J. Dairy Sci. 95:4114–4129
http://dx.doi.org/10.3168/jds.2011-5019

Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels

M. Erbe,*[1] B. J. Hayes,†‡§[1,2] L. K. Matukumalli,# S. Goswami,‖ P. J. Bowman,†‡ C. M. Reich,†‡ B. A. Mason,†‡ and M. E. Goddard†¶

*Department of Animal Sciences, Animal Breeding and Genetics Group, Georg-August-University Göttingen, 37075 Göttingen, Germany
†Biosciences Research Division, Department of Primary Industries, Bundoora, Victoria 3083, Australia
‡Dairy Futures Cooperative Research Centre, Victoria 3083, Australia
§La Trobe University, Bundoora, Victoria 3086, Australia
#Bovine Functional Genomics Laboratory, USDA, Beltsville, MD 20705
‖Bioinformatics and Computational Biology, George Mason University, Manassas, VA 20110
¶Faculty of Land and Food Resources, University of Melbourne, Parkville, Victoria, 3010, Australia

ABSTRACT

Achieving accurate genomic estimated breeding values for dairy cattle requires a very large reference population of genotyped and phenotyped individuals. Assembling such reference populations has been achieved for breeds such as Holstein, but is challenging for breeds with fewer individuals. An alternative is to use a multibreed reference population, such that smaller breeds gain some advantage in accuracy of genomic estimated breeding values (GEBV) from information from larger breeds. However, this requires that marker-quantitative trait loci associations persist across breeds. Here, we assessed the gain in accuracy of GEBV in Jersey cattle as a result of using a combined Holstein and Jersey reference population, with either 39,745 or 624,213 single nucleotide polymorphism (SNP) markers. The surrogate used for accuracy was the correlation of GEBV with daughter trait deviations in a validation population. Two methods were used to predict breeding values, either a genomic BLUP (GBLUP_mod), or a new method, BayesR, which used a mixture of normal distributions as the prior for SNP effects, including one distribution that set SNP effects to zero. The GBLUP_mod method scaled both the genomic relationship matrix and the additive relationship matrix to a base at the time the breeds diverged, and regressed the genomic relationship matrix to account for sampling errors in estimating relationship coefficients due to a finite number of markers, before combining the 2 matrices. Although these modifications did result in less biased breeding values for Jerseys compared with an unmodified genomic relationship matrix, BayesR gave the highest accuracies of GEBV for the 3 traits investigated (milk yield, fat yield, and protein yield), with an average increase in accuracy compared with GBLUP_mod across the 3 traits of 0.05 for both Jerseys and Holsteins. The advantage was limited for either Jerseys or Holsteins in using 624,213 SNP rather than 39,745 SNP (0.01 for Holsteins and 0.03 for Jerseys, averaged across traits). Even this limited and nonsignificant advantage was only observed when BayesR was used. An alternative panel, which extracted the SNP in the transcribed part of the bovine genome from the 624,213 SNP panel (to give 58,532 SNP), performed better, with an increase in accuracy of 0.03 for Jerseys across traits. This panel captures much of the increased genomic content of the 624,213 SNP panel, with the advantage of a greatly reduced number of SNP effects to estimate. Taken together, using this panel, a combined breed reference and using BayesR rather than GBLUP_mod increased the accuracy of GEBV in Jerseys from 0.43 to 0.52, averaged across the 3 traits.

Key words: genomic selection, multiple breeds

INTRODUCTION

To accurately predict genomic breeding values for selection candidates with no phenotype of their own, a very large reference population of genotyped and phenotyped individuals is required to derive the prediction equation (Goddard, 2009; VanRaden et al., 2009; Brøndum et al., 2011). Although this has been achieved for breeds such as Holstein-Friesian dairy cattle in some countries (e.g., Wiggans et al., 2011), for smaller breeds, assembling such large reference populations is likely to be very challenging (particularly for breeds with limited numbers of progeny-tested sires available for use in the reference population).

Received October 4, 2011.
Accepted February 27, 2012.
[1]These authors contributed equally to this manuscript.
[2]Corresponding author: Ben.hayes@dpi.vic.gov.au

An alternative is to use a multi-breed reference population, such that the total number of individuals in the reference set is large. For this strategy to actually increase the accuracy of genomic estimated breeding values (**GEBV**) within a breed requires 1) sufficiently dense markers such that the associations between the marker alleles and the alleles at the QTL affecting the traits are consistent across breed and 2) at least a proportion of the QTL segregating in several of the breeds. de Roos et al. (2008) demonstrated that associations between alleles of pairs of SNP (using 1 SNP as a surrogate for a QTL) were conserved across Holstein, Jersey, and Angus populations, provided that markers were <10 kb apart. They concluded that to find markers that are in linkage disequilibrium with QTL across diverged breeds, such as Holstein, Jersey, and Angus, would require approximately 300,000 markers. The Bovine HapMap Consortium (Gibbs et al., 2009) reached a similar conclusion, demonstrating that among *Bos taurus* breeds, associations between alleles at different SNP were 90% conserved across breed provided the SNP were less than 10 kb apart. In a simulated data set with the same level of linkage disequilibrium both within and across breeds as observed for real Holstein and Jersey populations, de Roos et al. (2009) demonstrated that the most accurate genomic predictions were achieved when phenotypes from all populations were combined in 1 reference set, provided the marker density was sufficiently high (equivalent to a marker every 10 kb).

In real data, marker density has been limited to a marker approximately every 60 kb (approximately 50,000 SNP genome wide, termed **50K**). In a multi-breed beef cattle population, Kizilkaya et al. (2010) demonstrated limited across population predictive ability using these 50K SNP. Hayes et al. (2009a) and Pryce et al. (2011) both demonstrated very limited or no increase in accuracy of genomic predictions using these SNP with combined Holstein Jersey, and Holstein, Jersey and Fleckvieh dairy cattle reference populations, respectively.

With the recent development of an approximately 777K bovine array [Illumina Bovine high density (HD); Illumina Inc., San Diego, CA], the hypothesis that the accuracy of genomic predictions for some breeds can be improved by using a multi-breed reference population, provided marker density is sufficiently high, can be tested.

One challenge here is that a very large number of animals have been genotyped with 50K, and are unlikely to be regenotyped with the approximately 777K SNP. In this study, we explore imputation of genotypes (e.g., Browning and Browning, 2009; Marchini and Howie, 2010) as an efficient strategy to derive a large reference set with 800K genotypes.

We then explore alternative methods for deriving the SNP prediction equation. A widely used method for genomic prediction is genomic BLUP (**GBLUP**; e.g., VanRaden, 2008; Goddard, 2009), in which the expected relationship matrix among the animals in the population is replaced with the realized relationship matrix (or genomic relationship matrix) derived from markers. An approach is outlined for calculating the genomic relationship matrix, which takes into account both the inbreeding since the breeds diverged from a common population, and the inbreeding that has occurred since the founders of the pedigree used to derive the expected relationship matrix. This allows the genomic relationship matrix and expected relationship matrix to be combined to maximize the accuracy of prediction. Further, with such dense SNP data, an efficient strategy may be to allow a proportion of SNP to be removed from the prediction model. We outline a new computationally efficient method that allows this.

MATERIALS AND METHODS

Data

The Illumina Bovine SNP50v2.0 and BovineHD chips were used to genotype the animals. The bovine BeadChips were processed by following the Infinium protocol from Illumina, and the BeadChips were scanned using the iScan scanner. The raw data was analyzed using GenomeStudio software.

Two genotype data sets were used in this study. The first was heifers and bulls genotyped with the Illumina High-Density Bovine SNP chip (which we will call the 800K panel). The second data set was 2,257 Holstein and 540 Jersey Bulls genotyped with the Illumina Bovine 50K array (which we will call the 50K panel; Matukumalli et al., 2009).

For the first genotype data set, 903 Holstein-Friesian heifers from a feed conversion efficiency trial (Pryce et al., 2012), 93 Holstein-Friesian key ancestor bulls, and 93 key ancestor Jersey bulls were genotyped with the Illumina High-Density Bovine SNP chip, which has 777,963 SNP markers. The SNP positions used were from UMD 3.1 (University of Maryland, College Park, MD). Stringent quality control procedures were applied to the data. These included the use of the Illumina GenCall score, which describes the performance of genotyping each SNP in each individual. From previous experience, genotype calls with GenTrain score (GenCall) >0.6 are high quality; below this value they were excluded. There were 650,934 SNP genotyped at Gen-

Call >0.6. Furthermore, 343 mitochondrial SNP, 1,124 Y chromosome SNP, and 1,735 unmapped SNP were excluded. Some 55 SNP with duplicate map positions were removed so 625,925 SNP remained. Forty-eight individuals with fewer than 90% of SNP genotyped at GenCall <0.6 were removed. Across the remaining samples, 99.6% of SNP were genotyped at GenCall >0.6. Animals with excess heterozygosity (>0.4) were removed, as this is a good indicator of sample contamination. Five animals were identified with heterozygosity above this threshold; however, all of these had already been removed in the step above (i.e., >90% of SNP genotyped). The final stage of filtering was for SNP with very low minor allele frequency (SNP with less than 10 copies of the rare allele in the population were removed). An additional filter was imposed to filter SNP with low imputation accuracy; this is described below.

