J. Dairy Sci. 102:3230–3240
https://doi.org/10.3168/jds.2018-15810

Genomic prediction of bull fertility in US Jersey dairy cattle

Fernanda M. Rezende,[1,2] Juan Pablo Nani,[1,3] and Francisco Peñagaricano[1,4]*
[1]Department of Animal Sciences, University of Florida, Gainesville 32611
[2]Faculdade de Medicina Veterinária, Universidade Federal de Uberlândia, Uberlândia MG 38410-337, Brazil
[3]Estación Experimental Agropecuaria Rafaela, Instituto Nacional de Tecnología Agropecuaria, Rafaela SF 22-2300, Argentina
[4]University of Florida Genetics Institute, University of Florida, Gainesville 32610

ABSTRACT

Service sire has a major effect on reproductive success in dairy cattle. Recent studies have reported accurate predictions for Holstein bull fertility using genomic data. The objective of this study was to assess the feasibility of genomic prediction of sire conception rate (SCR) in US Jersey cattle using alternative predictive models. Data set consisted of 1.5k Jersey bulls with SCR records and 95k SNP covering the entire genome. The analyses included the use of linear and Gaussian kernel-based models fitting either all the SNP or subsets of markers with presumed functional roles, such as SNP significantly associated with SCR or SNP located within or close to annotated genes. Model predictive ability was evaluated using 5-fold cross-validation with 10 replicates. The entire SNP set exhibited predictive correlations around 0.30. Interestingly, either SNP marginally associated with SCR or genic SNP achieved higher predictive abilities than their counterparts using random sets of SNP. Among alternative SNP subsets, Gaussian kernel models fitting significant SNP achieved the best performance with increases in predictive correlation up to 7% compared with the standard whole-genome approach. Notably, the use of a multi-breed reference population including the entire US Holstein SCR data set (11.5k bulls) allowed us to achieve predictive correlations up to 0.315, gaining 8% in accuracy compared with the standard model fitting a pure Jersey reference set. Overall, our findings indicate that genomic prediction of Jersey bull fertility is feasible. The use of Gaussian kernels fitting markers with relevant roles and the inclusion of Holstein records in the training set seem to be promising alternatives to the standard whole-genome approach. These results have the potential to help the dairy industry improve US Jersey sire fertility through accurate genome-guided decisions.

Received October 8, 2018.
Accepted November 29, 2018.
*Corresponding author: fpenagaricano@ufl.edu

Key words: biologically informed model, kernel-based prediction, multi-breed reference population, sire conception rate

INTRODUCTION

One of the major challenges of the dairy industry worldwide is to improve fertilization rates and minimize embryonic losses in order to optimize conception rates. Service sire has been recognized as an important factor affecting herd fertility in dairy cattle, influencing not only the fertilization process, but also the viability of the embryo during the early stages of development (Kropp et al., 2014). Recently, Ortega et al. (2018) proved that the reduced ability of low fertility bulls to establish pregnancy is multifactorial and encompasses sperm fertilizing ability, preimplantation embryonic development, and placenta and embryo development after conceptus elongation and pregnancy recognition. The US dairy industry has had access to a phenotypic evaluation of dairy bull fertility called sire conception rate (**SCR**) that is based on a large, nationwide database of confirmed pregnancy records (Kuhn and Hutchison, 2008). Notably, there is a remarkable variation in SCR with more than 10% conception rate difference between high-fertility and low-fertility bulls (Peñagaricano et al., 2012; Rezende et al., 2018). Therefore, it is clear that service sire fertility should be considered in breeding schemes aimed to improve the reproductive efficiency of dairy herds.

The prediction of unobserved genetic values or yet-to-be-observed phenotypes is relevant not only in animal breeding, but also in plants and personalized medicine. Several methodologies have been proposed for the estimation of genetic effects and the prediction of complex phenotypes using genomic data (Meuwissen et al., 2001; Gianola et al., 2006; VanRaden, 2008; Misztal et al., 2009). Additionally, various marker selection strategies have been suggested to reduce the number of variants tested and improve the accuracy of genomic predictions, such as selecting markers with large effects (Weigel et al., 2009), the use of markers

linked to genes associated with relevant gene-sets or pathways (Edwards et al., 2016; Abdollahi-Arpanahi et al., 2017), or markers with presumed functional roles (Koufariotis et al., 2014). Furthermore, multi-breed genomic prediction within and across-country was proposed to increase the number of individuals in the reference population as an effort to achieve more accurate predictions (Hayes et al., 2009). Although there is no consensus on the best prediction methodology, genomic prediction has become a routine procedure in livestock production. Its adoption in dairy cattle has doubled the annual rates of genetic gain for production traits, and has increased from 3- to 4-fold for lowly heritable traits, including health and female fertility (Weller et al., 2017). Indeed, genomic selection has dramatically changed the genetic trend for daughter pregnancy rate in US Holstein, from close to zero to large and favorable in a short period of time (García-Ruiz et al., 2016). On the other hand, and despite its importance, the genetic improvement of dairy bull fertility has received less attention. Our group has recently identified candidate genomic regions, individual genes, and biological processes underlying bull fertility in Holstein cattle (Han and Peñagaricano, 2016; Nicolini et al., 2018), as well as reported promising results for predicting Holstein SCR values using genomic data (Abdollahi-Arpanahi et al., 2017).

