Original Article

Unknown-parent groups in single-step genomic evaluation

I. Misztal¹, Z.G. Vitezica², A. Legarra³, I. Aguilar⁴ & A.A. Swan⁵

1 Department of Animal and Dairy Science, University of Georgia, Athens, GA, USA
2 Université de Toulouse, UMR 1289 TANDEM, INRA/INPT-ENSAI/ENVET F-31326, Castanet-Tolosan, France
3 INRA, Station d’Amélioration Génétique des Animaux, Castanet-Tolosan, France
4 Instituto Nacional de Investigación Agropecuaria, Las Brujas, Uruguay
5 Animal Breeding and Genetics Unit, University of New England, Armidale, NSW, Australia

Keywords
Bias, genomic evaluation; genomic relationships; unknown-parent groups.

Correspondence
I. Misztal, Department of Animal and Dairy Science, University of Georgia, Athens, GA, 30602-2771, USA. Tel: 706 542 0961; Fax: 706 583 0274; E-mail: ignacy@uga.edu

Received: 18 September 2012; accepted: 30 November 2012

Summary
In single-step genomic evaluation using best linear unbiased prediction (ssGBLUP), genomic predictions are calculated with a relationship matrix that combines pedigree and genomic information. For missing pedigrees, unknown selection processes, or inclusion of several populations, a BLUP model can include unknown-parent groups (UPG) in the animal effect. For ssGBLUP, UPG equations also involve contributions from genomic relationships. When those contributions are ignored, UPG solutions and genetic predictions can be biased. Options to eliminate or reduce such bias are presented. First, mixed model equations can be modified to include contributions to UPG elements from genomic relationships (greater software complexity). Second, UPG can be implemented as separate effects (higher cost of computing and data processing). Third, contributions can be ignored when they are relatively small, but they may be small only after refinements to UPG definitions. Fourth, contributions may approximately cancel out when genomic and pedigree relationships are constructed for compatibility; however, different construction steps are required for unknown parents from the same or different populations. Finally, an additional polygenic effect that also includes UPG can be added to the model.

Introduction
A genomic evaluation can be conducted with best linear unbiased prediction (BLUP) using either multi-step (VanRaden et al. 2009) or single-step methodology (Aguilar et al. 2010; Christensen & Lund 2010). Whereas a multi-step evaluation uses results of a regular pedigree BLUP for subsequent steps, a single-step genomic BLUP (ssGBLUP) evaluation uses BLUP with a relationship matrix that combines pedigree and genomic information. An ssGBLUP evaluation is potentially simpler and more accurate than a multi-step genomic evaluation (Chen et al. 2011; Vitezica et al. 2011), and it has the potential to avoid bias from selection of animals based only on genomic information (Patry & Ducroq 2011; VanRaden 2012).

The ssGBLUP method has been used for several large-scale analyses, including in dairy cattle (Tsutui et al. 2011; Harris et al. 2012), in pigs (Forni et al. 2011; Christensen et al. 2012) and in chickens (Chen et al. 2011). Although ssGBLUP was accurate and computationally efficient in most studies, problems were reported when the model included unknown-parent groups (UPGs). For commercial pig data, large re-ranking was observed compared with traditional BLUP, and the convergence rate deteriorated as the number of traits increased (S. Forni, personal communication). In an analysis with multiple breeds of sheep, A.A. Swan (personal communication) found
that UPG solutions were biased when UPGs were fitted in an expanded relationship matrix ($A^{-1}$; Quaas 1988), e.g. using the transformation of Quaas & Pollak (1981). However, biases were much smaller when UPGs were fitted as separate covariates, i.e. before QP transformation. No biases were observed for dairy cattle by Tsuruta et al. (2011) and Harris et al. (2012) with models including UPGs.