In the second set of animals (2,797 Holstein and Jersey progeny-tested bulls), genotyped for the 50K panel, quality filters were imposed as described in Hayes et al. (2009a). Further, SNP that were not on the 800K panel after quality control in the data set were removed, leaving 39,745 SNP of the 50K panel. Mendelian consistency checks were performed on both 50K and 800K data, and genotypes failing Mendelian consistency checking were set to missing.

Phenotype data for the 2,797 bulls were daughter trait deviations (**DTD**; e.g., VanRaden and Wiggans, 1991) for milk yield, fat yield, and protein yield, from single-trait models.

Imputation

Imputation of the 50K data set to 800K genotypes was performed with BEAGLE software (Browning and Browning, 2009). Prior to this step, cross-validation was used to assess the accuracy of imputation that could be achieved. The Holstein heifers that were genotyped for the 800K panel were split into 2 subsets at random. In the second split, the genotypes were cut down to the 39,745 SNP on the 50K panel. Imputation was then performed, and the accuracy of imputation was taken as the proportion of genotypes that were correctly imputed. This process was then repeated, but using the second split to impute into the first split. To assess the value of having key ancestors genotyped on the 800K panel for imputation, both runs were repeated with the key ancestors 800K genotypes added.

For the Jerseys, there were only 93 key ancestor bull genotypes for the 800K panel. The accuracy of imputation was assessed using cross-validation again, but dropping 20 bulls at random as the set with 39,745 SNP. This was performed 5 times.

It became obvious as a result of imputing 50K to 800K in the Holstein heifer data cross-validations that a small number of SNP (1,231) were imputed very inaccurately, with accuracy across animals below 80% (Figure 1). Accuracy here is defined as the proportion of genotypes that are correctly imputed. We postulated that these SNP could be mismapped. We attempted to remap the SNP using linkage disequilibrium information. For each of the 1,231 SNP, the R^2 with all the other 624,924 SNP was calculated using genotype frequencies as described by Zaykin et al. (2008). The weighted (by distance from the center of the window) average R^2 was calculated in 20 SNP windows across the genome. If the window with the highest average R^2 with the remapped SNP was greater than 1,000 kb different to the position in the original map file, the new position of the SNP being remapped was at the center of the window with the highest weighted average R^2 value. This algorithm is implemented in ldMapper, a program available from the authors.

The imputation was performed again using the proposed new positions of the SNP (Supplementary Material Table S1, available in the online version of this paper). This greatly improved the accuracy of imputation for 601 of the SNP; however, 630 of the SNP were still poorly imputed (Figure 1). These were removed from the data set, giving a final data set of 624,213 SNP for the 800K panel. The cross-validations described above were redone to get the final results. The 800K panel genotypes (actually 624,213 SNP) were then imputed into the 50K bull data set.

Finally, as the BEAGLE imputation as implemented here does not use pedigree information, we tested for Mendelian inconsistencies in the post-BEAGLE (imputed) 800K genotypes. We found that a small proportion of SNP genotypes were inconsistent in sire-son comparisons (e.g., opposing homozygotes), amounting to 0.6% of the genotypes.

Transcriptome Panel

To test both the hypothesis that mutations affecting quantitative traits reside in exons, introns, and regulatory regions, and to potentially reduce the computational demand when calculating genomic predictions, we tested another panel of SNP that were in the 624,213 above (800K panel) and also within or near the transcribed part of the genome. The start-stop positions of the transcribed part of the genome were as defined by L. K. Matukumalli (author on the current paper), plus SNP within 1 kb of these stop or start positions. The transcribed part of the genome was identified from a large collection of mRNA transcripts, mapped to the UMD 3.0 bovine assembly (http://www.cbcb.umd.

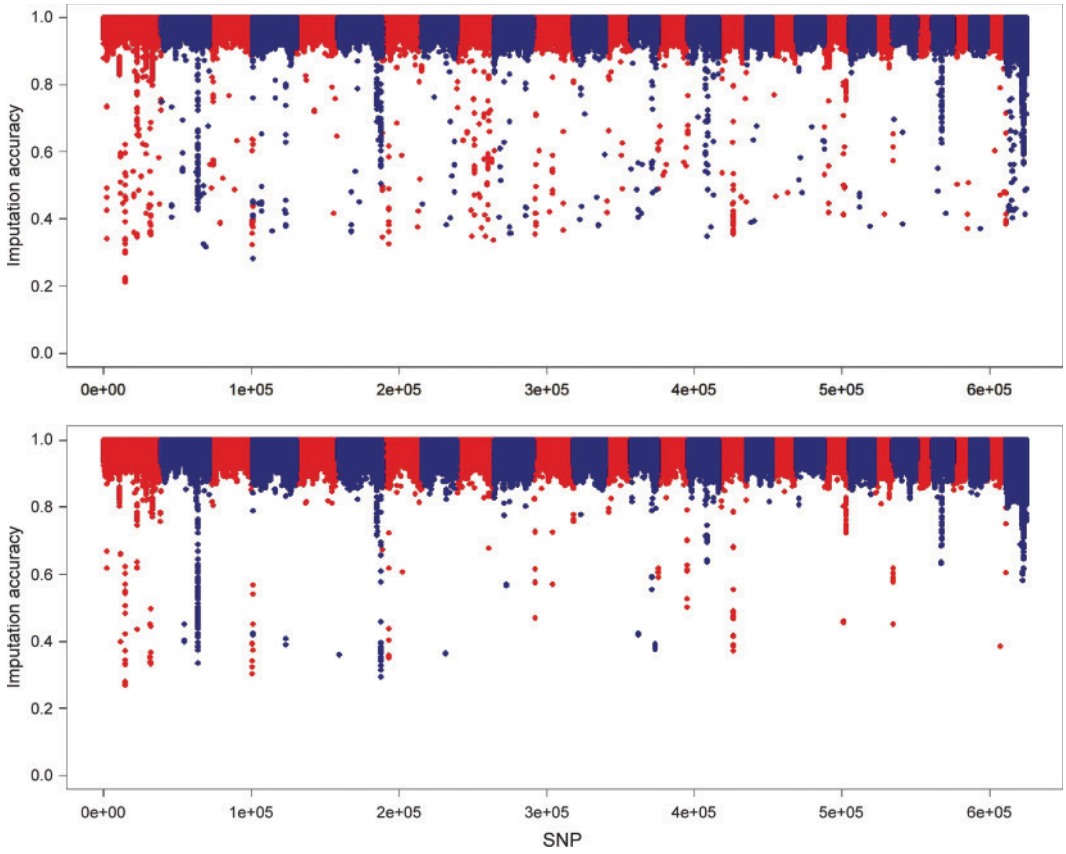

Figure 1. Accuracy of imputation by SNP using BEAGLE software (Browning and Browning, 2009), before and after remapping 1,231 SNP with <80% accuracy of imputation in the original data set. Single nucleotide polymorphisms with <80% accuracy of imputation were remapped using linkage disequilibrium (LD), with the new position taken as the position that gave the highest LD in a window of 20 SNP, with all genome positions considered. Color version available in the online PDF.

edu/research/bos_taurus_assembly.shtml). This panel (which we will call the transcriptome panel, **TRANS**) consisted of 58,532 SNP.

Methods for Genomic Prediction

The bulls in each breed were split into reference bulls (those progeny tested before 2007) and validation bulls (those progeny tested in 2007 or later). There were 1,897, 360, 454, 86 Holstein reference, Holstein validation, Jersey reference, and Jersey validation bulls, respectively. Unless otherwise described, the reference set combined both the Holstein and Jersey reference bulls. The surrogate used for accuracy of GEBV was the correlation of GEBV and DTD in the validation bulls. This surrogate was not corrected for the reliability of the DTD (which averaged 0.8 in the validation sets). The regression of GEBV on DTD was also calculated. For each method, the SNP subsets used were 50K, 800K, and TRANS panels.

The methods used to predict GEBV were as follows.

