Breeding Value Estimation for Fat Percentage Using Dense Markers on *Bos taurus* Autosome 14

A. P. W. de Roos,*¹ C. Schrooten,* E. Mullaart,* M. P. L. Calus,† and R. F. Veerkamp†

*HG, 6802 EB Arnhem, the Netherlands

†Animal Breeding and Genomics Centre, Animal Sciences Group, Wageningen University and Research Centre, 8200 AB Lelystad, the Netherlands

ABSTRACT

Prediction of breeding values using whole-genome dense marker maps for genomic selection has become feasible with the advances in DNA chip technology and the discovery of thousands of single nucleotide polymorphisms in genome-sequencing projects. The objective of this study was to compare the accuracy of predicted breeding values from genomic selection (GS), selection without genetic marker information (BLUP), and gene-assisted selection (GEN) on real dairy cattle data for 1 chromosome. Estimated breeding values of 1,300 bulls for fat percentage, based on daughter performance records, were obtained from the national genetic evaluation and used as phenotypic data. All bulls were genotyped for 32 genetic markers on chromosome 14, of which 1 marker was the causative mutation in a gene with a large effect on fat percentage. In GS, the data were analyzed with a multiple quantitative trait loci (QTL) model with haplotype effects for each marker bracket and a polygenic effect. Identical-by-descent probabilities based on linkage and linkage disequilibrium information were used to model the covariances between haplotypes. A Bayesian method using Gibbs sampling was used to predict the presence of a putative QTL and the effects of the haplotypes in each marker bracket. In BLUP, the haplotype effects were removed from the model, whereas in GEN, the haplotype effects were replaced by the effect of the genotype at the known causative mutation. The breeding values from the national genetic evaluation were treated as true breeding values because of their high accuracy and were used to compute the accuracy of prediction for GS, BLUP, and GEN. The allele substitution effect for the causative mutation, obtained from GEN, was 0.35% fat. The accuracy of the predicted breeding values for GS (0.75) was as high as for GEN (0.75) and higher than for BLUP (0.51). When some markers close to the QTL were omit-

ted from the model, the accuracy of prediction was only slightly lower, around 0.72. The removal of all markers within 8 cM from the QTL reduced the accuracy to 0.64, which was still much higher than BLUP. It is concluded that, when applied to 1 chromosome and if genetic markers close to the QTL are available, the presented model for GS is as accurate as GEN.

Key words: genomic selection, genetic marker, fat percentage

INTRODUCTION

Molecular genetic selection can lead to much higher genetic gains than conventional quantitative genetic selection, especially for traits with low heritability, phenotypes that are difficult to record, unfavorable genetic correlations, and genotype $\times$ environmental interactions (Meuwissen and Goddard, 1996; Dekkers and Hospital, 2002). Animal breeding programs have been using molecular genetic information for many years, but its effect has been less than initially expected (Dekkers, 2004). One of the reasons is the difficulty to find the causal mutations in QTL, or genetic markers that are in high population-wide linkage disequilibrium (LD) with a QTL. Many genetic markers that are in population-wide linkage equilibrium or low LD with a QTL have been found, but these are much more difficult to use in molecular genetic selection, because the linkage phase between the marker and the QTL needs to be estimated for each family (Dekkers, 2004).

Advances in DNA chip technology and the discovery of many thousands of single nucleotide polymorphisms (SNP) in genome sequencing projects have provided new opportunities to find markers in LD with QTL and to use them for selection (Andersson and Georges, 2004). Haley and Visscher (1998) predicted that the development of cheap and high-density marker maps would move the selection based on polygenes plus individual loci to effective total genomic selection (GS). This would greatly improve selection before phenotypic information from the animal or its progeny is available (for example, selection among young bulls before prog-

Received March 1, 2007.

Accepted June 22, 2007.

¹Corresponding author: roos.s@hg.nl

eny testing). Furthermore, it can enable selection among young animals or embryos, which may dramatically reduce the generation interval. Meuwissen et al. (2001) presented a method to predict breeding values using genome-wide dense marker maps. Using Bayesian statistics, the effects of 50,000 simulated haplotypes were estimated from only 2,200 phenotypic records. After that, the total genetic value of an animal was predicted with an accuracy of 0.85 by summing the estimated effects of the haplotypes of the animal for each marker bracket. This method, GS, attempts to explain all genetic variation by genetic markers without selection of markers that contribute to the genetic variance. It was concluded that GS can substantially increase the rate of genetic gain, especially if combined with reproductive techniques to shorten the generation interval. It can be argued that prediction of breeding values is not the same as making selection decisions, but because GS is the accepted name for the method proposed by Meuwissen et al. (2001), this term is used throughout the paper.

