SINGLE STEP, A GENERAL APPROACH FOR GENOMIC SELECTION

Andres Legarra[a*], Ole F Christensen[b], Ignacio Aguilar[c] and Ignacy Misztal[d]

[a]INRA, UMR1388 GenPhySE, BP52627, 31326 Castanet Tolosan, France
andres.legarra@toulouse.inra.fr

[b]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Blichers Alle 20, P.O. BOX 50, DK-8830 Tjele, Denmark
OleF.Christensen@agrsci.dk

[c]Instituto Nacional de Investigación Agropecuaria, Las Brujas 90200, Uruguay
iaguilar@inia.org.uy

[d]Department of Animal and Dairy Science, University of Georgia, Athens 30602-2771, USA
ignacy@uga.edu

*Correspondig author: andres.legarra@toulouse.inra.fr Phone:+33561285182 Fax: +33561285353

ABSTRACT

Genomic evaluation methods assume that the reference population is genotyped and phenotyped. This is most often false and the generation of pseudo-phenotypes is uncertain and inaccurate. However, markers obey transmission rules and therefore the covariances of marker genotypes across individuals can be modelled using pedigree relationships. Based on this, an extension of the genomic relationship matrix can be constructed in which genomic relationships are propagated to all individuals, resulting in a combined relationship matrix, which can be used in a BLUP procedure called the Single Step Genomic BLUP. This procedure provides so far the most comprehensive option for genomic evaluation. Several extensions, options and details are described: compatibility of genomic and pedigree relationships, Bayesian regressions, multiple trait models, computational aspects, etc. Many details scattered through a series of papers are put together into this abstract.

KEYWORDS

Genetic evaluation, genomic evaluation, markers, BLUP, relationship

1. INTRODUCTION: BRIEF EXCURSION INTO METHODS FOR GENOMIC EVALUATION

1.1 Marker information

Genetic progress by selection and mating is based on prediction of the ability of the parents to breed the most efficient descendants. This process of prediction is called genetic evaluation or prediction. Genetic evaluation in plants and livestock has, for the last century, been based on the use of phenotypes at the traits of interest, together with pedigree. In most cases, these evaluations ignore the physical base of heredity, i.e., DNA, and use a simplified conceptual representation of the transmission of genetic information from parents to offspring; namely, each parent passes on average half its genetic constitution, associated with an unknown sampling known as Mendelian sampling. Recent technical developments allow stepping further into biology and peeking at the genome in the form of single nucleotide polymorphisms, known as SNP markers. These markers depict, in an incomplete manner, the differences between DNA inherited by two individuals. They can be used in multiple ways; in this section we will present very briefly how they are typically used in genetic evaluation (or prediction or estimation of breeding values: EBV hereinafter) in a parametric framework. Most genomic evaluations follow the principle of estimating the *conditional expectation* of the breeding value in view of all information, which has optimal properties if the assumptions of the model hold (e.g., Fernando and Gianola 1986). This (parametric) paradigm has been extremely fruitful over the last decades, allowing for the development of BLUP, REML, Bayesian estimators and giving a coherent framework to solve many applied problems in animal breeding (e.g., Gianola and Fernando, 1986).

The notion of prediction or estimation of random effects is absent in many statistical textbooks (but check, for instance, Casella and Berger (1990)). However, it has been treated as early as Smith (1936) with key references e.g. in Cochran (1951), Henderson (1973) or Fernando and Gianola (1986). Based on those authors, the "correct" model of prediction consists in writing down the statistical association between phenotypes and breeding values, then derive the EBVs from the conditional distribution of breeding values given the phenotypes.

1.2 Bayesian regression

Typically, in genomic predictions, the phenotypes of a population are considered as a function of the breeding values, and the breeding value of individuals, $\boldsymbol{u}$ (or part of it) is decomposed in a sum of marker effects $\boldsymbol{a}$ (e.g., Meuwissen et al., 2001; VanRaden, 2008). These marker effects are summed according to the genotype of the individual, coded as (0,1,2) for the (AA, Aa, aa) genotypes. In matrix notation $\boldsymbol{u} = \boldsymbol{M}\boldsymbol{a}$. It follows that one way of estimating breeding values is to estimate marker effects and then use $\hat{\boldsymbol{u}} = \boldsymbol{M}\hat{\boldsymbol{a}}$. In order to estimate marker effects, one needs to assume a prior distribution for them. The process of estimation of marker effects using the statistical model for phenotypes $p(\boldsymbol{y}|\boldsymbol{a})$ and the prior for markers $p(\boldsymbol{a})$ is often called *Bayesian regression on markers*. A difficult decision is the choice of the prior for markers. An extensive literature in the subject shows improved value, for some traits and populations, of "heavy-tailed" a priori distributions (e.g., VanRaden et al., 2009).

1.3 RR-BLUP or GBLUP

If multivariate normality is assumed for the effect of markers, interesting things happen in the algebraic developments. The first one is that the estimator becomes what is called RR-BLUP. The second is the existence of closed forms for the estimators of marker effects in the form of Henderson's Mixed Model Equations; these estimators greatly simplify computations and can be easily extended, e.g. for multiple trait situations. The third is the existence of a so-called equivalent model, in which breeding values (and not marker effects) are directly computed by the use of a covariance matrix $\boldsymbol{ZD_aZ'}$ (VanRaden, 2008), where $\boldsymbol{Z} = \boldsymbol{M} - 2\boldsymbol{P}$ and $\boldsymbol{P}$ contains p_i, the allelic frequencies of markers. In the most common case it is assumed that $Var(\boldsymbol{a}) = \boldsymbol{D_a} = \boldsymbol{I}\sigma_u^2/2\Sigma p_i q_i$, where σ_u^2 is the genetic variance, so that $\boldsymbol{G} = \boldsymbol{ZZ'}/2\Sigma p_i q_i$. This is called the *genomic relationship matrix* and will frequently be referred to later. Properties of $\boldsymbol{G}$ for populations in Hardy-Weinberg equilibrium are an average diagonal of 1 and an average off-diagonal of 0. Genomic evaluation using $\boldsymbol{G}$ gives the same estimated breeding values as a marker-based RR-BLUP and has the additional advantage of fitting very well into ancient developments (e.g., for multiple trait) and current software. An interesting feature of the genomic relationship matrix is that it can be seen as an "improved" estimator of relationships based on markers instead of pedigrees (VanRaden, 2008; Hayes, 2009), and is closely related to estimators of relationships based on markers used in conservation genetics (Ritland, 1996; Toro et al., 2011).

2. THE PROBLEM OF MISSING GENOTYPES AND THE USE OF PSEUDO-DATA

Genotyping an individual is an expensive process that also requires the availability of a biological sample. Therefore, in most populations either the most recent or the most representative animals (e.g., sires in dairy cattle) have been genotyped. Some individuals are genotyped with low-density chips that genotype only some markers. From these, genotypes at

all markers can be efficiently imputed (e.g., VanRaden et al., 2012) and we will consider these individuals as genotyped. A non-genotyped individual is one for which *there is no genotype at any loci*. Therefore, the methods for genomic prediction described above cannot be applied directly, as there often is not phenotype for the individual genotyped and viceversa; this is particularly true for sex-limited traits (milk yield, fertility, prolificacy). Therefore, animal breeders have used pseudo-data or *pseudo-phenotypes*. A pseudo-phenotype is a projection of the phenotypes of individuals close to the genotyped one. In dairy cattle and sheep, pseudo-phenotypes typically used are corrected daughter performances (daughter yield deviations, VanRaden and Wiggans, 1991), whereas in other species de-regressed proofs are often used, with a variety of *ad hoc* adjustments (Garrick et al., 2009; Ricard et al., 2013).

This process is therefore clumsy and we call it *multiple step*. A regular evaluation based on pedigree is run first, and its results are used to create pseudo-performances. Then, a genomic evaluation model is used. This results in losses of information, inaccuracies and biases, whose importance depends on the species and data set. There are several possible problems:

1. The information of a close relative is ignored in the genomic prediction, for instance the dam of a bull if this dam has phenotype but not genotype.
2. The information of a close relative is ignored in the creation of pseudo-phenotypes, for instance a non-genotyped parent. This is serious if the progeny of the genotyped individual is scarce and therefore parental phenotypes are informative (see Ricard et al. (2013) for a discussion in a horse application).
3. Covariances among pseudo-phenoypes are not correctly modelled. For instance, the yield deviations of two unrelated cows in the same herd will be correlated (e.g., if the herd effect is underestimated they will be biased upwards). This is ignored in the genomic model, which acts as if pseudo-phenotypes were perfectly clean of environmental errors.