GBLUP. The following model was fitted to the data

$$\mathbf{y} = \mathbf{1_n}\mu + \mathbf{Zg} + \mathbf{e},$$

where $\mathbf{y}$ is a vector of phenotypes, $\mathbf{1_n}$ is a vector of 1s, μ is an overall mean, $\mathbf{Z}$ is a design matrix allocating records to breeding values, $\mathbf{g}$ is a vector of genomic breeding values, and $\mathbf{e}$ is a vector of random normal deviates with variance $V(\mathbf{e}) \sim N(0, \sigma_e^2)$, where σ_e^2 is the error variance. The variance of breeding values was $V(\mathbf{g}) = \mathbf{G}\sigma_g^2$, where $\mathbf{G}$ is the genomic relationship matrix derived as in Yang et al. (2010), with no consideration of breed, and σ_g^2 is a genetic variance. Then, breeding values for both phenotyped and nonphenotyped individuals can be predicted as

$$[\hat{\mathbf{g}}] = \left[\mathbf{Z'Z} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_g^2}\right]^{-1}\left[\mathbf{Z'}\left(\mathbf{y} - \mathbf{1}_n\hat{\mu}\right)\right],$$

where $\hat{\mathbf{g}}$ is a vector of EBV, $\mathbf{Z}'$ is the transpose of $\mathbf{Z}$, and $\hat{\mu}$ is an estimate of the mean.Variance components were estimated with ASReml software (Gilmour et al., 2002).

GBLUP_mod. Goddard et al. (2011) argued that EBV, and particularly the accuracy derived from the coefficient matrix from GBLUP, are biased due to sampling errors in elements of the genomic relationship matrix due to a finite number of markers, where the expectation of $\mathbf{G}$, with G defined as above, given the estimate of G ($\hat{\mathbf{G}}$), $E(\mathbf{G}|\hat{\mathbf{G}}) \neq \hat{\mathbf{G}}$. This also means, for example, that information from the expected relationship matrix ($\mathbf{A}$) derived from pedigree and $\mathbf{G}$ cannot be combined to maximize the accuracy of the EBV in a one-step approach (e.g., Misztal et al., 2009).

Goddard et al. (2011) suggested a new genomic relationship matrix that regressed elements of $\mathbf{G}$ toward $\mathbf{A}$ to account for sampling error in estimating coefficients of $\mathbf{G}$ to create a new matrix $\mathbf{G^*}$:

$$\mathbf{G}^* = [\mathbf{A} + b(\mathbf{G} - \mathbf{A})], \quad [1]$$

where

$$b = V(\mathbf{G})/[V(\mathbf{G}) + 1/m] \quad [2]$$

and $V(\mathbf{G})$ is the variance of the nondiagonal elements of $\mathbf{G}$ obtained with m markers; $V(\mathbf{G})$ can be obtained simply by taking all of the nondiagonal elements of $\mathbf{G}$, where $\mathbf{G}$ is calculated as in Yang et al. (2010) and calculating the variance of these elements.

To derive $\mathbf{G^*}$ for a multi-breed population, an appropriate base population relative to which $\mathbf{G}$ and $\mathbf{A}$ are both defined must be chosen. One logical base population in our situation is that immediately before the divergence of Holsteins and Jerseys.

First, a $\mathbf{G}$ matrix can be calculated, which records covariances relative to a base that is a composite breed (c) made up of a proportion of α Holsteins and (1 – α) Jerseys.

$$\mathbf{G}_c = \mathbf{WW}'/\mathbf{M},$$

where $\mathbf{W}$ is a centered matrix calculated as $\mathbf{W} = \mathbf{X} - 2\mathbf{p}$, with $\mathbf{p} = \alpha\mathbf{p}_{hol} + (1 - \alpha)\mathbf{p}_{jer}$, and $\mathbf{M} = 2\sum_{i=1}^{m} p_i\left(1 - p_i\right)$. Here, $\mathbf{p}_{hol}$ and $\mathbf{p}_{jer}$ are the average allele frequencies of the 2 allele in Holsteins and Jerseys, respectively; $\mathbf{X}$ is a matrix of animals by SNP, with SNP genotypes coded 0 = 11, 1 = 12, or 2 = 22; $\alpha = \dfrac{F_{jer}}{F_{jer} + F_{hol}}$, with F_{jer} and F_{hol} defined below; and p_i is the frequency of the 2 allele for the ith SNP. The calculation of $\mathbf{G}_c$ is similar to that described by VanRaden (2008) for a purebred population but with a modification to the allele frequencies to scale $\mathbf{G}$ to the composite base. Our approach is different from that of Harris and Johnson (2010), who also derived $\mathbf{G}$ for a multibreed population. They used the approach of partitioning the diagonals of the matrix into breed fractions to account for different variances among breeds and include segregation variances because of different allele frequencies among breeds. However, their approach will accommodate crossbred animals; ours would need to be extended to do this.

Then, in our approach $\mathbf{G}_c$ is adjusted for the inbreeding that has occurred in both breeds relative to the old base (the base at the divergence of Holsteins and Jerseys):

$$\mathbf{G} = \mathbf{G}_c\,(1 - F) + 2F,$$

where F is the inbreeding relative to an *F1* base:

$$F = \frac{F_{jer}F_{hol}}{F_{jer} + F_{hol}},$$

$$F_{jer} = 1 - \frac{\sum_{i=1}^{m} 2p_{jer,i}(1 - p_{jer,i})}{\sum_{i=1}^{m}\left[p_{hol,i}(1 - p_{jer,i}) + p_{jer,i}(1 - p_{hol,i})\right]},$$

and

$$F_{hol} = 1 - \frac{\sum_{i=1}^{m} 2p_{hol,i}(1 - p_{hol,i})}{\sum_{i=1}^{m}\left[p_{hol,i}(1 - p_{jer,i}) + p_{jer,i}(1 - p_{hol,i})\right]}.$$

The pedigree-derived $\mathbf{A}$ must also be converted to the old base (e.g., Powell et al. 2010). For the within-Holstein blocks, $\mathbf{A} = \mathbf{A}_{ped}\,[1 - (F - f_{hol})] + 2\,(F - f_{hol})$, where f_{hol} is the amount of inbreeding that has occurred since the base of the pedigree within Holsteins; we approximated this as the average of the off-diagonal elements of $\mathbf{A}_{ped}$. The within-Jersey block was constructed in the same way. All elements of the Holstein × Jersey block of A were 0, as no pedigree links existed between the breeds. Note that in practice, the estimate of f_{hol} and f_{jer} could be an underestimate due to the incompleteness of the pedigree. With an incomplete pedigree the base is less well defined.

Once **G** and **A** were constructed, the regression of **G** toward **A** to account for sampling errors in the genomic relationship coefficients (Equation 1) was determined. This was done separately for each breed, and the breed × breed block (e.g., Holstein × Jersey) by calculating the variance of the off-diagonal elements within each of these blocks.

BayesR. The GBLUP approaches assume that all markers have a small effect and that these effects are normally distributed (e.g., Habier et al., 2007; Hayes et al., 2009b). Given the large number of markers, a more appropriate prior may be that some of the markers are not in linkage disequilibrium with QTL, so have zero effect, whereas others have a small to moderate effect. This prior was proposed by Meuwissen et al. (2001). The challenge of implementing a method that uses such a mixture prior is computational efficiency—for example, in the BayesB of Meuwissen et al. (2001), sampling of SNP variances from their posterior distributions simultaneously with the SNP effects required a Metropolis Hastings algorithm. Verbyla et al. (2009) described a stochastic search variable selection (**BayesSSVS**) strategy, which maintained the same assumptions about the distributions of SNP effects while maintaining constant dimensionality, which allowed a Gibbs sampling scheme to be used to construct the posterior distributions of the parameters. However, one potential criticism of both BayesB and BayesSSVS is that the proportion of SNP in each distribution was not sampled appropriately, such that the means of the posterior distributions of the proportion of SNP with a zero or nonzero effect closely reflected the prior values of these proportions (e.g., "lack of Bayesian learning"; Habier et al., 2011). Here, both to overcome this drawback of BayesB and BayesSSVS, and for computational efficiency, we propose a new method that assumes that the true SNP effects are derived from a series of normal distributions, the first with zero variance, up to one with a variance of approximately 1% of the genetic variance.