Meuwissen et al. (2001) used the flanking markers of a putative QTL to define a haplotype, which means that all marker brackets that carry the same marker alleles are assumed to have the same effect, whereas in reality they may carry different QTL alleles. Furthermore, they did not include a matrix of identical-by-descent (IBD) probabilities between marker brackets, which means that covariances among different haplotypes were assumed to be zero. These assumptions were relaxed in the multiple-QTL mapping method presented by Meuwissen and Goddard (2004), which used the IBD probability matrix among haplotypes as described by Meuwissen and Goddard (2001). The multiple-QTL mapping method has been applied in QTL mapping studies (Olsen et al., 2005) but can also be applied as a method for GS by using a dense marker map with whole-genome coverage.

In the present literature, GS has not been applied to real data. The objective of this study was to validate the method of Meuwissen and Goddard (2004) for prediction of genomic breeding values on a dairy cattle data set for 1 chromosome and to compare the accuracy of prediction to a method without marker information and to a method in which the causative mutation underlying an important QTL is known. Furthermore, the effect of omitting markers close to the QTL on the accuracy of the prediction was analyzed.

MATERIALS AND METHODS

Materials

Data were obtained from a QTL mapping study using a granddaughter design comprising 1,300 progeny-

tested Holstein-Friesian bulls born from 1973 to 1994 (Farnir et al., 2002). Twenty-seven grandsires had at least 10 sons, which summed up to 1,135 sons in total, and, on average, 42 sons per grandsire for validation, as explained later. Estimated breeding values for fat percentage, obtained from the official August 2006 genetic evaluation for the Netherlands and Flanders, were used as phenotypic records. All bulls were genotyped for 32 markers on *Bos taurus* autosome 14. The marker set comprised 13 microsatellite markers and 19 SNP markers (Figure 1). The percentage of heterozygous animals was from 41 to 81 (64 on average) for the microsatellite markers and from 0 (for marker 14) to 65 (43 on average) for the SNP markers. Marker 7 was the K232A substitution in the acyl coenzyme A:diacylglycerol acyltransferase 1 (**DGAT1**) gene, which was shown to have a large effect on fat percentage (Grisart et al., 2002; Winter et al., 2002). Eleven grandsires were heterozygous for this marker, whereas 12 grandsires were homozygous for the A allele that was associated with low fat percentage, and 4 grandsires were homozygous for the K allele that was associated with high fat percentage.

The map was constructed based on the bovine composite map (www.livestockgenomics.csiro.au/perl/gbrowse.cgi/bosmap/). The centimorgan position of the markers that were not placed on the composite map were calculated by interpolation using their base pair positions on the National Center for Biotechnology Information bovine sequence map (version 3.1, www.ncbi.nlm.nih.gov) and the base pair position of neighboring markers on the composite map. Figure 1 shows the positions of the SNP and microsatellite markers relative to the causative mutation in the DGAT1 gene. Haplotypes were constructed from the marker genotypes by comparing the genotype of an animal to that of its sire (dams were not genotyped). This was informative in situations when the animal or its sire was homozygous. If both animal and sire were heterozygous but the animal had genotyped offspring, the linkage phase with the closest informative marker was assumed the same as in the majority of the offspring. For example, consider an animal with genotype A/a at locus 1 and B/b at locus 2 and its sire with genotypes A/a and B/B, respectively. For locus 1, it is unclear whether the A or a allele was inherited from the sire, whereas at locus 2, allele B was inherited from the sire. To infer the phase at locus 1, the genotypes of the progeny of the animal were considered: for progeny that were homozygous at both loci, their haplotypes could be determined. If the majority of the progeny of this animal inherited haplotype AB or ab, allele A was assumed paternal, whereas if the majority inherited haplotype aB or Ab, allele a was assumed paternal. Markers with

Figure 1. Positions of the microsatellite and SNP markers relative to the causative mutation in the DGAT1 gene (cM)

unknown phase were treated as missing for the particular animal.

Model

The data were analyzed with a multiple QTL model (Meuwissen and Goddard, 2004):

$$y_i = \mu + u_i + \sum_{j=1}^{31} (q_{ij1} + q_{ij2})v_j + e_i$$

where y_i = the EBV for fat percentage of sire i ; μ = the overall mean; u_i = the polygenic effect of sire i ; v_j = the direction of the QTL effects of the haplotypes at marker bracket j ; q_{ij1} (q_{ij2}) = the size of the QTL effect, expressed in units of v_j , for the paternal (maternal) haplotype of sire i at marker bracket j ; and e_i = the residual term for sire i (Meuwissen and Goddard, 2004). The covariance among polygenic effects (u_i) was modeled as $\mathbf{A} \times \sigma_G^2$, where $\mathbf{A}$ = the relationship matrix, which was based on the full known pedigree (4,688 animals, including females), and σ_G^2 = the polygenic variance. The model assumes a putative QTL is in the midpoint of a marker bracket, but it may also account for QTL that are on other positions in the marker bracket or even outside the marker bracket. The covariance among the haplotypes at bracket j (q_{ij}) was modeled as $\mathbf{H}_j$ (i.e., the matrix of estimated IBD probabilities among the haplotypes at the midpoint of bracket j). The variance of q_{ij} was assumed 1, because v_j distinguishes between