4. Many key parameters are difficult to obtain. One of them is precisions of pseudo-phenotypes, which are in most cases rough approximations.
5. There is no feedback. An improved estimation of the breeding value of the genotyped animal should go into the regular pedigree-based genetic evaluation and improve its global accuracy.
6. When genomic selection is applied, animals are selected as parents based on their DNA. The implication is that when phenotypes are obtained from a scheme that has used genomic selection, evaluation based on pedigree becomes biased and is no longer appropriate (Patry and Ducrocq, 2011). Hence, current approaches for constructing pseudo-phenotypes will also become inappropriate due to problems of bias.
7. The process is extremely difficult to generalize. For instance, the multiple-trait generalization of pseudo-phenotypes is basically non-existent, and the pseudo-phenotypes for maternal traits result in much less accurate multiple step predictions (Lourenco et al., 2013).

Some of these defaults can be palliated. VanRaden et al. (2009) used a selection index to *a posteriori* add information from non-genotyped dams to bull genomic evaluations. The procedures of creation of pseudo-phenotypes can be refined over and over. However, the existence of these problems calls for a unified procedure for prediction of genetic value. This paper will describe such a procedure: the *Single Step*.

3. DEVELOPMENT OF THE SINGLE STEP METHOD FOR GENOMIC EVALUATION

Legarra et al. (2009) and Christensen and Lund (2010) developed in parallel the basic theory for the Single Step. They started from two somehow different points of view that turned out to

result in the same formulation, and we will present both developments, starting with the latter one.

3.1 The Single Step as "imputing" missing genotypes

To some extent, missing genotypes can be deduced from existing genotypes, for instance a dam mated to a sire AA producing an offspring Aa is necessarily carrier of one allele a. In statistical theory, a way to deal with missing information is to augment the model with this missing information (*e.g.*, Tanner and Wong, 1987). This missing information needs to be inferred from the other data, and its joint distribution needs to be considered. This means that a "best guess" of missing information in view of observed data, as suggested by Hickey et al. (2012), who imputed genotypes for the complete ungenotyped population, is not correct enough. Even if one considers the uncertainty of individual "guesses" the across-individual uncertainty is extremely difficult to ascertain or deal with.

An example may clarify this point.. Assume a very long complex pedigree and the final generation genotyped for one locus, with allelic frequency $p = frequency(a)$. Due to only having one generation with genotypes and to the long and complex pedigree best guesses of genotypes in the base animals will be nearly identical and equal to $2p$, for all individuals. Therefore, using "best guess" of genotype without taking uncertainty into account, all base population individuals will be treated by the genomic evaluation as identical, which will force them to have the same estimated breeding value, which is paradoxical. For each individual the uncertainty can be assessed by noting that the distribution of genotypes in this case is approximately AA (with probability q^2), Aa (with probability $2pq$) and aa (with probability p^2), but the joint distribution of genotypes for individuals in the base population is much more difficult to characterize. In principle, incorporation of uncertainty can be done by

sampling all possible genotypic configurations of all individuals, e.g. by a Gibbs sampling procedure (e.g. Abraham et al., 2007) but this is computationally infeasible for data of the size used in practical genetic evaluations.

Christensen and Lund (2010), considered the problem as follows. Their objective was to create an extension of the genomic relationship matrix to ungenotyped animals. Following an idea of Gengler et al. (2007), they treated the genotypes as quantitative traits. This makes sense because genotypes are quantitative (0/1/2) and follows Mendelian transmissions. Therefore the covariance of the genotypes z of two individuals i and j is described by their relationship, i.e. $Cov(z_i, z_j) = A_{ij}2pq$ (e.g., Cockerham, 1969). This is less informative than considering the genotype as a union of two discrete entities following Mendelian rules (e.g., sometimes we can exactly deduce a genotype from close relatives) but makes the problem analytically tractable for all cases.

Christensen and Lund (2010) started by inferring the genomic relationship matrix for all animals using inferred (imputed) genotypes for ungenotyped animals; these can simply be obtained as $\hat{\mathrm{Z}}_1 = \mathrm{A}_{12}\mathrm{A}_{22}^{-1}\mathrm{Z}_2$, where 1 and 2 stand for nongenotyped and genotyped animals, respectively. This provides the "best guess" of genotypes. However, the missing data theory requires the joint distribution of these "guessed" genotypes. Assuming that multivariate normality holds for genotypes (this is an approximation, but very good when many genotypes are considered), the "best guess" $\mathrm{E}(\mathrm{Z}_1|\mathrm{Z}_2) = \hat{\mathrm{Z}}_1$, and the conditional variance expresses the uncertainty about the "guess"

$Var(\widehat{\boldsymbol{Z}}_1|\boldsymbol{Z}_2) = (\boldsymbol{A}_{11} + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}) \sum_k 2p_k q_k$,. These two results can be combined to obtain the desired augmented genomic relationships. For instance, for the nongenotyped animals,

$$Var(\boldsymbol{u}_1) = \frac{\widehat{\boldsymbol{Z}}_1\widehat{\boldsymbol{Z}}'_1}{2\Sigma p_k q_k} + (\boldsymbol{A}_{11} + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}),$$

which equals

$$Var(\boldsymbol{u}_1) = \boldsymbol{A}_{11} + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}$$

Finally, the augmented genomic relationship matrix is

$$Var\begin{pmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\end{pmatrix} = \boldsymbol{H} = \begin{pmatrix}\boldsymbol{A}_{11} + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G}\\ \boldsymbol{G}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{G}\end{pmatrix},$$

and with inverse

$$\boldsymbol{H}^{-1} = \boldsymbol{A}^{-1} + \begin{pmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{G}^{-1} - \boldsymbol{A}_{22}^{-1}\end{pmatrix}$$

assuming that $\boldsymbol{G}$ is invertible (this will be dealt with later). Therefore, by using an algebraic data augmentation of missing genotypes Christensen and Lund (2010) derived a simple expression for an augmented genomic relationship matrix and its inverse, without the need to explicitly augment, or "guess", all genotypes for all non-genotyped animals.

3.2 The Single Step as Bayesian updating of the relationship matrix

Legarra et al. (2009) arrived to the same expressions that of Christensen and Lund (2010) in a different manner. They also considered how to construct an extended relationship matrix. However, instead of dealing with individual markers, they dealt with overall breeding values

that can be written as $\boldsymbol{u_2 = Z_2 a}$. They reasoned as follows. Prior to observation of markers, the joint distribution of breeding values (assuming a genetic variance of 1 to simplify notation) is multivariate normal

$$p\begin{pmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\end{pmatrix} = N(\mathbf{0}, \boldsymbol{A})$$

with covariance matrix

$$Var\begin{pmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\end{pmatrix} = \boldsymbol{A} = \begin{pmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{pmatrix}$$

After observing the markers, this covariance matrix will change. The joint distribution above can be split into the product of a marginal and a conditional density; i.e. $p(\boldsymbol{u}_1, \boldsymbol{u}_2) = p(\boldsymbol{u}_1|\boldsymbol{u}_2)p(\boldsymbol{u}_2)$, where

$$p(\boldsymbol{u}_1|\boldsymbol{u}_2) = N(\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{u}_2, \boldsymbol{A}_{11} + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}).$$

Or, in other terms, $\boldsymbol{u}_1 = \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{u}_2 + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ and $\boldsymbol{u}_2$ are independent, and $Var(\boldsymbol{\epsilon}) = \boldsymbol{A}_{11} + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}$.

As discussed before, in presence of marker genotypes the genomic relationship matrix can be considered as fully informative about relationships of individuals, without the need to resort to pedigree or knowledge of previous, or future, nongenotyped individuals. Therefore, *after* observing the marker genotypes

$$p(\boldsymbol{u}_2|markers) = N(\mathbf{0}, \boldsymbol{G}).$$

Marker genotypes influence the relationships among nongenotyped individuals and relationships between nongenotyped and genotyped individuals indirectly. Assuming that these relationships are only influenced by marker genotypes through the genomic

relationships among genotyped individuals, and assuming that the statistical distribution is determined by these relationships, one can write that

$$p(\boldsymbol{u}_1|\boldsymbol{u}_2, markers) = p(\boldsymbol{u}_1|\boldsymbol{u}_2)$$

Therefore, the joint distribution of breeding values *after* observing the markers is:

$$p(\boldsymbol{u}_1, \boldsymbol{u}_2|markers) = p(\boldsymbol{u}_1|\boldsymbol{u}_2)p(\boldsymbol{u}_2|markers)$$

From these results, expressions for the covariance of breeding values are immediate. For instance, $Var(\boldsymbol{u}_1) = \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} + \boldsymbol{A}_{11} + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}$, where the part involving $\boldsymbol{G}$ is the variability associated to the conditional mean of breeding values of nongenotyped individuals given the genotyped ones; and the second part is the variability beyond this conditional mean. Finally, the result

$$Var\begin{pmatrix}\boldsymbol{u}_1\\ \boldsymbol{u}_2\end{pmatrix} = \boldsymbol{H} = \begin{pmatrix}\boldsymbol{A}_{11} + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G}\\ \boldsymbol{G}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{G}\end{pmatrix}$$

is obtained, in full agreement with Christensen and Lund (2010). The reason for this agreement is that in both cases a central assumption is that the influence of marker genotypes on nongenotyped individuals is via relationships determined by the numerator relationship matrix $\boldsymbol{A}$.

