# Bayesian Approach to Integrate Molecular Data into Genetic Evaluations

***N. Gengler[1,2] and C. Verkenne[1]***
*[1] Animal Science Unit, Gembloux Agricultural University, B-5030 Gembloux, Belgium*
*[2] National Fund for Scientific Research, B-1000 Brussels, Belgium*

## 1. Introduction

Integration of molecular data into breeding value estimation is an important issue. Recently about 60.000 SNPs (Single Nucleotide Polymorphisms) became available for cattle (Van Tassell *et al.*, 2007) and simulation showed that genomic selection based on these SNPs could improve greatly the selection of young bulls (Meuwissen *et al.*, 2001). However, even if a selection based on molecular information is in theory very promising, it has proven to be very difficult to implement in real life situations, especially if genetic evaluations are needed for extremely large populations as for dairy cattle, and if very few animals are genotyped. Most current ideas are based on methods where first phenotypic information is resumed and then integrated as daughter yield deviations or deregressed proofs in simplified mixed models. However, very few contributions addressed the issue of reintegration of these molecular data into large breeding value estimation systems. Also molecular data are not used to estimate breeding values for all evaluated animals. In the present study we propose a Bayesian approach to address both issues. We present an illustration for a single gene; however this approach could be easily extended to many SNPs effects and can thus be considered as a first step leading to the integration of genomic selection into national breeding value estimation systems.

## 2. Material and Methods

### *2.1. Equivalent Mixed Inheritance Model*

A generic mixed inheritance model combining fixed QTL or single gene effects (**g**) and random polygenic (**u**) effects can be written as $\mathbf{y} = \mathbf{X\beta} + \mathbf{ZQg} + \mathbf{Zu} + \mathbf{e}$ where **y** is a vector of data, $\boldsymbol{\beta}$ a vector of fixed effects, **X** and **Z** are incidence matrices linking effects and **y,** and **Q** is a matrix linking QTL effects and animals. The basic assumptions can be written as $\mathrm{E}\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}$ and $\mathrm{var}\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}$. A very simple equivalent model is $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu}^* + \mathbf{e}$ where $\mathbf{u}^* = \mathbf{Qg} + \mathbf{u}$ and the basic assumptions are modified to $\mathrm{E}\begin{bmatrix}\mathbf{u}^*\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{Qg}\\ \mathbf{0}\end{bmatrix}$ and $\mathrm{var}\begin{bmatrix}\mathbf{u}^*\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}$. Following Quaas (1988) and using the same strategy he used to integrate genetic groups, the following alternative mixed model equations (**MME**) allow the joint estimation of $\boldsymbol{\beta}$, $\mathbf{u}^*$ and **g**:

$$\begin{bmatrix}\mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} & \mathbf{0}\\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z+G^{-1}} & \mathbf{-G^{-1}Q}\\ \mathbf{0} & \mathbf{-Q'G^{-1}} & \mathbf{Q'G^{-1}Q}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}^*\\ \hat{\mathbf{g}}\end{bmatrix} = \begin{bmatrix}\mathbf{X'Ry}\\ \mathbf{Z'Ry}\\ \mathbf{0}\end{bmatrix}$$

These MME can be easily modified to accommodate *a priori* known single gene or QTL effects $\tilde{\mathbf{g}}$ estimation of $\boldsymbol{\beta}$ and $\mathbf{u}^*$:

$$\begin{bmatrix}\mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z}\\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z+G^{-1}}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}^*\end{bmatrix} = \begin{bmatrix}\mathbf{X'R^{-1}y}\\ \mathbf{Z'R^{-1}y+G^{-1}Q\tilde{g}}\end{bmatrix}$$

Final equations are similar to Bayesian prediction as described by Henderson (1984) and e.g., used by Schaeffer and Jamrozik (1996), where *a priori* knowledge contributes to the estimation of random effects. In a genomic selection setting, $\mathbf{Q\tilde{g}}$ could represent the sum of SNPs effects which one might call genomic EBVs (Estimated Breeding Values).

### *2.2. Integration into PCG Solver*

Solving of modified MME can be done very efficiently using a standard PCG solver (e.g.,

Stranden and Lidauer, 1998) however rearranging to adjust the computed differences between the observed right hand sides and the products of the original coefficient matrix **C** and the vector of current solutions **a** by adding $\mathbf{G}^{-1}\mathbf{Q}\tilde{\mathbf{g}}$ which is equivalent to $\mathbf{G}^{-1}\mathrm{E}(\mathbf{u}^*)$. If differences in right hand sides are updated during PCG iterations, this has to be done only once and modifications in the code consist in one single operation.