The UPGs were designed to accommodate missing pedigrees and data in the animal model (Quaas 1988), in particular when selection was practised in foreign countries and therefore imported animals had a different average than national populations (e.g. Thompson 1979). Use of UPGs also is common in populations for which assigning paternity is difficult, such as sheep, and UPGs are crucial in modelling multi-breed populations in beef (Legarra et al. 2007). Use of UPGs allows for more realistic estimates of genetic trends (Tieron et al. 2002). As $n$ UPGs are equivalent to having $n$ covariates in the model, UPGs need to be defined carefully to avoid confounding with other fixed effects (Sullivan & Schaeffer 1994). The purpose of this study was to investigate issues of UPGs specific to ssGBLUP. As the number of possible scenarios for UPGs in different populations is high, the issues are investigated analytically rather than experimentally.

**Material and Methods**

**Model**

Assume a model with explicit UPGs, ignoring all fixed effects other than the groups (Quaas 1988):

$$y = ZQs + Za + e,$$

where $s$ is the group effect, $a$ is the animal additive effect, $Z$ is a matrix that assigns records to animals, $Q$ is a matrix that assigns animals to groups and $e$ is the residual. More specifically, each row of $Q$ contains the genetic group composition of an animal. The variances are $\text{var}(a) = A\sigma_e^2$ and $\text{var}(e) = I\sigma_e^2$, where $A$ is the numerator relationship matrix and the variance components $\sigma_e^2$ and $\sigma_a^2$ are set to unity for convenience. The mixed model equations with explicit group effects are

$$\begin{bmatrix} Z'Z + A^{-1} & -A^{-1}Q \\ -Q'A^{-1} & Q'A^{-1}Q \end{bmatrix} \begin{bmatrix} s \\ a \end{bmatrix} = \begin{bmatrix} Z'y \\ 0 \end{bmatrix}.$$

Such equations can be easily constructed without creating $Q$ directly:

$$\begin{bmatrix} Z'Z + A \\ \end{bmatrix} \begin{bmatrix} s \\ a \end{bmatrix} = \begin{bmatrix} Z'y \\ 0 \end{bmatrix},$$

where $Z'$ includes additional rows with zeroes and $A$ is created using the following simple rules (Quaas 1988) or accounting for inbreeding (VanRaden 1992):

$$A' = \sum_{i} A_i', \quad \text{where} \quad A_i' = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ 0.5 & 0.5 & \vdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \ddots & 2 \end{bmatrix}, \quad \text{sire or UPG}$$

$$\begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ 0.5 & 0.5 & \vdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \ddots & 2 \end{bmatrix}, \quad \text{dam or UPG}$$

$$\begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ -1 & -1 & \ddots & 2 \end{bmatrix}, \quad \text{animal}$$

where the summation is over all animals and $m_i$ is accounting for Mendelian sampling for animal $i$. With inbreeding ignored, $m_i$ is a function of the number of known parents ($1$ if both known, $2/3$ if one known, $0.5$ if both unknown). With inbreeding accounted for, $m = 2/(2 - F_{\text{ sire}} - F_{\text{ dam}})$ with both known, $2/(3 - F_{\text{ sire}} - 2 F_{\text{ dam}})$ with dam known, $2/(3 - 2 F_{\text{ sire}} - F_{\text{ dam}})$ with sire known, $2/(4 - 2 F_{\text{ sire}} - 2 F_{\text{ dam}})$ with both unknown, where $F_{\text{ sire}}$ and $F_{\text{ dam}}$ are inbreeding coefficients of sire and dam, respectively.

**Analysis**

The ssGBLUP method includes a matrix $H$ that combines pedigree and genomic relationships (Legarra et al. 2009). When explicit UPGs are included,

$$\begin{bmatrix} Q'Z'Q & Q'ZZ \\ Z'Q & Z'Z + H^{-1} \end{bmatrix} \begin{bmatrix} s \\ a \end{bmatrix} = \begin{bmatrix} Q'Z'y \\ Z'y \end{bmatrix},$$

where $H^{-1}$ is the same as in Aguilar et al. (2010):

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

 above, $G$ is a genomic relationship matrix and $A_{22}$ is a numerator relationship matrix for animals in $G$.

