# Combining genomic and classical information in national BLUP evaluations

Vincent Ducrocq, Z. Liu

# Combining Genomic and Classical Information in National BLUP Evaluations

***V. Ducrocq[1] and Z. Liu[2]***
*[1] UMR1313 INRA, Génétique Animale et Biologie Intégrative, 78 352 Jouy-en-Josas, France*
*[2] VIT, Heideweg 1, D-27283 Verden, Germany*
*vincent.ducrocq@jouy.inra.fr*

## Abstract

Blending genomic information with classical performances into a joint BLUP analysis has some appealing features, in particular its simplicity and its potential ability to account for genomic preselection of young sires. A simple approach consists in computing specific genomic equivalent daughters contributions and genomic equivalent daughter performances. Two cases are presented here, depending on the way genomic evaluations are performed: using prediction equations or BLUP with a genomic relationship matrix. It is shown through a small example that genomic EDC should be computed with caution to avoid double-counting, especially when closely related animals are genotyped. Otherwise, bias results and inflated reliabilities are obtained.

## 1. Introduction

Despite of the rapid development of genomic selection in dairy cattle, there is a consensus on the need to maintain national and international evaluations (see Report on the Interbull technical workshop in Uppsala 2009, Bulletin 39, Interbull website). At both levels, it is desirable to combine the existing sources of information: pedigree, performance data and genomic data. Methods have been proposed to do so directly using all the data available (Misztal *et al.*, 2009) but these are computationally very demanding and not feasible at the international level. hence, there is a need for a robust approach to perform this blending in a way that everybody can understand and accept.

For many years at Interbull data, the concepts of (Daughter) Yield Deviation ((D)YD) and Equivalent Daughter Contribution (EDC) have been used to combine information from different countries. This study is an attempt to extend these tools to include the new genomic information source.

## 2. Material & Methods

### 2.1 General strategy

We propose the following three-step approach:

i) from genomic evaluations, "genomic EDC" ($\Psi_i^G$) are computed in a way reflecting the actual amount of information coming from the knowledge of the genome in "BLUP terms". Such genomic EDC have been published; see for example by Van Raden *et al.*, 2009).

Two distinct genomic evaluation strategies will be considered here:

- when Direct Genomic Values (DGV) are computed using prediction equations established from a training population (say, estimating relevant SNP effects using a PLS or Bayes B approach);
- when DGV are computed from mixed model equations involving a genomic relationship matrix between genotyped individuals.

ii) Once all genomic EDC ($\Psi_i^G$) are available, genomic "Equivalent Daughter Performances" ($EDP_i^G$) are calculated multiplying the coefficient matrix of the Mixed Model Solutions (MME) established based on the $\Psi_i^G$ and the standard relationship matrix **A**, i.e. the left hand side of:

$$\left[\begin{pmatrix} \Psi_1^G & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Psi_m^G \end{pmatrix} + \alpha \mathbf{A}^{-1}\right]\begin{bmatrix} \vdots \\ \hat{a}_i^G \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \Psi_i^G \ EDP_i^G \\ \vdots \end{bmatrix} \quad (1)$$

with the vector of estimated DGV $\hat{a}_i^G$. As a result, the BLUP solutions based on these computed genomic EDP and EDC weights are fully consistent with the DGV coming from a "pure" genomic evaluation.

iii) Finally, these EDP are added to the (inter)national evaluation with their associated weight $\Psi_i^G$ as if they were own records (or equivalently, average daughter deviations) of the genotyped individuals.

Such a strategy, if successful, is an easy-to-understand approach which allows :
- to combine sources of information in BLUP in a natural way and to propagate genomic information to related ungenotyped animals;
- to better correct for biases caused by genomic pre-selection because then, data on which genomic selection is based is included and so selection is accounted for in the conventional BLUP evaluation (Patry and Ducrocq, 2009). Note that this supposes that genomic EDP and genomic EDC are available on all culled animals;
- to compute reasonable reliabilities of GEBV, inverting the augmented MME or using the Harris and Johnson's (1998) information source method, considering genomic information as own record of the genotyped animals.

Note that the existing BLUP software can be used.

### 2.2.1 Why a special computation for genomic EDC ?

Consider the pedigree in figure 1 to illustrate the need for a proper computation of $\Psi_i^G$.