3.3 Genetic properties of the extended relationship matrix

Matrix $\boldsymbol{H}$ above can be seen as a modification of regular pedigree relationships to accommodate genomic relationships. For instance, too seemingly unrelated individuals will appear as related in **H** if their descendants are related in **G**. Accordingly, two descendants of

individuals that are related in **G** will be related in **H**, even if the pedigree disagrees. Indeed, it has been suggested (Sun et al., 2013) to use **H** in mating programs to avoid inbreeding.

Contrary to common intuition from BLUP or GBLUP, genotyped animals without phenotype or descendants *cannot* be eliminated from matrix **H**. The reason is that these animals potentially modify pedigree relationship across other animals, possibly notably their parents. For instance imagine two half-sibs, offspring of one sire mated to two nongenotyped, unrelated cows. If these two half sibs are virtually identical, **H** will include this information and the cows will be made related (even identical) in **H**.

3.4 Single Step Genomic BLUP

Because the Single Step relationship matrix provides an explicit and rather sparse (inverse of) the extended relationship matrix $\boldsymbol{H}$, its application to genomic evaluation is immediate. A full specification of the Single Step Genomic BLUP assumes the following model:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Wu} + \boldsymbol{e}$$

$$Var(\boldsymbol{u}) = \boldsymbol{H}\sigma_u^2; Var(\boldsymbol{e}) = \boldsymbol{I}\sigma_e^2$$

with $\boldsymbol{H}$ and its inverse as shown above. The logic of BLUP (Henderson, 1973 and many other publications) holds and the only change is to use $\boldsymbol{H}$ instead of the numerator relationship matrix. Genomic predictions estimating simultaneously all breeding values and using all available information are, for the single trait case, the solutions to the mixed model equations (e.g., Aguilar et al., 2010; Christensen and Lund, 2010):

$$\begin{pmatrix} \boldsymbol{X'X} & \boldsymbol{X'W} \\ \boldsymbol{W'X} & \boldsymbol{W'W} + \boldsymbol{H^{-1}}\lambda \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X'y} \\ \boldsymbol{W'y} \end{pmatrix}$$

where $\lambda = \sigma_e^2/\sigma_u^2$.

Note that any formulation using relationship matrix $\boldsymbol{A}$ can use $\boldsymbol{H}$ instead, and therefore there is also Single Step REML, for instance in Legarra et al. (2009) and Forni et al. (2011).

4. EXTENSIONS AND REFINEMENTS OF THE SINGLE STEP

As said above, any model that has been fit as BLUP can be fit as Single Step. We will describe a few of these extensions that are of interest.

4.1 Pseudo-Single Step.

Also called "blending" (e.g. Su et al., 2012) this has been used to include all males of a population with pseudo-phenotypes, where some are genotyped and some are not. This is a compromise between using all information (which might be complex) and ignoring pseudo-phenotypes of non-genotyped males, for instance sires of genotyped males. Accuracy increases, but less than with true Single Step (Baloche et al., 2014).

Multiple trait Extension to deal with multiple traits is immediate. The mixed model equations are, in the usual notation:

$$\begin{pmatrix} \boldsymbol{X'R^{-1}X} & \boldsymbol{X'R^{-1}W} \\ \boldsymbol{W'R^{-1}X} & \boldsymbol{W'R^{-1}W + H^{-1} \otimes G_0} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X'R^{-1}y} \\ \boldsymbol{W'R^{-1}y} \end{pmatrix}$$

Where $\boldsymbol{R} = \boldsymbol{I} \otimes \boldsymbol{R}_0$, $\boldsymbol{R}_0$ is the matrix of residual covariances across traits and $\boldsymbol{G_0}$ is the matrix of genetic covariances across traits. Extension to random regressions or maternal effect models is very similar.

287

288 *4.2 Marker effect estimates*

289 The GBLUP and other models based on genomic relationship matrices such as the Single Step
290 do not directly provide estimates of marker effects. These are of interest in order to spot major
291 gene (or QTL) localizations and also in order to provide a lees computationally demanding
292 evaluation of new born animals that are genotyped but without phenotypes. The marker
293 effects can be deduced from estimated breeding values of the genotyped individuals by using
294 their joint distribution (Henderson, 1973; Strandén and Garrick, 2009). Consider

$$Var\begin{pmatrix}\boldsymbol{u_2}\\ \boldsymbol{a}\end{pmatrix}=\begin{pmatrix}\boldsymbol{Z_2 D_a Z_2'} & \boldsymbol{Z_2 D_a}\\ \boldsymbol{D_a Z_2'} & \boldsymbol{D_a}\end{pmatrix}$$

295 where, usually, $\boldsymbol{D_a}=\boldsymbol{I}\sigma_u^2/2\Sigma p_i q_i$(this assumption will be relaxed later). Applying the
296 multivariate normality, $\widehat{\boldsymbol{u}}_{\boldsymbol{2}}|\widehat{\boldsymbol{a}}=\boldsymbol{Z_2}\widehat{\boldsymbol{a}}$ (the breeding value is the sum of marker effects) and

297 $\widehat{\boldsymbol{a}}|\widehat{\boldsymbol{u_2}}=\boldsymbol{D_a Z_2'}(\boldsymbol{Z_2 D_a Z_2'})^{-1}\widehat{\boldsymbol{u}}_2$ where $\boldsymbol{Z_2 D_a Z_2'}=\sigma_u^2\boldsymbol{G}$. This result has been used, e.g., by
298 Wang et al. (2012), and it will appear later in this paper.

299

300 *4.3 Extra polygenic effect*

301 It has been often argued that markers do not capture all genetic variation. This can be shown
302 by estimating variance assigned to markers and pedigree (e.g. Legarra et al., 2008) or because
303 some genomic evaluation procedures give better cross-validation results when an extra
304 polygenic term based exclusively on pedigree relationships is added. The GBLUP
305 (VanRaden, 2008) and the derivations in the Single Step can accommodate this very easily
306 (Aguilar et al., 2010; Christensen and Lund, 2010). Let us decompose the breeding values of
307 genotyped individuals in a part due to markers and a residual part due to pedigree, $\boldsymbol{u}_2=$

$\boldsymbol{u}_{m,2} + \boldsymbol{u}_{p,2}$ with respective variances $\sigma_u^2 = \sigma_{u,m}^2 + \sigma_{u,p}^2$. It follows that $Var(\boldsymbol{u}_2) = (\alpha\boldsymbol{G} + (1-\alpha)\boldsymbol{A}_{22})\sigma_u^2$ where $\alpha = \sigma_{u,m}^2/\sigma_u^2$. Therefore, the simplest way is to create a modified genomic relationship matrix $\boldsymbol{G}_w$ ($\boldsymbol{G}$ in Aguilar et al., 2010; $\boldsymbol{G}_w$in VanRaden, 2008 and Christensen and Lund, 2010) as $\boldsymbol{G}_w = \alpha\boldsymbol{G} + (1-\alpha)\boldsymbol{A}_{22}$ and to plug this relationship matrix in all the expressions before. This has the additional advantage of making $\boldsymbol{G}_w$ invertible, which is not guaranteed for $\boldsymbol{G}$. Equivalently, one can fit *two* random effects, one $\boldsymbol{u}_m$with covariance matrix $\boldsymbol{H}\sigma_{u,m}^2$ and another $\boldsymbol{u}_p$with covariance matrix $\sigma_{u,p}^2$.