After QP transformation, the mixed model equations become
\[
\begin{bmatrix}
z' Z \cdot H^{-1} & -H^{-1} Q \\
-Q' H^{-1} & Q' H^{-1} Q
\end{bmatrix} \begin{bmatrix}
0 \\
0
\end{bmatrix} = \begin{bmatrix}
z' y \\
0
\end{bmatrix}.
\]
(1)

No simple rules for UPGs exist for creation of the above expressions with the Q matrix. The left-hand side of equation (1) can be presented in detail as:

\[
\begin{bmatrix}
z' Z + A^{-1} + \begin{bmatrix}
0 & 0 \\
0 & G^{-1} - A_{22}^{-1}
\end{bmatrix} - A^{-1} + \begin{bmatrix}
0 & 0 \\
0 & G^{-1} - A_{22}^{-1}
\end{bmatrix}
\end{bmatrix}
\begin{bmatrix}
0 \\
0
\end{bmatrix}
\]

\[
\begin{bmatrix}
-Q' A^{-1} + \begin{bmatrix}
0 & 0 \\
0 & G^{-1} - A_{22}^{-1}
\end{bmatrix} Q' A^{-1} - \begin{bmatrix}
0 & 0 \\
0 & G^{-1} - A_{22}^{-1}
\end{bmatrix}
\end{bmatrix}
\begin{bmatrix}
0 \\
0
\end{bmatrix}
\]

and then split in two:

\[
\begin{bmatrix}
z' Z + A^{-1} + \begin{bmatrix}
0 & 0 \\
0 & G^{-1} - A_{22}^{-1}
\end{bmatrix} - A^{-1} Q
\end{bmatrix}
\begin{bmatrix}
0 \\
0
\end{bmatrix}
\begin{bmatrix}
0 & 0 \\
0 & G^{-1} - A_{22}^{-1}
\end{bmatrix}
\begin{bmatrix}
0 \\
0
\end{bmatrix}
\]

\[
\begin{bmatrix}
-Q' A^{-1} + \begin{bmatrix}
0 & 0 \\
0 & G^{-1} - A_{22}^{-1}
\end{bmatrix} Q' A^{-1} - \begin{bmatrix}
0 & 0 \\
0 & G^{-1} - A_{22}^{-1}
\end{bmatrix}
\end{bmatrix}
\begin{bmatrix}
0 \\
0
\end{bmatrix}
\]

(2)

where Q2 is a matrix assigning genotyped animals to groups. When UPGs in ssGBLUP are considered only through transformed A*, only the first part is created, and the second part is ignored, which can lead to biases. The biases to UPGs will be small when the additional contributions due to UPG: \((G^{-1} - A_{22}^{-1})Q_2\) and \(Q_2'(G^{-1} - A_{22}^{-1})Q_2\) are relatively small. The biases will cancel out if \((G^{-1} - A_{22}^{-1})Q_2 = 0\).

Different approaches to account for UPGs in ssGBLUP

Add contributions

Adding the contributions is relatively easy and inexpensive because Q2 is relatively small and the products involving Q2 are relatively easy and inexpensive to compute using recursion. However, adding contributions requires changes to software and may be hard to implement with strategies where \(G^{-1}\) and \(A_{22}^{-1}\) are not created explicitly (Legarra & Durocq 2012; VanRaden 2012). Also, the evaluations may still be poor if the UPGs are defined poorly.

Use explicit groups

When the number of UPGs is small, an easy approach may be to use these groups explicitly in the model as ZQs. However, when the number of groups is large, such a change would require adding a large number of effects to the model, potentially greatly increasing computation time. Also, Q needs to be explicitly created, although it can be recursively computed from the pedigree file (Quaas 1988). The explicit groups may have an advantage in multiple-trait models where some traits are sparsely recorded. Fitting all groups to all traits can lead to very poor estimates of UPGs. Fitting UPGs explicitly allows different definitions of UPGs per trait, including removal of UPGs from traits with few phenotypic records or under no selection.