The Jersey breed is the second most important dairy breed in the United States, representing at least 12% of the US dairy cow population. The proportion of Jersey semen sold domestically by National Association of Animal Breeders members increased from 6% in 2000 to 13% in 2016 (Dechow et al., 2018). Although Jersey cattle generally has higher conception rate than Holstein cattle, its reproductive performance remains suboptimal (Norman et al., 2009). As in the Holstein breed, cow fertility traits are routinely evaluated and included in US Jersey selection programs, whereas bull fertility has been largely ignored. Evidence is growing, however, that genetic factors explain part of the variation in Jersey sire fertility. Indeed, we recently identified individual genes and gene sets strongly associated with SCR in US Jersey bulls (Rezende et al., 2018).

To the best of our knowledge, no study to date has explored the feasibility of predicting service sire fertility in Jersey cattle using genomic data. Therefore, our first objective was to evaluate the genomic prediction of SCR in US Jersey bulls using almost 100k SNP spanning the whole genome. The use of biological information for prediction of complex traits is gaining ground in livestock genomics. As such, the second objective was to investigate the predictive performance of SNP subsets with presumed functional roles. Finally, given that the US Jersey data set is relatively small, we assessed the prediction of Jersey bull fertility using a multi-breed reference population including US Holstein records.

MATERIALS AND METHODS

Phenotypic and Genotypic Data Sets

Sire conception rate is a national phenotypic evaluation of bull fertility provided to the US dairy industry since 2008, initially by the Animal Improvement Programs Laboratory of the USDA and now by the Council on Dairy Cattle Breeding (**CDCB**). This evaluation is based on cow field data, and it is intended as phenotypic rather than a genetic evaluation, because the estimates reflect both genetic and nongenetic effects (Kuhn and Hutchison, 2008; Kuhn et al., 2008). Sire conception rate is defined as the expected difference in conception rate of a specific bull compared with the mean of all evaluated bulls.

Service sire conception records of 1,569 Jersey bulls from 29 consecutive evaluations performed between August 2008 and April 2018 were used in this study. For bulls with multiple evaluations, only the most reliable SCR value, that is, the SCR record with most breedings, was considered in the analyses. The Jersey SCR values ranged from −15.1 to +5.5, and the number of breedings per bull ranged from 200 to 26,617. The SCR records are freely available at the CDCB website (https://www.uscdcb.com/). The SCR reliabilities (REL) reported by CDCB are calculated based on the number of breedings (n) as $REL = 100 \times [n/(n + 260)]$, and ranged from 44 to 99.

A total of 107,371 SNP genotypes for all the Jersey bulls with SCR records were kindly provided by the Cooperative Dairy DNA Repository (Columbia, MO). Markers that mapped to the sex chromosomes, had minor allelic frequency below 1%, or had a call rate less than 90% were removed. After editing, a total of 95,434 SNP markers were retained for subsequent analyses.

The use of a multi-breed reference population has been proposed to increase the total number of individuals in the reference set, in an attempt to achieve more accurate predictions. In this study, we evaluated the feasibility of Jersey SCR prediction using a multi-breed reference population including the entire US Holstein SCR data set. Specifically, a total of 11,539 Holstein sires with SCR records and genotypes for the 95,434 SNP were available to include in the training population. As described for Jersey, only one record per Holstein bull, the most reliable SCR value, was considered in these multi-breed prediction analyses.

Alternative SNP Subsets

For the first objective, where the goal was to investigate the performance of genomic models for predicting SCR values, alternative predictive models using the whole SNP data set were evaluated. For the second objective, where the goal was to investigate the predictive ability of different SNP subsets, 3 different strategies were applied for marker selection, including the use of significant, gene ontology (**GO**), and genic SNP markers. For the third objective, where we assessed multibreed-based genomic prediction of Jersey sire fertility, both the entire SNP data set and the relevant SNP subsets were assessed.