marker brackets with no QTL, a small QTL, or a large QTL. Meuwissen and Goddard (2004) included v_j in their model as a direction vector comprising the effects of a putative QTL at marker bracket j on multiple traits to avoid the estimation of a covariance matrix among traits for every putative QTL. In a single-trait model, such as this study, v_j is just a scalar, and it may seem more logical to combine the effects of putative QTL into q_{ij} and remove v_j from the model. The model including v_j is presented here, however, because the method and software were developed for a multiple-trait application as described by Meuwissen and Goddard (2004). The IBD probabilities between haplotypes were calculated using the algorithm of Meuwissen and Goddard (2001), which combines LD with linkage information and, for each bracket j , considers all 32 surrounding markers and all available pedigree information. The effective population size was assumed 100, and the number of generations since an arbitrary founder population was also assumed 100, as in Meuwissen and Goddard (2004). To reduce the rank of the IBD matrix, base haplotypes (i.e., haplotypes that were inherited from a nongenotyped parent) were clustered with other base haplotypes into 1 combined haplotype when their IBD probability was above 0.95. The IBD probability of the combined haplotype with each other base haplotype was computed as the average of the IBD probabilities of the original haplotypes with that other haplotype. If the matrix of IBD probabilities among base haplotypes was not positive definite after clustering, the matrix

was bent by setting the negative eigenvalues to 0.01. The matrix was subsequently inverted by LU decomposition. The elements in $\mathbf{H}_j^{-1}$ for the descendant haplotypes were calculated using the algorithm of Fernando and Grossman (1989).

A Markov chain Monte Carlo method using Gibbs sampling was used to estimate the joint posterior probability density of the unknowns in the model (Meuwissen and Goddard, 2004). The effect of a putative QTL at bracket j , v_j , was sampled from a normal distribution, $N(0, \sigma_v^2)$, if a QTL was sampled in bracket j , whereas v_j was sampled from $N(0, \sigma_v^2/100)$ if no QTL was sampled in bracket j . The variance of the putative QTL effects, σ_v^2 , was sampled from an inverted χ^2 distribution with a prior variance of 0.000723% , which was calculated as the additive genetic variance divided by 100 (i.e., assuming 100 additive and unrelated QTL affecting the trait) across all chromosomes. The presence of a QTL in bracket j was sampled from a Bernoulli distribution with probability equal to

$$\frac{P(v_j | \sigma_v^2) \times \text{Pr}_j}{P(v_j | \sigma_v^2) \times \text{Pr}_j + P(v_j | \sigma_v^2/100) \times (1 - \text{Pr}_j)},$$

where $P(v_j | \sigma_v^2)$ = the probability of sampling v_j from

$$N(0, \sigma_v^2) \left(\text{i.e., } \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-\frac{v_j^2}{2\sigma_v^2}} \right) \text{ and } \text{Pr}_j = \text{prior probability of}$$

the presence of a QTL in bracket j . Pr_j was calculated as the length of bracket j divided by the total length of all 31 brackets, 29.07 cM. More details on the prior distributions and the fully conditional distributions can be found in Meuwissen and Goddard (2004). The Gibbs sampler was run for 25,000 iterations, and 5,000 iterations were removed as burn-in. Earlier studies revealed that posterior means after 25,000 iterations were hardly different from posterior means after 300,000 iterations (A. P. W. de Roos; unpublished data). The software was developed by the authors from earlier programs written by Theo Meuwissen (University of Life Sciences, Ås, Norway).

Alternative Methods

In this study, 3 methods for genetic evaluation and selection were compared using the data described above: BLUP, GS, and gene-assisted selection (GEN). The model for GS was as described above. The model for BLUP was as described for GS but without the effects of the haplotypes (i.e., $y_i = \mu + u_i + e_i$). The model for GEN was as described for GS, but the effects of the haplotypes were replaced by the effect of the marker

genotype at the causative mutation in the DGAT1 gene [i.e., $y_i = \mu + u_i + (q_{i1} + q_{i2}) v_j + e_i$], where q_{i1} (q_{i2}) = the effect of the paternal (maternal) marker of sire i at the causative mutation in the DGAT1 gene, expressed in units of v_j . For animals with unknown genotype ($n = 62$), a third marker allele was created to model their effect. The covariance among the markers (q) was assumed zero.

