Proposal for Computing Genomic EBV

L. R. Schaeffer
Centre for Genetic Improvement of Livestock
Department of Animal & Poultry Science
University of Guelph, Guelph, ON, Canada N1G 2W1

Revised: March 23, 2009

1 Introduction

A brief history of developments in using SNPs to estimate breeding values of animals will be the preface for this proposal.

1.1 Using SNP Haplotypes

MEUWISSEN, HAYES, and GODDARD (2001) proposed methods of predicting total genetic value using a genome-wide dense marker map from a limited number of phenotypic records using marker haplotypes. Simultaneous estimation of the effects of marker haplotypes was conducted, and then the sum of those effects over the entire genome yields the genomic EBV of the animal. The problems with this approach are that the sequence of markers needs to be known, and secondly, the inheritance of marker haplotypes from the sire needs to be determined. Positions of markers are not known exactly in all cases. Determination of the haplotype is also not clear in all situations depending on the genotypes of the sire and dam, which may or may not be available.

1.2 Using SNP Genotypes

To overcome the problems with haplotypes, SNP genotypes were used directly, and the simultaneous estimation of the effects of many thousands of SNPs from less than a thousand genotyped animals was applied. The locations of the SNPs did not matter, except that they be distributed evenly across the genome and that they be within 1 centiMorgan of QTLs. The inheritance from parent to progeny did not need to be known. The problem was that there were many thousands of unknowns to estimate from data on relatively few animals, i.e. overparameterization of the model. The methodology of estimation was also being refined from least square methods to several Bayesian methods.

1.3 Better Relationships

Stranden and Garrick(2009) and VanRaden(2008) showed that the thousands of SNPs could be used to derive a better, enhanced additive genetic relationship matrix among the animals that were genotyped. Let $\mathbf{A}^{++}$ be the improved relationship matrix. The SNPs indicated that some

animals were more related to each other than the values in the **A** matrix would normally predict, and others were less related (i.e. fewer SNP genotypes in common). The resulting genomic EBV (GEBV) would be more accurate than EBV based only on data (DEBV). While the inverse of **A** is easily calculated and used to obtain DEBV, the inverse of $\mathbf{A}^{++}$ is not possible to calculate for large numbers of animals. Methods have been proposed that do not require the inverse of $\mathbf{A}^{++}$, but this matrix is very dense compared to the inverse of **A**. This means that most elements (99%) in the matrix are not equal to zero. In comparison $\mathbf{A}^{-1}$ has only about 5% non-zero elements. Therefore, any multiplications using $\mathbf{A}^{++}$ will take a long time to calculate, if they can be computed at all.

1.4 Three Pronged Approach

The First Prong is the calculation of DEBV in the usual manner, ignoring all SNP information. The Second Prong is the calculation of GEBV on genotyped animals only. The Third Prong is the merger of DEBV and GEBV into a combined EBV (CEBV). This is simple and quick. The Third Prong can be achieved in a number of different ways. An alternative is to leave DEBV and GEBV as separate entities and to present both to the public. A second alternative is to only show the CEBV. Lastly, all three figures could be presented. If all animals were genotyped, then GEBV would be the only numbers needed.

2 Proposal

2.1 Are All SNPs Needed?

Hayes and Goddard (2001) showed that there are many genes of small effect (pigs and dairy cattle), and few genes with large effects. Also, the few genes of large effects account for the majority of genetic variation in a trait. Their figure showed that 50 genes accounted for 99% of the variation in dairy cattle, and about 95% in pigs. The question is whether all 50,000 SNPs are needed to get accurate GEBV. This proposal is based on the premise that only a small number of SNPs (say 200, or maybe up to 2000) are needed. A problem is that each trait to be evaluated will require a different set of 200 SNPs. Picking the best 200 SNPs for a trait will not take very much time. Because multiple trait systems are used for production and fertility traits, having different SNPs for each trait could complicate the programs for genetic evaluation, but this is not impossible to overcome.

Currently, there are 3000 or so bulls genotyped for 50K SNPs. Using these bulls and their EBVs, analyses would be run to determine the best 200 SNPs. At the same time there will be estimates of the regressions on these 200 SNP genotypes. These estimates can be used as prior information into the later analyses, much as in the MTP system for predicting lactation yields in dairy cattle milk recording.