The model fitted to the data was

$$\mathbf{y} = \mathbf{1_n}'\mu + \mathbf{Wu} + \mathbf{Zv} + \mathbf{e},$$

where $\mathbf{y}$ is a vector of n DTD for each trait; $\mathbf{W}$ is the $(n \times m)$ design matrix allocating records to the marker effects described above; vector $\mathbf{u}$ is a $(m \times 1)$ vector of SNP effects assumed normally distributed $\left[u_i \sim N\left(0, \sigma_i^2\right)\right]$; $\mathbf{e}$ is a vector of random deviates, where σ_e^2 is the error variance; v_j is the polygenic breeding value of the jth animal, $V\left(v\right) = \mathbf{A}\sigma_a^2$, where $\mathbf{A}$ is the average relationship matrix; σ_a^2 is the polygenic variance; and $\mathbf{Z}$ is a matrix that allocates records to animals.

The variance of the ith SNP effect had 4 possible values:

$$\sigma_1^2 = 0, \sigma_2^2 = 0.0001\sigma_g^2, \sigma_3^2 = 0.001\sigma_g^2, \sigma_4^2 = 0.01\sigma_g^2,$$

where σ_g^2 is the assumed total genetic variance, which was calculated as $\sigma_g^2 = r_{DTD}^2\sigma_{DTD}^r$, with r_{DTD}^2 being the assumed reliability of the DTD, and σ_{DTD}^2 the variance of the DTD. Using these variances results in shrinkage that allows the SNP effects themselves to range from zero effect to moderate effect. The proportions of the SNP in each distribution were pr1, pr2, pr3, and pr4, respectively, in a vector **pr**.

Bayesian estimation of the parameters was used. The prior distribution of the proportions of SNP in each distribution **pr** was the Dirichlet distribution, with $\boldsymbol{\alpha} = \mathbf{1}$ (where $\boldsymbol{\alpha}$ is a 4×1 vector of pseudo counts, all with value 1 to give an almost uninformative prior with the numbers of SNP used here). The Dirichlet distribution is a convenient choice of prior, as it is a conjugate before the multinomial distribution, such that the posterior distribution of **pr** is $\sim Dir(\boldsymbol{\alpha} + \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a vector containing the number of SNP in each distribution estimated from the data. To obtain these estimates, we first calculated 4 likelihoods assuming the considered SNP being in 1 of the 4 normal distributions at a time with the respective probability pr_k. The likelihood that SNP i is in distribution k is

$$LogL(i,k) = -0.5\log\left|\mathbf{V}\right| - 0.5(\mathbf{y}^{*\prime}\mathbf{y}^* - \mathbf{y}^{*\prime}\mathbf{Z}^*\hat{\mathbf{u}})/\sigma_e^2 + \log(pr_k),$$

where $\mathbf{y}^*$ is the vector of phenotypes corrected for all marker effects other than marker i, the overall mean, and the polygenic effects $\left(\hat{\mathbf{u}}\right)$; $\mathbf{Z}^*$ is a column vector containing the SNP genotypes of all animals for SNP i; $\mathbf{V}$ is the variance-covariance structure of a reduced model, including only the effect of the respective SNP and a residual effect; and $\log|\mathbf{V}|$ was calculated as $n\log\left(\sigma_e^2\right) + \log\left(\sigma_i^2\mathbf{Z}^{*\prime}\mathbf{Z}^*/\sigma_e^2 + 1\right)$, where $\mathbf{Z}^*$ contains only the information for the current SNP effect.

Then, the probability that SNP i is in distribution k is

$$\frac{1}{\sum_{l=1}^{4}\exp\left[L(i,l) - L(i,k)\right]}.$$

Based on these probabilities, we selected the normal distribution to sample the SNP effect from using a uniform random variate, using the probabilities of the SNP being in each of the distributions for the current iteration. Over all the SNP, we thus obtained estimates for the elements of $\boldsymbol{\beta}$.

The posterior of **pr** cannot be estimated directly, as it is conditional on both the estimates of the SNP effects (to calculate $\mathbf{y}^*$) and estimates of the polygenic effects $\hat{\mathbf{u}}$. A Gibbs sampling scheme was, therefore, used to sample from the posterior distributions of all parameters conditional on the other parameters.

Prior distributions for other parameters were as described by Verbyla et al. (2009). The Gibbs sampling scheme was similar to that described by Meuwissen et al. (2001) for BayesA, but with the addition of a polygenic effect, and with the SNP variances described above. At the end of each iteration, the proportion of SNP in each distribution was sampled from the posterior Dirichlet distribution as described above. We also compared r(GEBV, DTD) from GBLUP_mod and BayesR to those derived from SNP effects estimated by BayesA (Meuwissen et al., 2001).

RESULTS

Accuracy of Imputation

In the Holstein heifer data set, the accuracy of imputation of 50K to 800K was similar across the 2 cross-validations, with an average of 97.4% (Table 1). Adding the key ancestor 800K genotypes improved the accuracy of imputation by 0.5%, despite the limited number of these ancestors. The average accuracy of imputation in the Jersey cross-validations was lower, likely reflecting the much more limited number of animals genotyped for the 800K panel.

Comparison of GBLUP and GBLUP_mod

To check that the proposed modifications to the **G** matrix and **A** matrix in the GBLUP_mod method resulted in relationship matrices expressed relative to the same base population, before $\mathbf{G}^*$ was calculated, we checked the average of the diagonal elements for each breed, and the average off-diagonal elements within and across breeds. These were very close (Table 2). The regressor $\hat{b}$ of G toward A, which accounts for sampling error in estimating the coefficients of **G** is also given for each block. Within a breed, the value of $\hat{b}$ was only slightly less than 1; however, in the across-breed block, the value of $\hat{b}$ was lower at 0.89, reflecting the fact that across-breed genomic relationships are smaller in magnitude, and are estimated with lower precision than within-breed genomic relationships. However, the value of 0.89 is surprisingly high, and may reflect the fact that the Australian dairy herd was upgraded from a largely Jersey base, such that relatively large chromosome segments originating from Jerseys can still be found in cattle classified as Holstein.

Next, we evaluated the effect of using GBLUP_mod rather than GBLUP on the accuracy and bias of GEBV. For the 800K panel, the accuracy of GEBV [as indicated by the surrogate measure r(GEBV, DTD)] from GBLUP and GBLUP_mod was similar for the Holstein validation data set, but, on average, 0.03 higher for GBLUP_mod in the Jersey validation data set (Table 3). Regressions of DTD on GEBV were closer to 1 in

Table 1. Accuracy of imputation of 50,000 to 800,000 SNP (50K to 800K) in cross-validation of 940 Holstein and 93 Jersey genotypes[1]

Genotype	Cross-validation	Accuracy of genotype imputation (%)
Holstein		
Heifers only	1	97.4
	2	97.9
	Average	97.7 ± 0.01
Heifers + key ancestors	1	98.0
	2	98.0
	Average	98.0 ± 0.05
Jersey		
	1	96.1
	2	95.0
	3	97.0
	4	95.4
	5	94.2
	Average	95.6 ± 0.05

[1]Cross-validation in the Holstein data set involved splitting the 843 heifers in 2 approximately equal subsets, and then in silico reducing the numbers of genotypes to the 50K panel. In Jerseys, approximately 20 individuals at each cross-validation were assigned to have their genotypes reduced to the 50K panel.

Table 2. Average of elements of expected and realized relationship matrices (**A** and **G**, respectively), after rescaling to a base that was at the time of divergence of Holsteins and Jerseys[1]

Statistic	Matrix elements	Validation	**A**	**G**	$\hat{b}$
Average	Diagonal	Holstein	1.09	1.11	—
Average	Diagonal	Jersey	1.20	1.22	—
Average	Block	Holstein	0.20	0.19	0.96
Average	Block	Jersey	0.42	0.39	0.97
Average	Block	Across breed	0.00	0.01	0.89

[1]The regressor $\hat{b}$ of G toward A, which accounts for sampling error in estimating the coefficients of **G** is given for each block.

the Jersey validation data sets with GBLUP_mod than with GBLUP in all traits.

Comparison of GBLUP_mod, BayesR, and BayesA for the 800K Panel

Table 4 shows the results for BayesR, BayesA, and GBLUP_mod for the combined reference population and the 800K panel. The BayesR method gave higher r(GEBV, DTD) for both milk yield and fat yield than GBLUP_mod, whereas r(GEBV, DTD) for protein yield was similar. Averaged across the traits, the advantage of BayesR over GBLUP_mod was 0.05 in r(GEBV, DTD). This advantage was observed in both the Holstein and the Jersey validation data set. The regression of DTD on GEBV (Table 4) was similar for all methods.

To compare BayesR with a well-known Bayesian method, we also ran BayesA. BayesA gave similar, but very slightly lower r(GEBV, DTD) for milk yield than BayesR and similar results in terms of slope [b(DTD, GEBV)].