For GS, 6 alternatives were compared to study the effect of having fewer markers surrounding the causative mutation in the DGAT1 gene. This was done by using only markers m to 32 in the evaluation, where $m = 1, 7, 15, 18, 22$, and 25 for alternatives GS1, GS7, GS15, GS18, GS22, and GS25, respectively. This approach, as opposed to reducing the marker density across the whole chromosome segment, was chosen to show the effect of increasing the distance between markers and QTL. Note that in GS1 and GS7, the causative mutation in the DGAT1 gene, which was marker 7, was used in the evaluation, whereas it was not used in the other alternatives.

Comparison of Alternatives

The alternative methods were compared by the pseudo-accuracy of predicted breeding values of 1,135 sons that were sired by 27 grandsires. For each grandsire, the phenotype of 1 out of each 20 sons was omitted from the data, but the equations corresponding to the polygenic effect, the haplotype effects (for GS), and gene effects (for GEN) of these sons were kept in the model. After the evaluation, the breeding values of the sons whose phenotype was omitted from the data were predicted by summing the estimated mean, the polygenic effect, and the corresponding haplotype effects (for GS) or gene effects (for GEN). For each alternative method, 20 evaluations were performed, so the phenotype of each son was omitted from the data in 1 evaluation and used in the other 19 evaluations. After all 20 evaluations, the predicted breeding values of the 1,135 sons were compared with their EBV from the national genetic evaluation, which were regarded as true breeding values because of their high reliability (around 0.95).

RESULTS

Haplotype Clustering

The total number of haplotypes before clustering was twice the number of genotyped animals (i.e., 2,600 haplotypes, of which 1,350 haplotypes were base haplotypes). After clustering, the number of base haplotypes was from 44 to 239, depending on bracket, and the total number of haplotypes was from 305 to 514. In general, the number of (base) haplotypes was higher for larger

Table 1. Statistics of the true breeding values (TBV)¹ and the predictions from gene-assisted selection (GEN),² genomic selection (GS),³ and BLUP (n = 1,135)

Method	Mean	SD	Variance	Minimum	Maximum
TBV	0.017	0.364	0.133	-1.01	1.36
GEN	0.013	0.269	0.072	-0.52	0.85
GS1	0.013	0.277	0.077	-0.51	0.83
GS7	0.012	0.276	0.076	-0.54	0.83
GS15	0.012	0.267	0.071	-0.53	0.83
GS18	0.011	0.262	0.069	-0.52	0.82
GS22	0.012	0.260	0.067	-0.51	0.81
GS25	0.008	0.222	0.049	-0.47	0.76
BLUP	0.003	0.184	0.034	-0.40	0.57

¹TBV = national EBV for fat percentage-based progeny information.

²GEN = model using the genotypes at the causative mutation in the DGAT1 gene (marker 7).

³GS = model using haplotypes of markers, with GS_m using markers *m* to 32.

brackets. The strong clustering of base haplotypes was partly due to the use of pedigree data in the calculation of IBD probabilities (Meuwissen and Goddard, 2001) and the fact that many of the base haplotypes were haplotypes of related animals or even half-sibs.

Predicted Breeding Values

The variance of the true breeding values for fat percentage, for the 1,135 bulls whose phenotypic records were omitted from one of the evaluations, was 0.133%² (Table 1). The variance of the predicted breeding values was 0.077%² for GS1, 0.072%² for GEN, and 0.034%² for BLUP (Table 1). For all alternatives, the mean difference between the true breeding values and the predictions was close to zero (Table 2). The residual variance, however, was 0.058%² for GS1, 0.059%² for GEN, and 0.099%² for BLUP (Table 2). The residual variances for GS1 and GS7 were equal (0.058%²), whereas the residual variance for GS15, GS18, and GS22 were

Table 2. Statistics of the difference between true breeding values¹ and the predictions from gene-assisted selection (GEN),² genomic selection (GS),³ and BLUP (n = 1,135)

Method	Mean	SD	Variance	Minimum	Maximum
GEN	-0.005	0.243	0.059	-1.09	0.67
GS1	-0.004	0.240	0.058	-0.89	0.71
GS7	-0.006	0.241	0.058	-0.86	0.71
GS15	-0.006	0.249	0.062	-0.92	0.70
GS18	-0.006	0.252	0.064	-1.05	0.70
GS22	-0.005	0.255	0.065	-1.01	0.70
GS25	-0.009	0.279	0.078	-1.16	0.81
BLUP	-0.014	0.314	0.099	-1.02	0.81

¹National EBV for fat percentage-based progeny information.

²GEN = model using the genotypes at the causative mutation in the DGAT1 gene (marker 7).

³GS = model using haplotypes of markers, with GS_m using markers *m* to 32.

slightly higher, from 0.062 to 0.065%². The residual variance for GS25 was clearly higher than for the other GS alternatives (0.078%²) but still lower than BLUP (0.099%²). The variance of the predictions plus the residual variance was 0.13%² for all alternative methods, which equals the variance of the true breeding values.