4.3 Compatibility of genomic and pedigree relationships

This is a key issue in genomic evaluation that has received small attention beyond Single Step developers even though, as shown by Vitezica et al. (2011), it also affects multiple step methods. The derivations above of Single Step mixed model equations include terms such as $\boldsymbol{G} - \boldsymbol{A}_{22}$ and $\boldsymbol{G^{-1}} - \boldsymbol{A}_{22}^{-1}$. This suggests that $\boldsymbol{G}$ and $\boldsymbol{A}_{22}$, the genomic and relationship matrices, need to be compatible. It has been long known (e.g., Ritland 1996) that relationships estimated from markers need to use allelic frequencies at the base populations; otherwise a severe bias in the estimated relationships is observed (VanRaden 2008; Toro et al., 2011). However, typically base population frequencies are unknown because pedigree recording started before biological sampling of individuals. The two derivations of the Single Step assume, either implicitely or explicitely, that the base frequencies are known. In the derivation of Christensen and Lund (2010) the allele frequencies enter explicitely. In the derivation of Legarra et al. (2009) the hypothesis is that the breeding value of the genotyped population is 0. This hypothesis will be wrong if either there has been selection or drift, which is commonly the case; the average breeding value will change, and the genetic variance will reduce. These

problems were soon observed by analysis of real life data sets (Chen et al., 2011b; Forni, 2011; Christensen et al., 2012) and verified by simulation (Vitezica et al., 2011).

Several proposals exist so far to make pedigree and genomic relationships compatible. The three first proposals "tune" matrix $\boldsymbol{G}$ to make it compatible with $\boldsymbol{A}_{22}$, in the form $\boldsymbol{G}^* = a + b\boldsymbol{G}$, where a can be understood as an "overall" relationship and b as a change in scale (or genetic variance). VanRaden (2008) suggested a regression of observed on expected relationships, minimizing the residuals of $a + b\boldsymbol{G} = \boldsymbol{A}_{22} + \boldsymbol{E}$. This reflects the fact that over conceptual repetitions of our population (same pedigree but different meiosis and genotypes) $E(\boldsymbol{G}) = \boldsymbol{A}_{22}$ if $\boldsymbol{G}$ is the realized relationship and $\boldsymbol{A}_{22}$ is the expected relationship (VanRaden, 2008; Hayes et al., 2009). The distribution of $\boldsymbol{E}$ is not homoscedastic (Hill and Weir, 2011; Garcia-Cortes et al., 2013) and this precluded scholars from trying this approach because it would be sensible to extreme values (Christensen et al., 2012), e.g., if many far relatives are included, for which the deviations in $\boldsymbol{E}$ can be very large. A second approach is to model the distribution of the mean of genotyping individuals, i.e., to assume a unknown mean μ for genotyped individuals: $p(\boldsymbol{u}_2) = N(\mathbf{1}\mu, \boldsymbol{G})$. This is a random variable: the effect of selection or drift on the trait will vary from one conceptual repetition to another. One can equally write $p(\boldsymbol{u}_2) = N(\mathbf{0}, \boldsymbol{G} + \mathbf{1}\mathbf{1}'Var(\mu))$ with μ integrated out. An unbiased method forces the distribution of average values of breeding values ($\overline{\boldsymbol{u}}_2$) to be identical and therefore, the adjustment uses $\boldsymbol{G}^* = a + b\boldsymbol{G}$ with $b = 0$ and $a = \overline{\boldsymbol{A}}_{22} - \overline{\boldsymbol{G}}$. Although this models correctly the change due to genetic trend, it does not consider the fact that there is a reduction in genetic variance from the base population to the genotyped individuals considered in $\boldsymbol{A}_{22}$ but not in $\boldsymbol{G}$; this reduction is considered in $\boldsymbol{A}_{22}$. This has been tackled twice. The first manner is to consider genotyped individuals as a subpopulation of all individuals in the population and to

use Wright's fixation index theory, which allows putting relationships in any scale (Cockerham, 1969, 1973). Translated to our context (Powell et al., 2010) this implies $a = \overline{\boldsymbol{A}}_{22} - \overline{\boldsymbol{G}}$ and $b = 1 - a/2$ (Vitezica et al., 2011). The value of a can be understood as an overall within-population relationship within the genotyped individuals, for a random mating population. This overall relationship cannot be estimated by $\boldsymbol{G}$ for lack of base allele frequencies. The value of $a/2$ can be understood as the "extra" decrease in genetic variance in a random mating population of average relationship $\overline{\boldsymbol{A}}_{22}$. Christensen et al. (2012) remarked that the hypothesis of random mating population is not likely for the group of genotyped animals, since they would born in different years and some being descendants of others, and suggested to infer a based on drift of the mean of the population (as in Vitezica et al., 2011) and b based on the expected genetic variance, which is encapsulated on the average inbreeding observed in $\boldsymbol{G}$ and $\boldsymbol{A}_{22}$. More formally, the empirical variance of breeding values: $S^2_{u_2} = \overline{(\boldsymbol{u}_2'\boldsymbol{u}_2)} - \overline{(\boldsymbol{u}_2)}^2$ has an expectation $\left(\frac{tr(\boldsymbol{A}_{22})}{n} - \overline{\boldsymbol{A}}_{22}\right)\sigma^2_u$ or $\left(\frac{tr(\boldsymbol{G}^*)}{n} - \overline{\boldsymbol{G}}^*\right)\sigma^2_u$ where n is the number of individuals. Forcing unbiasedness implies $\frac{tr(\boldsymbol{G})}{n}b + a = \frac{tr(\boldsymbol{A}_{22})}{n}$ and $+b\overline{\boldsymbol{G}} = \overline{\boldsymbol{A}}_{22}$. In random mating populations in Hardy-Weinberg equilibrium (for instance in dairy cattle and sheep), it turns out that $b = 1 - a/2$ as in Vitezica et al. (2011**)**. If restricting the group of animals for which compatibility is required to those that are born in a certain generation, the assumption of random mating among those genotyped animals is not unreasonable to assume in many livestock species. All these corrections utilize some estimate of the allelic frequencies, and using observed allele frequencies (either based on all genotyped animals, or based on a subset born in a certain generation) is usually done.

Finally, Christensen (2012) suggested the opposite point of view, to "tune" $\boldsymbol{A}_{22}$ to $\boldsymbol{G}$ instead of the opposite. Pedigrees are arbitrary and depend on the start of pedigree, whereas genotypes at the markers are absolute. Allele frequencies, though, change all the time. He

379 modelled the likelihood of markers given the pedigree as a quantitative trait and then
380 integrated over the uncertain allele frequencies. This amounts to fix allele frequencies at 0.5
381 and introduce two extra parameters, γ and s. The γ parameter can be understood as the overall
382 relationship across the base population such that current genotypes are more likely, and
383 integrates the fact that the assumption of unrelatedness at the base population is false in view
384 of genomic results (two animals who share alleles at markers are related even if the pedigree
385 is not informative). More precisely, he devised a new pedigree relationship matrix, $\boldsymbol{A}(\gamma)$
386 whose founders have a relationship matrix $\boldsymbol{A}_{bas} = \gamma + \boldsymbol{I}(1 - \gamma/2)$. Parameter s can be
387 understood as the counterpart of $2\Sigma p_k q_k$ (heterozygosity of the markers) in the base
388 generation. Both parameters can be deduced from maximum likelihood. This model is the
389 only one which introduces all the complexities of pedigrees (former ones are based on
390 average relationships) but it has not been tested with real data so far (Christensen, 2012).

391

392 *4.4 Computational algorithms*

393 The use and development of the Single Step has been possible through the use of several state
394 of the art algorithms. Construction and inversion of matrix $\boldsymbol{G}$ are respectively quadratic and
395 cubic, and are much optimized by the use of efficient algorithms and parallel computations
396 (Aguilar et al., 2011). Construction of matrix $\boldsymbol{A}_{22}$ has been possible, for very large pedigrees,
397 by the algorithm of Colleau (2002) which uses Henderson's decomposition of $\boldsymbol{A} = \boldsymbol{TDT}'$ to
398 devise a "solving" that allows easy multiplication of $\boldsymbol{w} = \boldsymbol{Av}$ and computation of $\boldsymbol{A}_{22}$ in
399 linear time (Aguilar et al., 2011).

400 Further, the use of the solver known as preconditioned conjugated gradients (PCG) allows an
401 easy programming to solve the Single Step mixed model equations. PCG proceeds by
402 repeated multiplications $(\boldsymbol{LHS})\boldsymbol{sol}$. In practice, this product is split into products

$$\begin{pmatrix} \boldsymbol{X'X} & \boldsymbol{X'W} \\ \boldsymbol{W'X} & \boldsymbol{W'W + A^{-1}\lambda} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{u}} \end{pmatrix}$$

For which very efficient algorithms already exist (e.g. Strandén and Lidauer, 1999) and a part

$$(\boldsymbol{G^{-1}} - \boldsymbol{A_{22}^{-1}})\lambda\, \widehat{\boldsymbol{u}}_{\boldsymbol{2}}$$

which can be done very efficiently, in particular using parallelization.