2.2 Simulation Study

A population of 10,000 individuals was generated. For each individual 5,000 independent SNPs were simulated. The effects of each gene were simulated from a Poisson distribution, such that the heritability of the overall trait was 0.25. Five generations of random selection and matings were conducted to give the 10,000 individuals. All animals of both sexes were observed for the same trait. The animal genetic merit was the sum of the genotypic effects of all 5,000 loci. Observations were created by adding a random residual effect to each animal's genetic merit.

The data were analyzed by a simple animal model with an overall mean as the only fixed factor. The **A** matrix was used in the analysis. The residual variance was estimated from this model.

Data were analyzed by a second model in which the regression on a single loci genotype was included along with the animal additive genetic effect. All 5,000 loci were put in the model, one at a time, and the locus giving the smallest residual variance was determined. The other loci were ordered by the residual variance given by each (lowest to highest).

The model was augmented by keeping the best locus from the previous runs, and adding a second locus to the model. All 4,999 remaining loci were added to the model, one at a time, until the locus giving the smallest residual variance was found.

This process was repeated until 70 loci had been added to the model. The decreases in residual variance are presented in Table 1.

Table 1.
Residual Variance After Addition of Another Locus To Model.

No. Loci	Residual Var.	No. Loci	Residual Var.
0	763.22	15	726.12
1	758.14	20	716.25
2	755.55	25	707.38
3	753.02	30	698.71
4	750.63	35	690.79
5	748.12	40	683.36
6	745.69	45	675.61
7	743.22	50	668.72
8	741.09	55	661.90
9	738.96	60	655.45
10	736.87	70	643.33

With 20 loci added to the model, the reduction in residual variance was 6.15%, and with 70 loci, the reduction was 15.71%. The decreases in residual variance are becoming smaller, but should continue up to 200 loci. This is the way that the SNPs are proposed to be chosen for inclusion into the model. A better, faster search strategy is needed to find the minimal set of SNPs.

The total genetic merit of an animal would be the sum of the animal additive polygenic

solution plus the regressions on the animal's genotypes for the selected loci in the model. An example will follow.

2.3 Polygenic SNP Model

The proposed system will require many females to be genotyped, and nearly all males. Assume a single trait observed on over a million females. Initially there would be only a few hundred females genotyped, but this should be increased to 10,000 or more over time. The 200 SNPs that influence the trait are assumed to have been determined from analyses of bull proofs and bull genotypes using a model similar to that which will be described here.

Let $\mathbf{y}$ be the observations on individual cows. The animal additive genetic effects, a_i, will be partitioned into two parts,

$$a_i \;=\; p_i + \sum_{k=1}^{m} c_k g_{ki},$$

where p_i is a "polygenic" portion of the genetic merit of animal i, g_{ki} is the genotype (1, 2, or 3) for the k^{th} SNP of animal i, and c_k are the regression coefficients to be estimated, which are constant over all animals in the population. The number of SNPs included was m, assumed to be 200 or fewer loci. The covariance matrix of p_i is assumed to be $\mathbf{A}\sigma_p^2$. Also, the regressions are taken to be random (as in the Bayes methods) with covariance matrix equal to $\mathbf{I}\sigma_c^2$. In practice, this variance would need to be estimated, or a separate variance could be estimated for each SNP locus in the model, as in the Bayes B methods. The polygenic and SNP genetic parts are assumed to be independent of each other, and σ_p^2 reflects the remaining polygenic variance after accounting for the 200 SNP genotypes. The $\mathbf{A}$ matrix is assumed to be the approximately appropriate correct covariance matrix for the remaining genetic effects. The model equation is

$$y_i = \mu + \sum_{k=1}^{m} (c_k g_{ki}) + p_i + e_i.$$

2.4 Numerical Example

Assume just 4 SNPs for this example data. The animals and their genotypes are given in Table 2.

Table 2.
Example Data for Proposal.