Significant SNP. The association between each SNP and SCR was assessed fitting a single-marker linear model with the SNP allele count as linear covariate and the USDA-CDCB SCR evaluation as a categorical variable (class effect with 29 levels). The SNP markers with nominal P-value ≤ 0.05 were considered as significant SNP (top SNP).

Gene Ontology SNP. Gene ontology terms can be defined as groups of genes that are involved in the same biological process or molecular function. In our previous studies, using Fisher's exact test, we identified 2 GO terms, namely calcium ion binding (GO: 0005509) and pyrophosphatase activity (GO: 0016462), significantly enriched with genes associated with SCR in both Jersey and Holstein breeds (Peñagaricano et al., 2013; Rezende et al., 2018). Here, we assessed the set of SNP linked to genes in these 2 relevant GO terms. First, the list of genes involved in each of these 2 GO terms was retrieved using the Ensembl *BioMart* database (http://www.ensembl.org/biomart) based on the information provided by the UMD3.1 bovine genome assembly (Zimin et al., 2009). Then, a given SNP marker was assigned to a particular GO gene if it was located within or at most 15 kb either upstream or downstream of the gene.

Genic SNP. This SNP subset consisted of (1) SNP markers mapped near genes (5 kb either upstream or downstream from the gene), (2) SNP markers mapped in regulatory regions within the gene [5′ untranslated region (**UTR**) or 3′ UTR], and (3) SNP markers within coding exons including synonymous, missense, and stop codon variants. The mapping of SNP markers into these 7 functional classes was performed using a gene annotation file downloaded from the Genome Browser's Variant Annotation Integrator database (http://genome.ucsc.edu/cgi-bin/hgVai) from the University of California Santa Cruz, based on Bos taurus UMD3.1 genome assembly.

It is worth noting that the predictive performance of each functional SNP subset was compared with the performance exhibited by a random set of markers (i.e., a SNP subset with the same number of markers but randomly sampled across the genome). The idea was to investigate the benefits of using SNP with presumed functional roles beyond simply accounting for population structure (genomic relationships).

Predictive Models

The predictive ability of either the entire SNP data set or alternative functional SNP subsets, either using only Jersey records or combining Jersey and Holstein records in the reference population, was evaluated using Bayesian reproducing kernel Hilbert spaces (**RKHS**) regression models (Gianola and van Kaam, 2008; Morota and Gianola, 2014). In particular, we investigated the performance of single-kernel models fitting one set of SNP per time using either linear or Gaussian kernels. All these analyses were implemented under the general kernel-based regression model, $\mathbf{y} = \mathbf{Xb} + \mathbf{K}\boldsymbol{\alpha} + \mathbf{e}$, where $\mathbf{y}$ is the vector of SCR records, $\mathbf{b}$ is the vector of fixed effects including a general intercept and the USDA-CDCB SCR evaluation class effect with 29 levels, $\mathbf{X}$ is the incidence matrix linking the fixed effects to SCR records, $\mathbf{K}$ is an n × n kernel matrix constructed from observed SNP genotypes, $\boldsymbol{\alpha}$ is the vector of RKHS regression coefficients to be inferred that minimizes the following objective function, $l(\boldsymbol{\alpha}|\lambda) = (\mathbf{y} - \mathbf{K}\boldsymbol{\alpha})'(\mathbf{y} - \mathbf{K}\boldsymbol{\alpha}) + \lambda\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}$ is a penalty on model complexity, which is taken to be the square of the norm function, and λ is a regularization parameter $\left(\lambda = \sigma_e^2 / \sigma_g^2\right)$. The 2 random components of the model, $\boldsymbol{\alpha}$ and $\mathbf{e}$, were distributed as $\boldsymbol{\alpha} \sim \left(\mathbf{0}, \mathbf{K}^{-1}\sigma_g^2\right)$ and $\mathbf{e} \sim \left(\mathbf{0}, \mathbf{R}^{-1}\sigma_e^2\right)$, where σ_g^2 and σ_e^2 are the genetic and residual variances, respectively, and $\mathbf{R}$ is an identity matrix.

On the one hand, the linear kernel is expected to capture genetic signals through genomic relationships under additive inheritance, which is equivalent to the ridge regression model with additive genomic relationship matrix (the so-called GBLUP) proposed by VanRaden (2008). The linear kernel ($\mathbf{K_L}$) is obtained by setting $\mathbf{K_L} = \mathbf{SS}'/p$, where $\mathbf{S}$ is a matrix of centered and standardized SNP genotypes and p represents the number of SNP. On the other hand, nonlinear kernels, such as the Gaussian kernel, are able to capture complex interactions of the genome, including nonadditive effects that are important for accurate phenotypic predictions. In this study, a Gaussian kernel ($\mathbf{K_G}$) was

evaluated in the average squared-Euclidean distance between genotypes as follows:

$$\mathbf{K_G}\left(w_i, w_{i'}\right) = \exp\left\{-h \times \frac{\sum_{k=1}^{p}\left(w_{ik} - w_{i'k}\right)^2}{p}\right\},$$

where w_i and $w_{i'}$ are the genotypes centered and standardized of bull i and i', k is equal to the number of SNP, and h is the bandwidth parameter (set as 0.5) chosen over a grid of values to maximize the prediction accuracy (De los Campos et al., 2010).

