Genomic Selection Based on Flexible Haplotypes Applied to Norwegian Red Cattle Pedigrees

Xijiang Yu[*] and T.H.E. Meuwissen

Introduction

Genomic selection (GS) technology is currently being adopted by the dairy cattle breeding industries around the world. Genome-wide breeding value (GWEBV) prediction plays a pivotal role for this new technology. Its accuracy depends on the statistical methods used, genome and population structure. Although the originally genomic selection was applied on haplotypes (**?**), several studies found that GWEBV based on individual SNP effects were as accurate or more accurate than GWEBV based on haplotype effects (**?**; **?**; **?**). However, haplotype effects are expected to be in a stronger linkage disequilibrium than single SNPs, and thus explain more of the QTL variance (**?**). Thus, in principle haplotypes can explain more genetic variance and yield higher accuracy, but they do not always achieve that in simulation studies, probably because the number of effects to be estimated becomes too large. Our hypothesis is that this is due to a too rigid definition of haplotypes: if haplotypes are defined a flexible and judicious way the problem of estimating too many effects is circumvented. This is supported by **?**, who found superior GWEBV accuracies for large haplotypes when an IBD (Identity-by-Descent) matrix was used to reduce the 'effective' number of effects to be estimated. However, fitting e.g., 50,000 IBD matrices in a genomic selection model is impractical. The aim of this paper is to present a novel 'flexible haplotypes approach', called 'FlexHap', and compare it to other single SNP based GWEBV prediction in a real Norwegian Red Cattle (NRF) pedigree.

Material and methods

Flexible haplotypes. Figure 1 gives an example of a complicate haplotype structure, where the first haplotype is arbitrarily colored yellow. At the same locus, other haplotypes obtain the same color if they are 'similar' or a different color otherwise. In the FlexHap approach, all the different 'color by position' haplotypes are fitted in the statistical model, and the model equation is shown for each of the haplotypes on the right of Figure 1, where position are identified by letters (A-F). The result is that the big yellow haplotype 1 is split in to 6 effects (1A-1F), but this is equivalent to fitting one big effect for the yellow haplotype with 6 times more variance (since it is 6 times longer than the smallest haplotype).

[*]Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, 1432 Ås

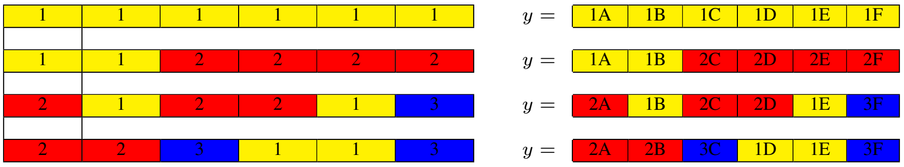

Figure 1. Example of a complex haplotype structure, and how it is fitted by the statistical model. (A-F) denote positions.

The question remains when do we call two haplotypes similar at a position and when not. Based on experiences with IBD matrix calculations, it was decided that the two haplotypes are similar if $L + R \geq 10$ and $\min(L, R) > 1$, where L (R) is the number of markers to the left (right) of the position that is are identical until a break occurs, i.e. a position where the markers are not identical. In other words, we require a segment of at least 10 markers with identical alleles that contains the position, and the position is not right at the border of this segment. For historical reasons, we defined the region between two markers, i.e. the marker bracket, as a ‘position’.

The statistical model for the flexible haplotypes is thus:

$$y_i = \mu + \sum_{j=1}^{L-1} (\text{hap}_{ij_1} + \text{hap}_{ij_2})$$

where $L - 1$ is the number of marker brackets ($L = 500$ here); hap_{ij_1} (hap_{ij_2}) is the paternal (maternal) haplotype of animal i at bracket j, and haplotypes are defined by Figure 1 as ‘color by position’ identifiers. In a SNP based model, hap_{ij_1} and hap_{ij_2} are replaced by the alleles of SNP$_j$ in animal i: SNP$_{ij_1}$ and SNP$_{ij_2}$, and summation extends from $j = 1$ to $j = L$, i.e. to the last SNP.