The correlation between true breeding values and predictions from a certain alternative method can be regarded as the pseudo-accuracy of that method (i.e., how well it can predict the EBV that is based on the performance of a large progeny group). The pseudo-accuracy for GEN was slightly lower than for GS1, 0.746 vs. 0.752, because of animals with unknown genotype at the causative mutation in the DGAT1 gene (Table 3). These animals had poor predictions in GEN and better predictions in GS1, because GS1 uses information from all markers. When considering only animals with known genotype at the causative mutation in the DGAT1 gene (n = 1,091), the pseudo-accuracy was 0.763 for GEN, whereas for the other alternative methods, the pseudo-accuracy was as presented in Table 3. Among the GS alternatives, the pseudo-accuracy was highest for GS1 (0.752) and also high for GS7, GS15, GS18, and GS22 (0.715 to 0.751). The pseudo-accuracy for GS25 was clearly lower (0.643), and it was lowest for BLUP (0.508). Note that the pseudo-accuracy for BLUP is lower than for parent averages in many other BLUP evaluations, because the data set included only phenotypic information on genotyped bulls. The correlations among the predictions were higher than 0.97 between GS1 and GS7 and among GS15, GS18, and GS22, whereas the correlations between GS25 and the other GS alternatives were from 0.85 to 0.90. Correlations between BLUP and the other methods were from 0.67 to 0.83.

For GEN, the average effect of marker genotype AA and KK was -0.351 and +0.352, respectively, which can be regarded as the estimated allele substitution effect (Table 4). The standard deviation within each genotype was very low (0.009), which indicates that the estimates were very consistent across the 20 evaluations for GEN. For GS1 and GS7, the sum of the haplotype effects over all brackets was on average very similar to the estimates from GEN, but the standard deviation within a genotype was higher, from 0.045 to 0.118 (Table 4). This shows that, on average, animals with a given genotype at the QTL get the same effect from their haplotypes as in GEN, but some animals were underestimated, whereas others were overestimated. For GS15, GS18, and GS22, the means were closer to zero, and the standard deviations were higher. For GS25, the means were reduced further (i.e., -0.153 for genotype AA and +0.198 for genotype KK). The correlation between the sum of the allele effects, as estimated in GEN,

Table 3. Correlations among true breeding values (TBV)¹ and predictions from gene-assisted selection (GEN),² genomic selection (GS),³ and BLUP (n = 1,135)

Method	GEN	GS1	GS7	GS15	GS18	GS22	GS25	BLUP
TBV	0.746 ⁴	0.752	0.751	0.731	0.722	0.715	0.643	0.508
GEN		0.944	0.942	0.881	0.876	0.871	0.832	0.694
GS1			0.983	0.933	0.924	0.916	0.859	0.674
GS7				0.936	0.927	0.917	0.861	0.680
GS15					0.981	0.972	0.882	0.700
GS18						0.989	0.888	0.709
GS22							0.895	0.717
GS25								0.826

¹TBV = national EBV for fat percentage-based progeny information.

²GEN = model using the genotypes at the causative mutation in the DGAT1 gene (marker 7).

³GS = model using haplotypes of markers, with GS_m using markers *m* to 32.

⁴The correlation between TBV and GEN was 0.763 when only considering bulls with known genotype.

and the estimated haplotype effects, summed over all brackets, was 0.937 for GS1, 0.935 for GS7, 0.829 for GS15, 0.821 for GS18, 0.812 for GS22, and 0.748 for GS25. This shows that the estimated haplotype effects of GS1 and GS7 corresponded well with the allele effects estimated in GEN, whereas the accumulated haplotype effects of GS25 had the greatest differences with GEN.

Table 4. Mean and standard deviation of the estimated haplotype effects summed over all brackets,¹ according to the genotype at the causative mutation in the DGAT1 gene

Method	Mean ²		
	AA (431 bulls)	AK (516 bulls)	KK (144 bulls)
GEN ³	-0.351	0.000	0.352
GS1 ⁴	-0.353	0.000	0.343
GS7	-0.343	0.000	0.350
GS15	-0.281	0.000	0.308
GS18	-0.265	0.000	0.303
GS22	-0.259	0.000	0.288
GS25	-0.152	0.000	0.198
Method	SD		
	AA (431 bulls)	AK (516 bulls)	KK (144 bulls)
GEN	0.009	0.009	0.009
GS1	0.047	0.114	0.079
GS7	0.045	0.118	0.053
GS15	0.137	0.137	0.099
GS18	0.133	0.136	0.101
GS22	0.132	0.135	0.112
GS25	0.095	0.104	0.106

¹Each animal had 1 observation, obtained from the evaluation in which its phenotype was omitted from the data.

²Expressed as deviation from genotype AK.

