J. Dairy Sci. 101:1–16
https://doi.org/10.3168/jds.2017-13366

Multibreed genomic prediction using multitrait genomic residual maximum likelihood and multitask Bayesian variable selection

M. P. L. Calus,*1 M. E. Goddard,†‡ Y. C. J. Wientjes,* P. J. Bowman,‡§ and B. J. Hayes§#
*Wageningen University & Research, Animal Breeding and Genomics, PO Box 338, 6700 AH Wageningen, the Netherlands
†Faculty of Veterinary and Agricultural Science, University of Melbourne, Melbourne, Victoria 3010, Australia
‡Agriculture Research, Department of Economic Development, Jobs, Transport and Resources, Melbourne, Victoria 3083, Australia
§School of Applied Systems Biology, La Trobe University, Bundoora, Victoria 3083, Australia
#Queensland Alliance for Agriculture and Food Innovation, Centre for Animal Science, The University of Queensland, St. Lucia, Queensland 4072, Australia

ABSTRACT

Genomic prediction is applicable to individuals of different breeds. Empirical results to date, however, show limited benefits in using information on multiple breeds in the context of genomic prediction. We investigated a multitask Bayesian model, presented previously by others, implemented in a Bayesian stochastic search variable selection (BSSVS) model. This model allowed for evidence of quantitative trait loci (QTL) to be accumulated across breeds or for both QTL that segregate across breeds and breed-specific QTL. In both cases, single nucleotide polymorphism effects were estimated with information from a single breed. Other models considered were a single-trait and multitrait genomic residual maximum likelihood (GREML) model, with breeds considered as different traits, and a single-trait BSSVS model. All single-trait models were applied to each of the 2 breeds separately and to the pooled data of both breeds. The data used included a training data set of 6,278 Holstein and 722 Jersey bulls, as well as 374 Jersey validation bulls. All animals had genotypes for 474,773 single nucleotide polymorphisms after editing and phenotypes for milk, fat, and protein yields. Using the same training data, BSSVS consistently outperformed GREML. The multitask BSSVS, however, did not outperform single-trait BSSVS, which used pooled Holstein and Jersey data for training. Thus, the rigorous assumption that the traits are the same in both breeds yielded a slightly better prediction than a model that had to estimate the correlation between the breeds from the data. Adding the Holstein data significantly increased the accuracy of the single-trait GREML and BSSVS in predicting the Jerseys for milk and protein, in line with estimated correlations between the breeds of 0.66 and 0.47 for milk and protein yields, whereas only the BSSVS model significantly improved the accuracy for fat yield with an estimated correlation between breeds of only 0.05. The relatively high genetic correlations for milk and protein yields, and the superiority of the pooling strategy, is likely the result of the observed admixture between both breeds in our data. The Bayesian model was able to detect several QTL in Holsteins, which likely enabled it to outperform GREML. The inability of the multitask Bayesian models to outperform a simple pooling strategy may be explained by the fact that the pooling strategy assumes equal effects in both breeds; furthermore, this assumption may be valid for moderate- to large-sized QTL, which are important for multibreed genomic prediction.
Key words: genomic prediction, multibreed, Bayesian variable selection

INTRODUCTION

One of the benefits of genomic prediction is that it can use information across groups of individuals, such as different livestock breeds, which are not connected through any recent pedigree links. Considering the hypothesis that genomic prediction relies on linkage disequilibrium (**LD**) between SNP and QTL (Meuwissen et al., 2001), the expectation was that genomic prediction across breeds would be possible if the SNP density was large enough. This expectation was supported by the supposition that genomic prediction across Holsteins and Jerseys would be possible if the number of SNP was greater than 300,000 (de Roos et al., 2008). This, however, was based on simulations that assumed the QTL underlying the traits of interest are the same and have the same effects among different breeds.