Comparison of Different Marker Panels

For genomic predictions within a pure breed, there was no advantage of either the 800K or TRANS panel over the 50K panel when GBLUP_mod was used (Table 5). When BayesR was used, there was only a very small advantage (and not significant), given the sample size used, in r(GEBV, DTD) of using the 800K or the TRANS panel over the 50K panel in some cases (Table 6). This was of the order of 0.01 averaged across traits for Holsteins, comparing the 800K to the 50K panel, and 0.02 for Jerseys comparing the TRANS panel to the 50K panel (Table 7).

Some improvement for prediction across breeds occurred using only the other breed as the reference when BayesR was used with either the 50K or the 800K panel, compared with the GBLUP_mod results. For

Table 3. Correlations of daughter trait deviations (DTD) and genomic EBV {GEBV; [r(GEBV, DTD)]} and regressions of DTD on GEBV slopes [b(DTD, GEBV)] from genomic BLUP (GBLUP) and GBLUP_mod methods[1]

		Trait			
Method	Validation	Milk yield	Fat yield	Protein yield	Average
r(GEBV, DTD)					
GBLUP	Holstein	0.58	0.58	0.56	0.57
	Jersey	0.33	0.46	0.40	0.40
GBLUP_mod	Holstein	0.58	0.58	0.56	0.57
	Jersey	0.36	0.49	0.44	0.43
b(DTD,GEBV)					
GBLUP	Holstein	1.04	1.19	0.94	1.05
	Jersey	0.53	0.86	0.71	0.70
GBLUP_mod	Holstein	1.04	1.16	0.93	1.04
	Jersey	0.69	1.00	0.94	0.88

[1]The GBLUP_mod method uses a rescaled genomic relationship matrix, and regresses the **G** matrix toward the **A** matrix to account for the error in estimating realized relationship coefficients due to a finite number of markers.

Table 4. Accuracy of prediction [expressed as r(GEBV, DTD)] and slopes [b(DTD, GEBV)] of the regression of daughter trait deviations (DTD) on predicted genomic EBV (GEBV) for GBLUP_mod, BayesA, and BayesR with a multi-breed reference population and the 800,000-SNP (800K) panel (the result averaged across traits is also given)

Method[1]	Validation	Milk yield	Fat yield	Protein yield	Average
r(GEBV,DTD)					
GBLUP_mod	Holstein	0.58	0.58	0.56	0.57
	Jersey	0.36	0.49	0.44	0.43
BayesA	Holstein	0.61	0.66	0.58	0.62
	Jersey	0.48	0.49	0.46	0.48
BayesR	Holstein	0.62	0.66	0.57	0.62
	Jersey	0.51	0.49	0.46	0.49
b(DTD,GEBV)					
GBLUP_mod	Holstein	1.04	1.16	0.93	1.04
	Jersey	0.69	1.00	0.94	0.88
BayesA	Holstein	1.04	1.12	0.94	1.03
	Jersey	0.82	0.91	0.86	0.86
BayesR	Holstein	0.99	1.12	0.91	1.01
	Jersey	0.84	0.92	0.86	0.88

[1]The GBLUP_mod method uses a rescaled genomic relationship matrix, and regresses the **G** matrix toward the **A** matrix to account for the error in estimating realized relationship coefficients due to a finite number of markers; BayesR is a Bayesian method for deriving the prediction equation that assumes SNP effects follow a series of normal distributions; and BayesA is a Bayesian method for deriving the prediction equation that assumes SNP effects follow a Student's t distribution. Complete descriptions are given in the text.

the TRANS panel, the accuracy for predicting Jersey GEBV from a Holstein-only reference looked promising (0.24 average across traits; Table 7). Interestingly, the r(GEBV, DTD) for milk yield was much higher (0.40 and 0.30, respectively) with both methods when the TRANS panel was used compared with both other panels (Tables 5 and 6).

When a combined reference set was used, BayesR clearly outperformed GBLUP_mod across all scenarios and traits, especially with prediction of fat yield in Holstein (up to 0.08 higher) and milk yield (0.15 higher) and protein yield (0.05 higher) in Jersey in all panels.

The best results for predicting the minor breed (Jerseys) were obtained with a combined reference set, BayesR and the TRANS panel [r(GEBV, DTD) = 0.52; Table 7]. This was 0.09 higher than that obtained using GBLUP_mod, the combined reference set, and the 800K panel (Table 4).

Distribution of SNP Effects

For BayesR, we could calculate the number of SNP in each distribution (explaining 0, 0.01, 0.1, or 1% of the genetic variance). This was achieved by calculating the posterior mean of the sampled proportions of SNP in each of the 4 distributions over all post burn-in iterations, and multiplying them by the total number of SNP. The results show that, on average, only between 7 and 14% (depending on trait) of all SNP contribute to the prediction of genomic breeding value with the

Table 5. Accuracy of genomic prediction [r(GEBV, DTD)] from GBLUP_ mod[1] using different marker panels and either single-breed or combined reference populations[2]

		Milk yield			Fat yield			Protein yield		
Reference	Validation	50K	800K	TRANS	50K	800K	TRANS	50K	800K	TRANS
Holstein	Holstein	0.61	0.58	0.62	0.58	0.57	0.57	0.57	0.56	0.55
	Jersey	−0.07	−0.01	0.30	−0.24	−0.16	−0.05	−0.31	−0.21	0.05
Jersey	Holstein	0.04	−0.03	0.03	0.16	0.18	0.11	0.14	0.16	0.08
	Jersey	0.38	0.37	0.39	0.49	0.48	0.47	0.43	0.43	0.42
Combined	Holstein	0.60	0.58	0.62	0.58	0.58	0.57	0.57	0.56	0.55
	Jersey	0.35	0.36	0.45	0.47	0.49	0.44	0.40	0.44	0.48

[1]The GBLUP_mod method uses a rescaled genomic relationship matrix, and regresses the **G** matrix toward the **A** matrix to account for the error in estimating realized relationship coefficients due to a finite number of markers.

[2]GEBV = genomic EBV; DTD = daughter trait deviations; 50K = 50,000-SNP panel; 800K = 800,000-SNP panel; TRANS = transcriptome panel.

Table 6. Accuracy of genomic prediction [r(GEBV, DTD)] from BayesR[1] using different marker panels and either single-breed or combined reference populations[2]

		Milk yield			Fat yield			Protein yield		
Reference	Validation	50K	800K	TRANS	50K	800K	TRANS	50K	800K	TRANS
Holstein	Holstein	0.62	0.63	0.63	0.64	0.65	0.63	0.55	0.57	0.56
	Jersey	0.27	0.24	0.40	0.12	0.21	0.12	−0.05	0.05	0.21
Jersey	Holstein	0.19	0.03	0.15	0.29	0.29	0.18	0.13	0.10	0.12
	Jersey	0.49	0.48	0.53	0.48	0.46	0.47	0.42	0.41	0.43
Combined	Holstein	0.61	0.62	0.62	0.65	0.66	0.64	0.56	0.57	0.57
	Jersey	0.45	0.51	0.57	0.50	0.49	0.45	0.43	0.46	0.53

[1]BayesR is a Bayesian method for deriving the prediction equation that assumes SNP effects follow a series of normal distributions.
[2]GEBV = genomic EBV; DTD = daughter trait deviations; 50K = 50,000-SNP panel; 800K = 800,000-SNP panel; TRANS = transcriptome panel.

50K panel. Similar absolute numbers of SNP were in distribution 2, 3, and 4 with the 800K panel; that is, the majority of SNP with this panel (over 99%) were estimated to be in the first distribution, which had zero variance (Table 8).

When a combined (Holstein and Jersey) reference set was used, for all traits, the number of SNP in the 0.01 distribution was lower than or similar to the purebred Holstein scenario. For distribution 3, the number of SNP was clearly lower than when a single breed reference set was used, whereas it was usually higher for distribution 2. Possible reasons for this are proposed in the discussion.

In most cases, the number of SNP in distribution 1 and 2 was clearly lower for fat yield than for both of the other traits with all SNP panels. With the Jersey reference set, more SNP were assumed to explain larger parts of the total variance than with the Holstein reference set. For the TRANS panel, the number of SNP in distribution 1 and 2 could be expected to be higher, as the SNP for this panel were all located in or near transcribed regions. However, we did not observe this trend.

DISCUSSION

In this study, we tested 3 hypotheses: 1) the accuracy of genomic estimated breeding values would be increased using denser marker panels, when the validation animals and reference animals were the same breed, 2) the advantage of using a denser marker panel would be even greater when the validation animals and reference animals were from different breeds, or a combined breed reference set was used, and 3) a method for deriving the prediction equation that could result in a large number of SNP effects being set to zero (e.g., excluded from the prediction model) would result in the greatest advantage from increasing the density of the marker panel.