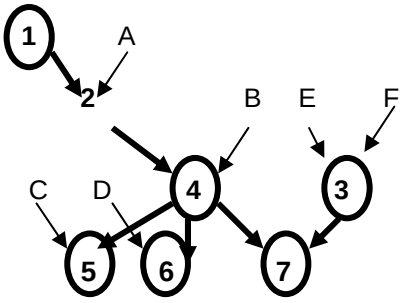


**Figure 1.** Pedigree used for the numerical example (circled individuals are the genotyped ones).

Only animals 1 to 7 are included in the analysis. They are all genotyped, except animal 2. Animals A to F are not genotyped and are ignored here because they do not bring any extra genomic information and are not linking any genotyped animals.

Consider that for each genotyped animal i, the amount of genomic information is equivalent to $n_{GS_i}$ =10 additional daughters. For a trait with heritability equal to 0.25, this leads to a reliability of a genotyped animal of $R_i = n_{GS_i} / (n_{GS_i} + \alpha) = 0.40$ with $\alpha$=15.

With a BLUP approach,, DGV reliabilities of all animals could be derived from system (1) as:

$$R_i = 1 - PEV_i / \sigma_a^2$$

where $PEV_i$ is approximated using the $i^{th}$ diagonal element of the inverse of:

$$\mathbf{B} = \left[ \begin{pmatrix} \Psi_1^G & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Psi_m^G \end{pmatrix} + \alpha \mathbf{A^{-1}} \right]$$

or equivalently:

$$R_i = i^{th} \text{ element of } (\mathbf{I} - \alpha \mathbf{B^{-1}}) \qquad (2)$$

Columns 2 and 3 (under "Iteration 1") of table 1 report the reliabilities for the numerical example computed either from (2) or using Harris and Johnson's approach, and assuming $\Psi_i^G = n_{GS_i}$. In such a case, it is seen that the actual reliabilities are overestimates of the expected value $R_i$ = 0.40, in particular for animal 4 (+27%) and its sons, e.g. animal 7 (+22%) which has both its parents genotyped. In fact, if the number of genotyped progeny of animal 4 increases, its reliability tends to 1, which does not make sense (the prediction equations do not have an $R^2$ of 1). This illustrates a problem of double-counting: when the complete genotype of one individual is known, the genotype of related animals is irrelevant as far as prediction equations are concerned, in contrast with what the MME assume.

In practice, a way to alleviate this inconsistency would be to take $\Psi_i^G < n_{GS_i}$ when i has genotyped related animals. A method to compute such $\Psi_i^G$ is proposed in the next section.

**Table 1.** Reliabilities of the animals from the numerical example, with naive EDC (columns 2 and 3) or with the strategy proposed (column 5). Column 4 indicates the final genomic EDC at convergence (starting value = 10).

| Animal | Iteration 1 | | Genomic EDC $\Psi_i^G$ | Final reliability (Harris & Johnson) |
|---|---|---|---|---|
| | True Reliability | Harris & Johnson | | |
| 1 | 0.411 | 0.414 | 9.56 | 0.400 |
| 2 | 0.203 | 0.224 | 0 | 0.191 |
| 3 | 0.439 | 0.438 | 8.78 | 0.400 |
| 4 | 0.508 | 0.508 | 5.33 | 0.400 |
| 5 | 0.445 | 0.445 | 8.51 | 0.400 |
| 6 | 0.445 | 0.445 | 8.51 | 0.400 |
| 7 | 0.487 | 0.487 | 6.61 | 0.400 |

### 2.3 Computation of $\Psi_i^G$ when DGV come from prediction equations

Let A be the training population used to construct the prediction equations, B the validation population and C the set of other genotyped animals (bulls and cows). A prediction equation based on A is used to obtain DGV for all genotyped animals in B and C.

