuracy when the number of core animals is sufficient but influences the convergence rate; the random choice is preferable (Bradford et al., 2017). Lately, Pocrnic et al. (2019a) found that accuracies obtained with N core animals are like those obtained with **G** ignoring all but the largest N eigenvalues. This explains why the accuracy with the algorithm for proven and young (**APY**) using 25% of the optimal number of core animals is almost the same, as 25% of important eigenvalues explain 90% of the genetic variation in **G**. In fact, the recursion acts not on individual chromosome segments but on their clusters.

Inverse by singular value decomposition

The inverse of **G** can be derived from the eigenvalue decomposition:

$$\mathbf{G} = \mathbf{UDU}'$$

where **U** is a matrix of eigenvectors and **D** is a matrix of eigenvalues. If all eigenvalues are positive, the inverse of **G** is:

$$\mathbf{G}^{-1} = \mathbf{U}'\,\mathbf{D}^{-1}\mathbf{U}$$

If **G** has small rank, only a small fraction of eigenvalues will be meaningful. Let $\mathbf{D}_t$ indicate a fraction of **D** with non-negligible eigenvalues, and let $\mathbf{U}_t$ be the corresponding eigenvectors. Then:

$$\mathbf{G}_t^{-1} = \mathbf{U}_t'\,\mathbf{D}_t^{-1}\mathbf{U}_t$$

While eigenvalue decomposition of **G** requires creating **G** explicitly and can be very expensive, a less expensive alternative, when there are more genotyped animals than SNP, is the

singular value decomposition (**SVD**) of the matrix of SNP content (**Z**), where $\mathbf{Z} = \mathbf{UD}^{0.5}\mathbf{V}$. The SVD for a matrix of 720k animals by 60k SNP takes less than a day (Y. Masuda, University of Georgia, Athens, GA, personal communication). The SVD concept can be applied separately for each chromosome (Ødegård et al., 2018)

Inverse by the Woodbury formula

Mäntysaari et al. (2017) proposed an inverse of $\mathbf{G} = \mathbf{ZZ}' + \mathbf{I}\varepsilon$ based on the Woodbury formula to overcome computing challenges when the number of genotyped animals is greater than the number of SNP:

$$\mathbf{G}^{-1} = \frac{1}{\varepsilon}\mathbf{I} - \frac{1}{\varepsilon}\mathbf{Z}(\frac{1}{\varepsilon}\mathbf{Z}'\mathbf{Z} + \mathbf{I})^{-1}\mathbf{Z}'\frac{1}{\varepsilon}$$

where $\mathbf{Z}'\mathbf{Z}$ is the design matrix of SNP BLUP and **I** is an identity matrix with the same dimension. The formula is an exact inversion but is based on an arbitrary value of ε (i.e., 0.05**I**, $0.05\mathbf{A}_{22}$), without which **G** could not be full rank. The "Woodbury" $\mathbf{G}^{-1}$ is dense and is not used explicitly. Its use is only for PCG systems in which only a product of this matrix by a vector is desired, being reformulated as:

$$\mathbf{G}^{-1}\mathbf{q} = \frac{1}{\varepsilon}\left\{\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \mathbf{I}\varepsilon)^{-1}\mathbf{Z}'\right\}$$
$$q = \frac{1}{\varepsilon}\left\{\mathbf{I} - \mathbf{Z}(\mathbf{UDU}')^{-1}\mathbf{Z}'\right\}$$
$$xq = \frac{1}{\varepsilon}\{\mathbf{I} - \mathbf{SS}'\}q,$$

with

$$\mathbf{S} = \mathbf{ZU}'\mathbf{D}^{-1/2}$$

Matrix **S** has dimensions equal to the number of animals by the number of SNP. In practice, the SNP BLUP design-matrix $\mathbf{Z}'\mathbf{Z}$ is not full rank, and one dimension can be reduced to the actual rank (5k to 15k for one breed) by truncating **U** and **D** to eliminate small eigenvalues.

Single-step Bayesian Regression

If 50k SNP are enough for predictions, an alternative idea was to impute genotypes of non-genotyped animals, resulting in the same 50k SNP effects to estimate regardless of the number of genotyped animals. Let $\mathbf{u}_2$, the vector of breeding values for genotyped animals be equal to **Za**, where **a** is a vector of SNP effects. Legarra et al. (2009) showed that the conditional distribution of breeding values for non-genotyped and genotyped animals has an expectation equal to $\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2$. Replacing $\mathbf{u}_2$ by **Za**:

$$\mathbf{u}_1 = \mathrm{E}(\mathbf{u}_1 \mid \mathbf{u}_2) + \varepsilon = \mathbf{A}_{12}(\mathbf{A}_{22})^{-1}\mathbf{Za} + \varepsilon = \mathbf{Ta} + \varepsilon$$

where **T** can be called an imputation matrix for non-genotyped animals and ε can be called an imputation error. Then, the breeding values in an animal model can be replaced by:

$$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} T \\ Z \end{bmatrix}\mathbf{a} + \begin{bmatrix} \varepsilon \\ 0 \end{bmatrix}$$

Regardless of the number of animals, the number of "genomic" unknowns is equal to the number of SNP, although there is an additional uncorrelated effect ε with a simple relationship structure (Fernando et al., 2014). The model was reformulated for economy of memory by Fernando et al. (2016) who called it "hybrid model," although the same model had already been proposed by Legarra and Ducrocq (2012). As the imputation was expensive, and the model is conceived to use Gibbs sampling methods, the implementation of ssBR in the BOLT software used graphical processing units, that is, **GPU** (Garrick et al., 2018). Compared with ssGBLUP, ssBR allows the user to estimate SNP effects directly but the implementation of complex models (e.g., correlated maternal effects with multiple traits) is quite complex. The method of ssBR was used for a multi-breed evaluation done by the Simmental association for more than 10 breeds (Golden et al., 2018) but they decreased the number of SNPs from 50k to about 2.5k preselected SNP, contrary to all other species who abandoned the idea of preselecting markers because an optimal subset of markers may not be optimal a few generations later. A more general (and simple) formulation of the ssBR model was given by Taskinen et al. (2017).

Other approaches

Legarra and Ducrocq (2012) developed an assymetric method where **G** was not inverted, but the method did not scale up well. Also, both Legarra and Ducrocq (2012) and Liu et al. (2014) proposed methods that used SNP effects estimated for genotyped animals. Vandenplas et al (2019) showed that such models when solved by the PCG algorithm require a special preconditioner for convergence.

Preselection bias

Under genomic selection, BLUP becomes biased (Patry and Ducrocq, 2011a, 2011b) due to preselection on Mendelian sampling; for instance, only offspring that has received the "good" alleles from a sire gets to be recorded. This has an impact on multistep methods which use BLUP as a first step, because they will tend to penalize genomically selected animals and, therefore, to underestimate the genetic trend. The bias can be corrected for (Wiggans et al., 2011, 2012), but the corrections need to be reevaluated as genotyping increases. Single-step GBLUP is expected to be resistant to selection bias (VanRaden et al., 2012; Legarra et al., 2014) as it considers all available information jointly. Masuda et al. (2018b) ran evaluations with BLUP and ssGBLUP for production traits in U.S. Holsteins. They found that the trends for BLUP level off, when they should actually increase, whereas trends for ssGBLUP were consistent. Based on the work at UGA in dairy and in pigs (unpublished), typical trends for genotyped animals by BLUP and ssGBLUP indicating preselection in BLUP are shown in Figure 1. The preselection bias can intensify when more animals are genotyped.