The support for hypothesis 1) was limited. The r(GEBV, DTD) for the Holstein population did increase when the 800K panel was used rather than the 50K panel, but only by 0.01 averaged across traits, and only when BayesR was used. For Jersey (using Jersey reference to predict GEBV in a Jersey validation set), the average r(GEBV, DTD) actually decreased by 0.01

Table 7. Accuracy of genomic prediction [r(GEBV, DTD)] from BayesR[1] using different marker panels and either single-breed or combined reference populations, averaged across traits

		Panel		
Reference	Validation	50K	800K	TRANS
Holstein	Holstein	0.61	0.62	0.61
	Jersey	0.11	0.17	0.24
Jersey	Holstein	0.20	0.14	0.15
	Jersey	0.46	0.45	0.48
Combined	Holstein	0.61	0.62	0.61
	Jersey	0.46	0.49	0.52

[1]BayesR is a Bayesian method for deriving the prediction equation that assumes SNP effects follow a series of normal distributions.
[2]GEBV = genomic EBV; DTD = daughter trait deviations; 50K = 50,000-SNP panel; 800K = 800,000-SNP panel; TRANS = transcriptome panel.

Table 8. Average number of SNP in the 4 normal distributions modeled with BayesR[1]

	Reference								
	Jersey			Holstein			Combined		
Panel[2]	Milk (kg)	Fat (kg)	Protein (kg)	Milk (kg)	Fat (kg)	Protein (kg)	Milk (kg)	Fat (kg)	Protein (kg)
50K									
1st	35,730	34,201	36,179	34,991	35,917	35,844	34,245	34,558	34,880
2nd	3,677	5,276	3,268	4,612	3,598	3,798	5,410	5,040	4,820
3rd	315	255	287	134	222	93	81	139	36
4th	24	13	10	8	8	10	9	7	8
800K									
1st	620,151	620,026	619,488	620,570	620,544	620,151	620,372	619,526	619,650
2nd	3,727	3,828	4,462	3,390	3,528	3,538	3,579	4,467	4,478
3rd	306	339	254	245	227	122	251	210	77
4th	29	20	9	9	13	9	11	10	8
TRANS									
1st	54,742	54,850	54,242	54,144	55,233	54,953	53,317	54,121	54,272
2nd	3,480	3,210	4,039	4,264	3,064	3,480	5,145	4,257	4,206
3rd	276	455	241	116	225	93	63	143	48
4th	34	17	10	7	11	7	7	11	6

[1]The average number of SNP was calculated as the mean proportion of SNP in the distribution times the total number of SNP. BayesR is a Bayesian method for deriving the prediction equation that assumes SNP effects follow a series of normal distributions.

[2]50K = 50,000-SNP panel; 800K = 800,000-SNP panel; TRANS = transcriptome panel.

when the 800K panel was used rather than the 50K. In contrast to humans where a very large number of SNP are necessary for accurate genomic predictions due to a large effective population size (e.g., Wray et al., 2007), in modern dairy cattle breeds effective population sizes are sufficiently small that linkage disequilibrium (**LD**) between SNP and potential QTL is captured even with the 50K panel, and increasing this LD by using a denser panel does not have much effect. Evidence for this is that the proportion of the genetic variance captured by the 50K panel is only slightly lower than that from the 800K panel (Table 9; Haile-Mariam et al., accepted), regardless of which method is used. In sharp contrast to what is observed in human populations, we were able to capture almost 90% of the heritability of our phenotype (DTD) estimated from pedigree with the markers; in human populations this figure is more like 56% for a trait such as human height (Yang et al., 2010). Interestingly, the proportion of variance unexplained with BayesR was greatest with fat yield. One explanation for this may be that the largest distribution from which SNP effects are sampled has a variance of 1%, resulting in overshrinking of the effect of DGAT1, such that less variance is explained.

For Jerseys, we must point out that our reference population was small; therefore, any potential advantage in using denser panels may be obscured by the estimation error associated with the greatly increased number of SNP. Further, for Jerseys, the imputation reference set (for imputation of 800K from 50K) com-

Table 9. Proportion of genetic variance (estimated from pedigree) unaccounted for by SNP markers, using the Holstein-only reference set[1]

		Trait		
Method[2]	Panel	Milk yield	Fat yield	Protein yield
GBLUP_MOD	50K	0.12	0.13	0.17
	800K	0.11	0.12	0.15
BayesR	50K	0.08	0.22	0.12
	800K	0.08	0.18	0.10

[1]For BayesR, this was calculated as the estimated polygenic variance from the model divided by the total genetic variance; for GBLUP_mod, it was calculated as the variance explained by the modified **G** matrix divided by the genetic variance estimated from a model with only a polygenic effect with co(variance) matrix the expected relationship matrix (**A**).

[2]GBLUP = genomic BLUP.

prised only 93 key ancestors, which led to clearly lower imputation accuracies than in Holsteins (Table 1). Inaccurate genotype imputation would have reduced the possible advantages of using the 800K panel (and a multi-breed reference population) for Jerseys.

Support for hypothesis 2) was a little more convincing; the average of r(GEBV, DTD) across traits in the Jersey validation set, with Holsteins used as the reference, increased from 0.11 (50K) to 0.17 (800K) when BayesR was used(Table 7). With 800K SNP, the persistence of phase among SNP and QTL alleles should be consistent across *B. taurus* breeds (Gibbs et al., 2009). However, this assumes the same QTL are segregating in the different breeds, whereas our results suggest this is only true in a proportion of cases, as discussed below.

There was some support for hypothesis 3). The greatest increase in r(GEBV, DTD) from using the 800K panel rather than the 50K panel were observed when BayesR was used rather than GBLUP_mod (for example, for prediction of Jersey GEBV from the combined reference population). These results suggest that to take advantage of the increased marker density, methods that either explicitly remove SNP from the model or set their effect to zero (2 ways of achieving the same thing) are necessary.

One possible explanation for our results (especially the limited gains in r(GEBV, DTD) from using 800K compared with 50K) is that we have greatly increased the number of SNP effects to be estimated, without increasing the number of records. Particularly the Jersey population is small, so that the effect of the large increase in the number of estimation errors could erode the accuracy of GEBV. An alternative to using all 800K SNP would be to select a much smaller subset that may be a priori more relevant, thus avoiding the need to estimate a very large number of SNP effects. For our TRANS panel, we selected a subset of SNP from the 800K that was included the transcribed portion of the genome (L. K. Matukumalli, author on the current paper). The TRANS panel worked reasonably well for all traits and led to similar or even better (e.g., in milk yield with BayesR) results than with both the other SNP panels. The average r(GEBV, DTD) for Jerseys was highest using this panel, and accuracies of across breed prediction using the other breed as reference set were quite promising.

Our results for the increase in accuracy for the minor breed (Jerseys) using a combined reference and the 800K panel can be compared with the simulated results from de Roos et al. (2009). The simulation those authors used to generate marker associations within and across breeds was based on actual LD within and across similar populations to those considered here. If the divergence time between Holsteins and Jerseys is taken at approximately 300 generations (e.g., de Roos et al., 2008), then their simulation results would suggest that the increase in the accuracy of genomic EBV for Jerseys, as a result of using the 800K panel and combining the reference populations, should have been considerably greater than was observed here. Some of the explanation may be due to too few records to accurately estimate the 800K marker effects, as described above, and imperfect imputation of 800K from 50K, particularly in Jerseys. However, de Roos et al. (2009) also simulated QTL that were segregating in both breeds in most cases. Our results suggest that only some of the QTL segregate across breed. For example, for milk yield, the 9 SNP in Holstein that explained 1% of the genetic variance according to their posterior mean from BayesR (Table 8) were tightly clustered in 3 regions, on chromosome 14 (DGAT1), chromosome 5, and chromosome 11. Although the QTL on chromosome 14 and chromosome 5 were detected in Jerseys (as evidenced by clusters of SNP in the fourth distribution of BayesR, explaining 1% of the variance, using a Jersey-only reference population), no evidence indicated that the QTL on chromosome 11 was segregating in Jerseys. Further, in Jerseys, QTL were affecting milk yield segregating on chromosomes 23 and 16 (again tracked by SNP with posterior means in the fourth distribution of BayesR), and these were not segregating in Holstein. This is a subject for further investigation, but these preliminary results suggest that roughly half the QTL explaining 1% of the genetic variance segregate across Jerseys and Holsteins.

