# The Unified Approach For Genomic Prediction: Issues Related To SNP Allele Frequencies, Scaling And Effects Of Selection

*O. F. Christensen*[*]

## Introduction

For genomic prediction, not all animals in the population are genotyped; i.e. for cattle, bulls (and perhaps bull mothers) are genotyped, but the entire cow population is not genotyped. The strategy used in practice for genomic prediction has therefore been a multi-step procedure, where first, the pedigree information is used to to create pseudo-data like deregressed proofs for the genotyped animals, and second, genomic prediction is based on these pseudo-data and the SNP-markers for the genotyped animals. For such multi-step procedure, having selection (both selective genotyping and selective phenotyping) in the system causes biases in the procedure which are difficult to characterise and could be of great importance for breeding programs with genomic selection. As an alternative, a unified approach to genomic prediction has recently been developed (Legarra *et al.* (2009);Aguilar *et al.* (2010);Christensen and Lund (2010)) where no such creation of pseudo-data is needed, and instead a full model for all animals (including animals that are not genotyped) is used. Since the model can incorporate the data used for selection (i.e. markers for selection candidates and phenotypes used for selection) there are in principle no such problems with biases from selection. However, the approach requires that SNP allele frequencies are known for the unselected base population, and as discussed in Christensen and Lund (2010) the possible effects of estimating these allele frequencies in scenarios with selection have not been sufficiently investigated. In Aguilar *et al.* (2010) the use of several different allele frequencies was compared for genomic prediction of US Holstein cattle where the conclusion was that using all allele frequencies equal to $1/2$ performed the best but also that some scaling of genomic relationship matrix was needed. Here, for the unified approach some issues in relation to allele frequencies, scaling and effects of selection are investigated.

## Material and methods

### Model

The unified approach (Legarra *et al.* (2009);Aguilar *et al.* (2010);Christensen and Lund (2010)) is based on the following model

$$y = X\beta + Zg + e, \tag{1}$$

[*] Faculty of Agricultural Sciences, Aarhus University, Denmark

where $y$ is phenotype, $X$ and $Z$ are incidence matrices, $\beta$ denotes fixed effects, $e$ is error, and $g \sim N(0, \sigma_g^2 \tilde{G}_w)$ is the genetic effect. Here the extended genomic relationship matrix $\tilde{G}_w$ has the inverse

$$(\tilde{G}_w)^{-1} = \left[ \begin{array}{cc} G_w^{-1} - A_{11}^{-1} & 0 \\ 0 & 0 \end{array} \right] + A^{-1}, \tag{2}$$

where

$$G_w = (1-w)G(m^{obs}) + wA_{11}$$

corresponds to the genotyped animals, $A$ is the usual pedigree derived additive relationship matrix, and the genomic relationship matrix based on genotyped animals is

$$G(m^{obs}) = (m^{obs} - p)(m^{obs} - p)^{\mathrm{T}}/s \tag{3}$$

with $m_{ij} = -1$, when the SNP $j$ of individual $i$ is 11, $m_{ij} = 0$ for 12, and $m_{ij} = 1$ for 22, scaling $s = \sum_j 2\rho_j(1-\rho_j)$, and $p_j$ is the column with entries $2\rho_j - 1$. The parameters $\rho_1, \ldots, \rho_p$ are allele frequencies of the second allele, and the parameter $w$ is the relative weight on the polygenic effect (part of genetic effect not captured by markers). Parameter estimation and prediction is based on AI-REML and the mixed model equations - further details can be found in Christensen and Lund (2010).

In the model above, the allele frequencies $\rho_1, \ldots, \rho_p$ should represent allele frequencies in the unselected base population in order for the genomic relationship matrix (3) to be on the same scale as the pedigree derived relationship matrix $A_{11}$ for the genotyped animals. In the analysis, $\rho_1, \ldots, \rho_p$ are assumed to be known, but in practice they have to be computed from the observed markers. Christensen and Lund (2010) note that the effect of estimating these allele frequencies needs further investigation, in particular in scenarios with selection. Aguilar *et al.* (2010) used $\rho_j = 1/2$ but with a different scaling $s \neq 2\sum \rho_j(1-\rho_j)$. Therefore the choice of scaling $s$ is also an issue that needs attention.

**Simulation study**

Here we compare two scenarios, one with selection on phenotype, and one with random selection. In both scenarios, we assume 10 chromosomes each 160 cM long, and use a panel of $p = 5000$ equidistant SNP markers. It is assumed that 500 QTLs affect the phenotype, and the size of these effects is simulated from a Gamma$(5.4, 0.42)$-distribution. First, a base population consisting of 400 boars and 4000 sows is generated by assuming random mating for 50 generations in a population with an effective population size of 100. Then the 400 boars are mated with the 4000 sows to produce 40000 offspring in generation 1 (half of them males). It is assumed that phenotypes of all boars in generation 1 are available. For genotyping, 400 boars are selected either based on the highest value of own phenotype (scenario 1) or randomly (scenario 2).

The use of three different genomic relationship matrices is compared. For the first two matrices, allele frequencies are estimated as averages of observed allele frequencies, where for the first matrix $s = \sum_j 2\rho_j(1-\rho_j)$ and for the second matrix $s = mean(d_1, \ldots, d_n)$ with $d_i = ((m^{obs} - p)(m^{obs} - p)^{\mathrm{T}})_{ii}/(A_{11})_{ii}$. For the third matrix, all allele frequencies are assumed to be 0.5 and $s = mean(d_1, \ldots, d_n)$ where $d_i = ((m^{obs})(m^{obs})^{\mathrm{T}})_{ii}/(A_{11})_{ii}$.