Differences between trends from BLUP and ssGBLUP can be used indirectly as a measure of the effectiveness of the genomic section. If the trend by ssGBLUP is increasing and the trend by BLUP is lower, genomic selection is successful. If the trends by both methods are identical, genomic selection does not have an impact over the regular selection. If in an extreme case, the trends by ssGBLUP decrease, it means either poor implementation of genomic selection or a change in the selection objectives.

Validation of genomic predictions

Genomic evaluations are validated by realized accuracies or reliabilities computed from predictions based on incomplete data to predictions/phenotypes based on complete data —see review by Daetwyler et al. (2013) and Legarra and Reverter (2018). Several types of validations are currently applied, and each one is suitable for a different data structure. The k-fold cross-validation depends on splitting the population into *n* samples

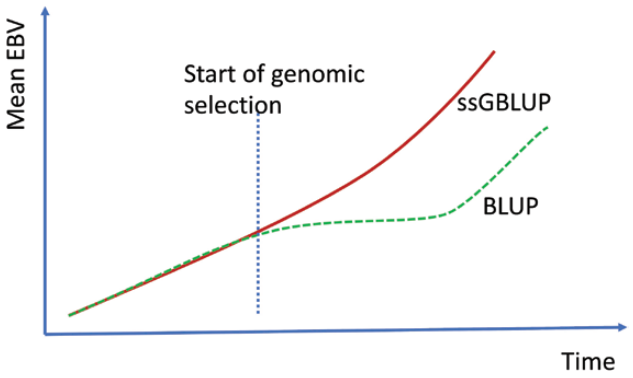

Figure 1. Trend of (G)EBV with ssGBLUP (solid) and BLUP (dashed) indicating preselection bias in BLUP.

and predicting phenotypes of one sample from the remaining samples (Saatchi et al., 2011). It is primarily used for small data sets when only one generation is genotyped or when other methods cannot be applied. As it follows from the decomposition of GEBV (VanRaden and Wright, 2013; Lourenco et al., 2015a), accuracies by clustering methods, such as the k-fold, depend on the algorithm for creating the clusters. In particular, BLUP may emerge as the best method if most animals in each cluster are progeny of the same ancestors.

The validation needs to consider that the breeder wants to predict the next generation from former ones, in other words, forward and not backward or "sideways." Thus, another validation method is based on a comparison of pseudo-observations of sires (DYD or DRP) with their (G)EBV obtained without their daughter's information (VanRaden et al., 2009). This validation type is only realistic for populations with sires that have large progeny sizes and phenotype recording is mainly on progeny, such as in dairy cattle. If pseudo-observations are computed by BLUP under genomic selection, this validation may be biased by preselection (Masuda et al., 2018b). If pseudo-observations are computed by ssGBLUP, the bias can be avoided but there is a danger of double counting of the genomic information, especially if progeny sizes are small. Yet another validation method that is called predictive ability or predictivity can be used when the validation animals have only their own records but not progeny (Legarra et al., 2008). It is based on correlations between GEBV obtained without a phenotype and the phenotype adjusted for fixed effects. However, it can only be computed for simple models and depends on the quality of adjustments (Legarra and Reverter, 2018). Accuracies based on validation are depressed by selection and, therefore, are lower than individual theoretical accuracies based on prediction error variance (**PEV**; Bijma, 2012; Lourenco et al., 2015a).

A completely different approach to validation was taken by Legarra and Reverter (2018) in a method called **LR**, which stands for linear regression. The method LR examines regressions and correlations of (G)EBV using complete and partial data sets while accounting for the relatedness of animals in the validation and additive variances under selection. The advantage of method LR is the ability to support any model and any data structure. For example, Bermann et al. (2020) were able to calculate the accuracy of evaluations for a threshold model. However, the method requires the additive variance for the validation population, which may be hard to estimate as typically these are a subset of selected animals. Without such a variance, only relative comparisons among methods are possible, although they are useful to rank methods.

Individual theoretical accuracies

Individual accuracies are published with (G)EBV as a measure of precision and they are based on true or approximated PEV derived from mixed model equations (Henderson, 1984). The PEV can be obtained either via efficient matrix inversion, for example, by REML with sparse matrix package YAMS (Masuda et al., 2015) or via Gibbs sampling (Tsuruta et al., 2017; Garrick et al., 2018). This is affordable for up to ~100K individuals genotyped. The last option can support larger data sets if the computation is by GPU (Garrick et al., 2018). For complex models and large populations, the computation of PEV is usually too expensive and approximations are used instead. With genomic information, the PEV for the ith animal can be approximated as (Misztal et al., 2013a):

$$\mathrm{PEV}_i \quad \sigma_e^2 \approx \frac{1}{\frac{\sigma_e^2}{\sigma_a^2} + d_i^r + d_i^p + d_i^g}$$

where d are contributions (in terms of effective daughters or observations) due to pedigrees (r), phenotypes (p), and genomic information (g), and σ_a^2 and σ_e^2 are additive and residual variances, respectively. Approximate contributions due to pedigree and phenotypic information were determined by earlier studies (Misztal and Wiggans, 1988; Meyer, 1989; VanRaden and Wiggans, 1991).

With the multistep SNP model, the contribution due to genomic information could be calculated by inversion for any number of genotyped animals (VanRaden et al., 2011; Liu et al., 2017b). To avoid double counting, the calculations exclude the genomic information that is already included in the pedigree information. In ssGBLUP, the genomic contribution can be calculated by combined differences between genomic and pedigree relationships (Misztal et al., 2013a). Edel et al. (2019) provided formulas for avoiding double counting in ssBR. Efficient computation of genomic accuracies for any model and data set is still a research topic but not a hot one because when the models are too large for direct inversion, genomic predictions are accurate enough for selection.

Genetic parameters under genomic selection

Plant breeders estimate variance components at each genetic evaluation, partly because they have several random effects (e.g., blocks) and partly because their data sets are small. In contrast, animal breeders tend to use either once-in-a-while estimates or to use pedigree-based estimates for genomic evaluation purposes, for example, as in VanRaden (2008). Genetic parameters can be estimated with genomic information using ssGBLUP and normal tools such as REML or Monte Carlo Markov Chain via Gibbs sampling. The use of the genomic information increases the costs of computations because the inverse of **G** is usually dense, whereas non-genomic mixed model equations are sparse. Masuda et al. (2014) developed a sparse matrix package that recognizes and processes dense blocks rapidly. A four-trait single-step AIREML model with 15k genotyped animals took less than 1 h with the new package (Masuda et al., 2015); the computations increase cubically with the number of genotyped animals and of traits.