In addition, some implementations of the Single Step have used unsymmetric equations to avoid inversion of $\boldsymbol{G}$ (Misztal et al., 2009; Aguilar et al., 2013), with solution by the Bi-Conjugate Gradient Stabilized algorithm. Legarra and Ducrocq (2012) reviewed and suggested implementations of the Single Step with view towards very large data sets such as in dairy cattle. Problems of these data sets are complexities of the already existing software for regular BLUP and genomic evaluations and size of the data sets, which preclude inversion (and even construction) of $\boldsymbol{G}$. They suggested two main alternatives: a non-symmetric system of equations with non-inverted $\boldsymbol{A}_{22}$ and $\boldsymbol{G}$, and an iterative procedure similar to the Multiple Step but in which results from genomic evaluations would be reintroduced in the regular BLUP evaluation, and results from regular BLUP would be "data" for the genomic evaluations. The non-symmetric system shows slow convergence on large data sets (Aguilar et al., 2013), whereas the iterative method is still untested on large data sets. This is still an active field of research.

4.5 Bayesian regressions in the Single Step

Bayesian or non-linear regressions with non-normal priors for marker effects are certainly more efficient for some traits and species, with the most known example being contents in dairy cattle (VanRaden et al., 2009). This has inspired the search for its integration.

Bayesian regressions can be understood as inferring the variances associated to each marker in the expression $Var(\boldsymbol{a}) = \boldsymbol{D_a}$, i.e. the elements $\sigma^2_{a,k}$ in the diagonal of $\boldsymbol{D_a}$ being SBP specific. Zhang et al. (2010) and Legarra et al. (2011) checked that running a full Bayesian regression to estimate breeding values, or using it to infer variances in $\boldsymbol{D_a}$ to use $\boldsymbol{G} = \boldsymbol{Z_2 D_a Z_2'}$ in a GBLUP gave essentially the same solution. Legarra et al. (2009) suggested to use such $\boldsymbol{G}$ with precomputed variances in the Single Step procedures. This has not been attempted so far. In a similar spirit, Wang et al. (2012) suggested to compute variances in $\boldsymbol{D_a}$ in an iterative manner within the Single Step. They obtained the marker effects from the expression $\widehat{\boldsymbol{a}}|\widehat{\boldsymbol{u_2}} = \boldsymbol{D_a Z_2'}(\boldsymbol{Z_2 D_a Z_2'})^{-1}\widehat{\boldsymbol{u}}_2$, to later infer the i-th marker variance as (proportional to) $\hat{a}_i^2$ (Sun et al., 2013). Note that this estimate is severely biased (it ignores the uncertainty in the estimation of $\hat{a}_i$) and therefore an empirical correction needs to be applied, which is not the case in true Bayesian or maximum likelihood procedures (De los Campos et al., 2009; Shen et al., 2013). After computation of a new $\boldsymbol{G}$ Single Step GBLUP is rerun and markers are re-estimated, and the procedure is iterated a few times. Their simulation showed an increased accuracy of this method for traits with large QTLs.

Legarra and Ducrocq (2012) suggested two ways of dealing with Bayesian regressions. The first one was to use an equivalent set of mixed model equations including marker effects:

$$\begin{pmatrix} \boldsymbol{X'X} & \boldsymbol{X_1'W_1} & \boldsymbol{X_2'W_2Z_2} \\ \boldsymbol{W_1'X} & \boldsymbol{W_1'W_1' + A^{11}\lambda} & \boldsymbol{A^{12}\lambda} \\ \boldsymbol{Z_2'W_2X_2} & \boldsymbol{Z_2'\,A^{12}\lambda} & \boldsymbol{Z_2'W_2W_2Z_2 + Z_2'\left(A^{22} - A_{22}^{-1}\right)Z_2\lambda + D_a^{-1}\sigma_e^2} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{u}}_1 \\ \boldsymbol{a} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X'y} \\ \boldsymbol{W_1'y_1} \\ \boldsymbol{Z_2'W_2'y_2} \end{pmatrix}$$

This system of equations could be solved by a Bayesian procedure such as the Gibbs sampler. In the second option, an equivalent iterative procedure can iterate between solutions to regular BLUP and (Bayesian) genomic predictions; the results of one would be introduced into the

443 other. Because this system does not infer marker variances *per se*, it does not suffer from the
444 bias in variance estimation of Wang et al (2012). Tuning markers to be in the same scale as
445 pedigree in this set of equations would include an extra unknown for the parameter μ in
446 Vitezica et al. (2011).

447

448 In addition, Fernando et al. (2013) recently presented another system of equations explicit on
449 marker solutions. Equations include marker effects for *all* individuals, imputed following
450 Gengler's method, and residual pedigree-based EBV for nongenotyped animals, $\boldsymbol{\epsilon}$. This $\boldsymbol{\epsilon}$ is
451 what remains of the breeding value after we fit (imputed) SNP effects to nongenotyped
452 individuals. Therefore total genetic value: $\boldsymbol{u} = \begin{pmatrix} \widehat{\mathbf{Z}}_1 \\ \mathbf{Z}_2 \end{pmatrix} \boldsymbol{a} + \begin{pmatrix} \boldsymbol{\epsilon} \\ 0 \end{pmatrix} = \widehat{\boldsymbol{Z}}\boldsymbol{a} + \begin{pmatrix} \boldsymbol{\epsilon} \\ 0 \end{pmatrix}$.

453 Their final Single Step mixed model equations are

454 $$\begin{pmatrix} \boldsymbol{X'X} & \boldsymbol{X'W\widehat{Z}} & \boldsymbol{X_1'W_1} \\ \boldsymbol{\widehat{Z}'W'X} & \boldsymbol{\widehat{Z}'W'W\widehat{Z}} + \mathbf{I}\frac{\sigma_e^2}{\sigma_a^2} & \boldsymbol{\widehat{Z}_1'W_1'W_1} \\ \boldsymbol{W_1'X_1} & \boldsymbol{W_1'W_1\widehat{Z}_1} & \boldsymbol{W_1'W_1} + \boldsymbol{A}^{11}\frac{\sigma_e^2}{\sigma_g^2} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{a}} \\ \widehat{\boldsymbol{\epsilon}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X'y} \\ \boldsymbol{\widehat{Z}'W'y} \\ \boldsymbol{W_1'y} \end{pmatrix},$$

455 in which a Gibbs sampler can iterate to obtain Bayesian estimates. These equations are
456 simpler than previous ones but at the cost of a very dense and large system of equations.

457 All these methods for Bayesian regressions in Single Step are largely untested, and only
458 Wang et al. (2011) method is efficiently implemented and has been used in real data sets
459 (Dikman et al., 2013), for which no alternative currently exists.

460

461 *4.6 Unknown parent groups*

462 Missing genealogy and/or crosses are ubiquitous in animal breeding. Typical solutions are fit
463 of unknown parent groups, which model different means across groups of founders well
464 identified, i.e. belonging to different generations or breeds, and estimation is done using a
465 special version of $\boldsymbol{A^{-1}}$ (Quaas, 1988). Unfortunately, the Single Step Mixed Model equations
466 do not accommodate this well, because of the additional matrices $(\boldsymbol{G^{-1}} - \boldsymbol{A_{22}^{-1}})$. The problem
467 was explained in detail by Misztal et al. (2013b) who showed that proper equations would
468 imply complex terms of the form $\boldsymbol{Q_2'(G^{-1} - A_{22}^{-1})Q_2}$, implying matrix $\boldsymbol{Q_2}$ with fractions of
469 each unknown parent group for each genotyped animal. These modifications are difficult to
470 compute and program. Current alternatives involve ignoring the term (often with negligible
471 results) or using the original Westell-Robinson model, which is in the form

$$\boldsymbol{y = Xb + Qg + Wu + e}$$

472 (Quaas, 1988) and fitting unknown parent groups as covariates. This is satisfactory and
473 involves no approximations, but cumbersome to implement and of slow convergence.