We used a model where ‘big haplotypes’ were assumed to have a normal prior with large variance and ‘small haplotypes’ had a normal prior with small variance. The size of the large and small variance was decided by the Pareto Principle (http://en.wikipedia.org/wiki/Pareto_principle), which states that $X\%$ of the haplotypes explain $(100 - X)\%$ of the variance, e.g. 10% of the SNPs explain 90% of the variance. This model is called FlexHapP. We calculated X here as $(N_{\text{QTL}}/L) \times 100\%$, where N_{QTL} is the number of QTL. The same approach was used for the SNP based model, which is called MixP since it is a mixture model where the prior mixing proportions are based on the Pareto Principle. In case of the SNP based model, also a model was used that that assumed that the SNP effects had a normal prior distribution with all variances being equal, which is also known as GWBLUP.

Norwegian Red Cattle. The Norwegian Red Cattle pedigree used in this study consisted 19,523 individuals spread over 8 generations as kindly provided by GENO AS (http://www.geno.no). This pedigree included 104 imported bulls. The data contained also an identifier for whether a bull was genotyped or not and 2,165 were genotyped, which includes the 104 imported bulls. We will base our simulations on this real pedigree and will assume that the same animals are genotyped as in the real life situation. The cattle population was sorted first to make sure parents appear before their offspring. The 1,915 oldest genotyped bulls were marked as the training set, and the youngest 250 bulls were marked as the evaluation set, i.e. whose EBV are predicted.

Genome structure. The parents of unknown origin individuals were sampled from an ideal population of effective size 200 in each scenario, which was simulated for 10,000 generations to achieve a mutation drift balance and linkage disequilibrium between the loci. The genome consisted of 1 Morgan / 10^8 base-pairs. The mutation rate was 10^{-8} per base-pair per meiosis. Markers and QTL loci were randomly selected amongst those with minor allele frequency 0.05. The number of markers was 500 whereas various numbers of QTL and heritabilities were simulated. QTL effects are sampled from a Laplace distribution with mean 0 and scale parameter 1. The genotypes of the last generation were gene-dropped into the sorted real population pedigree. No more mutation occurs at this stage. The 1,915 training bulls were genotyped and phenotyped, and the 250 evaluation bulls were only genotyped. The heritabilities, h^2_{chr}, used here should be interpreted as per chromosome heritabilities, i.e. they are about 30 times smaller than the total trait heritabilities (or reliabilities in case of daughter-yield-deviations).

Results and discussion

Table 1 shows that the flexible haplotype approach is superior when the number of QTL is small, but very similar to the other approaches when the number of QTL is large. Application of the Pareto Principle deteriorated the accuracies of the SNP based approaches, since MixP performed worse than GWBLUP. It seems that the Pareto Principle as applied here puts too much weight on the big genes. The latter may be remedied by using higher values of X than was used here, e.g. 20% of the genes explain 80% of the variance, but this was not attempted here.

It may also be noted that the accuracies of GWBLUP are very much independent of the number of QTL, which is expected since GWBLUP does not give extra weight to the SNPs with big effects. Hence, it does not benefit from few genes having a big effect. This can also be seen from the formula for the accuracy of GWBLUP, as shown by **?** and **?**, which depends on the size of the number of independent segments in the genome, but not on the actual number of QTL.

The FlexHapP approach did seem to take advantage of a reduction of the number of QTL, but so do other SNP based approaches such as BayesB. The BayesB approach as described by **?** yielded accuracies of 0.47, 0.65, and 0.71 for h^2_{chr} values of .01, .02 and .03, respectively. Thus, BayesB, which is a SNP based approach, seems equally able to take advantage of a reduction of the number of QTL. In future work we intent to investigate the performance of FlexHapP with different values of X, including $X = 50\%$ which implies that all haplotypes have a priori equal variance (as in GWBLUP). Furthermore, we intent to test the approach on real phenotypic data.

Acknowledgement: We would like to thank GENO AS for kindly providing the Norwegian Red Cattle pedigree.

Table 1: Accuracies of GWEBV obtained by GWBLUP, MixP and FlexHapP when the chromosomal heritability (h^2_{chr}) and number of QTL (N_{QTL}) was varied.

h^2_{chr}	N_{QTL}	GWBLUP	MixP	FlexHapP
.01	5	.39	.31	.46
.02	5	.53	.41	.62
.03	5	.54	.43	.68
.01	20	.41	.37	.40
.02	20	.50	.46	.53
.03	20	.56	.51	.61
.01	200	.39	.39	.39
.02	200	.52	.52	.52
.03	200	.57	.57	.57