The objective is to find $\Psi_i^G$ such that reliabilities computed using (2) – or equivalently, using Harris and Johnson's approximation – are consistent with the initial expectation of $R_i$ = 0.40 for genotyped animals. For this, a modification of Harris and Johnson (HJ)'s approach is proposed (see the appendix for details). In the first steps of the HJ algorithm, $\Psi_i^G = n_{GS_i}$ is used. Only the last step is modified: instead of combining the reliability due to own and progeny information $R_{oi+prog(i)}$ with reliability $R_{pedig(i)}$ from pedigree to obtain the final reliability $R_i$, both $R_i$ and $R_{pedig(i)}$ are considered known and used to derive first $R_{oi+prog(i)}$, then $R_{oi}$ (considering $R_{prog(i)}$ known ) and finally $\Psi_i^G = \frac{\alpha\, R_{oi}}{1\text{-}R_{oi}}$. However, $R_{pedig(i)}$ and $R_{prog(i)}$ were initially computed using an incorrect $\Psi_i^G = n_{GS_i}$. Hence, all steps are repeated with the new $\Psi_i^G$. In a few iterations (4 in our numerical example), convergence is reached.

The last two columns of table 1 show the final values of $\Psi_i^G$ for our numerical example and the corresponding final reliabilities, which are conform to expectation. The final values of $\Psi_i^G$ are smaller than $n_{GS}$, sometimes substantially (5.33 instead of 10 for animal 4).

This example was extended by adding 7 new sons of animal 4 (leading to 10 genotyped sons). Then, the initial overestimation of the reliability of all animals is larger than before, especially for animal 4 (+61%). With the proposed approach, the final values of $\Psi_4^G$ for animal (4) is 0 (and would be less than -5, without setting a constraint $\Psi_i^G > 0$ !). Despite of this, the final genomic reliability of animal (4) is still far too large! This is clearly undesirable. But overall the results are much more satisfying for the youngest animals: candidates to selection have correct reliabilities.

### 2.4 Computation of $EDP_i^G$

Knowing all $\Psi_i^G$ and $\hat{a}_i^G$, this is trivial. From (1):

$$\begin{bmatrix} \vdots \\ EDP_i^G \\ \vdots \end{bmatrix} = \begin{pmatrix} \Psi_1^G & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Psi_m^G \end{pmatrix}^{-1} \left[ \begin{pmatrix} \Psi_1^G & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Psi_m^G \end{pmatrix} + \alpha \mathbf{A}^{-1} \right] \begin{bmatrix} \vdots \\ \hat{a}_i^G \\ \vdots \end{bmatrix} \quad (3)$$

### 2.5 Combination of classical and genomic information

In the classical genetic evaluation, "own" genomic EDP for all genotyped animals (in B and C populations, not in A) are added assuming the following model:

$$EDP_i^G = cg_k + a_i^G + e \qquad (4)$$

with Var(e) = $\left(\Psi_i^G\right)^{-1}\sigma_e^2$ where $cg_k$ is a contemporary group for genotyped animals (for example, animals evaluated with the same prediction equation).

The solution of the resulting MME leads to combined (GEBV) estimates $\hat{a}_i$. The inverse of these MME or the information source method of Harris and Johnson can be used to derive realistic reliabilities for these GEBV.

### 2.6 Computation of $\Psi_i^G$ when DGV are computed using a genomic relationship matrix between genotyped individuals

In such a case, for animals that are both genotyped and phenotyped, GEBV are currently computed combining results from three evaluations (e.g., Van Raden *et al.*, 2009): EBV from a conventional evaluation $E_1$, DGV from genomic evaluation $E_2$ using a genomic relationship matrix and $EBV_{subset}$ from a conventional BLUP evaluation $E_3$ using only the genotyped animals. This special evaluation $E_3$ is important because it is used to correct for the difference in pedigree information between evaluations $E_1$ and $E_2$, as we know the genomic evaluation $E_2$ has much fewer ancestors included than the classical evaluation $E_1$. Selection index theory is used for this purpose.

The evaluation $E_3$ is problematic: because not all the performance records are included, the effects of selection on genetic trend and on reduction of genetic variance are not properly accounted for. This leads to biased GEBV as well as overestimated reliabilities. As an alternative, we suggested to simply consider the evaluation $E_3$ as a tool to calculate $\Psi_i^G$, as follows:

If a BLUP approach is to be chosen for integrating genomic information into conventional evaluation, the extra EDC contributed by genomic information only can be derived from the EDC used in the two evaluations $E_2$ and $E_3$.