Implementation

Kernel models were implemented in a Bayesian framework via Markov chain Monte Carlo. For each model, a single chain was run with a total of 100,000 iterations. Inferences were based on 14,000 mildly correlated samples obtained after discarding the first 30,000 samples as burn-in and using a thinning interval equal to 5. Runs lasted between 1 and 24 h depending on the kernel, number of markers, and reference population size. Convergence of the chain was checked by visual inspection of trace plots of some parameters, such as variance components. All of these analyses were carried out using the R package Bayesian Generalized Linear Regression (version 1.0.5; Pérez and de los Campos, 2014).

Evaluation of Model Predictive Ability

To evaluate the ability of the different RKHS regression models to predict yet-to-be-observed SCR values, a 5-fold cross-validation approach was applied. For the first and second objectives (i.e., the prediction of SCR values using either all SNP or only markers with presumed function roles), the entire Jersey data set was randomly split into 5 disjoint subsets of approximately equal size. In each iteration of the cross-validation, 4 of the 5 subsets were used as a training set (*train*) to estimate the solutions of fixed and random effects, and the remaining set was used as testing set (*test*) to evaluate model predictive ability. The prediction of SCR in the testing set $\left(\hat{\mathbf{y}}_{test}\right)$ is given by $\hat{\mathbf{y}}_{test} = \mathbf{X}_{test}\hat{\mathbf{b}}_{train} + \hat{\mathbf{g}}_{test}$ with $\hat{\mathbf{g}}_{test} = \mathbf{K}_{test,train}\mathbf{K}^{-1}_{train}\hat{\mathbf{g}}_{train}$, where $\mathbf{X}_{test}$ is the design matrix, $\hat{\mathbf{b}}_{train}$ is the vector of fixed effects, $\mathbf{K}_{test,train}$ is a rectangular kernel matrix of genomic relationships between training and testing bulls, $\mathbf{K}_{train}$ is genomic relationship between bulls in the training set, and $\hat{\boldsymbol{g}}_{train}$ is the vector of predicted genomic values of bulls in the training set. This procedure was repeated until the 5 subsets were used as testing sets. For the third objective, multi-breed genomic prediction, all the available US Holstein SCR records were used, with an imposed restriction that Holstein bulls were only included in the training set along with 4/5 of the Jersey bulls. All the cross-validation procedures were repeated 10 times; therefore, the output of each analysis was the average of 50 estimates. Additionally, in those analyses involving SNP selected at random across the bovine genome, the sampling of the SNP was repeated 10 times, so each analysis resulted in a total of 500 estimates.

Model predictive ability was also evaluated using the leave-one-out cross-validation. Note that this method is a particular case of the k-fold cross-validation with k equal to the number of observations. Given the computational demand of this analysis, only predictive models fitting Jersey records using all of the SNP were evaluated.

The ability to predict yet-to-be-observed SCR values was assessed using the Pearson product-moment correlation coefficient (CORR) and the mean-squared error of prediction (**MSEP**). The correlation between observed SCR values (y) and predicted SCR values $(\hat{y})$, and the MSEP, defined as the average of the squared differences between y and $\hat{y}$, were calculated in each cross-validation testing fold.

RESULTS AND DISCUSSION

Reproductive efficiency has a large effect on the profitability of the dairy industry. Service sire has been recognized as an important factor affecting herd fertility in dairy cattle. We recently reported accurate predictions for service sire fertility, measured as SCR, using genomic data in US Holstein cattle (Abdollahi-Arpanahi et al., 2017). This study was specially conducted to evaluate the feasibility of genomic prediction of SCR in US Jersey bulls. We evaluated the prediction of Jersey service sire fertility using either the entire SNP data set or only subsets of markers with presumed functional roles. We also investigated the genomic prediction of Jersey bull fertility using a multi-breed reference population combining Holstein and Jersey SCR records.