Ignore genomic contributions while potentially refining definitions of UPGs

When contributions due to the genomic part in UPGs in (2) are relatively small, ignoring the contributions may result in negligible changes to solutions of UPGs. Such a case is likely for large populations where only a small fraction of the population is genotyped, the distributions of groups within and outside of the genotyped subpopulations are not greatly different, and the groups are assigned carefully to avoid near confounding. Another case to ignore the contributions is when all UPG solutions change, but their estimable functions change very little (i.e. they are shifted by a constant term). Ignoring the contributions may be possible only after refining the UPG definitions, eliminating groups with few animals and merging groups that no longer have any effect on predictions of current animals. Other options are to...
treat UPG as random effects or even apply time series priors to obtain smooth UPG, e.g. as in Legarra et al. (2007).

An obvious although imperfect test to see whether the contributions can be ignored is similarity of UPG solutions of ssGLUP (with contributions ignored) and regular (pedigree-based) BLUP.

Remove UPGs from the model
When the goal of an evaluation is to predict the best EBVs of the youngest generation, removing UPGs from the model may have little effect on the accuracy of those EBVs. This is more likely in cases where UPGs are at least one to two generations removed from the youngest populations or when UPGs are assigned mechanically without examining the effects of possible confoundings among different UPGs. Removing UPGs from the model is an obvious choice when including UPGs decreases the accuracy of the evaluation in BLUP (Phocas & Laloe 2004), although EBVs may become biased.

Changes to relationship matrices
Modifying \( G, A_{22} \) or both may be possible, so that their contributions to UPGs cancel out or are sufficiently small to be ignored. Such modifications could additionally minimize biases of genomic EBVs (GE-BVs) that are present without UPGs. In general, the genomic and pedigree relationships should be matched. Normally, matching involves changes to \( G \) (VanRaden 2008; Vitezica et al. 2011), although similar effects may be obtained with changes to \( A \) (Christensen 2012). Which matrix to change could be a matter of convenience for ssGLUP (e.g. to ensure positive-definiteness of the matrices) provided that the difference \( G^{-1} - A_{22}^{-1} \) is optimal in some sense (e.g. for unbiased EBVs). The exact type of modifications depends on the structure of UPGs and is discussed in the following section.

Empirical refinement to ssGLUP
In some cases, biases to UPG could be excessive only because of deficiency in the \( H \) matrix. This matrix was developed under many assumptions, many of which may not hold in practice. Those assumptions include the same genetic parameters in the genotyped sample as in the complete population, the existence of complete data to account for selection bias and all additive variability explained by the genomic component. Several studies (Tsutsumi et al. 2011; Christensen et al. 2012; Harris et al. 2012) found that better accuracies and lower biases of GE-BVs can be achieved by fine-tuning \( z, \beta, \Omega, \varphi \) and \( \omega \) in \( H^{-1} \) parameterized as

\[
H^{-1} - A_{22}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & z(zG+\beta A_{22}+\varphi J)^{-1} - \omega A_{22}^{-1} \end{bmatrix}
\]

where \( J \) is a matrix of 1's. In general, \( z \approx 0.8, \beta = 1 - z, \varphi = 0 \) and \( \Omega = \omega = 1 \). Harris et al. (2012) used \( 1 = \omega \) as low as 0.3. Low values for \( \Omega \) and \( \omega \) decrease the effect of the genomic information on animals with phenotypic information while limiting the effect on EBVs of young animals (Misztal et al. 2010). Lower weights on genomic components also reduce the effect of unaccounted contributions due to UPGs. The value of \( \varphi \) can be chosen to reduce the bias of EBVs of genotyped animals as in Vitezica et al. (2011) or Chen et al. (2011).