Comparing genomic and pedigree-based estimates of variance components relies on the compatibility of genomic and pedigree information (Legarra, 2016). Without selection and with a complete pedigree, the estimates of variance components ignoring or using the genomic information are usually similar, although with the genomic information they have lower standard errors (Forni et al., 2011). Under strong

selection, the estimates ignoring the genomic information are biased (Gao et al., 2019; Hidalgo et al., 2020). The computed bias due to preselection depends on the accuracy of modeling and intensity of selection. For example, the popular QMSim program for simulation of genomic data (Sargolzaei and Schenkel, 2009) only performs BLUP selection. If various types of single-step methods show different results despite being equivalent models, the actual variances are affected by small details in the models (e.g., Gao et al., 2019)

Before genomic selection, the genetic parameters were thought to be generally stable, but this was not studied in depth. Under genomic selection, there are indications of rapidly changing parameters, perhaps due to the Bulmer effect (Van Grevenhof et al., 2012; Hidalgo et al., 2020). For instance, bias for U.S. dairy genomic evaluations decreased when heritability was reduced to about 70%—50% of the original value (Wiggans et al., 2012; Misztal et al., 2017), which is an indicator of overestimated heritability. Hidalgo et al. (2020) used a Gibbs sampling approach to analyze the changes in genetic parameters for growth and fitness traits in pigs. To make the computations possible, analyses were done in time slices of 3 yr, and genotypes were restricted to parents and animals with records. Over time, heritabilities for growth were reduced by one half and the antagonistic genetic correlations between growth and fitness traits became almost twice as strong. Estimates without genomic information were quite different. The aforementioned study illustrated the tradeoffs in parameter estimation under genomic selection. Without genomic information, the estimates may be biased, and with all the genomic information available the computation are expensive. Cesarani et al. (2019) reported biased variance components estimates under genomic selection when the genomic information was truncated or too few generations were used. A modest compromise is to restrict genotypes only to those animals on which selection was more intense and to remove genotypes of all young animals and possibly of nonparents. Genetic parameter estimation with a large number of genotypes can be possible in GBLUP when the APY algorithm is applied. However, in ssGBLUP, $\mathbf{A}_{22}^{-1}$is relatively dense and using it in computations eliminates most of the gains due to using a sparse $\mathbf{G}_{APY}^{-1}$.

Stability of GEBV

Under BLUP, the evaluation of an animal depends nearly only on its phenotype, parents, and progeny. Therefore, EBV for animals with no new information are stable even if the accuracy is low (and PEV high). In genomic evaluations, all genotyped animals are connected through **G**. It means that information on new genotyped animals affects all the other genotyped animals, causing fluctuations. Changing core animals in the APY algorithm also causes fluctuations in GEBV even though the accuracy is not affected (Misztal et al., 2019). When short-term fluctuations are undesirable, for example, for merchandising, one solution is to use full model genomic prediction (by SSGBLUP or multistep methods) periodically (say once a month), compute SNP effects, and run interim (e.g., weekly) indirect predictions based on backsolved SNP effects. While with small data the indirect predictions can be inaccurate due to ignoring parent average, in large populations, the fraction of parent average in GEBV is small and indirect predictions have similar accuracy to complete predictions (Lourenco et al., 2015b; Garcia et al, 2020). To mitigate risk associated with potential rank changes of young bulls, semen from a team of bulls may be marketed instead of semen from individual bulls (e.g., https://www.dairynz.co.nz/animal/animal-evaluation/bull-team/).

Using sequence data for genomic predictions

As sequencing is becoming less expensive, there is an interest in exploiting sequence information in animal genetics. If all causative variants and their substitution effects could be identified, genomic prediction would be perfect (i.e., selection accuracy = 1.0). If those effects were conserved across breeds, accurate multi-breed evaluation would be possible (Goddard, 2017). But substitution effects may vary from breed to breed, even at the QTL level, due to gene–environment interaction and to nonadditive gene action (Duenk et al., 2020). Sequence data are available through selective sequencing of key animals across species (e.g., 1000 Bull Genomes Project; http://www.1000bullgenomes.com/; Hayes and Daetwyler, 2019) and imputation for the remaining animals (Ros-Freixedes et al., 2020). For a successful incorporation of potential causative SNP, they need to be very close to the actual causative SNP, and their a priori variance in a model need to be large as otherwise their value is strongly regressed toward 0 (Brøndum et al., 2015).

Practical results using sequence data from large data sets yielded mixed results. Some studies have found no improvement (Erbe et al., 2012) and some showed a small improvement, in particular Moghaddar et al. (2019) who found an increase in the accuracy of "distant" animals of ~0.10 using selected sequence variants. In a study that yielded up to 5% improvement in reliability across traits, VanRaden et al. (2017) partly used a bin concept, where they eliminated most of the SNP close to SNP with the largest effects. The bin concept, popular in plants, recognizes that QTLs are nested in chromosome segments and attempts to locate at most a few SNP per segment (Xu, 2013); fewer SNP reduce the impact of priors and reduce shrinkage of causative SNP. Fragomeni et al. (2017) showed that ssGBLUP can account for causative SNP if they have a large weight in a weighted **G**. In a study on stature in U.S. Holstein using the potential causative SNP identified by VanRaden et al. (2017), Fragomeni et al. (2019) found that the addition of potential causative SNP to the current SNP panel increased reliabilities in GBLUP but not in ssGBLUP, and reliabilities from ssGBLUP were the highest. Similar results were found in Belgian Blue cattle (J.L. Gualdrón-Duarte, University of Liege, Belgium, personal communication) and for health traits in dairy cattle (S. DeNise, Zoetis, Kalamazoo MO, personal communication). A few real, validated major genes (as identified by molecular genetics) explaining up to 10% of the genetic variance have been found and included in ssGBLUP evaluations, either as correlated traits (Legarra and Vitezica, 2015) or as weighting the G matrix appropriately. In general, any of these two strategies work and result in small, but less than expected, improvements on accuracy (Carillier-Jacquin et al., 2016; Teissier et al., 2018; Oget et al., 2019). Possibly, the causative SNP are already accounted for by the values of chromosome segments with large data. Some improvement with the causative SNP could be due to imperfect modeling by GBLUP with pseudo-data such as DRP or DYD instead of records.

There is a dilemma whether causative SNP with large effect, if found, should be used in selection programs for strongly selected traits. With long-term selection, most likely genes with positive effect for most traits are fixed or close to fixation, and genes that still have a large effect but are not fixed are likely to show undesirable pleiotropy. A chromosomal deletion in pigs increased growth but decreased fertility (Derks et al., 2018). Manhattan plots for mortality and milk yield, using a two-trait analysis, in U.S. Holstein showed the same peaks on chromosome 14 (Tsuruta et al., 2017). Georges et al. (2019) cite many studies indicating pleiotropy as a result of balancing

selection, for example, where disruptive variants in genes increase muscularity but affect the viability of fitness. Negative effects of pleiotropy on low heritability traits may be hard to identify but can be important in the long run.

Balancing selection resulting in intermediate gene frequencies may be unlikely in cases where selection indices are utilized even with pleiotropy. While causative SNPs with large effects are not likely to be fixed after years of pedigree BLUP, the trend toward fixation will be faster with genomic selection. In the extreme, the fixation will negatively affect low heritability or sparsely recorded traits.

Genome-wide association studies

A standard tool for traditional genome-wide association studies (**GWAS**) is a model where one marker is analyzed at a time as fixed effect (Kennedy et al., 1992), for example, an efficient mixed-model association expedited—EMMAX (Kang et al., 2010). To reduce spurious signals due to a population structure, an animal effect using a pedigree or **G** is added to the model (Kennedy et al., 1992). Alternatively, many studies use Bayesian methods such as BayesB or BayesR with all SNP considered jointly, interpreting large signals as markers to nearby QTLs. While the former studies determine SNP significance using *P*-values, the latter usually estimate fractions of explained variance per segment of the genome, for example, 1 Mb.