Genomic Prediction of Jersey Bull Fertility

The predictive ability of linear and Gaussian kernel-based models fitting the entire set of SNP is shown in Figure 1. Note that the linear kernel model fitting all the SNP is equivalent to the standard GBLUP model. Predictive performance was evaluated using 5-fold cross-validation repeated 10 times, so each analysis resulted in 50 estimates. The Gaussian kernel model

exhibited a slightly better predictive ability than the linear kernel model, showing higher average predictive correlation (0.299 vs. 0.292) and lower average mean squared error of prediction (4.665 vs. 4.686). These results suggest that most of the predictive power is driven by additive genomic relationships under the classical

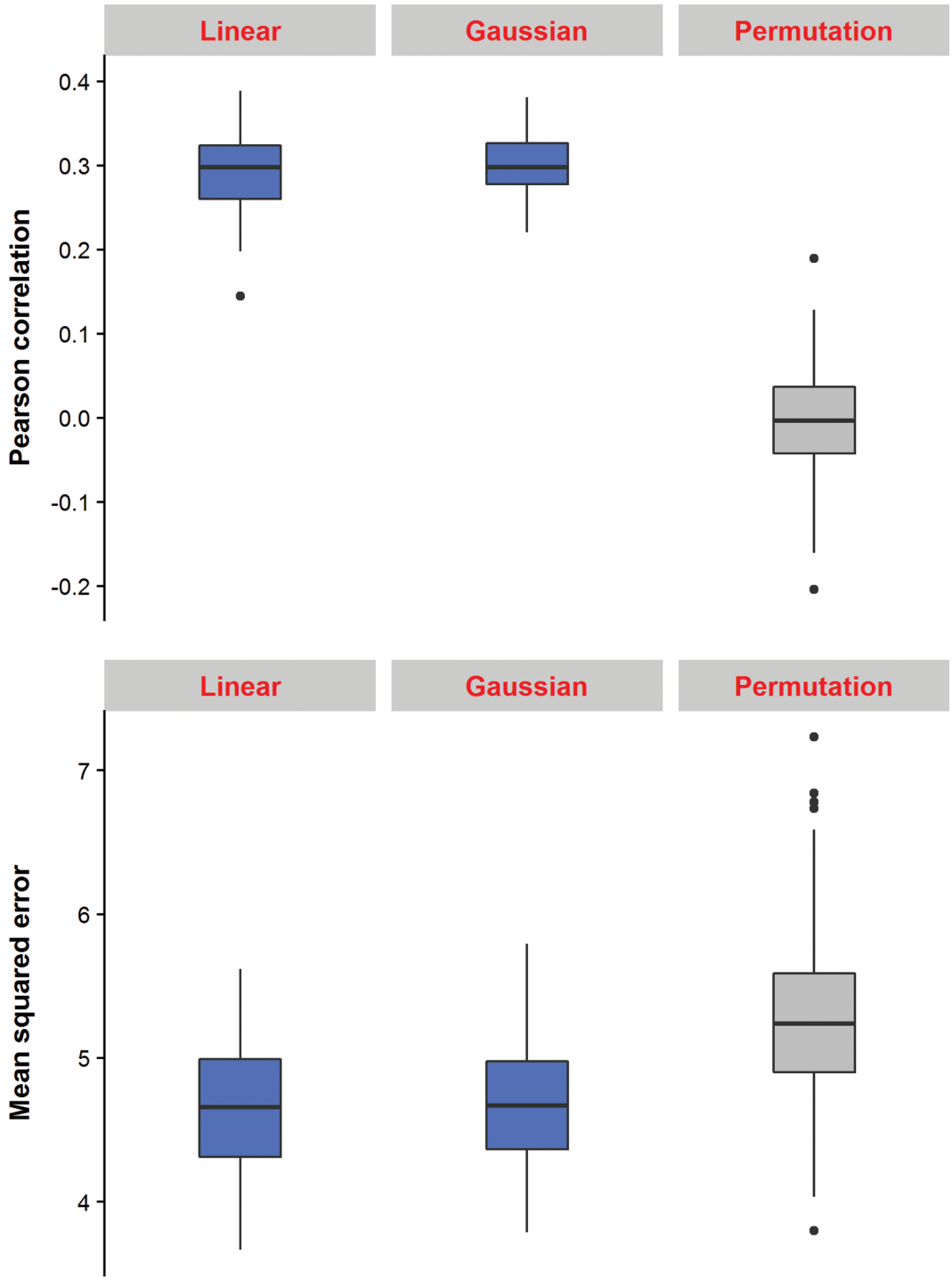

Figure 1. Predictive ability of kernel-based models fitting the entire SNP data set. Predictive correlation (top) and mean squared error of prediction (bottom) were calculated for both linear and Gaussian kernel-based models using 5-fold cross-validation with 10 replicates (blue). Permutation analysis was performed using the linear kernel-based model in a 5-fold cross-validation procedure with 100 replicates (gray). The bottom and top of the box represent first and third quartiles; the horizontal line denotes the median; the whiskers correspond to 1.5× interquartile distance; and dark dots are outliers.

infinitesimal model, although gene-by-gene and other forms of genetic interactions, captured by the Gaussian kernel, may also have a small contribution to Jersey sire fertility.