### *2.3. Comparison with Other Models – Quantitative Example*

*2.3.1. Data*

Data provided by the Walloon Breeding Association (AWE) for the routine evaluation for milk production of January 2007 were used and included 13,992,889 test-day (**TD**) records for 778,923 dairy and dual-purpose cows in production. The pedigree file contained 1,429,939 animals (cows with production records and ancestors). A total of 1417 dual purpose Belgian Blue (**DP-BBB**) individuals in the pedigree, from which 1183 cows with records, were genotyped for the myostatin (**mh**) gene.

*2.3.2. Comparison of models*

The model used for the routine evaluation for milk production in the Walloon Region (Auvray and Gengler, 2002) is the following multi-trait multi-lactation (3 traits x 3 lactations) model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=0}^{2}\mathbf{W}_i\mathbf{h}_i + \sum_{i=0}^{2}\mathbf{Z}_i^*\mathbf{p}_i + \sum_{i=0}^{2}\mathbf{Z}_i\mathbf{u}_i + \mathbf{e}$$

(Model 1)

where **y** is a vector of milk, fat and protein test-day records, **β** is a vector of fixed effects, $\mathbf{h}_i$ is a vector of herd × period of calving random regression coefficients for polynomials of order i, $\mathbf{p}_i$ is a vector of permanent environmental random regression coefficients for polynomials of order i, $\mathbf{u}_i$ is a vector of random polygenic additive effects for polynomials of order i, **e** is a vector of random residuals, **X, W, Z** and $\mathbf{Z}^*$ are incidence matrices linking **y** and the different effects. (Co)variance components were the same as those used during routine genetic evaluation, which are based on those obtained by Gengler *et al.* (1999).

Joint estimation of single gene and of the other effects is theoretically ideal as long as all genotypes are known. Therefore, first of all, all genotypes, known and estimated with the method described Gengler *et al.* (2007), were used in the evaluation. As we found out in a preliminary research that the dominance effect was not significant (Table 1), this effect was removed from the model and only one additional regression, on gene content (the number of copies of the mh allele for each animal), was introduced in Model 1 to give the following model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_0\begin{bmatrix}\mathbf{q}\\ \hat{\mathbf{q}}\end{bmatrix}\mathrm{g_s} + \sum_{i=0}^{2}\mathbf{W}_i\mathbf{h}_i + \sum_{i=0}^{2}\mathbf{Z}_i^*\mathbf{p}_i + \sum_{i=0}^{2}\mathbf{Z}_i\mathbf{u}_i + \mathbf{e}$$

(Model 2)

where **q** is a vector of known and $\hat{\mathbf{q}}$ is a vector of estimated gene content, $\mathrm{g_s}$ is the allelic substitution effect and $\mathbf{Z}_0$ is the incidence matrix linking **y** to animals, which is the same as the incidence matrix for the constant polynomial. In order to increase the proportion of known genotypes, we applied the following rules: 1) All meat BBB (M-BBB) sires used in Artificial Insemination and born after 1985 (830 individuals) were supposed mh/mh as the mutated mh allele was already fixed for this breed at that time. 2) All non-BBB animals (659,971 individuals) were supposed +/+ as the probability that these animals were carrying the mh gene was considered very small. A total of 662,218 individuals were supposed to have a known genotype. For the remaining 768,291 animals, gene contents were estimated using the method described by Gengler *et al*. (2007), with some improvement based on genetic groups, according to the breed, the herd-book type of the animal, its breeding activity and its birth date.

Model 3 was the equivalent mixed inheritance model using Bayesian prediction described above. This model was developed by writing Model 1 differently:

$$\mathbf{y} = \mathbf{X\beta} + \sum_{i=0}^{2} \mathbf{W}_i \mathbf{h}_i + \sum_{i=0}^{2} \mathbf{Z}_i^* \mathbf{p}_i + \mathbf{Z}_0 \mathbf{u}_0^* + \sum_{i=1}^{2} \mathbf{Z}_i \mathbf{u}_i + \mathbf{e}$$

(Model 3)