An important question, given our results, is whether further increasing marker density (for example, through whole genome sequencing) will lead to more accurate genomic predictions than from the 50K panel. This question can only be answered once sufficient individual cattle genomes have been sequenced. However, a simulation study (Meuwissen and Goddard, 2010) did show that sequence data, where the actual mutation causing trait variation was included in the data set, led to an increase in the accuracy of GEBV of 3 to 5% over the densest marker panel they simulated. Perhaps even more importantly, the authors demonstrated that in their simulation, prediction equations derived from whole-genome sequence data will lead to a slower decrease in the accuracy of GEBV as the reference population and selection candidates are separated by more generations. This is in contrast to the accuracies of GEBV from the 50K panel in dairy cattle, which decrease rapidly with genetic distance of the target population from the reference population (Habier et al., 2010). A reduced decay in accuracy may also be

Table 10. Accuracy [r(GEBV, DTD)] for milk yield from BayesR and GBLUP_mod in the Holstein validation set bulls grouped according to whether or not they had a sire in the Holstein reference population[1]

	Method	
Panel[2]	BayesR	GBLUP_mod
50K with sire	0.64	0.61
50K without sire	0.55	0.56
800K with sire	0.64	0.60
800K without sire	0.57	0.51

[1]DTD = daughter trait deviations; GEBV = genomic EBV; GBLUP = genomic BLUP. BayesR is a Bayesian method for deriving the prediction equation that assumes SNP effects follow a series of normal distributions. The GBLUP_mod method uses a rescaled genomic relationship matrix, and regresses the **G** matrix toward the **A** matrix to account for the error in estimating realized relationship coefficients due to a finite number of markers.

[2]50K = 50,000-SNP panel; 800K = 800,000-SNP panel.

achieved with the 800K panel. We do not have the data to test this hypothesis. However, if we divide our validation data set into those bulls that do and do not have a sire in the Holstein reference population, and then compare r(GEBV, DTD) for milk yield for these 2 sets from the 50K and 800K panels, a slightly reduced decay in accuracy for the 800K panel compared with the 50K panel, for bulls with and without a sire (Table 10), was only observed when BayesR was used to derive the prediction equation. Results were similar for protein yield; however, for fat yield accuracies were actually higher for the group of validation bulls without sires in the reference. This could have been partially an effect of the DGAT1 mutation—closer inspection showed that the SNP tracking this mutation was at more intermediate frequency in the validation bulls with no sires in the reference, compared with those with sires in the reference. Our results here are only suggestive and would not be significant; more investigation of the effect of increasing marker density, with a greater range of relationship to the reference set, on the rate of decay of prediction accuracy is required.

Another potential advantage of using whole-genome resequencing data in prediction of GEBV may be the potential to capture low-frequency mutations that contribute to genetic variation. Allele frequencies of the SNP on the 50K panel are more or less distributed uniformly (i.e., it is a selection where SNP with very low minor allele frequency are underrepresented; e.g., Matukumalli et al., 2009). This is also true for the 800K data (data not shown). For high and stable LD between SNP and QTL, similar allele frequencies of the loci are necessary. Quantitative trait loci with low minor allele frequencies may thus not be in sufficient LD with a SNP and their variance cannot be captured. This may be one explanation why the difference in proportion of unaccounted genetic variance is small between the 50K and the 800K panel (Table 9). Note that for the 800K panel, animals in the reference set were not genotyped themselves, but imputed. Imputation of SNP with low minor allele frequency is more difficult than for SNP with moderate allele frequencies, which can also result in less accurate estimation of SNP effects and, consequently, missing parts of genetic variance. Whether or not resequencing allows some of these low-frequency variants to be captured will depend on how many animals are sequenced before imputation of sequence data in the reference population.

Regarding the 50K panel, several authors have presented studies analyzing real data sets with different methods for the estimation of the SNP effects. In most studies, accuracies achieved with BLUP approaches were very similar to those achieved with Bayesian methods (e.g., VanRaden et al., 2009). For prediction of a breed from a multi-breed reference set, BayesR performed best in our study. As described in previous studies (e.g., Hayes et al., 2010), the superiority of Bayesian approaches is generally greater in traits that are strongly influenced by a few moderate to large genes, which was also observed in our study (compare fat to protein). With GBLUP_mod, the variance assumed to be explained is the same for each SNP. Therefore, if more and more markers are used in the model, the expected variance per SNP will be smaller. When modeling traits with 1 or more underlying genes with larger effects, this can be the disadvantage when using GBLUP_mod in comparison to a Bayesian method (Meuwissen and Goddard, 2010). This theory would lead to the assumption that prediction with GBLUP will be even more disadvantageous when even more SNP are modeled simultaneously. In our study, we saw clearly better results with BayesR than with GBLUP_mod for the traits fat yield and milk yield, for all marker panels. However, we did not observe that the difference in accuracy between the methods was larger for the 800K panel.

There were generally fewer SNP in the third and fourth posterior distributions from the BayesR analysis, those with the largest variance, when a combined-breed reference was used compared with single-breed reference sets (Table 8). This may reflect the fact that many SNP are not in the same phase with QTL across breeds. Then, it could be expected that only the SNP having the same LD structure with the QTL in both breeds would have a moderate effect when the combined reference is used. Pryce et al. (2011) found that a more concentrated set of SNP or even a single SNP captured the effect of DGAT1 in a multi-breed reference population compared with pure-breed reference sets. Following the results of BayesR, which showed a

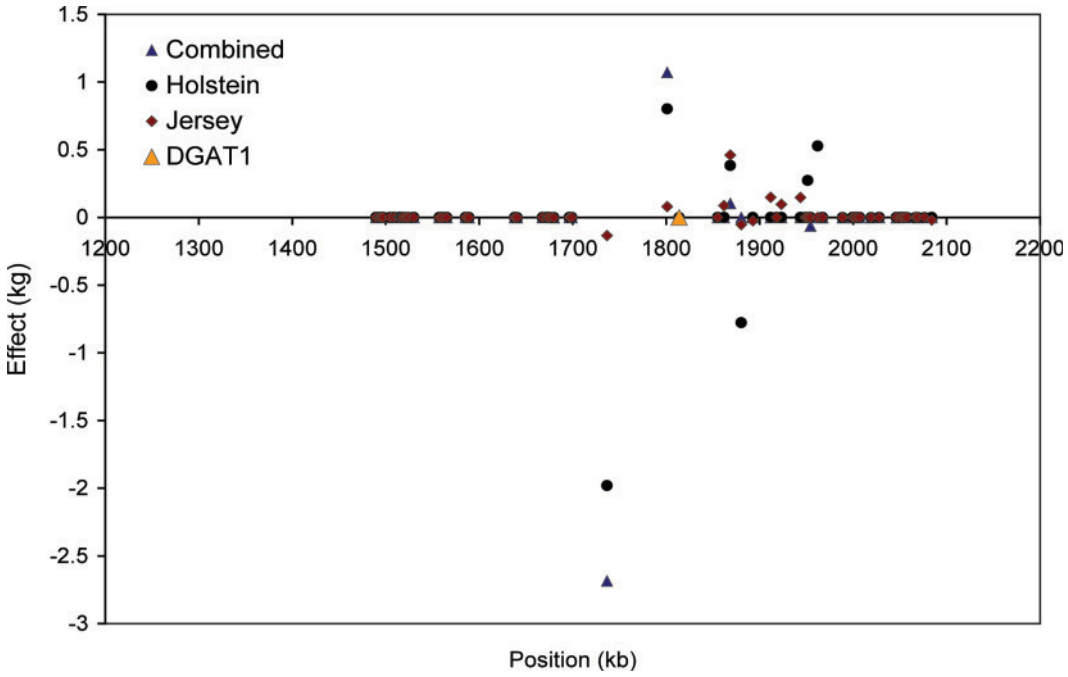

Figure 2. The effect of SNP on fat yield as estimated by a new method (BayesR), which used a mixture of normal distributions as the prior for SNP effects, including one distribution that set SNP effects to zero, from different reference populations in the DGAT1 region. Color version available in the online PDF.

decreased number of SNP explaining moderate parts of the variance in the multi-breed reference set for all traits, we also investigated the DGAT1 region and did find a decreased number of SNP capturing the DGAT1 effect when a combined reference set was used (Figure 2). Hayes et al. (2009a) concluded that a SNP capturing an effect in a multi-breed reference population must be very close to the potential QTL, as they have to be in high LD across breeds. Assuming that the more concentrated set of SNP with moderate effects implies the SNP are closer located to the QTL, the prediction accuracy will be more persistent over generations than with a purebred reference.