³GEN = model using the genotypes at the causative mutation in the DGAT1 gene (marker 7); sum of paternal and maternal marker effect is presented [instead of sum of haplotype effects for genomic selection (GS)].

⁴GS = model using haplotypes of markers, with GS_m using markers *m* to 32.

Posterior QTL Probability

In 18 out of 20 evaluations for GS1, 1 bracket out of the first 5 brackets had a posterior QTL probability above 0.95 (i.e., in more than 95% of the Gibbs samples in the stationary phase, a QTL was sampled in that marker bracket), whereas the other 4 brackets had a posterior QTL probability below 0.01. The flanking markers of the first 5 brackets were very close to each other (0.01 cM) but more than 3 cM apart from the causative mutation in the DGAT1 gene. Brackets 6 and 7, which have the causative mutation in the DGAT1 gene as a flanking marker, had posterior QTL probabilities lower than 0.08 in all evaluations. The posterior QTL probability was on average 0.34 in bracket 23 and 0.18 in bracket 10. In the larger brackets (i.e., where the distance between the flanking markers was higher than 0.10 cM), the posterior QTL probability was lower than 0.01 in all evaluations. Figure 2 has the posterior QTL probability, averaged over 20 evaluations, for alternative methods GS1, GS7, GS15, GS18, GS22, and GS25. In 19 out of 20 evaluations for GS7, 1 bracket out of brackets 7 to 14 had a high posterior QTL probability (>0.80). For GS15, the average posterior QTL probability was high in brackets 15 and 23. The posterior QTL probabilities for GS18 and GS22 only showed a peak in bracket 23, whereas GS25 only had a peak in bracket 31.

DISCUSSION

The estimated allele substitution effect at the causative mutation of the DGAT1 gene was 0.35% fat. Grisart et al. (2002) estimated an allele substitution effect of $0.17 \pm 0.012\%$ fat using daughter yield deviations of Dutch Holstein-Friesian bulls, which are approximately half breeding values when the number of daughters is large. After multiplication with 2, their estimate

Figure 2. Posterior QTL probabilities, averaged over 20 evaluations, for genomic selection (GS) alternatives GS1, GS7, GS15, GS18, GS22, and GS25, where GS m used markers m to 32.

is very similar to that obtained in this study. Thaller et al. (2003) found an allele substitution effect of 0.28% fat in German Holstein cattle, which is lower than in our study, but an allele substitution effect of 0.35% fat in German Fleckvieh, which does correspond to our result. Bennewitz et al. (2004) found an allele substitution effect of 0.26% fat in German Holstein-Friesian cattle but also concluded that the K232A mutation in the DGAT1 gene is not responsible for all genetic variation at the proximal end of bovine chromosome 14. Kühn et al. (2004) found allele substitution effects of 0.34 and 0.28% fat based on German Holstein-Friesian bulls with heterozygous (A/K) and homozygous (A/A) sires, respectively (effects were multiplied with 2, because daughter yield deviations were used). Kühn et al. (2004) also found a significant effect of the variable number of tandem repeat polymorphism in the promoter region of the DGAT1 gene for animals that were homozygous (A/A) for the K232A mutation in the DGAT1 gene. One allele, with an allele frequency of 16% in the German Holstein-Friesian population (based on maternally inherited alleles), was only found in combination with allele A at the K232A mutation, and this third allele had an allele substitution effect of 0.06% fat, besides the effect of the K232A mutation (effect was multiplied by 2). Kühn et al. (2004) argued that the variable number of tandem repeat polymorphism is a causative mutation with a direct effect on the expression of the

DGAT1 gene. The bulls in our study were not genotyped for this polymorphism, but given the similarity of the population history, it is expected that this third allele in the DGAT1 gene is also present in our data. In GS, the effects of IBD chromosome segments are estimated, which may account for more than 2 QTL alleles, whereas GEN assumes that the K232A mutation is the only source of variation in the DGAT1 gene. Kaupe et al. (2007) found an allele substitution effect of 0.28% fat in German Holstein cattle but also found evidence for a second QTL on chromosome 14 with an allele substitution effect of 0.04% fat. This second QTL, which is in the CY11B1 gene, may be modeled by the haplotypes of bracket 23 in our study, where the average posterior QTL probability was 0.34 in GS1 and above 0.70 in GS15, GS18, and GS22. According to the National Center for Biotechnology Information bovine sequence map, version 3.1 (www.ncbi.nlm.nih.gov), the locations of the DGAT1 gene and the CY11B1 gene differ 2.80 Mbp, which corresponds to 3.5 cM. The flanking markers of bracket 23, however, are positioned at 4.18 and 4.19 cM of the DGAT1 gene (see Figure 1).