Difference with Model 1 lies in the expectations on $\mathbf{u}_0^*$ $E[\mathbf{u}_0^*] = \begin{bmatrix} \mathbf{q} \\ \mathbf{0} \end{bmatrix} \tilde{g}_s$ where $\tilde{g}_s$ represents the allele substitution effect (+ allele replaced by mh allele) that is considered known *a priori*. For animals with known genotypes **q** is a vector of known gene content. Non genotyped animals were assigned a zero expectation. The general hypothesis is that estimates of $\tilde{g}_s$ have been provided by *a priori* research. In the case of the use of SNPs the expectations on $\mathbf{u}_0^*$ could be the sum of SNPs effects obtained from genomic selection. In this illustration, required estimates of allele substitution effects were obtained by a preliminary study using a model with the same effects as in Model 1 but with additional genotype class effects which allowed the estimation of allelic substitution effect based on genotyped animals only. Sampling errors for coefficients were estimated using mixed model conjugate gradient normal equations as described by Croquet *et al.* (2006). Significance of allele substitution effects was tested using an approximate t-test with N-rank(X) degrees of freedom, where N is the number of test-day records.

**Table 1.** Allelic substitution (α), dominance effects (d) and sampling error (SE) in kg for each lactation (305 days) for milk, fat and protein used in the illustration.

| Trait | | Lactation | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | | | 2 | | | 3 | | |
| | | Effect | SE | t-value | Effect | SE | t-value | Effect | SE | t-value |
| Milk | α | -190.4 | 39.1 | 4.87*** | -207.0 | 50.1 | 4.13*** | -172.7 | 65.5 | 2.64** |
| | d | 40.2 | 54.1 | 0.74NS | -27.6 | 70.8 | 0.39NS | 91.2 | 90.6 | 1.01NS |
| Fat | α | -7.5 | 1.63 | 4.60*** | -9.2 | 2.16 | 4.26*** | -7.2 | 2.90 | 2.48* |
| | d | 2.18 | 2.25 | 0.97NS | -1.82 | 3.02 | 0.60NS | 1.02 | 8.47 | 0.12NS |
| Protein | α | -6.2 | 1.17 | 5.30*** | -6.7 | 1.65 | 4.06*** | -5.9 | 2.03 | 2.91** |
| | d | 2.21 | 1.65 | 1.34NS | 0.04 | 2.20 | 0.02NS | 1.60 | 2.80 | 0.57NS |

NS : non significant, *: P > 0.975, **: P > 0.995, ***: P > 0.9995

Model 4 differed from Model 3 in the expectations on $\mathbf{u}_0^*$ $E[\mathbf{u}_0^*] = \begin{bmatrix} \mathbf{q} \\ \hat{\mathbf{q}} \end{bmatrix} \tilde{g}_s$ where $\hat{\mathbf{q}}$ is a vector of estimated gene content.

The four models were compared according to their predictive ability for breeding values and for the nine production traits. Using the production data set, a subset was created by removing all TD records from 2003 to 2007. This partial subset (p) was then analysed with the four models. For each deleted TD record of genotyped animals without records before 2003 and each model, expectations of performances $\hat{\mathbf{y}}_p$ were computed as the sum of the solutions, weighted by the respective regression coefficients. Values were compared to observed yields $\mathbf{y}_f$ in the full dataset using correlations and mean square errors (MSE).

For each animal without records before 2003, expectations of total breeding values (polygenic effects and if necessary single gene effect added) were computed for 305-d yields based on partial data, and were compared to equivalent breeding values obtained from the full data set. Comparisons were based on correlations and prediction error variances (PEV), defined as variance of the differences between both estimates.

## 3. Results and Discussion

Results illustrating the prediction of TD yields are shown in Table 2. Correlations and mean square errors were very similar for all models. This means that the prediction ability of the different models were similar. In terms of correlations there was a slight advantage for

Models 3 and 4. For MSE Models 1 and 4 showed the best results. In general these results were expected given the general properties of linear models, which minimize residual variances and maximize correlations between observed and predicted yields.