Genomic evaluation under a marker model is equivalent to a genomic evaluation using genomic relationship matrix of genotyped animals. For evaluation $E_2$, genomic evaluation MME are:

$$\left[\begin{pmatrix}\Psi_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \Psi_m\end{pmatrix}+\alpha\mathbf{G}^{-1}\right]\begin{bmatrix}\vdots\\ \hat{a}_i^G\\ \vdots\end{bmatrix}=\begin{bmatrix}\vdots\\ \Delta_i\\ \vdots\end{bmatrix} \qquad (5)$$

where $\Psi_i$ is the EDC of genotyped animal i obtained from conventional evaluation $E_1$,
**G** is the realised genomic relationship matrix,
$\hat{a}_i^G$ is DGV of animal i, and
$\Delta_i$ is the right-hand side for animal i.

Let $\mathbf{C}=\left[\begin{pmatrix}\Psi_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \Psi_m\end{pmatrix}+\alpha\mathbf{G}^{-1}\right]$

then the reliability of DGV of animal i is obtained as:

$$R_i^{E_2} = i^{th} \text{ element of } (\mathbf{I} - \alpha\mathbf{C^{-1}}) \qquad (6)$$

The original MME system for the genotyped animals only, in the subset evaluation $E_3$, is:

$$\left[\begin{pmatrix}\Psi_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \Psi_m\end{pmatrix}+\alpha\mathbf{A}^{-1}\right]\begin{bmatrix}\vdots\\ \hat{a}_i^C\\ \vdots\end{bmatrix}=\begin{bmatrix}\vdots\\ \Delta_i\\ \vdots\end{bmatrix} \qquad (7)$$

then the reliability of the animal i is again of the form:

$$R_i^{E_3} = i^{th} \text{ element of } (\mathbf{I} - \alpha\mathbf{B^{-1}}) \qquad (8)$$

The additional EDC contributed by genomic information only can then be calculated by converting reliability gain, $R_i^{E_2} - R_i^{E_3}$, as:

$$\Psi_i^G = \alpha\frac{R_i^{E_2} - R_i^{E_3}}{1-(R_i^{E_2} - R_i^{E_3})} \qquad (9)$$

Note that this extra EDC were requested by Interbull for the genomic MACE test run of April 2009.

From here, the same approach as in section 2.4 above to generate $EDP_i^G$ can be applied to the genotyped animals (equation (3)).

## Conclusion

Post-processing of genomic DGV and classical genetic evaluations within a BLUP framework is conceptually not difficult and has several appealing benefits, such as the automatic correction of bias due to genomic preselection, the potential extension to international GMACE and the use of existing software. However, it has to be done with care. Otherwise, genomic reliabilities are overestimated and resulting GEBV may be biased, because of overconfidence or overweighing of DGV results. An evaluation jointly including genomic information and performance records is highly desirable, but may be too demanding computationally (as in Misztal *et al.*, 2009) or may require strategic changes (e.g., use of marker-assisted evaluation as in Ducrocq *et al.*, 2009, instead of pure genomic evaluation). As a simple (simplistic?) alternative, we propose an approach to attenuate this drawback. But thismay not be sufficient because BLUP implicitly assumes that genomic information on a close parent increases the amount of information available to compute DGV in the same way as performance records do. Further improvements should be considered, e.g. including a "residual" covariance between genotyped individuals.

## Acknowledgements

This study was motivated by and discussed within the Task Force on the role of genomic information in genetic evaluations. Comments from its members is gratefully acknowledged.

## References

Ducrocq, V., Fritz, S., Guillaume, F. & Boichard. 2009. French report on the use of genomic evaluation. *Interbull Bulletin* 39, 17-22.

Harris, B. & Johnson, D. 1998. Approximate reliabilities of genetic evaluations under an animal model. *J. Dairy Sci. 81,* 2723-2728.

Misztal, I., Legarra, A. & Aguilar, I. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information. *60th Annual Meeting EAAP. Book of Abstracts 15,* 298.

Patry C. & Ducrocq V. 2009. Bias due to selection. *Interbull Bulletin 39,* 77-82.

Van Raden P.M., *et al.* 2009. Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci. 92,*16–24.