It is well established that the larger the reference population, the more accurate the predictions of unobserved genetic values or yet-to-be-observed phenotypes are. In this context, leave-one-out cross-validation has the advantage of maximizing the training population size ($N-1$). This cross-validation approach, sometimes very computational and time consuming, minimizes the testing bias but yields estimates with high variance (Breiman and Spector, 1992). Here, we evaluated the predictive performance of alternative whole-genome kernel-based models using the leave-one-out cross-validation procedure. The idea was to investigate the potential of the entire Jersey data set to predict the yet-to-be observed SCR value of a single individual. Interestingly, higher predictive correlations (0.308 and 0.315) and lower MSEP values (4.622 and 4.595) were observed compared with the 5-fold cross-validation, for both linear and Gaussian kernels. This represents an increase in prediction ability of about 5%, regardless of the kernel used. In practice, these findings suggest that the use of the entire Jersey SCR data set can be very useful for predicting male fertility of a single newborn bull calf.

One legitimate question that might arise is if the predictions of Jersey sire fertility are entirely driven by the genotype data or, instead, if some nongenetic hidden factors explain part of the model predictive power. One way to answer this question is by using a permutation test, also called a randomization test (Churchill and Doerge, 1994). The random shuffling of the phenotypes across genotypes does not preserve any genotype-phenotype link, and hence, the shuffled data sets correspond to the null hypothesis of no relationship between SNP data and SCR values. In this case, the predictive ability of the models, if any, would be driven by nongenetic variables. Here, we generated 100 shuffled data sets by randomly permuting SCR values and genotype data, and then assessed whole-genome SCR prediction using 5-fold cross-validation (a total of 500 estimates). Notably, this resampling procedure yielded an average predictive correlation equal to zero (Figure 1, gray boxplot), illustrating that our predictions of Jersey bull fertility are entirely driven by the genotype data.

Predictive Performance of Alternative SNP Subsets

Figure 2 shows the predictive ability of alternative kernel-based models fitting different subsets of SNP markers. Of the 95,434 SNP evaluated in this study, about 15k SNP were marginally associated with SCR (top SNP subset). It should be emphasized that the significance of each SNP was systematically evaluated in each iteration of the 5-fold cross-validation in the training set, and only the markers with P-value ≤ 0.05 were then used to predict unobserved SCR values in the testing set. Note that our goal was to predict a yet-to-be-observed phenotype instead of pinpointing causal mutations, and hence, controlling type I error was not a major priority. Interestingly, the top SNP subsets outperformed their counterparts using random sets of SNP regardless of the type of kernel used (Figure 2). Previous studies reported that SNP subsets containing markers with large effects achieved comparable or even higher predictive performance than that obtained using all the SNP (Weigel et al., 2009; Moser et al., 2010; Vazquez et al., 2010; Weller et al., 2014). Here, top SNP subsets outperformed the entire SNP set showing higher average predictive correlation for both linear (0.307 vs. 0.292) and Gaussian (0.312 vs. 0.299) kernels. This represents an increase in predictive correlation up to 7%. These results suggest that predictive models fitting only markers with large effects can be an alternative to the standard whole-genome approach for predicting bull fertility in US Jerseys.

A total of 4,534 SNP (GO SNP subset) were located within or near 1,191 genes linked to GO terms calcium ion binding (GO: 0005509) and pyrophosphatase activity (GO: 0016462), 2 terms significantly associated with dairy bull fertility (Peñagaricano et al., 2013; Rezende et al., 2018). The predictive performance of kernel-based models fitting GO SNP subsets did not outperform their counterparts with random SNP (data not shown). Therefore, we should conclude that the predictive ability exhibited by these SNP subsets is not driven by their functional role, but rather by accounting for genomic relationships. We recently investigated the predictive ability of alternative gene-set SNP subsets for predicting SCR in Holstein cattle, and as in this current study, gene-set markers did not improve prediction compared with the standard whole-genome approach (Abdollahi-Arpanahi et al., 2017). A better understanding of the genetic mechanisms underlying bull fertility plus a more complete bovine genome annotation might provide new opportunities for predicting service sire fertility using gene-set data.