Several empirical studies have shown that the accuracy of multibreed, compared with single-breed, genomic prediction is, at best, slightly higher, but often remains unchanged or is even slightly lower when breeds are

Received June 20, 2017.
Accepted January 4, 2018.
[1]Corresponding author: mario.calus@wur.nl

distantly related (Erbe et al., 2012; Karoui et al., 2012; Olson et al., 2012; Lund et al., 2014; Zhou et al., 2014). In situations where breeds are closely related, increases in accuracy from multibreed genomic prediction are more easily obtained (Brøndum et al., 2011), especially if the initial training data of the predicted breed is small (Hozé et al., 2014b). One possible explanation for the limited success of multibreed genomic prediction is that the genetic basis of traits has evolved, at least to a partially different extent, in the breeds involved, whereas the genomic prediction model is not flexible enough to accommodate these differences. Differences in genetic backgrounds may be due, for instance, to only a partial overlap between loci affecting a trait across breeds, to interactions with the genetic background of the breed, and to differences in allele frequencies and LD patterns of loci, which do affect any traits in different breeds.

One proposed strategy to accommodate these differences between breeds is to use multitrait (**MT**) models, where trait-by-breed combinations are treated as different but correlated traits (Karoui et al., 2012; Olson et al., 2012; Huang et al., 2014; Zhou et al., 2014). All these studies applied an MT genomic (**G**)BLUP type of model. One important assumption underlying this model is that, across the genome, 1 single genetic correlation between breeds is considered, which assumes for each SNP, a priori, the same covariance structure between effects among different breeds. An alternative model, which has been proposed recently, is the so-called multitask Bayesian learning model for multibreed genomic prediction (Chen et al., 2014), which does not consider the same covariance structure between breeds across the genome. This is effectively a Bayesian variable selection model, which uses the data on all breeds to decide whether or not a variable is selected into the model. In other words, this model accumulates evidence across breeds to determine whether or not a SNP is linked to a QTL. The SNP effects are subsequently estimated separately within each breed, using only phenotypic information on the breed itself. The implementation, as presented by Chen et al. (2014), however, does not explicitly accommodate SNP linked to a breed-specific QTL. That said, there are indications that modeling both breed-specific and common QTL is beneficial for multibreed genomic prediction (van den Berg et al., 2016b).

The objective of the current study, therefore, was to expand the multitask Bayesian learning model to allow for SNP linked to a breed-specific QTL to obtain a large effect in one breed and a small effect in another, as well as to compare this to the originally proposed multitask Bayesian learning model and several other models. These other models include single-trait (**ST**) and MT genomic BLUP-type models, and a ST Bayesian variable selection model. In all ST models, either phenotypes of only 1 of the breeds were used or phenotypes of different breeds were pooled and analyzed simultaneously, as if the same trait was involved. Analyses were performed on a data set including Holsteins, with a moderate size of training set, and Jerseys, with a small-sized size of training set. Validation was, in all cases, only performed for the Jersey breed.

MATERIALS AND METHODS

Data

Phenotypic Data. The data used in our study contained 7,994 Holstein and 1,378 Jersey bulls with both genotypes and phenotypes available. The Holstein bulls originated from Australia (35%), New Zealand (15%), and the Netherlands (50%), whereas the Jersey bulls originated from Australia (43%) and New Zealand (57%). The phenotypes were deregressed proofs (**DRP**) for milk, fat, and protein yields, which were derived from international multiple trait across-country evaluation EBV as computed by Interbull and converted to the Australian scale. Each DRP had a weight computed as effective daughter equivalents (**EDC**), which was derived from the corresponding multiple trait across-country evaluation EBV. Average reliabilities of the DRP for the Holstein training bulls, as computed from the EDC, were 0.81, 0.77, and 0.76, respectively, for milk, fat, and protein yields. Average reliabilities of the DRP for the Jersey training bulls were 0.84 for milk, fat, and protein yields.

As the Jersey data set was considerably smaller than the Holstein data set, we only expected improvement in genomic prediction accuracy by adding information from the other breed for Jerseys, whereas validation of the models described in the next section was only performed using Jersey validation bulls. The data were split into groups of training and validation bulls by assigning all bulls born before January 2004 to the training data set. This yielded an initial training data set containing 6,278 Holstein and 1,004 Jersey bulls and a validation data set containing 374 Jersey bulls. Analysis of the data revealed that those 374 Jersey bulls had strong relationships with the Jersey training bulls, which likely reduced the potential effect of adding the Holstein training data to a considerable extent. To reduce the relationship with the training data set, close relatives of the 374 Jersey validation bulls were removed from the training data; this included 93 sires, 105 paternal half-sibs (i.e., sons of sires of validation bulls), 4 maternal half-sibs (i.e., sons of dams of valida-