## Appendix: An adaptation of the information source method (Harris and Johnson, 1998) to compute $\Psi_i^G$

The "Information source method" is based on one fundamental equation to compute the reliability $R^{x+y}$ of an evaluation combining two sources x and y of completely independent information. If $R^x$ and $R^y$ represent the reliability of each source, we have:

$$R^{x+y} = \frac{R^x + R^y - 2\,R^x R^y}{1 - R^x R^y} \qquad \text{(A1)}$$

which can also be written as:

$$R^x = \frac{R^{x+y} - R^y}{R^{x+y} R^y + 1 - 2R^y} \qquad \text{(A2)}$$

The key point here is that the final reliability $R_i$ of any genotyped animal is supposed to be known and equal to $R_i$. $R_i$ may depend for example on the prediction equation used. What we want is to "distribute" this reliability into three independent sources (pedigree, progeny and own=genomic ($R_{oi}$)) such that, using the own contribution $R_{oi}$ as basic ingredient in the Harris and Johnson algorithm, one gets back $R_i$ for genotyped animals. Therefore, we want to compute $R_{oi} = \frac{\Psi_i^G}{\Psi_i^G + \alpha}$ such that $\Psi_i^G = \frac{\alpha\, R_{oi}}{1\text{-}R_{oi}}$ is the equivalent daughter contribution of the genomic information of i. This will be done iteratively, initially considering $\Psi_i^G = n_{GS}$, any realistic starting value. The Harris and Johnson's approach consists of 4 steps. The first three will be applied at each iteration as in the initial version and the fourth one will be used to determine a new $\Psi_i^G$

i) for each animal i, from the youngest to the oldest: Compute the contribution $R_{oi}$ coming from the own (=genomic) performance of animal i and cumulate this contribution to its sire ($R_{prog(sire)}$) and dam ($R_{prog(dam)}$) reliability coming from progeny information. Here, $R_{prog(sire)}$ and $R_{prog(dam)}$ are initially 0. Equation (A1) above is used for example to combine the current $R_{prog(sire)}$ and the reliability of the sire coming from the performance of progeny i (=0.25 $R_{oi}$) into a new $R_{prog(sire)}$.

ii) At the end of step i), all $R_{prog(sire)}$ and $R_{prog(dam)}$ are known. They include only information from sons and daughters. To include information from grand progeny and further generations down, the 0.25 $R_{prog(sire\ or\ dam)}$ are cumulated into their own parents' one, again from the youngest to the oldest animal.

iii) At the same time progeny ($R_{prog(i)}$) and own ($R_{oi}$) contribution are combined into a reliability $R_{oi+prog(i)}$ again using equation (A1).

iv) Finally, the pedigree information is added, from the oldest animal to the youngest, in the following way: first, for each individual i, the pedigree information and the own + progeny information are made independent from each other by "subtracting" from $R_{sire}$ and $R_{dam}$ reliability of its parents the information coming from i (=0.25$R_{oi+prog(i)}$) using equation (A2). This gives $R_{sire}^{(-i)}$ and $R_{dam}^{(-i)}$. The pedigree information for i free of the information of i has reliability $R_{pedig(i)} = \frac{R_{sire}^{(-i)} + R_{dam}^{(-i)}}{4}$.

This is where the proposed algorithm differs from Harris and Johnson's original approach. In their case, they combine $R_{pedig(i)}$ and $R_{oi+prog(i)}$ using equation (A2) once more to get the final $R_i$.

In our case, for a genotyped animal, we know this final result: $R_i$. This $R_i$ is the combination of pedigree information free of i and the own (=genomic) + progeny information. Then we have two cases:

- if i is not genotyped, we proceed as before to get $R_i$ combining $R_{pedig(i)}$ and $R_{oi+prog(i)}$ using equation (A1).
- if i is genotyped, we know that $R_i = \Re$. We use equation (A2) to get $R_{oi+prog(i)}$ (removing the contribution $R_{pedig(i)}$). Then we use this same equation (A2) again to get $R_{oi}$ by removing the information coming from the progeny. This $R_{oi}$ is used to compute a new $\Psi_i^G$ specific to i.

At the end of these 4 steps, we get new $\Psi_i^G < n_{GS}$ for each genotyped animals. It will be seen that in some extreme cases, $\Psi_i^G < 0$. Of course, it must be bounded to 0 then.

So far, the $\Psi_i^G$ terms were computed assuming incorrect $\Psi_j^G$ for the progeny j of i (in particular $\Psi_j^G$ for the first round). So the whole steps 1-4 are repeated again, and again until the $\Psi_i^G$ do not move any more.