Animal	Sire	Dam	Genotypes				Obs.
1			1	2	3	2	
2			2	2	1	3	
3			3	1	2	1	
4							
5			3	2	1	3	
6	1	5	2	1	2	2	135
7	1	4	1	3	3	1	91
8	2	4	1	2	2	2	28
9	2	6	3	2	1	3	153
10	3	5					90
11	3	6					74
12	3	7	2	2	3	1	100
13	1	8					71
14	2	6	2	1	2	2	147
15	3	9	3	1	1	2	98

Note that some animals do not have genotypes, and some animals do not have observations, or both. For animals with genotypes the residual variance for those records will be $0.85(\sigma_e^2)$, and for animals without genotypes the residual variance for their records will be σ_e^2, assuming that the four genotypes account for 15% of a reduction in residual variance over a simple animal model.

The first step is to predict the missing SNP genotypes for animals not actually genotyped. Animals 10, 11, and 13 can be easily predicted because the genotypes of both parents are known. For animal 10, for example, the genotypes should be 3, 1.5, 1.5, and 2, respectively, or an average of the genotype values of the parents. Animals 11 and 13 are determined in the same manner giving 2.5, 1, 2, 1.5 for animal 11, and 1, 2, 2.5, and 2 for animal 13. The genotype of animal 4 must be deduced from its progeny, animals 7 and 8. Looking at animals 4, 7, and 1, then animal 4 contributed alleles 1, 2, 2, and unknown to the genotype of progeny 7. For progeny 8, animal 4 contributed alleles 1, unknown, 2, and 1. The most likely genotypes for animal 4 are then 1, 2, 2, and 2. If there is an animal with unknown genotypes, and without parents or progeny having genotypes, then the average genotype values of animals that have been genotyped, can be used, which are 2.03, 1.70, 1.93, and 1.97, respectively.

2.4.1 Setting Up Mixed Model Equations

Only the genotypes (actual and predicted) of the animals with records go into the data analysis. In dairy cattle, this will be the cows or females of the population. Thus, genotypes on cows will be crucial for this proposal to work well in the long run.

The **X** matrix for the overall mean in the model is a column vector of 10 ones. The **Z** matrix

is divided into two components, one for the regressions on SNP genotypes, i.e.

$$\mathbf{Z}_1 = \begin{pmatrix} 2 & 1 & 2 & 2 \\ 1 & 3 & 3 & 1 \\ 1 & 2 & 2 & 2 \\ 3 & 2 & 1 & 3 \\ 3 & 1.5 & 1.5 & 2 \\ 2.5 & 1 & 2 & 1.5 \\ 2 & 2 & 3 & 1 \\ 1 & 2 & 2.5 & 2 \\ 2 & 1 & 2 & 2 \\ 3 & 1 & 1 & 2 \end{pmatrix},$$

and the other part for the animal polygenic effects, $\mathbf{Z}_2$ matrix is $(\mathbf{0}_{10x5} \;\; \mathbf{I}_{10x10})$, and the residual covariance matrix is diagonal,

$$\mathbf{R} \;=\; \text{diag}\Big(\; .85 \quad .85 \quad .85 \quad .85 \quad 1 \quad 1 \quad .85 \quad 1 \quad .85 \quad .85 \;\Big).$$

The relationship matrix inverse is

$$\mathbf{A}^{-1} \;=\; \frac{1}{2}\begin{pmatrix} 5 & 0 & 0 & 1 & 1 & -2 & -2 & 1 & 0 & 0 & 0 & 0 & -2 & 0 & 0 \\ 0 & 5 & 0 & 1 & 0 & 2 & 0 & -2 & -2 & 0 & 0 & 0 & 0 & -2 & 0 \\ 0 & 0 & 6 & 0 & 1 & 1 & 1 & 0 & 1 & -2 & -2 & -2 & 0 & 0 & -2 \\ 1 & 1 & 0 & 4 & 0 & 0 & -2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 4 & -2 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 \\ -2 & 2 & 1 & 0 & -2 & 7 & 0 & 0 & -2 & 0 & -2 & 0 & 0 & -2 & 0 \\ -2 & 0 & 1 & -2 & 0 & 0 & 5 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 0 \\ 1 & -2 & 0 & -2 & 0 & 0 & 0 & 5 & 0 & 0 & 0 & 0 & -2 & 0 & 0 \\ 0 & -2 & 1 & 0 & 0 & -2 & 0 & 0 & 5 & 0 & 0 & 0 & 0 & 0 & -2 \\ 0 & 0 & -2 & 0 & -2 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 \\ -2 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & -2 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 4 \end{pmatrix}.$$

The mixed model equations are

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_2 \\ \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_1 + \mathbf{I}(50) & \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_2 \\ \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{Z}_2 + \mathbf{A}^{-1}\alpha \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{\mathbf{c}} \\ \hat{\mathbf{p}} \end{pmatrix} \;=\; \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix},$$

where $\alpha = 3$ for this example.