A subset of 4,870 SNP (genic SNP subset) of the 95,434 available markers mapped in regulatory regions outside but close to genes (5 kb upstream or downstream), regulatory regions within genes (5′ UTR and 3′ UTR) and coding exons. Notably, the genic SNP subset showed better predictive power than random

SNP, with increases in predictive correlations of about 4% regardless of the type of kernel (Figure 2). In addition, both linear and Gaussian kernel-based models fitting genic SNP showed slightly higher predictive correlations (0.296 and 0.302 vs. 0.292 and 0.299) and lower MSEP (4.680 and 4.655 vs. 4.686 and 4.665) than their counterparts fitting the entire SNP data set. Previous studies have investigated the contribution of genic and nongenic regions to additive genetic variance and model predictive performance with somewhat

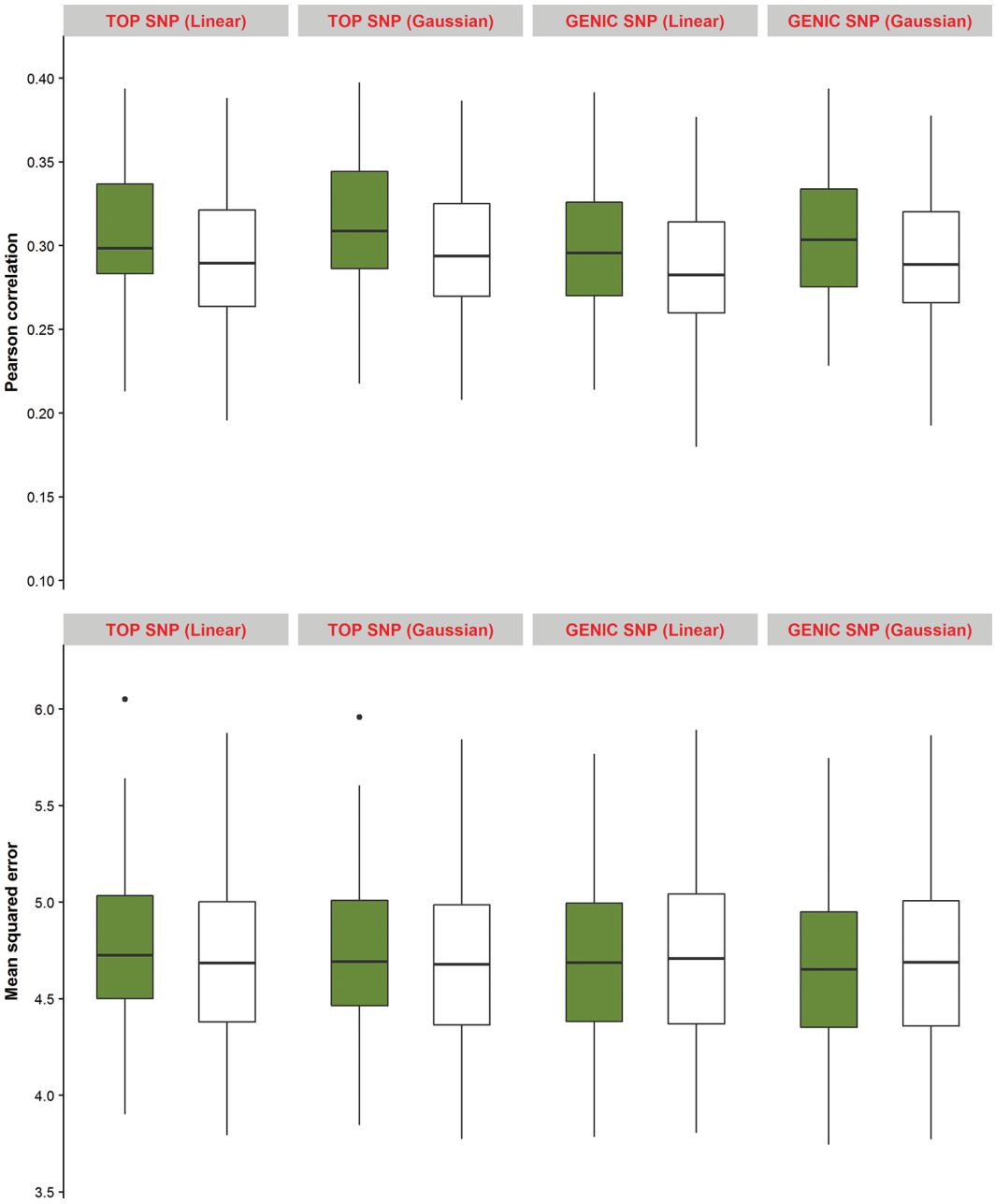

Figure 2. Predictive ability of kernel-based models fitting different SNP subsets. Predictive correlation (top) and mean squared error of prediction (bottom) were calculated for both linear and Gaussian kernel-based models using 5-fold cross-validation with 10 replicates. Each analysis was performed using either the functional SNP class of interest (green) or a set of SNP with the same size but randomly sampled from the genome (white). Top SNP: set of SNP markers with a nominal P-value ≤ 0.05; genic SNP: set of SNP markers located within or near annotated genes. The bottom and top of the box represent first and third quartiles; the horizontal line denotes the median; the whiskers correspond to 1.5× interquartile distance; and dark dots are outliers.