474

475 *4.7 Accuracies*

476 Individual accuracies can be obtained in principle from the inverse of the Single Step mixed
477 model equations. This is impossible in practice for medium to large data sets. Therefore,
478 Misztal et al. (2013a) suggested extending known approximations in the estimation of
479 accuracy to the Single Step case. Modifications involve use known approximations for the
480 pedigree-based BLUP and add extra information from $(\boldsymbol{G^{-1}} - \boldsymbol{A_{22}^{-1}})$ to each animal; then to
481 iterate the procedure. This procedure is accurate in dairy species, as attested by Misztal et al.
482 (2013a) and in Manech dairy sheep (Baloche et al., unpublished) where correlations between

483 approximate accuracies and exact accuracies from inverse of the Mixed Model Equations
484 were found to equal 0.95 .

485

486 **5. FUTURE DEVELOPMENTS**

487 Among important possible extensions, we will mention two: crosses and fit of dominance
488 effects.

489 *5.1 Crosses.*

490 Development of the Single Step has been done for purebred populations, in which heterosis is
491 absent, genetic variance is assumed constant throughout the generations, and matings are
492 (close to being) at random. In classical theory (e.g., Lo et al., 1997) populations involved in
493 crossing are assumed completely unrelated; this is subject to discussion depending on the
494 genetic architecture of the trait. For instance, Ibañez-Escriche et al. (2009) obtained the same
495 accuracy fitting markers with the same or different effects across breeds. Recently,
496 Christensen et al. (2014, in press) presented a Single Step in these lines, where the value of a
497 crossbred animals is a sum of gametic effects, each with a different within-pure breed
498 extended relationship matrix. These aspects need to be further derived. Also, testing in real
499 data sets is most necessary because simulations are unreliable for such complex cases.
500 However, crossbred data sets with genomic information are scarce so far.

501

502 *5.2 Dominance.*

503 Genomic predictions including dominance (e.g., Toro and Varona, 2010; Wellmann and
504 Bennewitz, 2012) are much easier than their pedigree counterparts, which are notoriously

505 difficult, in particular if inbreeding is involved (DeBoer and Hoeschele, 1993). Dominance
506 versions of GBLUP have been proposed (Su et al., 2012; Vitezica et al., 2013) and real data
507 analysis, done (Su et al., 2012; Ertl et al., 2013). However, these methods need that genotyped
508 animals have a phenotype. There are no methods to generate pseudo-phenotypes including
509 dominance, because all methods to generate pseudo-data involve additive relationships only.
510 For instance, computation of DYD's in dairy cattle will average to zero dominance deviations
511 of the offspring. Therefore Single Step methods for dominance are highly relevant, yet a
512 simple combination of pedigree-based and marker-based methods is difficult because the
513 pedigree-based method is already difficult.

514

515 **6. OBSCURE POINTS AND LIMITS**

516 *6.1 Treatment of linkage.*

517 Markers are physically linked and their co-ocurrence is correlated. However, most genomic
518 prediction models, including Bayesian Regressions and the Single Step, assume markers to be
519 unlinked. In addition, the pedigree-based matrix $\boldsymbol{A}$ assumes loci as unlinked as well.
520 Meuwissen et al. (2011) suggested a modified $\boldsymbol{H}$ matrix in which pedigree relationships
521 would be computed using pedigree and markers, in the spirit of Fernando and Grossmann
522 (1989), by means of iterative peeling. However, this has not been tested in real life data sets
523 and it is unknown if this methods generalizes well to large pedigrees.

524 *6.2 Convergence of solvers.*

525 The convergence rate with regular Single Step when solved by PCG iteration depends on
526 species. The rate is similar to BLUP and poses no problem with complete pedigree and a
527 uniform base population (e.g., chicken). The rate is also good when high-accuracy genotyped

528 animals (dairy bulls). The rate can be poor with complex models when the pedigree contains
529 many generations of animals without phenotypes. In such a case, restricting the pedigree to
530 fewer old animals improves the rate. Poor convergence rate in some models is due to
531 incompatibility between $\boldsymbol{G}$ and $\boldsymbol{A}_{22}$ when the pedigree has missing animals across generations
532 (Misztal et al., 2013). When $\boldsymbol{G}$ is scaled for an average $\boldsymbol{A}_{22}$, elements of $\boldsymbol{A}_{22}^{-1}$ due to animals
533 with very long pedigree are larger. Solutions to this problem include modifications to $\boldsymbol{A}$ (e.g.,
534 as in Christensen, 2012), or pedigree or even phenotype truncations. Lourenco et al (2014, in
535 press) investigated the effect of cutting pedigrees and phenotypes on accuracy for the
536 youngest generation. Use of data beyond 2 generations of phenotypes and 4 generations of
537 pedigree did not improve the accuracy while increasing computing costs.

538 In large data sets with many genotyped individuals (e.g., with genotyped cows) there are
539 reports of lack of, or very slow, convergence (Harris et al., 2013; VanRaden, unpublished).
540 This raises the question if the typical form of the mixed model equations for single-Step,
541 including $\boldsymbol{G}$ and $\boldsymbol{A}_{22}$ is the most appropriate, or alternative forms based on marker effects such
542 as those presented by Legarra and Ducrocq (2012) or Fernando et al. (2013) are better. A limit
543 to testing these approaches is the availability of very general software for BLUP. General
544 software (multiple trait, multiple effects, etc.) does not exist for marker-based methods.

545 *6.3 Computational limits.*

546 Computing and inverting $\boldsymbol{G}$ and and $\boldsymbol{A}_{22}$ is challenging and of cubic cost, which will
547 eventually preclude its use for, say, >100,000 animals, and alternatives have been suggested
548 (Legarra and Ducrocq, 2012; Fernando et al., 2013) but not thoroughly tested. These
549 alternatives would be either highly parallelizable or use indirect representations avoiding
550 explicit computations. However, so far, problems of convergence seem more limiting than
551 size.

7. CURRENT STATE AND PRACTICAL EXPERIENCES

7.1 Dairy sheep.

In France, the Lacaune, Manech and Basco-Bearnaise genomic evaluations use Single Step in its typical form, with corrections of $\boldsymbol{G}$ to match $\boldsymbol{A}_{22}$ and with the fit of unknown parent groups as covariates. Preliminary research did not show an added accuracy of Bayesian predictors (Duchemin et al., 2012). Single step results in higher accuracy than GBLUP with pseudo-phenotypes (Baloche et al., 2014) and in a much simpler implementation. Single Step will be the method for genomic prediction in the future Lacaune dairy sheep genomic selection scheme.

7.2 Dairy goat.

In France, the dairy goat population is testing genomic selection procedures with the Single Step as the evaluation tool (Carillier et al., 2013) although it is very soon to establish its impact.

7.3 Pigs.

In Denmark, routine genetic evaluation of the three DanBred breeds Duroc, Landrace and Yorkshire has since October 2011 been made by Single-Step in its typical form, with corrections of $\boldsymbol{G}$ to match $\boldsymbol{A}_{22}$. The implementation of genomic evaluation via Single-Step was straight-forward and it has resulted in increased accuracy compared to the traditional genetic evaluation. Breeding companies PIC and ToPigs use Single Step for genomic predictions.

7.4 Dairy cattle.

National evaluations are based on multiple step procedures, but most countries are willing to change to Single Step, and many are experimenting (e.g., VanRaden, unpublished; Harris et al., 2013). The reason for this change is the conceptual and practical simplicity of the Single Step, and its ability to account for genomic preselection (Patry and Ducrocq, 2011). Due to abundance of data and completeness of genotyping, tests show equivalent accuracies of Single Step and multiple step procedures (e.g., Aguilar et al., 2010). However, for several milkability traits, ssGBLUP was always more accurate than GBLUP (Gray et al., 2012). Also, Pribyl (2011) showed higher accuracy of the Single Step for Check Republic data.

7.5 Beef cattle

There are no studies on the application of Single Step to real data sets. However, in a simulation study by Lourenco et al. (2013), accuracies of genomic predictions with ssGBLUP were always higher than with BLUP, which was not the case with BayesC. This was particularly true for maternal traits.

7.6 Chicken

In studies on decay of genomic prediction over generations (Wolc et al., 2011), BayesB was more accurate than single-trait GBLUP but less accurate than 2-trait GBLUP; in that study, GBLUP was applied to a reduced animal model and was equivalent to ssGBLUP. Chen et al. (2011a,b) also showed higher accuracies of Single Step than with Bayesian regressions.