2.4.2 Solutions and GEBV

Solutions for two models will be compared. The first model is a simple animal model with only a mean and additive genetic effect in the model. The residual variances were assumed the same for all observations. Genotype information was not included. The second model is the proposed model. The solutions, in table form, were

Effect	Simple Animal Model	Proposed Model	Genomic Part	Combined EBV
μ	96.34	93.61		
c_1		1.85	1.46	
c_2		-0.60	-0.99	
c_3		-0.66	-1.05	
c_4		0.97	0.58	
a_1	5.43	6.77	-2.50	4.26
a_2	2.24	1.99	1.62	3.61
a_3	-4.52	-5.40	1.87	-3.53
a_4	-10.42	-10.77	-1.46	-12.22
a_5	7.27	7.41	3.08	10.49
a_6	14.72	15.90	0.99	16.94
a_7	-2.46	-1.38	-4.07	-5.45
a_8	-14.54	-15.77	-1.46	-17.23
a_9	15.14	15.68	3.08	18.76
a_{10}	0.27	-0.46	2.47	2.02
a_{11}	1.18	1.13	1.43	2.55
a_{12}	-2.47	-2.03	-1.62	-3.67
a_{13}	-7.52	-7.23	-1.98	-9.21
a_{14}	14.51	15.64	0.99	16.63
a_{15}	4.79	3.99	3.49	7.49

The regression coefficient estimates were made to sum to 0 by subtracting the simple mean of all regression coefficient estimates. These are shown under "Genomic Part" in the table above. The genetic contribution of the SNP genotypes for each animal is shown under "Genomic Part" for the animals. The contribution is the animal's genotype times the corresponding adjusted regression coefficients. For animal 1 for example,

$$\begin{aligned} \text{part}_1 &= 1.46(1) - 0.99(2) - 1.05(3) + 0.58(2) \\ &= -2.50. \end{aligned}$$

Because genotypes were deduced for all animals that were not genotyped, the SNP genetic contribution could be calculated for all animals. The accuracy of those contributions needs to be determined, given that a genotype may be a most probable genotype rather than an actual genotype. Genomic EBV are the sum of the polygenic portion and the "Genomic Part". Together they are labelled as "Combined EBV" in the table above, i.e. the last column. The combined or CEBV are better for ranking animals. The solutions from the simple animal model were used to predict $\mathbf{y}$, then $\hat{\mathbf{e}} = (\hat{\mathbf{y}} - \mathbf{y})$ was calculated and the sum of squares of residuals, (SSR), was computed. The same was computed for the proposed model. SSR(Animal model) = 116,083, and SSR(Proposed model) = 6,665. The proposed model fit the data better, but there were only 10 observations.

2.4.3 Young Animals

In practice, a young calf (animal 16) will be genotyped when it is born and a CEBV will be requested. Suppose the young calf was a progeny of animals 12 and 15 (from the example), and that the calf's genotypes were 3, 1, 2, and 2, respectively. The parent average of the polygenic part is $0.5(-2.03 + 3.99) = 0.98$, and the "Genomic Part" is

$$\begin{aligned} \text{part}_{16} &= 1.46(3) - 0.99(1) - 1.05(2) + 0.58(2) \\ &= 2.45, \end{aligned}$$

then the combined EBV would be 3.43.