Two animal effects
Assume that the additive variation can be partitioned as polygenic and genomic. Then, the model in this study will change to include two additive effects:

\[
y = Zu_p + Za + e,
\]

\( u_p = QS + a_p \) is a polygenic effect that includes the UPG effect, \( a_p \) is the other genetic effect that accounts for the genetic information, but ignores the UPG effect, \( var(a_p) = A_p\sigma^2_p \), \( var(a) = H(1 - \eta)\sigma^2_g \), and \( \eta \) is the fraction of the polygenic variance. In general, \( \eta \) could be slightly lower than the optimal \( \beta \) in the model above. For example, if \( x = 0.8 \) and \( \beta = 0.2 \) in a model without UPG, the new parameters could be \( \eta = 0.15, x = 0.94 \) and \( \beta = 0.06 \); the last parameter should be > 0 to guarantee positive-definite \( H \). In a study in pigs, Christensen et al. (2012) found minimal variations in ssGLUP accuracy when the fraction of polygenic variance varied from 5 to 30%. Therefore, the value of \( \eta \) between 0.1 and 0.25 may be acceptable. As \( s \) is a fixed effect, the solution to UPG effects should not depend on \( \eta \) directly. The use of two animal effects would require no changes to the model if two animal effects can be accommodated by software. However, the convergence rate when solving by iterative methods can be reduced especially for small \( \eta \).

UPGS and changes to relationship matrices
The UPGs can arise from several situations. Each situation will require a different modification to \( G \) or \( A_{22} \). Some of the situations and possible choices follow.

Missing parent(s) in single population
Assume a single population under selection with a base population from many generations ago. Assume that pedigree recording is incomplete and therefore some parents are not identified although they are in
the pedigree (e.g. in beef cattle or sheep). Also, assume that missing parents are replaced by UPGs that correspond to their expected generation number (for discrete generations) or as a function of a birth year of the animal. In such a case, $G$ is complete and correctly describes relationships among individuals (e.g. VanRaden 2008) regardless of pedigree completeness; however, $A$ is incomplete. The average relationship of the animals with missing parents in $A$ to the rest of the population may be 0 (if both parents are missing) or half the true relationship if one parent is missing.

At least two solutions exist in such a case. The first solution is to reconstruct pedigrees in $A$ based on $G$. However, such reconstruction (imputation) is possible only for individuals with a sufficient number of genotyped ancestors or descendants and will be difficult or impossible for non-genotyped animals. The second solution is to construct $A$ by fitting unknown parents as related; this assumes that animals from the same UPG are not unrelated, but that all animals share the same common average relationship, which is referred to as the inbreeding coefficient of the UPG and was initially suggested by VanRaden (1992) and refined by Colleau & Sargolzaei (2011). Suppose that animals with unknown parents born at time $t$ come from a population similar to their contemporaries at time $t$ rather than from a large base population mating at random. Average co-ancestry of their (unknown) parents, which is the inbreeding at generation $t$, should be the same for pedigreed or non-pedigreed individuals. An algorithm by VanRaden (1992) could recover most of the original inbreeding when the fraction of missing parents was $\leq 20\%$ (Luijten et al. 1999). An efficient algorithm for calculation of such inbreeding based on the recursive algorithm was developed by Aguilar & Misztal (2008). Such inbreeding could potentially be derived from $G$ (e.g. VanRaden 2008).

Missing parents from different breeds
Assume that UPGs are assigned based on breeds in a multi-breed population. In such a case, given some assumptions, $A$ is constructed properly (i.e. animals from different breeds are considered unrelated). If $G$ is constructed with breed differences ignored, animals from different breeds will have genomic covariances much different than 0 (Simeone et al. 2012). Therefore, $G$ needs to be rescaled to be compatible with $A$ (Harris & Johnson 2010; VanRaden et al. 2011; Harris et al. 2012). If UPGs reflect both breed differences as well as missing pedigrees within a breed, both $G$ and $A$ may have to be modified for compatibility. Correctly modelling the case of several breeds is important for genomic multi-breed evaluation even if the number of crossbreeds is small (Hayes et al. 2009; Ibañez-Escriche et al. 2009; Pryce et al. 2011).

Missing parents from external populations
In some cases, parents of animals from external populations of the same breed are modelled with UPGs. The external populations could be different lines of the same breed with those lines separated by long independent selection (e.g. independent pig or chicken lines) or could be from lines that share a large number of common males (e.g. national populations of dairy and beef). Assuming small predictivity across long separated lines (e.g. the Danish dairy population, which does not have strong genetic links with Swedish and Finnish populations (Brondum et al. 2011)), such lines can be treated as different populations or different breeds. External lines with common ancestors can be treated as the same population, an external population or a mixture of both.