8. Software

To our knowledge, the only publicly available software package which can directly run Single Step evaluations is the BLUPF90 family of programs (Misztal et al., 2002; http://nce.ads.uga.edu/wiki/doku.php?id=start) in which it is fully implemented including

regular BLUP, REML, Gibbs samplers, threshold models and iteration on data for very large data sets, and several options (most of them mentioned above). Other public packages such as DMU, (Madsen and Jensen, 2012, http://www.dmu.agrsci.dk/), Wombat (Meyer, 2013; http://didgeridoo.une.edu.au/km/wombat.php) can include covariance matrices computed externally, and therefore matrix $\boldsymbol{H}^{-1}$ needs to be computed with an external tool and then fit into the model.

9. Conclusion: overall benefits of the Single Step

The Single Step provides a simple method to combine all information in a simple manner, with the additional advantage of requiring little changes to existing software. Accuracy is usually as high as, if not greater than, any other method. Some studies concerning accuracy of the Single Step have been gathered in Table 1. Beyond its extra accuracy, it has the following interesting properties:

1. Automatic accounting of all relatives of genotyped individuals and their performances.
2. Simultaneous fit of genomic information and estimates of other effects (e.g., contemporary groups). Therefore not loss of information.
3. Feedback: the extra accuracy in genotyped individuals is transmitted to all their relatives (*e.g.* Christensen et al., 2012).
4. Simple extensions. Because this is a linear BLUP-like estimator, the extension to more complicated models (multiple trait, threshold traits, test day records) is immediate. Any model fit using relationship matrices can be fit using combined relationship matrices.
5. Analytical framework. The Single Step provides an analytical framework for further developments. This is notoriously difficult with pseudo-data.

621

622 TABLE 1 HERE

623

624

625

626 **Table 1. Accuracy of Single Step versus other methods in some species**

Authors	Single Step	Multiple step	Pedigree BLUP	Species, trait
Aguilar et al., 2010	0.70	0.70	0.60	Dairy cattle, final score
Baloche et al., 2013	0.47	0.43	0.32	Milk yield, dairy sheep
Chen et al., 2011b*	0.36		0.20	Breast meat, chicken
Chen et al., 2011a	0.37	0.09	0.28	Leg Score, chicken
Christensen et al., 2012*	0.35	0.35	0.18	Daily gain, pigs
Aguilar et al., 2011	0.39		0.26	Conception rate at first parity

627 *predictive abilities: $r(y, \hat{u})$

628

629

630 REFERENCES

631 Abraham, K.J., Totir, L.R., Fernando, R.L., 2007. Improved techniques for sampling complex
632 pedigrees with the Gibbs sampler. Gen Sel Evol 39, 27-38.
633 Aguilar, I., Legarra, A., Tsuruta, S., Misztal, I., 2013. Genetic evaluation using unsymmetric
634 single step genomic methodology with large number of genotypes. Interbull Bulletin 47.
635 Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., Lawlor, T.J., 2010. Hot topic:
636 a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic
637 evaluation of Holstein final score. J Dairy Sci 93, 743-752.
638 Aguilar, I., Misztal, I., Legarra, A., Tsuruta, S., 2011. Efficient computations of genomic
639 relationship matrix and other matrices used in the single-step evaluation. J Anim Breed Genet
640 128, 422-428.
641 Baloche, G., Legarra, A., Sallé, G., Larroque, H., Astruc, J.M., Robert-Granié, C., Barillet, F.,
642 2014. Assessment of accuracy of genomic prediction for French Lacaune dairy sheep. J Dairy
643 Sci 97, 1107-1116.
644 Carillier, C., Larroque, H., Palhière, I., Clément, V., Rupp, R., Robert-Granié, C., 2013. A
645 first step toward genomic selection in the multi-breed French dairy goat population. J Dairy
646 Sci 96, 7294-7305.
647 Casella, G., Berger, R.L., 1990. Statistical inference. Duxbury Press Belmont, CA.
648 Chen, C., Misztal, I., Aguilar, I., Tsuruta, S., Aggrey, S., Wing, T., Muir, W., 2011a.
649 Genome-wide marker-assisted selection combining all pedigree phenotypic information with
650 genotypic data in one step: An example using broiler chickens. J Anim Sci 89, 23-28.
651 Chen, C.Y., Misztal, I., Aguilar, I., Legarra, A., Muir, W.M., 2011b. Effect of different
652 genomic relationship matrices on accuracy and scale. J Anim Sci 89, 2673-2679.
653 Christensen, O., Madsen, P., Nielsen, B., Ostersen, T., Su, G., 2012. Single-step methods for
654 genomic evaluation in pigs. Animal 6, 1565-1571.

655 Christensen, O.F., 2012. Compatibility of pedigree-based and marker-based relationship
656 matrices for single-step genetic evaluation. Gen Sel Evol 44, 37.
657 Christensen, O.F., Lund, M.S., 2010. Genomic prediction when some animals are not
658 genotyped. Gen Sel Evol 42, 2.
659 Cochran, W., 1951. Improvement by means of selection. Second Berkeley Symposium on
660 Mathematical Statistics and Probability, pp. 449-470.
661 Cockerham, C.C., 1969. Variance of gene frequencies. Evolution 23, 72-84.
662 Cockerham, C.C., 1973. Analyses of gene frequencies. Genetics 74, 679.
663 Colleau, J.J., 2002. An indirect approach to the extensive calculation of relationship
664 coefficients. Gen Sel Evol 34, 409-422.
665 De Boer, I., Hoeschele, I., 1993. Genetic evaluation methods for populations with dominance
666 and inbreeding. Theor Appl Gen 86, 245-258.
667 de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K.,
668 Cotes, J.M., 2009. Predicting quantitative traits with regression models for dense molecular
669 markers and pedigree. Genetics 182, 375-385.
670 Dikmen, S., Cole, J.B., Null, D.J., Hansen, P.J., 2013. Genome-wide association mapping for
671 identification of quantitative trait loci for rectal temperature during heat stress in Holstein
672 cattle. PLoS ONE 8, e69202.
673 Duchemin, S., Colombani, C., Legarra, A., Baloche, G., Larroque, H., Astruc, J.-M., Barillet,
674 F., Robert-Granié, C., Manfredi, E., 2012. Genomic selection in the French Lacaune dairy
675 sheep breed. J Dairy Sci 95, 2723-2733.
676 Ertl, J., Legarra, A., Vitezica, Z.G., Varona, L., Edel, C., Reiner, E., Götz, K.-U., 2013.
677 Genomic analysis of dominance effects in milk production and conformation traits of
678 Fleckvieh cattle. Interbull Bulletin 47.
679 Fernando, R., Gianola, D., 1986. Optimal properties of the conditional mean as a selection

680 criterion. Theor Appl Gen 72, 822-825.
681 Fernando, R.L., Garrick, D.J., Dekkers, J.C.M., 2013. Bayesian regression method for
682 genomic analyses with incomplete genotype data. European Federation of Animal Science.
683 Wageningen Press, Nantes, France, p. 225.
684 Fernando, R.L., Grossman, M., 1989. Marker assisted prediction using best linear unbiased
685 prediction. Gen Sel Evol 21, 467-477.
686 Forni, S., Aguilar, I., Misztal, I., 2011. Different genomic relationship matrices for single-step
687 analysis using phenotypic, pedigree and genomic information. Gen Sel Evol 43, 1.
688 Garcia-Cortes, L.A., Legarra, A., Chevalet, C., Toro, M.A., 2013. Variance and Covariance of
689 Actual Relationships between Relatives at One Locus. PLoS ONE 8, e57003.
690 Garrick, D.J., Taylor, J.F., Fernando, R.L., 2009. Deregressing estimated breeding values and
691 weighting information for genomic regression analyses. Gen Sel Evol 41, 44.
692 Gengler, N., Mayeres, P., Szydlowski, M., 2007. A simple method to approximate gene
693 content in large pedigree populations: application to the myostatin gene in dual-purpose
694 Belgian Blue cattle. animal 1, 21-28.
695 Gianola, D., Fernando, R.L., 1986. Bayesian Methods in Animal Breeding Theory. J Anim
696 Sci 63, 217.
697 Gray, K.A., Cassady, J.P., Huang, Y., Maltecca, C., 2012. Effectiveness of genomic
698 prediction on milk flow traits in dairy cattle. Gen Sel Evol 44:24
699 Harris, B.L., Winkelman, A.M., Johnson, D.L., 2013. Impact of including a large number of
700 female genotypes on genomic selection. Interbull Bulletin 47.
701 Hayes, B.J., Visscher, P.M., Goddard, M.E., 2009. Increased accuracy of artificial selection
702 by using the realized relationship matrix. Genet Res 91, 47-60.
703 Henderson, C., 1973. Sire evaluation and genetic trends. In Proceedings of the animal
704 breeding and genetics symposium in honor of Dr. Jay L. Lush pp. 10-41.