Many studies, especially using small data, detect a large number of "large" markers, interpreting those as close to a QTL; however, the overlap of those markers across multiple populations or generations in a population under selection is minimal. This suggests that many detected associations are spurious (Fragomeni et al., 2014; Liu et al., 2017a). Studies using BayesB often show very high peaks, sometimes explaining >10% of the additive variance, especially with small data sets. As genomic selection with small data works on large clusters of chromosome segments (Pocrnic et al., 2019a), it is possible that some peaks may be tags to those clusters.

Many of these signals in GWAS are, therefore, probably false positives and can probably be explained by small data sets. If pedigree relationships are incomplete (e.g., ancestors not included), they would not account for population structure. In addition, *P*-values or False Discovery Rate are rarely reported.

Classical GWAS in EMMAX is conceived for a set of individuals that are genotyped and phenotyped. When genotyped animals have only records from progeny or other relatives, this method is only applicable in a multiple-step manner, that is, creating pseudo-phenotypes such as DRP or DYD as it was typically the case in dairy cattle (Boichard et al., 2003), but this is difficult to generalize to other species, where progeny sizes are smaller and many genotyped have phenotypes (e.g., weights) but not genotypes. However, Gualdrón-Duarte et al. (2014) and Bernal Rubio et al. (2016) showed the equivalence of *P*-values in GBLUP-based models with *P*-values in single-marker fixed regressions with a polygenic effect. Lu et al. (2018) extended the theory to ssGBLUP, and Aguilar et al. (2019) added this concept to the BLUPF90 package (Misztal et al., 2014b), with a successful implementation using 1 million birth weight phenotypes for American Angus, almost 2 million animals in the pedigrees, and 1,424 genotyped sires. The GWAS with *P*-values from ssGBLUP accounts for population structure, considers phenotypes from both genotyped and non-genotyped animals without additional steps, and allows for arbitrarily complex models. At this time, the method is limited to models where the left-hand side of the mixed model equations can be inverted, which sets a soft limit of perhaps ~100K genotyped animals.

Conclusions

Genomic selection methodology has been widely embraced by the animal breeding industry as evidenced by the scale of genotyping. The evaluation in most species except dairy cattle is by single-step methods, which consider all sources of information jointly, with methodology refined sufficiently to provide relatively unbiased evaluation for any data size, and easily accommodating causal genes. The dairy industry plans to move to single step are hampered by distributed ownership of phenotypic and genomic data. Most evaluations use <100K SNP chips without SNP selection or weighting, indirectly acknowledging that the prediction acts mostly on chromosome segments and less on markers of QTLs. Whether accurate determination of causative SNPs will lead to substantially increased accuracy of selection also across breeds is unclear. While the validations methods are less than perfect, they illustrate higher accuracy of evaluation with the genomic information. An important concern in long-term genomic selection may be a serious reduction of the additive variance that may limit future gains, especially given that the parameter estimation with the genomic information is difficult.

Acknowledgments

We gratefully acknowledge the very helpful comments by the two anonymous reviewers. This research was primarily supported by grants from American Angus Association, Cobb-Vantress, Genus PIC, Holstein Association USA, Smithfield Premium Genetics, Zoetis, and U.S. Department of Agriculture's National Institute of Food and Agriculture (Agriculture and Food Research Initiative competitive grant number 2015-67015-22936). This paper is presented at the ASAS 2019 Symposium.

Conflict of interest statement

The authors declare no real or perceived conflicts of interest.

Literature Cited

Aguilar, I., A. Legarra, F. Cardoso, Y. Masuda, D. Lourenco, and I. Misztal. 2019. Frequentist *p*-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. *Genet. Sel. Evol.* **51**:28. doi:10.1186/s12711-019-0469-3

Aguilar, I., and I. Misztal. 2008. Technical note: recursive algorithm for inbreeding coefficients assuming nonzero inbreeding of unknown parents. *J. Dairy Sci.* **91**:1669–1672. doi:10.3168/jds.2007-0575

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 score. *J. Dairy Sci.* **93**:743–752. doi:10.3168/jds.2009–2730

Aguilar, I., I. Misztal, A. Legarra, and S. Tsuruta. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *J. Anim. Breed. Genet.* **128**:422–428. doi:10.1111/j.1439-0388.2010.00912.x

Bermann, M., A. Legarra, M. K. Hollifield, Y. Masuda, D. Lourenco, and I. Misztal. 2020. Validation of genomic and pedigree predictions from threshold models using the linear regression (LR) method: an application in chicken mortality. *Genet. Sel. Evol.* (under review)

Bernal Rubio, Y. L., J. L. Gualdrón Duarte, R. O. Bates, C. W. Ernst, D. Nonneman, G. A. Rohrer, A. King, S. D. Shackelford, T. L. Wheeler, R. J. Cantet, *et al.* 2016. Meta-analysis of

genome-wide association from genomic prediction models. *Anim. Genet.* **47**:36–48. doi:10.1111/age.12378

Bijma, P. 2012. Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. *J. Anim. Breed. Genet.* **129**:345–358. doi:10.1111/j.1439-0388.2012.00991.x

Boichard, D., C. Grohs, F. Bourgeois, F. Cerqueira, R. Faugeras, A. Neau, R. Rupp, Y. Amigues, M. Y. Boscher, and H. Levéziel. 2003. Detection of genes influencing economic traits in three French dairy cattle breeds. *Genet. Sel. Evol.* **35**:77–101. doi:10.1186/1297-9686-35-1-77

Bradford, H. L., Y. Masuda, J. B. Cole, I. Misztal, and P. M. VanRaden. 2019a. Modeling pedigree accuracy and uncertain parentage in single-step genomic evaluations of simulated and US Holstein datasets. *J. Dairy Sci.* **102**:2308–2318. doi:10.3168/jds.2018-15419

Bradford, H. L., Y. Masuda, J. B. Cole, I. Misztal, and P. M. VanRaden. 2019b. Modeling pedigree accuracy and uncertain parentage in single-step genomic evaluations of simulated and US Holstein datasets. *J. Dairy Sci.* **102**:2308–2318. doi:10.3168/jds.2018-15419

Bradford, H. L., I. Pocrnić, B. O. Fragomeni, D. A. L. Lourenco, and I. Misztal. 2017. Selection of core animals in the algorithm for proven and young using a simulation model. *J. Anim. Breed. Genet.* **134**:545–552. doi:10.1111/jbg.12276

Brøndum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *J. Dairy Sci.* **98**:4107–4116. doi:10.3168/jds.2014-9005

Carillier-Jacquin, C., H. Larroque, and C. Robert-Granié. 2016. Including α s1 casein gene information in genomic evaluations of French dairy goats. *Genet. Sel. Evol.* **48**:54. doi:10.1186/s12711-016-0233-x

Cesarani, A., I. Pocrnic, N. P. P. Macciotta, B. O. Fragomeni, I. Misztal, and D. A. L. Lourenco. 2019. Bias in heritability estimates from genomic restricted maximum likelihood methods under different genotyping strategies. *J. Anim. Breed. Genet.* **136**:40–50. doi:10.1111/jbg.12367

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. doi:10.2527/jas.2010-3555

Christensen, O. F. 2012. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. *Genet. Sel. Evol.* **44**:37. doi:10.1186/1297-9686-44-37

Christensen, O. F., A. Legarra, M. S. Lund, and G. Su. 2015. Genetic evaluation for three-way crossbreeding. *Genet. Sel. Evol.* **47**:98. doi:10.1186/s12711-015-0177-6

Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.* **42**:2. doi:10.1186/1297-9686-42-2

Christensen, O. F., P. Madsen, B. Nielsen, and G. Su. 2014. Genomic evaluation of both purebred and crossbred performances. *Genet. Sel. Evol.* **46**:23. doi:10.1186/1297-9686-46-23

Cuyabano, B. C., G. Su, and M. S. Lund. 2015. Selection of haplotype variables from a high-density marker map for genomic prediction. *Genet. Sel. Evol.* **47**:61. doi:10.1186/s12711-015-0143-3

Daetwyler, H. D., M. P. Calus, R. Pong-Wong, G. de Los Campos, and J. M. Hickey. 2013. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* **193**:347–365. doi:10.1534/genetics.112.147983

Derks, M. F. L., M. S. Lopes, M. Bosse, O. Madsen, B. Dibbits, B. Harlizius, M. A. M. Groenen, and H. J. Megens. 2018. Balancing selection on a recessive lethal deletion with pleiotropic effects on two neighboring genes in the porcine genome. *PLoS Genet.* **14**:e1007661. doi:10.1371/journal.pgen.1007661

Duenk, P. P. Bijma, M. P. L. Calus, Y. C. J. Wientjes, and J. H. J. van der Werf. 2020. The impact of non-additive effects on the genetic correlation between populations. *G3 (Bethesda)* **10**:783–795. doi:10.1534/g3.119.400663

Edel, C., E. C. G. Pimentel, M. Erbe, R. Emmerling, and K. U. Götz. 2019. Short communication: calculating analytical reliabilities for single-step predictions. *J. Dairy Sci.* **102**:3259–3265. doi:10.3168/jds.2018-15707

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* **95**:4114–4129. doi:10.3168/jds.2011-5019

Fernando, R. L., H. Cheng, and D. J. Garrick. 2016. An efficient exact method to obtain GBLUP and single-step GBLUP when the genomic relationship matrix is singular. *Genet. Sel. Evol.* **48**:80. doi:10.1186/s12711-016-0260-7

Fernando, R. L., J. C. Dekkers, and D. J. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genet. Sel. Evol.* **46**:50. doi:10.1186/1297-9686-46-50

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. doi:10.1186/1297-9686-43-1

Fragomeni, B. O., D. A. L. Lourenco, A. Legarra, P. M. VanRaden, and I. Misztal. 2019. Alternative SNP weighting for single-step genomic best linear unbiased predictor evaluation of stature in US Holsteins in the presence of selected sequence variants. *J. Dairy Sci.* **102**:10012–10019. doi:10.3168/jds.2019-16262

Fragomeni, B. O., D. A. L. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2017. Incorporation of causative quantitative trait nucleotides in single-step GBLUP. *Genet. Sel. Evol.* **49**:59. doi:10.1186/s12711-017-0335-0

Fragomeni, B. O., D. A. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T. J. Lawlor, and I. Misztal. 2015. Hot topic: use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. *J. Dairy Sci.* **98**:4090–4094. doi:10.3168/jds.2014-9125

Fragomeni, Bde. O., I. Misztal, D. L. Lourenco, I. Aguilar, R. Okimoto, and W. M. Muir. 2014. Changes in variance explained by top SNP windows over generations for three traits in broiler chicken. *Front. Genet.* **5**:332. doi:10.3389/fgene.2014.00332

Gao, H., P. Madsen, G. P. Aamand, J. R. Thomasen, A. C. Sørensen, and J. Jensen. 2019. Bias in estimates of variance components in populations undergoing genomic selection: a simulation study. *BMC Genomics* **20**:956. doi:10.1186/s12864-019-6323-8

Garcia, A. L. S., Y. Masuda, S. Tsuruta, S. Miller, I. Misztal, and D. Lourenco. 2020. Indirect predictions with a large number of genotyped animals using the algorithm for proven and young. *J. Anim. Sci.* (in press)

Garcia-Baccino, C. A., A. Legarra, O. F. Christensen, I. Misztal, I. Pocrnic, Z. G. Vitezica, and R. J. Cantet. 2017. Metafounders are related to F st fixation indices and reduce bias in single-step genomic evaluations. *Genet. Sel. Evol.* **49**:34. doi:10.1186/s12711-017-0309-2

García-Ruiz, A., J. B. Cole, P. M. VanRaden, G. R. Wiggans, F. J. Ruiz-López, and C. P. Van Tassell. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proc. Natl. Acad. Sci. U. S. A.* **113**:E3995–E4004. doi:10.1073/pnas.1519061113

Garrick, D. J., D. P. Garrick, and B. L. Golden. 2018. *An introduction to BOLT software for genetic and genomic evaluations*. In: Proceedings of the 11th World Congress on Genetics Applied to Livestock Production; February 11 to 16, 2018; Auckland (New Zealand); p 973.

Garrick, D. J., J. F. Taylor, and R. L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evol.* **41**:55. doi:10.1186/1297-9686-41-55

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. doi:10.3168/jds.2007-0231

Georges, M., C. Charlier, and B. Hayes. 2019. Harnessing genomic information for livestock improvement. *Nat. Rev. Genet.* **20**:135–156. doi:10.1038/s41576-018-0082-2

Goddard, M. E. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**:245–257. doi:10.1007/s10709-008-9308-0

Goddard, M. E. 2017. Can we make genomic selection 100% accurate? *J. Anim. Breed. Genet.* **134**:287–288. doi:10.1111/jbg.12281

Golden, B. L., M. L. Spangler, W. M. Snelling, and D. J. Garrick. 2018. *Current single-step national beef cattle evaluation models used by the American Hereford Association and International Genetic Solutions, computational aspects, and implications of marker selection.* Proceedings of the Beef Improvement Federation 11th Genetic Prediction Workshop Refining Genomic Evaluation and Selection Indices; December 12 to 13, 2018; Kansas City (MO); p. 14–22.

Gualdrón Duarte, J. L., R. J. Cantet, R. O. Bates, C. W. Ernst, N. E. Raney, and J. P. Steibel. 2014. Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. *BMC Bioinformatics* **15**:246. doi:10.1186/1471-2105-15-246

Hayes, B. J., and H. D. Daetwyler. 2019. 1000 Bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes. *Annu. Rev. Anim. Biosci.* **7**:89–102. doi:10.1146/annurev-animal-020518-115024

Henderson, C. R. 1976. A simple method for computing the inverse of a relationship matrix used in prediction of breeding values. *Biometrics* **32**:69.

Henderson, C. R. 1984. *Applications of linear models in animal breeding*. Guelph (ON), Canada: University of Guelph.

Hidalgo, J., S. Tsuruta, D. Lourenco, Y. Masuda, Y. Huang, K. A. Gray, and I. Misztal. 2020. Changes in genetic parameters for fitness and growth traits in pigs under genomic selection. *J. Anim. Sci.* (under review)

Howard, J. T., T. A. Rathje, C. E. Bruns, D. F. Wilson-Wells, S. D. Kachman, and M. L. Spangler. 2018. The impact of truncating data on the predictive ability for single-step genomic best linear unbiased prediction. *J. Anim. Breed. Genet.* **135**:251–262. doi:10.1111/jbg.12334

Hsu, W. L., D. J. Garrick, and R. L. Fernando. 2017. The accuracy and bias of single-step genomic prediction for populations under selection. *G3 (Bethesda)*. **7**:2685–2694. doi:10.1534/g3.117.043596

Jónás, D., V. Ducrocq, M. N. Fouilloux, and P. Croiseau. 2016. Alternative haplotype construction methods for genomic evaluation. *J. Dairy Sci.* **99**:4537–4546. doi:10.3168/jds.2015–10433

Kachman, S. 2008. *Incorporation of marker scores into national cattle evaluations*. Proceedings of the 9th Genetic Prediction Workshop; December 10 to 11, 2008; Kansas City (MO): Beef Improvement Federation; p. 92–98.