Finally, computer processing times for BayesR were reasonable, at 35 h and 20 min for BayesR with the multi-breed reference and the 800K panel (Table 11). Using the TRANS panel greatly decreased processing time for all methods, such that this could be applied in national evaluations for dairy cattle. A multi-threaded implementation of the construction of the **G** matrix for

Table 11. Processing time (clock time) for multi-breed reference population (2,351 bulls) with 3 SNP panels[1]

	SNP panel[3]		
Method[2]	50K	800K	TRANS
GBLUP_mod			
Build and invert **G**	2 min	39 min	3 min
ASReml (1 trait)	20 min	20 min	20 min
BayesA		30 h 55 min	
BayesR	1 h 54 min	35 h 50 min	3 h 5 min

[1]Processors were Intel Xeon X5670. For GBLUP_mod, multi-threading was used in the construction and inversion of the **G** matrix, across 10 threads.

[2]GBLUP = genomic BLUP; ASReml = ASReml software (Gilmour et al., 2002). The GBLUP_mod method uses a rescaled genomic relationship matrix, and regresses the **G** matrix toward the **A** matrix to account for the error in estimating realized relationship coefficients due to a finite number of markers; BayesR is a Bayesian method for deriving the prediction equation that assumes SNP effects follow a series of normal distributions; and BayesA is a Bayesian method for deriving the prediction equation that assumes SNP effects follow a Student's t distribution. Complete descriptions are given in the text.

[3]50K = 50,000-SNP panel; 800K = 800,000-SNP panel; TRANS = transcriptome panel.

GBLUP_mod decreased computing time from several days to 3 min.

CONCLUSIONS

In this study, we investigated different marker panels and methods for prediction of genomic breeding values within and across breeds. Two new or modified methods were presented: GBLUP_mod, which scales the genomic relationship matrix to an appropriate base and regresses **G** toward **A** to account for sampling error in estimation of within- and across-breed genomic relationships, and BayesR, which assumes that SNP effects follow a mixture of normal distributions, including a distribution with zero variance. Although the GBLUP_mod method resulted in less biased breeding values than using an unmodified **G** matrix, the BayesR method performed best in terms of r(GEBV, DTD) in most studied scenarios, and gave regressions of DTD on GEBV of close to 1. In addition to having the best predictive ability, BayesR also presents the possibility of using the results (splitting of SNP into different classes of explained variance) directly for further analyses of, for example, genetic architecture or for SNP selection of less computationally demanding subsets. An additional benefit of the denser marker set of the 800K panel could be seen neither for within- nor for across-breed prediction directly in terms of significant increase of accuracy. However, the 800K panel was the basis for an informative subset of SNP in transcribed parts of the genome, which may be a good alternative to modeling the large number of SNP directly from the 800K panel, balancing the extra genomic information from the 800K with the effect of increased estimation errors from a very large number of SNP in our admittedly small data sets. This panel (TRANS) in combination with BayesR and a combined reference set gave the highest accuracies of prediction in Jerseys, the minor breed in this study.

ACKNOWLEDGMENTS

Parts of the analyses were carried out during a research stay of M. Erbe at the Department of Primary Industries in Victoria, Australia. This research was funded by the German Federal Ministry of Education and Research (Bonn, Germany) within the AgroClustEr "Synbreed–Synergistic plant and animal breeding" (Funding identification: 0315526).

REFERENCES

Brøndum, R. F., E. Rius-Vilarrasa, I. Strandén, G. Su, B. Guldbrandtsen, W. F. Fikse, and M. S. Lund. 2011. Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. J. Dairy Sci. 94:4700–4707.

Browning, B. L., and S. R. Browning. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84:210–223.

de Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of genomic predictions across multiple populations. Genetics 183:1545–1553.

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus Cattle. Genetics 179:1503–1512.

Gibbs, R. A., J. F. Taylor, C. P. Van Tassell, W. Barendse, K. A. Eversole, C. A. Gill, R. D. Green, D. L. Hamernik, S. M. Kappes, S. Lien, L. K. Matukumalli, J. C. McEwan, L. V. Nazareth, R. D. Schnabel, G. M. Weinstock, D. A. Wheeler, P. Ajmone-Marsan, P. J. Boettcher, A. R. Caetano, J. F. Garcia, O. Hanotte, P. Mariani, L. C. Skow, T. S. Sonstegard, J. L. Williams, B. Diallo, L. Hailemariam, M. L. Martinez, C. A. Morris, L. O. C. Silva, R. J. Spelman, W. Mulatu, K. Zhao, C. A. Abbey, M. Agaba, F. R. Araujo, R. J. Bunch, J. Burton, C. Gorni, H. Olivier, B. E. Harrison, B. Luff, M. A. Machado, J. Mwakaya, G. Plastow, W. Sim, T. Smith, M. B. Thomas, A. Valentini, P. Williams, J. Womack, J. A. Woolliams, Y. Liu, X. Qin, K. C. Worley, C. Gao, H. Jiang, S. S. Moore, Y. Ren, X.-Z. Song, C. D. Bustamante, R. D. Hernandez, D. M. Muzny, S. Patil, A. San Lucas, Q. Fu, M. P. Kent, R. Vega, A. Matukumalli, S. McWilliam, G. Sclep, K. Bryc, J. Choi, H. Gao, J. J. Grefenstette, B. Murdoch, A. Stella, R. Villa-Angulo, M. Wright, J. Aerts, O. Jann, R. Negrini, M. E. Goddard, B. J. Hayes, D. G. Bradley, M. Barbosa Da Silva, L. P. L. Lau, G. E. Liu, D. J. Lynn, F. Panzitta, and K. G. Dodds. 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science 324:528–532.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, S. J. Welham, and R. Thompson. 2002. ASReml User Guide. Release 1.0. VSN International Ltd., Hemel Hempstead, UK.

Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica 136:245–257.

Goddard, M. E., B. J. Hayes, and T. H. E. Meuwissen. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128:409–421.

Habier, D., R. L. Fernando, and J. C. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42:5.

Haile-Mariam, M., G. J. Nieuwhof, K. T. Beard, K. V. Konstantinov, and B. J. Hayes. Comparison of heritabilities of dairy traits in Australian Holstein-Friesian cattle from genomic and pedigree data and implications for genomic evaluations. J. Anim. Breed. Genet. (accepted).

Harris, B. L., and D. L. Johnson. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J. Dairy Sci. 93:1243–1252.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009a. Accuracy of genomic breeding values in multibreed dairy cattle populations. Genet. Sel. Evol. 41:51.

Hayes, B. J., J. Pryce, A. J. Chamberlain, P. J. Bowman, and M. E. Goddard. 2010. Genetic architecture of complex traits and accuracy of genomic prediction: Coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genet. 6:e1001139.

Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009b. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. (Camb.) 91:47–60.

Kizilkaya, K., R. L. Fernando, and D. J. Garrick. 2010. Genomic prediction of simulated multibreed and purebred performance using

observed fifty thousand single nucleotide polymorphism genotypes. J. Anim. Sci. 88:544–551.

Marchini, J., and B. Howie. 2010. Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 11:499–511.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4:e5350.

Meuwissen, T., and M. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185:623–631.

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.

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92:4648–4655.

Powell, J. E., P. M. Visscher, and M. E. Goddard. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Genet. 11:800–805.

Pryce, J. E., J. Arias, P. J. Bowman, S. R. Davis, K. A. Macdonald, G. C. Waghorn, W. J. Wales, Y. J. Williams, R. J. Spelman, and B. J. Hayes. 2012. Accuracy of genomic predictions of residual feed 14 intake and 250 day bodyweight in 15 growing heifers using 625,000 SNP markers. J. Dairy Sci. 95:2108–2119. http://dx.doi.org/10.3168/jds.2011-4628.

Pryce, J. E., B. Gredler, S. Bolormaa, P. J. Bowman, C. Egger-Danner, C. Fuerst, R. Emmerling, J. Sölkner, M. E. Goddard, and B. J. Hayes. 2011. Short communication: Genomic selection using a multi-breed, across-country reference population. J. Dairy Sci. 94:2625–2630.

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

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24.

VanRaden, P. M., and G. R. Wiggans. 1991. Derivation, calculation, and use of national animal model information. J. Dairy Sci. 74:2737–2746.

Verbyla, K. L., B. J. Hayes, P. J. Bowman, and M. E. Goddard. 2009. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genet. Res. (Camb.) 91:307–311.

Wiggans, G. R., P. M. VanRaden, and T. A. Cooper. 2011. The genomic evaluation system in the United States: Past, present, future. J. Dairy Sci. 94:3202–3211.

Wray, N. R., M. E. Goddard, and P. M. Visscher. 2007. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 17:1520–1528.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42:565–569.

Zaykin, D. V., A. Pudovkin, and B. S. Weir. 2008. Correlation-based inference for linkage disequilibrium with multiple alleles. Genetics 180:533–545.