**Table 2.** Average correlations and mean square errors (MSE) computed between predicted and observed TD yields for the three lactations.

| Model | Correlation | | | MSE | | |
|---|---|---|---|---|---|---|
| | Milk | Fat | Protein | Milk | Fat | Protein |
| (1) | 0.79 | 0.72 | 0.77 | 1112 | 199 | 107 |
| (2) | 0.79 | 0.72 | 0.77 | 1162 | 202 | 110 |
| (3) | 0.80 | 0.73 | 0.78 | 1192 | 207 | 114 |
| (4) | 0.80 | 0.73 | 0.78 | 1114 | 196 | 106 |

**Table 3.** Correlations and Prediction Error Variances (PEV) computed between breeding values estimated from the partial and full production data sets.

| Model | Correlation | | | PEV (x $10^6$) | | |
|---|---|---|---|---|---|---|
| | Milk | Fat | Protein | Milk | Fat | Protein |
| (1) | 0.60 | 0.60 | 0.60 | 6.74 | 1.21 | 0.67 |
| (2) | 0.56 | 0.54 | 0.54 | 6.72 | 1.21 | 0.67 |
| (3) | 0.77 | 0.77 | 0.78 | 5.65 | 1.02 | 0.56 |
| (4) | 0.76 | 0.75 | 0.76 | 6.02 | 1.10 | 0.60 |

Results illustrating the prediction of breeding values are shown in Table 3. For correlation, Models 3 and 4 were clearly better (Table 3), an expected result because these models included influence of mh on total breeding values. PEV were larger for Models 1 and 2 compared to 3 and 4. Behaviour of Model 2 was disappointing. The most likely reason was that the joint estimation using estimated gene content and population wide estimated substitution effects introduced a systematic bias. Models 3 and 4, using the Bayesian integration method, were the best. There was a slight advantage for Model 3 in PEV, most likely because for Model 4, use of estimated gene content was deteriorating the prediction ability of EBVs. However, Model 4 behaved better when its ability to predict future yields was considered.

## 4. Conclusions and Implications

Bayesian prediction was only demonstrated in this study. It proved functional and even superior to a mixed inheritance model, at least when large numbers of genotypes were unknown. However, the method is much more versatile and has a lot of potential if one wants to integrate genomic EBVs in the regular genetic evaluation systems by considering them known *a priori*. There are several hidden implications. First, by considering different expectations for animal EBV according to the known molecular data, we do no longer try to model polygenic background and known genes, markers or SNPs, separately. Bayesian prediction provides one single EBV which could be called combined EBV. The second implication is that integration of molecular data will most likely need prediction of genomic EBVs also for non-genotyped animals (our Model 4). This prediction might be problematic for single gene effects because of the lack of normality. However, genomic EBV obtained from large numbers of SNPs effects can easily be predicted for non-genotyped individuals using standard methods (e.g., selection index).


## Acknowledgements

Nicolas Gengler, who is Research Associate of the National Fund for Scientific Research (Brussels, Belgium), acknowledges his support. The authors gratefully acknowledge the financial support of the Ministry Agriculture of the Walloon Region of Belgium (MRW-DGA) (Namur, Belgium). Additional support was provided through grants 2.4507.02F (2) and F.4552.05 of the National Fund for Scientific Research.

## References


Auvray, B. & Gengler, N. 2002. Feasibility of a Walloon test-day model and study of its potential as tool for selection and management. *Interbull Bulletin 29,* 123-127.

Croquet, C., Mayeres, P., Gillon, A., Vanderick, S. & Gengler, N. 2006. Inbreeding depression for global and partial economic indexes, production, type, and functional traits. *J. Dairy Sci. 89,* 2257–2267.

Gengler, N., Mayeres, P. & Szydlowski, M. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal 1,* 21-28.

Henderson, C.R. 1984. *Linear models in animal breeding.* University of Guelph Press, Guelph, Canada.

Liu, A., Reinhardt, F., Szyda, J., Thomsen, H. & Reents, R. 2004. A marker assisted genetic evaluation system for dairy cattle using a random QTL model. *Interbull Bulletin 32,* 170 -174.

Meuwissen, T., Hayes, B. & Goddard, M. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics 157,* 1819-1829.

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

Schaeffer, L.R. & Jamrozik, J. 1996. Multiple-trait prediction of lactation yields for dairy cows. *J. Dairy Sci. 79,* 2044-2055.

Stranden, I. & Lidauer, M. 1999. Solving large mixed linear models using preconditioned conjugate gradient iteration. *J. Dairy Sci. 82,* 2779-2787.

Van Tassell, C., Matukumalli, L., Taylor, C., Smith, T., Sonstegard, T., Schnabel, R., De Silva, M., Wiggans, G., Liu, G., Moore S. & Taylor, J. 2007. Construction and application of a bovine high-density SNP assay. *2007 Joint Annual Meeting,* San Antonio, Texas, July 8-12, 2007.