For each scenario, ten independent simulations were made. For every simulation and each of the three different genomic relationships matrices, the parameter $w$ is estimated in order to investigate how sensitive that estimation is to selection; a discrete set of values $w = 0.01, 0.02, \ldots, 0.09, 0.10, 0.20, \ldots, 1.00$ are used. Diagonal elements and off-diagonal elements of $G(m^{obs})$ and correlations between estimated breeding values and true breeding values for genotyped and non-genotyped animals are also presented.

## Results and discussion

Table 1 shows the results from the two scenarios and with the use of three different $G(m^{obs})$ matrices. For the scenario 1 with selection we see for all ten simulations that $\hat{w} = 1$ for the two matrices based on estimated allele frequencies and that $\hat{w} = 0.01$ for the matrix with allele frequencies equal to $1/2$, whereas for the scenario 2 without selection we see $\hat{w} \approx 0.05$ (with some fluctuation between simulations) for all three matrices. To summarise, sensible estimates of $w$ are obtained in the scenario with random selection for all three matrices and in the scenario with selection for the matrix with allele frequencies $1/2$, whereas meaningless results are obtained in the scenario with selection for the two matrices based on estimated allele frequencies. In the scenario with selection, the difference in $\hat{w}$ between using estimated allele frequencies and using all allele frequencies equal to $1/2$ is striking, and is likely due to the allele frequencies not to the scale $s$, since similar results are seen for the two choices of scale $s$ when using estimated allele frequencies. We conclude that the estimation of $w$ is clearly sensitive to allele frequencies in a scenario with selection.

**Table 1: Estimates $\hat{w}$, summaries of $G(m^{obs})$ and correlation between estimated and true breeding values for the two scenarios and using the three different $G(m^{obs})$ matrices ($\hat{\rho}$, $s(\hat{\rho})$: $\hat{\rho}_j$ is estimated allele frequency of observed markers and $s = 2\hat{\rho}_j(1-\hat{\rho}_j)$; $\hat{\rho}$, $s$-aver.: $\hat{\rho}_j$ is estimated allele frequency of observed markers and $s = mean(d_1, \ldots, d_n)$; $1/2$, $s$-aver.: $\rho_j = 1/2$ and $s = mean(d_1, \ldots, d_n)$). Column four and five shows the means of the diagonal elements and the off-diagonal elements of $G(m^{obs})$, respectively. The last two columns show correlation between estimated genomic breeding values (using $w = 0.01$) for the genotyped and non-genotyped animals, respectively. Throughout the table averages of ten simulations are shown, and with exception of $\hat{w}$ for the random selection scenario the results did not show substantial variability between the simulations.**

| Scenario | $G$ type | $\hat{w}$ | diag-G | off-diag-G | cor-geno | cor-nongeno |
|---|---:|---:|---:|---:|---:|---:|
| select | $\hat{\rho}$, $s(\hat{\rho})$ | 1 | 0.999 | -0.00250 | 0.470 | 0.576 |
| select | $\hat{\rho}$, $s$-aver. | 1 | 1 | -0.00251 | 0.470 | 0.576 |
| select | $1/2$-$s$-aver. | 0.01 | 1 | 0.380 | 0.375 | 0.580 |
| random | $\hat{\rho}$, $s(\hat{\rho})$ | 0.053 | 0.999 | -0.00250 | 0.611 | 0.590 |
| random | $\hat{\rho}$, $s$-aver. | 0.051 | 1 | -0.00250 | 0.611 | 0.590 |
| random | $1/2$, $s$-aver. | 0.064 | 1 | 0.364 | 0.607 | 0.590 |

In Table 1 we also see a large difference in the off-diagonal elements between using the estimated allele frequencies and using allele frequencies $1/2$. For this simulation the off-diiagonal elements of $A_{11}$ are on average $0.0010$ for the scenario with selection and $0.0007$ for the scenario with random selection, and we see that using estimated allele frequencies gives a genomic relationship matrix much more similar to $A_{11}$ than when using allele frequencies $0.5$.

Finally, for the scenario with selection we see that the correlation between estimated breeding values and true breeding values is higher when using estimated allele frequencies compared to when using allele frequencies equal to $1/2$. This increase in correlation is mainly for the genotyped animals, whereas for the non-genotyped animals there is only a minor increase. The results show that even though the estimated $\hat{w} = 1$ does suggest some fundamental problems with the use of estimated allele frequencies, then when choosing a small value $w = 0.01$ the use of estimated allele frequencies actually provided more accurate estimated breeding values compared to the use of allele frequencies $1/2$. For comparison, for the scenario with random selection the correlations between estimated breeding values and true breeding values are almost the same for the three matrices.

## Conclusion

The conclusions from this study are not entirely clear-cut. Clearly, when using allele frequencies $1/2$ the off-diagonal elements of the genomic relationship matrix do not resemble the off-diagonal elements of the pedigree derived relationship matrix. On the other hand, when using estimated allele frequencies these off-diagonal elements are more similar. For the scenario without selection, there are no large differences in the parameter estimate $\hat{w}$ and the accuracy of genomic breeding values between using estimated allele frequencies and using allele frequencies $1/2$. However, for the scenario with selection some greater differences were seen: using estimated allele frequencies resulted in the spurious estimate of $\hat{w} = 1$ for the polygenic weight, and on the other hand the accuracy of genomic breeding values was higher than when and using allele frequencies $1/2$. The study therefore demonstrates that the choice of allele frequencies is an issue that needs consideration when there is selection, but the study does not provide much insight on what to do in a practice. Further research is needed on how to address this properly.

## References

Aguilar, I., Misztal, I., Johnson, D. L. *et al.* (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J Dairy Sci* **93**, 743–752.

Christensen, O. F. and Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genet Sel Evol* **42**, 2.

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