Kachman, S. D., M. L. Spangler, G. L. Bennett, K. J. Hanford, L. A. Kuehn, W. M. Snelling, R. M. Thallman, M. Saatchi, D. J. Garrick, R. D. Schnabel, *et al.* 2013. Comparison of molecular breeding values based on within- and across-breed training in beef cattle. *Genet. Sel. Evol.* **45**:30. doi:10.1186/1297-9686-45-30

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**:348–354. doi:10.1038/ng.548

Karaman, E., H. Cheng, M. Z. Firat, D. J. Garrick, and R. L. Fernando. 2016. An upper bound for accuracy of prediction using GBLUP. *PLoS One*. **11**:e0161054. doi:10.1371/journal.pone.0161054

Kennedy, B. W., M. Quinton, and J. A. van Arendonk. 1992. Estimation of effects of single genes on quantitative traits. *J. Anim. Sci.* **70**:2000–2012. doi:10.2527/1992.7072000x

Legarra, A. 2016. Comparing estimates of genetic variance across different relationship models. *Theor. Popul. Biol.* **107**:26–30. doi:10.1016/j.tpb.2015.08.005

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* **92**:4656–4663. doi:10.3168/jds.2009-2061

Legarra, A., O. F. Chistensen, I. Aguilar, and I. Misztal. 2014. Single step, a general approach for genomic selection. *Livest. Prod. Sci.* **166**:54–65. doi:10.1534/genetics.115.177014

Legarra, A., O. F. Christensen, Z. G. Vitezica, I. Aguilar, and I. Misztal. 2015. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. *Genetics* **200**:455–468. doi:10.1016/j.livsci.2014.04.029

Legarra, A., and V. Ducrocq. 2012. Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. *J. Dairy Sci.* **95**:4629–4645. doi:10.3168/jds.2011-4982

Legarra, A., and A. Reverter. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genet. Sel. Evol.* **50**:53. doi:10.1186/s12711-018-0426-6

Legarra, A., C. Robert-Granié, E. Manfredi, and J. M. Elsen. 2008. Performance of genomic selection in mice. *Genetics* **180**:611–618. doi:10.1534/genetics.108.088575

Legarra, A., and Z. G. Vitezica. 2015. Genetic evaluation with major genes and polygenic inheritance when some animals are not genotyped using gene content multiple-trait BLUP. *Genet. Sel. Evol.* **47**:89. doi:10.1186/s12711-015-0165-x

Liu, Z., M. E. Goddard, F. Reinhardt, and R. Reents. 2014. A single-step genomic model with direct estimation of marker effects. *J. Dairy Sci.* **97**:5833–5850. doi:10.3168/jds.2014-7924

Liu, Z., P. M. VanRaden, M. H. Lidauer, M. P. Calus, H. Benhajali, H. Jorjani, and V. Ducrocq. 2017b. Approximating genomic reliabilities for national genomic evaluation. *Interbull Bull.* **51**:75–85.

Liu, A., Y. Wang, G. Sahana, Q. Zhang, L. Liu, M. S. Lund, and G. Su. 2017a. Genome-wide association studies for female fertility traits in Chinese and Nordic Holsteins. *Sci. Rep.* **7**:8487. doi:10.1038/s41598-017-09170-9

Lourenco, D. A. L., B. O. Fragomeni, H. L. Bradford, I. R. Menezes, J. B. S. Ferraz, I. Aguilar, S. Tsuruta, and I. Misztal. 2017. Implications of SNP weighting on single-step genomic predictions for different reference population sizes. *J. Anim. Breed. Genet.* **134**:463–471. doi:10.1111/jbg.12288

Lourenco, D. A., B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach, R. J. Hawken, A. Legarra, and I. Misztal. 2015a. Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. *Genet. Sel. Evol.* **47**:56. doi:10.1186/s12711-015-0137-1

Lourenco, D. A., I. Misztal, S. Tsuruta, I. Aguilar, T. J. Lawlor, S. Forni, and J. I. Weller. 2014. Are evaluations on young genotyped animals benefiting from the past generations? *J. Dairy Sci.* **97**:3930–3942. doi:10.3168/jds.2013-7769

Lourenco, D. A., S. Tsuruta, B. O. Fragomeni, C. Y. Chen, W. O. Herring, and I. Misztal. 2016. Crossbreed evaluations in single-step genomic best linear unbiased predictor using adjusted realized relationship matrices. *J. Anim. Sci.* **94**:909–919. doi:10.2527/jas.2015-9748

Lourenco, D. A., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. K. Bertrand, T. S. Amen, L. Wang, D. W. Moser, *et al.* 2015b. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. *J. Anim. Sci.* **93**:2653–2662. doi:10.2527/jas.2014-8836

Lourenco, D. A. L., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, S. Miller, D. Moser, and I. Misztal. 2018.

Single-step genomic BLUP for national beef cattle evaluation in US: from initial developments to final implementation. *Proc. World. Cong. Appl. Livest. Prod.* **11**:495.

Lu, Y., M. J. Vandehaar, D. M. Spurlock, K. A. Weigel, L. E. Armentano, E. E. Connor, M. Coffey, R. F. Veerkamp, Y. de Haas, C. R. Staples, *et al.* 2018. Genome-wide association analyses based on a multiple-trait approach for modeling feed efficiency. *J. Dairy Sci.* **101**:3140–3154. doi:10.3168/jds.2017-13364

Lutaaya, E., I. Misztal, J. K. Bertrand, and J. W. Mabry. 1999. Inbreeding in populations with incomplete pedigrees. *J. Anim. Breed. Genet.* **116**:475–480. doi:10.1046/j.1439-0388.1999.00210.x

MacNeil, M. D., J. D. Nkrumah, B. W. Woodward, and S. L. Northcutt. 2010. Genetic evaluation of Angus cattle for carcass marbling using ultrasound and genomic indicators. *J. Anim. Sci.* **88**:517–522. doi:10.2527/jas.2009-2022

Makgahlela, M. L., I. Strandén, U. S. Nielsen, M. J. Sillanpää, and E. A. Mäntysaari. 2014. Using the unified relationship matrix adjusted by breed-wise allele frequencies in genomic evaluation of a multibreed population. *J. Dairy Sci.* **97**:1117–1127. doi:10.3168/jds.2013-7167

Mäntysaari, E. A., R. D. Evans, and I. Strandén. 2017. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. *J. Anim. Sci.* **95**:4728–4737. doi:10.2527/jas2017.1912

Masuda, Y., I. Aguilar, S. Tsuruta, and I. Misztal. 2015. Technical note: acceleration of sparse operations for average-information REML analyses with supernodal methods and sparse-storage refinements. *J. Anim. Sci.* **93**:4670–4674. doi:10.2527/jas.2015-9395

Masuda, Y., T. Baba, and M. Suzuki. 2014. Application of supernodal sparse factorization and inversion to the estimation of (co) variance components by residual maximum likelihood. *J. Anim. Breed. Genet.* **131**:227–236. doi:10.1111/jbg.12058

Masuda, Y., I. Misztal, A. Legarra, S. Tsuruta, D. A. Lourenco, B. O. Fragomeni, and I. Aguilar. 2017. Technical note: avoiding the direct inversion of the numerator relationship matrix for genotyped animals in single-step genomic best linear unbiased prediction solved with the preconditioned conjugate gradient. *J. Anim. Sci.* **95**:49–52. doi:10.2527/jas.2016.0699

Masuda, Y., I. Misztal, P. VanRaden, and T. Lawlor. 2018a. *Pre-selection bias and validation method in single-step GBLUP for production traits in US Holstein*. In: Proceedings of the 11th World Congress on Genetics Applied to Livestock Production; February 11 to 16, 2018; Auckland (New Zealand); p. 540.