2.4.4 Reliabilities

Let $\mathbf{C}_c$ represent the inverse elements of the mixed model coefficient matrix corresponding to the regressions on SNP genotypes. In the example this would be of order 4. Let $\mathbf{f}$ be the vector of values of the genotypes for an animal. For example, for animal 1, $\mathbf{f} = (1 \;\; 2 \;\; 3 \;\; 2)$, and for animal 10 would be $\mathbf{f} = (3 \;\; 1.5 \;\; 1.5 \;\; 2)$. For this example,

$$\mathbf{C}_c = \begin{pmatrix} .01841621 & .00059527 & .00098123 & -.00046843 \\ .00059527 & .01881381 & -.00049214 & .00018887 \\ .00098123 & -.00049214 & .01879366 & .00081384 \\ -.00046843 & .00018887 & -.00081384 & .01912397 \end{pmatrix},$$

then the variance of prediction error of the "Genomic Part" would be

$$\begin{aligned} \mathbf{C}_{gp} &= \mathbf{f}\mathbf{C}_c\mathbf{f}'\hat{\sigma}_e^2 \\ &= .3510765\hat{\sigma}_e^2 \end{aligned}$$

for animal 1, and would be $.3392277\hat{\sigma}_e^2$ for animal 10.

The off-diagonals between the genomic part and the polygenic parts will be very small and close to zero, if there are enough data. Thus, the two pieces could be considered to be nearly independent. This will need to be checked with a large data set before proceeding under the assumption of independence.

The variance of prediction error of the polygenic part comes from the diagonals of the inverse of the mixed model equations. For animals 1 and 10 the diagonal elements of the inverse were .3041048 and .2723333, respectively. Hence, the variance of prediction error of the combined EBV would be $.6551813\hat{\sigma}_e^2$ for animal 1, and $.611561\hat{\sigma}_e^2$ for animal 10.

To convert from variance of prediction error (VPE), the total genetic variance (VG) is needed. In the proposed model the total genetic variance is

$$VG = \frac{\sigma_e^2}{50} + \frac{\sigma_e^2}{\alpha A_{ii}},$$

where the first term represents the genetic variance of the "Genomic Part", the second term is the remaining polygenic variation, and A_{ii} is the diagonal of the **A** matrix for an animal (i.e. 1 plus the inbreeding coefficient). To convert to a reliability (REL) then calculate

$$REL = (VG - VPE)/VG$$

and multiply by 100. One could also do Gibbs sampling on the mixed model equations keeping the variances constant. About 100 samples would be needed. The variance of the sample values of the combined EBV would give VPE for each animal, and this will include the off-diagonals between genomic and polygenic parts.

3 Conclusions

The proposed model has some advantages and disadvantages. The advantages are

- Up to 200 SNPs out of 50,000 are utilized per trait, computations are reduced. If necessary, more SNPs could be used but would likely not be necessary.
- The model includes a polygenic part and a genomic part and these are estimated simultaneously.
- A complicated genetic relationship matrix is avoided as well as its inverse.
- Not all animals need to be genotyped, but genotypes need to be deduced from parents and progeny if an animal is not genotyped.
- Predictions for new animals can be derived easily.
- Reliabilities can be obtained readily by different approaches.

The disadvantages are

- Every trait will need a different set of SNPs.
- To work well in the long run, many more females will need to be genotyped, at least in dairy cattle.
- A large number of SNPs are not utilized at all, but at the same time their contribution is expected to be very small.

A chip panel could be made to only have those SNPs that contribute to each trait. If there are 100 traits, then if each trait has a different set of 200 SNPs, then a 20,000 SNP panel might be needed. However, some SNPs may overlap between traits. By having fewer SNPs on a panel, perhaps producers would be better able to afford to genotype their cows. If the cost is the same as the 50K chip, then only the 50K chip needs to be used.

There have been no comparisons to other proposed methods. Which method is more accurate is not known. The proposed model will give more accurate combined EBVs compared to the usual animal models, as evidenced by the smaller residual variance of the new model. The proposed method is simple and requires only minor modification of existing software to include another random factor in the model for each trait. Computing time will not be increased greatly over existing systems, but revisions to software may need time to implement this proposal.

This proposal was similar to comments made by Ben Hayes and Mike Goddard at the recent Interbull Meeting in Uppsala. There is little time to research this proposal thoroughly, however, before "something" has to be in place for CDN this fall. This comes from trying to out-race the competition, but the comment was that Canadian producers want to make sure the method was correct first. They did not indicate which method was correct or better, nor has anyone else.

4 References

Hayes, B. J., M. E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution 33:209-229.

Meuwissen, T. H. E, B. J. Hayes, M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Stranden, I., D. J. Garrick. 2009. Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. (submitted).

VanRaden, P. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423.