Methods GS1 and GS7 explained the effect of the mutation in the DGAT1 gene as well as method GEN, as shown by the variance of the predicted breeding values, the residual variance, the correlation between the predicted breeding values and the phenotypes, and the average sum of the haplotype effects for bulls with

identical genotype at the causative mutation of the DGAT1 gene. In many evaluations of GS1 or GS7, a QTL was modeled in a bracket that did not have the causative mutation in the DGAT1 gene as a flanking marker. From this result, it may be concluded that for GS, it is not necessary to find the QTL that causes the variation. Furthermore, if the distance between the causative mutation in the DGAT1 gene and the closest marker was from 0.5 to 4.1 cM (GS15, GS18, and GS22), the correlation between predicted breeding value and phenotype was only reduced by 3 to 5%, and the residual variance was only increased by 7 to 12%. This suggests that for prediction of breeding values using whole-genome markers, the effect of a QTL may still be picked up even if the closest marker is not within 0.5 cM of distance. However, that conclusion may not hold in other situations, for example, in which the effect of the QTL is much smaller or when only SNP markers are used instead of a combination of SNP and microsatellite markers or when the marker density is lower. Therefore, we do not draw conclusions with respect to the number of markers needed for GS.

To test whether the results shown in this study would also be found if the QTL had a much smaller effect on the trait, the analyses GEN and GS1 were also performed for protein production instead of fat percentage, because the effect of the causative mutation in the DGAT1 gene, relative to the total genetic variance, is much smaller for this trait. The allele substitution effect (A to K), obtained from GEN, was -4.6 kg of protein, which is consistent with other studies [e.g., Grisart et al. (2002) -5.6 kg, Thaller et al. (2003) -4.8 to -5.2 kg, Bennewitz et al. (2004) -4.9 kg, and Kaupe et al. (2007) -3.8 kg]. The correlation between predicted and true breeding values was 0.59 in GEN, 0.59 in GS1, and 0.56 in BLUP, so also for kilograms of protein GS performed as well as GEN. This shows that the results found in this study also apply to QTL with much smaller effects. The correlation is lower than for fat percentage, because the causative mutation in the DGAT1 gene explains less of the genetic variance for kilograms of protein than for fat percentage (Bennewitz et al., 2004).

The bulls, whose true breeding values were omitted from 1 of the 20 evaluations, were sons of 27 grandsires with, on average, 42 sons per grandsire. Their breeding values were predicted, whereas the true breeding values of 19 out of each 20 paternal half-sibs were still used in the model. In a more practical situation, young selection candidates will not have paternal half-sibs with breeding values based on daughter performance. Based on linkage and LD information, however, the paternal haplotypes may have high IBD probabilities with haplotypes of other animals that do have pheno-

typic information, so their effects can still be estimated accurately.

Genomic selection aims at capturing all genetic variance using dense markers across the whole genome. In this study, however, we used only markers from 1 chromosome segment that contained an important QTL. To explain all genetic variance, or at least a large proportion as in Meuwissen et al. (2001), dense markers are needed across the whole genome. Furthermore, to also capture the variation from many very small QTL, many phenotypes were required, because their effects may otherwise be too small to detect. The results on fat percentage and protein production, however, show that our conclusions hold for QTL with intermediate to large effects.

The model of Meuwissen and Goddard (2004) uses matrices of IBD probabilities among haplotypes based on linkage and LD information, whereas the Bayesian approach in Meuwissen et al. (2001) uses 2-marker haplotypes with IBD matrices. When the extent of LD between markers and QTL is not very high, it is advantageous to use haplotypes with more markers and IBD matrices, because then haplotypes that have identical markers but are not IBD at the QTL can obtain different estimates. Second, linkage information can be used by including pedigree information in the estimation of IBD probabilities. If the extent of LD between markers and QTL is very high, however, the advantage of using IBD matrices based on linkage and LD becomes small, as shown in a simulation study [M. P. L. Calus, A. P. W. de Roos, R. F. Veerkamp, and T. H. E. Meuwissen (Univ. Life Sciences, Dept. Anim. Aquacult. Sci., Ås, Norway); unpublished data]. The disadvantage of using IBD matrices is the computation time for estimating the IBD probabilities among base haplotypes. In this study, the average time to calculate 1 IBD matrix was 2 min on a Sun Fire V40z server, but for a data set with 2,446 animals, 2,393 base haplotypes, and 198 markers on 1 chromosome, this increased to 57 min per marker bracket (A. P. W. de Roos; unpublished data). Although there are strategies to limit the computation time, it will be very challenging to implement the model of Meuwissen and Goddard (2004) for GS with tens of thousands of markers and thousands of animals. The computation time for 25,000 iterations with the Gibbs sampler was, on average, 18 min in this study and 20 h for a data set with 2,446 animals and 2,755 genome-wide markers (A. P. W. de Roos; unpublished data). An extension to tens of thousands of markers is expected to increase the computing time for the Gibbs sampler to a few days, because the size of the IBD matrices can be reduced further by clustering when a higher marker density is used.