Masuda, Y., I. Misztal, P. M. VanRaden, and T. J. Lawlor. 2018b. *Differing genetic trend estimates from traditional and genomic evaluations of genotyped animals as evidence of preselection bias in US Holsteins*. In: ADSA Annual Meeting, Knoxville (TN); p. 5194–5206.

Matilainen, K., M. Koivula, I. Stranden, G. P. Aamand, and E. A. Mantysaari. 2016. Managing genetic groups in single-step genomic evaluations applied on female fertility traits in Nordic Red Dairy Cattle. *Interbull Bull.* **50**:71–75.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. Smith, T. S. Sonstegard, *et al.* 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One.* **4**:e5350. doi:10.1371/journal.pone.0005350

Meuwissen, T. H., 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., J. Odegard, I. Andersen-Ranberg, and E. Grindflek. 2014. On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. *Genet. Sel. Evol.* **46**:49. doi:10.1186/1297-9686-46-49

Meuwissen, T. H., M. Svendsen, T. Solberg, and J. Ødegård. 2015. Genomic predictions based on animal models using genotype imputation on a national scale in Norwegian Red cattle. *Genet. Sel. Evol.* **47**:79. doi:10.1186/s12711-015-0159-8

Meyer, K. 1989. Approximate accuracy of genetic evaluation under an animal model. *Livest. Prod. Sci.* **21**:87–100. doi:10.1016/0301-6226(89)90041–9

Meyer, K., B. Tier, and A. Swan. 2018. Estimates of genetic trend for single-step genomic evaluations. *Genet. Sel. Evol.* **50**:39. doi:10.1186/s12711-018-0410-1

Misztal, I. 2016. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics* **202**:401–409. doi:10.1534/genetics.115.182089

Misztal, I., H. L. Bradford, D. A. L. Lourenco, S. Tsuruta, Y. Masuda, A. Legarra, and T. J. Lawlor. 2017. Studies on inflation of GEBV in single-step GBLUP for type. *Interbull Bull.* **51**:38–42.

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. doi:10.3168/jds.2009-2064

Misztal, I., A. Legarra, and I. Aguilar. 2014a. Using recursion to compute the inverse of the genomic relationship matrix. *J. Dairy Sci.* **97**:3943–3952. doi:10.3168/jds.2013-7752

Misztal, I., S. Tsuruta, I. Aguilar, A. Legarra, P. M. VanRaden, and T. J. Lawlor. 2013a. Methods to approximate reliabilities in single-step genomic evaluation. *J. Dairy Sci.* **96**:647–654. doi:10.3168/jds.2012-5656

Misztal, I., S. Tsuruta, D. A. L. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. Vitezica. 2014b. Manual for BLUPF90 family of programs. Available from http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all7.pdf

Misztal, I., S. Tsuruta, I. Pocrnic, and D. Lourenco. 2019. *Changes in predictions when using different core animals in the APY algorithm*. Proceedings of the 70th Annual Meeting EAAP; August 26 to 30, 2019; Ghent, Belgium; p. 593.

Misztal, I., Z. G. Vitezica, A. Legarra, I. Aguilar, and A. A. Swan. 2013b. Unknown-parent groups in single-step genomic evaluation. *J. Anim. Breed. Genet.* **130**:252–258. doi:10.1111/jbg.12025

Misztal, I., and G. R. Wiggans. 1988. Approximation of prediction error variance in large-scale animal models. *J. Dairy Sci.* **71**:27–32. doi:10.1016/S0022-0302(88)79976-2

Moghaddar, N., M. Khansefid, J. H. J. van der Werf, S. Bolormaa, N. Duijvesteijn, S. A. Clark, A. A. Swan, H. D. Daetwyler, and I. M. MacLeod. 2019. Genomic prediction based on selected variants from imputed whole-genome sequence data in Australian sheep populations. *Genet. Sel. Evol.* **51**:72. doi:10.1186/s12711-019-0514-2

Muir, W. M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *J. Anim. Breed. Genet.* **124**:342–355. doi:10.1111/j.1439-0388.2007.00700.x

Ødegård, J., U. Indahl, I. Strandén, and T. H. E. Meuwissen. 2018. Large-scale genomic prediction using singular value decomposition of the genotype matrix. *Genet. Sel. Evol.* **50**:6. doi:10.1186/s12711-018-0373-2

Oget, C., M. Teissier, J. M. Astruc, G. Tosser-Klopp, and R. Rupp. 2019. Alternative methods improve the accuracy of genomic prediction using information from a causal point mutation in a dairy sheep model. *BMC Genomics* **20**:719. doi:10.1186/s12864-019-6068-4

Patry, C., and V. Ducrocq. 2011a. Accounting for genomic pre-selection in national BLUP evaluations in dairy cattle. *Genet. Sel. Evol.* **43**:30. doi:10.1186/1297-9686-43-30

Patry, C., and V. Ducrocq. 2011b. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. *J. Dairy Sci.* **94**:1011–1020. doi:10.3168/jds.2010-3804

Plieschke, L., C. Edel, E. C. Pimentel, R. Emmerling, J. Bennewitz, and K. U. Götz. 2015. A simple method to separate base population and segregation effects in genomic relationship matrices. *Genet. Sel. Evol.* **47**:53. doi:10.1186/s12711-015-0130-8

Pocrnic, I., D. A. L. Lourenco, C. Y. Chen, W. O. Herring, and I. Misztal. 2019b. Crossbred evaluations using single-step genomic BLUP

and algorithm for proven and young with different sources of data. *J. Anim. Sci.* **97**:1513–1522. doi:10.1093/jas/skz042

Pocrnic, I., D. A. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2016a. The dimensionality of genomic information and its effect on genomic prediction. *Genetics* **203**:573–581. doi:10.1534/genetics.116.187013

Pocrnic, I., D. A. Lourenco, Y. Masuda, and I. Misztal. 2016b. Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species. *Genet. Sel. Evol.* **48**:82. doi:10.1186/s12711-016-0261-6

Pocrnic, I., D. A. L. Lourenco, K. Y. Masuda, and I. Misztal. 2019a. Accuracy of genomic BLUP when considering a genomic relationship matrix based on the number of the largest eigenvalues: a simulation study. *Genet. Sel. Evol.* **51**:75.

Quaas, R. L. 1988. Additive genetic model with groups and relationships. *J. Dairy Sci.* **71**:1338–1345.

Quaas, R. L., and E. J. Pollak. 1981. Modified equations for sire models with groups. *J. Dairy Sci.* **64**:1868–1872.