Hickey, J.M., Kinghorn, B.P., Tier, B., van der Werf, J.H., Cleveland, M.A., 2012. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Gen Sel Evol 44.

Hill, W.G., Weir, B.S., 2011. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet Res (Camb), 1-18.

Ibáñez-Escriche, N., Fernando, R.L., Toosi, A., Dekkers, J.C.M., 2009. Genomic selection of purebreds for crossbred performance. Gen Sel Evol 41, 12.

Legarra, A., Aguilar, I., Misztal, I., 2009. A relationship matrix including full pedigree and genomic information. J Dairy Sci 92, 4656-4663.

Legarra, A., Ducrocq, V., 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.

Legarra, A., Misztal, I., 2008. Technical note: Computing strategies in genome-wide selection. J Dairy Sci 91, 360-366.

Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F., Fritz, S., 2011. Improved Lasso for genomic selection. Genet Res (Camb) 93, 77-87.

Legarra, A., Robert-Granié, C., Manfredi, E., Elsen, J.-M., 2008. Performance of genomic selection in mice. Genetics 180, 611-618.

Lo, L., Fernando, R., Grossman, M., 1997. Genetic evaluation by BLUP in two-breed terminal crossbreeding systems under dominance. J Anim Sci 75, 2877-2884.

Lourenco, D., Misztal, I., Wang, H., Aguilar, I., Tsuruta, S., Bertrand, J., 2013. Prediction accuracy for a simulated maternally affected trait of beef cattle using different genomic evaluation models. J Anim Sci 91, 4090-4098.

Madsen, P., & Jensen, J. (2000). A user's guide to DMU. A package for analysing multivariate mixed models. Version 6, 1-33.

730 Meuwissen, T., Luan, T., Woolliams, J., 2011. The unified approach to the use of genomic
731 and pedigree information in genomic evaluations revisited. J Anim Breed Genet 128, 429-
732 439.
733 Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using
734 genome-wide dense marker maps. Genetics 157, 1819-1829.
735 Meyer, K. (2007). WOMBAT—A tool for mixed model analyses in quantitative genetics by
736 restricted maximum likelihood (REML). Journal of Zhejiang University Science B, 8(11),
737 815-821.
738 Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., and Lee, D. H. (2002). BLUPF90
739 and related programs (BGF90). In Proceedings of the 7th World Congress on Genetics
740 Applied to Livestock Production, Montpellier, France, August, 2002. Session 28. (pp. 1-2).
741 Institut National de la Recherche Agronomique (INRA).
742 Misztal, I., Legarra, A., Aguilar, I., 2009. Computing procedures for genetic evaluation
743 including phenotypic, full pedigree, and genomic information. J Dairy Sci 92, 4648-4655.
744 Misztal, I., Tsuruta, S., Aguilar, I., Legarra, A., VanRaden, P., Lawlor, T., 2013a. Methods to
745 approximate reliabilities in single-step genomic evaluation. J Dairy Sci. 96, 647-654
746 Misztal, I., Vitezica, Z., Legarra, A., Aguilar, I., Swan, A., 2013b. Unknown-parent groups in
747 single-step genomic evaluation. J Anim Breed Genet. 130, 252-258.
748 Patry, C., Ducrocq, V., 2011. Evidence of biases in genetic evaluations due to genomic
749 preselection in dairy cattle. J Dairy Sci 94, 1011-1020.
750 Powell, J.E., Visscher, P.M., Goddard, M.E., 2010. Reconciling the analysis of IBD and IBS
751 in complex trait studies. Nat Rev Genet 11, 800-805.
752 Přibyl, J., Madsen, P., Bauer, J., Přibylová, J., Šimečková, M., Vostrý, L., Zavadilová, L.,
753 2013. Contribution of domestic production records, Interbull estimated breeding values, and
754 single nucleotide polymorphism genetic markers to the single-step genomic evaluation of

755 milk production. J Dairy Sci 96, 1865-1873.
756 Quaas, R.L., 1988. Additive genetic model with groups and relationships. J Dairy Sci 71,
757 1338-1345.
758 Ricard, A., Danvy, S., Legarra, A., 2013. Computation of deregressed proofs for genomic
759 selection when own phenotypes exist with an application in French show-jumping horses. J
760 Anim Sci 91, 1076-1085.
761 Ritland, K., 1996. Estimators for pairwise relatedness and individual inbreeding coefficients.
762 Genetical research 67, 175-185.
763 Shen, X., Alam, M., Fikse, F., Rönnegård, L., 2013. A novel generalized ridge regression
764 method for quantitative genetics. Genetics 193, 1255-1268.
765 Smith, H.F., 1936. A discriminant function for plant selection. Annals of Eugenics 7, 240-
766 250.
767 Strandén, I., Garrick, D.J., 2009. Technical note: Derivation of equivalent computing
768 algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci 92, 2971-
769 2975.
770 Strandén, I., Lidauer, M., 1999. Solving large mixed linear models using preconditioned
771 conjugate gradient iteration. J Dairy Sci 82, 2779-2787.
772 Su, G., Christensen, O.F., Ostersen, T., Henryon, M., Lund, M.S., 2012. Estimating additive
773 and non-additive genetic variances and predicting genetic merits using genome-wide dense
774 single nucleotide polymorphism markers. PLoS ONE 7, e45293.
775 Su, G., Madsen, P., Nielsen, U.S., Mäntysaari, E.A., Aamand, G.P., Christensen, O.F., Lund,
776 M.S., 2012. Genomic prediction for Nordic Red Cattle using one-step and selection index
777 blending. J Dairy Sci 95, 909-917.
778 Sun, C., Van Raden, P., 2013. Mating programs including genomic relationships. J Dairy Sci
779 96, 653.

780 Sun, X., Qu, L., Garrick, D.J., Dekkers, J.C., Fernando, R.L., 2012. A Fast EM Algorithm for
781 BayesA-Like Prediction of Genomic Breeding Values. PLoS ONE 7, e49157.
782 Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data
783 augmentation. J Amer Stat Assoc 82, 528-540.
784 Toro, M.Á., García-Cortés, L.A., Legarra, A., 2011. A note on the rationale for estimating
785 genealogical coancestry from molecular markers. Gen Sel Evol 43, 27.
786 Toro, M.A., Varona, L., 2010. A note on mate allocation for dominance handling in genomic
787 selection. Gen Sel Evol 42, 33.
788 VanRaden, P., Null, D., Sargolzaei, M., Wiggans, G., Tooker, M., Cole, J., Sonstegard, T.,
789 Connor, E., Winters, M., van Kaam, J., 2012. Genomic imputation and evaluation using high-
790 density Holstein genotypes. J Dairy Sci.
791 VanRaden, P., Wiggans, G., 1991. Derivation, calculation, and use of national animal model
792 information. J Dairy Sci 74, 2737-2746.
793 VanRaden, P.M., 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91,
794 4414-4423.
795 VanRaden, P.M., Tassell, C.P.V., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor,
796 J.F., Schenkel, F.S., 2009. Invited review: reliability of genomic predictions for North
797 American Holstein bulls. J Dairy Sci 92, 16-24.
798 Vitezica, Z., Aguilar, I., Misztal, I., Legarra, A., 2011. Bias in genomic predictions for
799 populations under selection. Genetics Research 93, 357-366.
800 Vitezica, Z.G., Varona, L., Legarra, A., 2013. On the additive and dominant variance and
801 covariance of individuals within the genomic selection scope. Genetics 195, 1223-1230.
802 Wang, H., Misztal, I., Aguilar, I., Legarra, A., Muir, W., 2012. Genome-wide association
803 mapping including phenotypes from relatives without genotypes. Genetics Research 94, 73-
804 83.

Wellmann, R., Bennewitz, J., 2011. The contribution of dominance to the understanding of quantitative genetic variation. Genetics Research 93, 139.

Wellmann, R., Bennewitz, J., 2012. Bayesian models with dominance effects for genomic evaluation of quantitative traits. Genetics Research 94, 21.

Wolc, A., Stricker, C., Arango, J., Settar, P., Fulton, J.E., O'Sullivan, N.P., Preisinger, R., Habier, D., Fernando, R., Garrick, D.J., Lamont, S.J., Dekkers, J.C.M., 2011. Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. Gen Sel Evol 43, 5.

Zhang, Z., Liu, J., Ding, X., Bijma, P., de Koning, D.-J., Zhang, Q., 2010. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS ONE 5, e12648.