Effect of incomplete pedigrees on biases and convergence rate
Incomplete pedigrees may cause biases in ssGBLUP even when UPGs are not in the model. Assume a single population under selection with all animals having complete and long pedigrees down to the same base population. When $G$ is scaled to match averages of $A_{22}$, EBVs are unbiased (Vieitez et al. 2011). When $G$ is scaled the same way, but pedigrees are incomplete, the elements of $G$ will be smaller, on average, than in $A_{22}$ for animals with long pedigrees, and larger, on average, for animals with short pedigrees. Based on findings of Chen et al. (2011), EBVs of animals with short (long) pedigrees will be biased upwards (downwards).

The problem of incomplete pedigrees can be addressed several ways. First, some pedigrees can be restored based on genotypes. Second, the additive relationships for animals with incomplete pedigrees can be adjusted to the oldest base generation as proposed by VanRaden (1992). The simplest approach is to reduce the number of generations in the pedigree. Then, the effect of missing parents for the oldest generations is eliminated.

Incomplete pedigrees can cause decreased convergence rate when ssGBLUP is solved by iterative methods. With incomplete pedigrees, the diagonal elements of $A_{22}$ will be larger for animals with longer pedigrees, which causes many elements of $(G^{-1} - A_{22})$ to be negative. Subsequently, $H^{-1}$ as computed can become close to non-positive definite. Commonly used
iterative methods such as Gauss-Seidel or preconditioned conjugate gradients (PCGs) require a semi-positive definite left-hand side. Truncation of pedigrees can drastically improve the convergence rate. S. Forni (personal communication) looked at convergence rate of ssGBLUP in a pig population when solving by PCG iteration. Initially, the convergence rate was very slow. When the number of generations of parents without phenotypes was reduced from >18 to 2, the convergence rate improved by 15 times; the realized accuracy for the last genotyped population was not affected.

Biases of EBVs for populations or the magnitude of convergence deceleration with incomplete pedigrees greatly depends on the population structure and the strategy for genotyping. When the training population includes a large number of progeny-tested sires (as in dairy cattle), the effect of the missing pedigree is small. This is because the contributions to $G^{-1}$ in $H^{-1}$ for progeny-tested animals are small compared with contributions from $A^{-1}$, and the effect of scaling of $G$ for young animals partially cancels out (Misztal et al. 2010).

Large differences between $G$ and $A_{22}$ also occur when identical twins or clones are genotyped. This problem is not directly connected to UPG, but is similar because $G$ contains extra information about relationships that is missing in $A$. Models might account for identicals by retaining only a single copy, or by making matrix $A$ singular to match the singular $G$.

Conclusions
When the model includes UPGs and evaluation is by ssGBLUP, UPG solutions may be biased if the UPGs are only considered in the creation of $A^*$, but not in the creation of genomic contributions. The bias may be small for large populations with few genotyped animals, but could be large in other populations. The bias can be removed by making appropriate contributions to mixed model equations, by treating UPGs explicitly in the model, by refinements to definitions of UPGs or by adding an additional animal effect to the model. Further research will determine whether the need for the contributions can be eliminated by appropriate modifications of $G$ and $A_{22}$. The modifications need to reflect different origins of UPGs: missing pedigrees in a closely selected population, multiple breeds, external lines or combinations of origins.

Acknowledgements
This study originated during IM stay at Université de Toulouse and would not be possible without IM stay at the University of New England. This project was supported by Agriculture and Food Research Initiative Competitive Grants no. 2009-55205-05665 and 2010-55205-20366 from the US Department of Agriculture’s National Institute of Food and Agriculture as well as by Holstein Association USA (Brattleboro, VT), BARD and PIC (Hendersonville, TN). Manuscript review by S. M. Hubbard of the Animal Improvement Programs Laboratory is gratefully acknowledged. The authors appreciate very useful comments by the two anonymous reviewers.

References