Ros-Freixedes, R., A. Whalen, C. Y. Chen, G. Gorjanc, W. O. Herring, A. J. Mileham, and J. M. Hickey. 2020. Accuracy of whole-genome sequence imputation using hybrid peeling in large pedigreed livestock populations. *Genet. Sel. Evol.* **52**:17. doi:10.1186/s12711-020-00536-8

Saatchi, M., M. C. McClure, S. D. McKay, M. M. Rolf, J. Kim, J. E. Decker, T. M. Taxis, R. H. Chapple, H. R. Ramey, S. L. Northcutt, *et al.* 2011. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genet. Sel. Evol.* **43**:40. doi:10.1186/1297-9686-43-40

Sargolzaei, M., and F. S. Schenkel. 2009. QMSim: a large-scale genome simulator for livestock. *Bioinformatics* **25**:680–681. doi:10.1093/bioinformatics/btp045

Stam, P. 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genet. Res.* **35**:131–155.

Steyn, Y., D. A. L. Lourenco, and I. Misztal. 2019. Genomic predictions in purebreds with a multibreed genomic relationship matrix1. *J. Anim. Sci.* **97**:4418–4427. doi:10.1093/jas/skz296

Strandén, I., and O. F. Christensen. 2011. Allele coding in genomic evaluation. *Genet. Sel. Evol.* **43**:25. doi:10.1186/1297-9686-43-25

Strandén, I., K. Matilainen, G. P. Aamand, and E. A. Mäntysaari. 2017. Solving efficiently large single-step genomic best linear unbiased prediction models. *J. Anim. Breed. Genet.* **134**:264–274. doi:10.1111/jbg.12257

Taskinen, M., E. A. Mäntysaari, and I. Strandén. 2017. Single-step SNP-BLUP with on-the-fly imputed genotypes and residual polygenic effects. *Genet. Sel. Evol.* **49**:36. doi:10.1186/s12711-017-0310-9

Teissier, M., H. Larroque, and C. Robert-Granié. 2018. Weighted single-step genomic BLUP improves accuracy of genomic breeding values for protein content in French dairy goats: a quantitative trait influenced by a major gene. *Genet. Sel. Evol.* **50**:31. doi:10.1186/s12711-018-0400-3

Tsuruta, S., D. A. L. Lourenco, Y. Masuda, I. Misztal, and T. J. Lawlor. 2019a. Controlling bias in genomic breeding values for young genotyped bulls. *J. Dairy Sci.* **102**:9956–9970. doi:10.3168/jds.2019-16789

Tsuruta, S., D. A. L. Lourenco, Y. Masuda, I. Misztal, and T. J. Lawlor. 2019b. Validation of genomic predictions for linear type traits in US Holsteins using over 2 million genotyped animals. *J. Dairy Sci.* **102** (Suppl. 1):397.

Tsuruta, S., D. A. L. Lourenco, I. Misztal, and T. J. Lawlor. 2017. Genomic analysis of cow mortality and milk production using a threshold-linear model. *J. Dairy Sci.* **100**:7295–7305. doi:10.3168/jds.2017-12665

Tsuruta, S., I. Misztal, I. Aguilar, and T. J. Lawlor. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. *J. Dairy Sci.* **94**:4198–4204. doi:10.3168/jds.2011-4256

Tsuruta, S., I. Misztal, D. A. Lourenco, and T. J. Lawlor. 2014. Assigning unknown parent groups to reduce bias in genomic evaluations of final score in US Holsteins. *J. Dairy Sci.* **97**:5814–5821. doi:10.3168/jds.2013-7821

Vandenplas, J., M. P. L. Calus, H. Eding, and C. Vuik. 2019. A second-level diagonal preconditioner for single-step SNPBLUP. *Genet. Sel. Evol.* **51**:30. doi:10.1186/s12711-019-0472-8

Vandenplas, J., J. J. Windig, and M. P. L. Calus. 2017. Prediction of the reliability of genomic breeding values for crossbred performance. *Genet. Sel. Evol.* **49**:43. doi:10.1186/s12711-017-0318-1

Van Grevenhof, E. M., J. A. Van Arendonk, and P. Bijma. 2012. Response to genomic selection: the Bulmer effect and the potential of genomic selection when the number of phenotypic records is limiting. *Genet. Sel. Evol.* **44**:26. doi:10.1186/1297-9686-44-26

VanRaden, P. M. 1992. Accounting for inbreeding and crossbreeding in genetic evaluation of large populations. *J. Dairy Sci.* **75**:3136–3144.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**:4414–4423. doi: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. doi:10.1186/1297-9686-43-10

VanRaden, P. M., M. E. Tooker, T. C. S. Chud, H. D. Norman, J. H. Megonigal Jr, I. W. Haagen, and G. R. Wiggans. 2020. Genomic predictions for crossbred dairy cattle. *J. Dairy Sci.* **103**:1620–1631. doi:10.3168/jds.2019-16634

VanRaden, P. M., M. E. Tooker, J. R. O'Connell, J. B. Cole, and D. M. Bickhart. 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet. Sel. Evol.* **49**:32. doi:10.1186/s12711-017-0307-4

VanRaden, P. M., M. E. Tooker, J. R. Wright, C. Sun, and J. L. Hutchison. 2014. Comparison of single-trait to multi-trait national evaluations for yield, health, and fertility. *J. Dairy Sci.* **97**:7952–7962. doi:10.3168/jds.2014-8489

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* **92**:16–24. doi:10.3168/jds.2008-1514

VanRaden, P. M., and G. R. Wiggans. 1991. Derivation, calculation, and use of national animal model information. *J. Dairy Sci.* **74**:2737–2746. doi:10.3168/jds.S0022-0302(91)78453-1

VanRaden, P. M., and J. R. Wright. 2013. Measuring genomic pre-selection in theory and in practice. *Interbull Bull.* **47**:147–150.

VanRaden, P. M., J. R. Wright, and T. A. Cooper. 2012. Adjustment of selection index coefficients and polygenic variance to improve regressions and reliability of genomic evaluations. *J. Dairy Sci.* **95**(Suppl. 2):446–447

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. doi:10.1017/S001667231100022X

Westell, R. A., R. L. Quaas, and L. D. V. Vleck. 1988. Genetic groups in an animal model. *J. Dairy Sci.* **71**:1310–1318.

Wiggans, G. R., T. A. Cooper, P. M. Vanraden, and J. B. Cole. 2011. Technical note: adjustment of traditional cow evaluations to improve accuracy of genomic predictions. *J. Dairy Sci.* **94**:6188–6193. doi:10.3168/jds.2011-4481

Wiggans, G. R., P. M. Vanraden, and T. A. Cooper. 2012. Technical note: adjustment of all cow evaluations for yield traits to be comparable with bull evaluations. *J. Dairy Sci.* **95**:3444–3447. doi:10.3168/jds.2011-5000

Xiang, T., O. F. Christensen, and A. Legarra. 2017. Technical note: genomic evaluation for crossbred performance in a single-step approach with metafounders. *J. Anim. Sci.* **95**:1472–1480. doi:10.2527/jas.2016.1155

Xiang, T., B. Nielsen, G. Su, A. Legarra, and O. F. Christensen. 2016. Application of single-step genomic evaluation for crossbred performance in pig. *J. Anim. Sci.* **94**:936–948. doi:10.2527/jas.2015-9930

Xu, S. 2013. Genetic mapping and genomic selection using recombination breakpoint data. *Genetics* **195**:1103–1115. doi:10.1534/genetics.113.155309