The posterior QTL probabilities showed that once a QTL was sampled in a certain bracket, it hardly ever moved to another bracket, whereas another Gibbs chain may sample the QTL in another bracket. Furthermore, the posterior QTL probabilities were not a reliable estimate for the position of the QTL. For GS that may not be a problem, because 1 or more neighboring brackets may absorb more of the QTL variance than the bracket that actually contains the QTL. This study shows that finding a causative mutation underlying a trait is very complicated, which underscores the benefit of GS, which does not require that the QTL are known, as opposed to GEN.

CONCLUSIONS

The multiple QTL model of Meuwissen and Goddard (2004) was successfully applied to a real dairy cattle data set on 1 chromosome as a method to predict breeding values for GS. The accuracy of predicted breeding values was equal between GEN and GS, even though GS did not reveal the exact position of the actual QTL in this study. A small reduction in accuracy was observed when the markers closest to the QTL were omitted from the model, but this reduction was larger when the distance between the QTL and the closest marker was more than 8 cM. It is concluded that GS is an attractive alternative to GEN for breeding programs, because it does not require the discovery of the causative QTL.

REFERENCES

- Andersson, L., and M. Georges. 2004. Domestic-animal genomics: Deciphering the genetics of complex traits. *Nat. Rev. Genet.* 5:202–212.
- Bennewitz, J., N. Reinsch, S. Paul, C. Looft, B. Kaupe, C. Weimann, G. Erhardt, G. Thaller, Ch. Kühn, M. Schwerin, H. Thomsen, F. Reinhardt, R. Reents, and E. Kalm. 2004. The DGAT1 K232A mutation is not solely responsible for the milk production quantitative trait locus on the bovine chromosome 14. *J. Dairy Sci.* 87:431–442.
- Dekkers, J. C. M. 2004. Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. *J. Anim. Sci.* 82(E. Suppl.):E313–E328.
- Dekkers, J. C. M., and F. Hospital. 2002. The use of molecular genetics in the improvement of agricultural populations. *Nat. Rev. Genet.* 3:22–32.
- Farnir, F., B. Grisart, W. Coppieters, J. Riquet, P. Berzi, N. Cambisano, L. Karim, M. Mni, S. Moiso, P. Simon, D. Wagenaar, J. Vilkkki, and M. Georges. 2002. Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: Revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. *Genetics* 161:275–287.
- Fernando, R. L., and M. Grossman. 1989. Marker-assisted selection using best linear unbiased prediction. *Genet. Sel. Evol.* 21:467–477.
- Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation of the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Res.* 12:222–231.
- Haley, C. S., and P. M. Visscher. 1998. Strategies to utilize marker-quantitative trait loci associations. *J. Dairy Sci.* 81(Suppl. 2):85–97.
- Kaupe, B., H. Brandt, E.-M. Prinzenberg, and G. Erhardt. 2007. Joint analysis of the influence of CYP11B1 and DGAT1 genetic variation on milk production, somatic cell score, conformation, reproduction, and productive lifespan in German Holstein cattle. *J. Anim. Sci.* 85:11–21.
- Kühn, C., G. Thaller, A. Winter, O. R. P. Bininda-Emonds, B. Kaupe, G. Erhardt, J. Bennewitz, M. Schwerin, and R. Fries. 2004. Evidence for multiple alleles at the DGAT1 locus better explains a quantitative trait locus with major effect on milk fat content in cattle. *Genetics* 167:1873–1881.
- Meuwissen, T. H. E., and M. E. Goddard. 1996. The use of marker haplotypes in animal breeding schemes. *Genet. Sel. Evol.* 28:161–176.
- Meuwissen, T. H. E., and M. E. Goddard. 2001. Prediction of identity by descent probabilities from marker-haplotypes. *Genet. Sel. Evol.* 33:605–634.
- Meuwissen, T. H. E., and M. E. Goddard. 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet. Sel. Evol.* 36:261–279.
- Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829.
- Olsen, H. G., S. Lien, M. Gautier, H. Nilsen, A. Roseth, P. R. Berg, K. K. Sundsaasen, M. Svendsen, and T. H. E. Meuwissen. 2005. Mapping of a milk production trait locus to a 420-kb region on bovine chromosome 6. *Genetics* 169:275–283.
- Thaller, G., W. Krämer, A. Winter, B. Kaupe, G. Erhardt, and R. Fries. 2003. Effects of DGAT1 variants on milk production traits in German cattle breeds. *J. Anim. Sci.* 81:1911–1918.
- Winter, A., W. Krämer, F. A. O. Werner, S. Kollers, S. Kata, G. Durstewitz, J. Buitkamp, J. E. Womack, G. Thaller, and R. Fries. 2002. Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA:diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content. *Proc. Natl. Acad. Sci. USA* 99:9300–9305.