dissimilar results (Yang et al., 2011; Li et al., 2012; Koufariotis et al., 2014; Morota et al., 2014; Do et al., 2015; Abdollahi-Arpanahi et al., 2016; Ni et al., 2017). For instance, Ni et al. (2017) reported slightly higher predictive ability using SNP within or around genes compared with whole-genome SNP data in laying chickens. Contrary, Morota et al. (2014) and Abdollahi-Arpanahi et al. (2016) found that prediction based on genomic regions is trait dependent in broiler chickens, and in general the whole-genome approach provided better predictive ability than functional classes considered individually. Similarly, Do et al. (2015) concluded that functional SNP classes yielded similar predictions than random sets of SNP for feed efficiency in pigs. Our results suggest that the majority of the genetic variants that affect Jersey sire fertility are located within or near annotated genes, which leads genic SNP to have similar or even better predictive power than the entire SNP data set.

Both significant and genic SNP subsets achieved better predictive performance than random sets of SNP. This suggests that the predictive power of these functional SNP classes is driven in part by their biological roles and not simply by accounting for population structure. In addition, these 2 functional SNP subsets exhibited comparable or even better predictive ability than all the SNP. Indeed, Gaussian kernel-based models fitting genic SNP and top SNP outperformed the standard whole-genome approach with predictive gains between 3 and 7%. These findings suggest that considering nonlinear effects together with the use of functional SNP classes would benefit the prediction of Jersey bull fertility.

Incorporation of Holstein Records in the Reference Population

Figure 3 shows the predictive performance of alternative models using a multi-breed Holstein and Jersey reference population. When a linear kernel-based model was used, adding the entire US Holstein SCR data set to the training population did not improve model predictive performance. In fact, considering the entire SNP data set, slightly lower predictive correlations (0.289 vs. 0.292) and higher MSEP values (4.694 vs. 4.686) were obtained using a multi-breed compared with a pure Jersey reference set. In general, multi-breed genomic prediction models assume that the genetic variants underlying the trait of interest are the same and have the same effects among breeds (de Roos et al., 2009). However, our group has shown that most of the genomic regions and individual genes that affect SCR in Jersey are nonsignificant in Holstein, and vice versa (Han and Peñagaricano, 2016; Rezende et al., 2018). This may be due to multiple causes including that the major variants affecting bull fertility in Jersey are not segregating in Holstein or these causative variants are indeed segregating in Holstein but are not in high linkage disequilibrium with the SNP markers. In fact, in a simulation study, de Roos et al. (2008) reported that successful genomic prediction across Holstein and Jersey breeds would require at least 300,000 markers, and most of the bull fertility studies have been performed using medium-density SNP data.

Of particular interest, increasing the training population size with Holstein records improved Jersey SCR prediction when the SNP were fit using Gaussian kernels (Figure 3). Indeed, using a multi-breed reference population, Gaussian kernel-based models fitting either all SNP or the genic SNP subset exhibited the best predictive performance with increases in predictive correlations between 2 and 4% compared with their counterparts using only Jersey records. The Gaussian kernel was introduced in quantitative genetics with the aim of predicting the total genetic value of an individual by exploiting nonlinear relationships between genotypes and phenotypes (Gianola et al., 2006; Gianola and van Kaam, 2008). In fact, the relatedness encoded as spatial genetic distance between individuals is expected to capture some of the complexity of the genome, including nonadditive effects (Morota and Gianola, 2014). It should be noted that Nicolini et al. (2018) and Rezende et al. (2018) have identified significant nonadditive effects underlying SCR in both Holstein and Jersey breeds. The better performance achieved by the Gaussian kernel can be due to its ability to capture the heterogeneous nature of this multi-breed reference population, including nonoverlapping major additive effects and important nonlinear effects.

Prediction Accuracy

Predictive ability can also be evaluated in terms of prediction accuracy, defined as the correlation between the true and the predicted breeding values, and often calculated by dividing the predictive correlation by the squared root of the trait heritability (Legarra et al., 2008). In our analysis, the predictive correlation obtained with the linear kernel-based model is a good estimate of the correlation between observed phenotypic values and predicted breeding values because, by definition, the mean of the SCR values per evaluation is zero, and hence, the term evaluation has a negligible effect on prediction. Therefore, if we divide the estimated predictive correlation by the square root of SCR heritability ($h^2 \approx 0.28$), we get a predictive accuracy

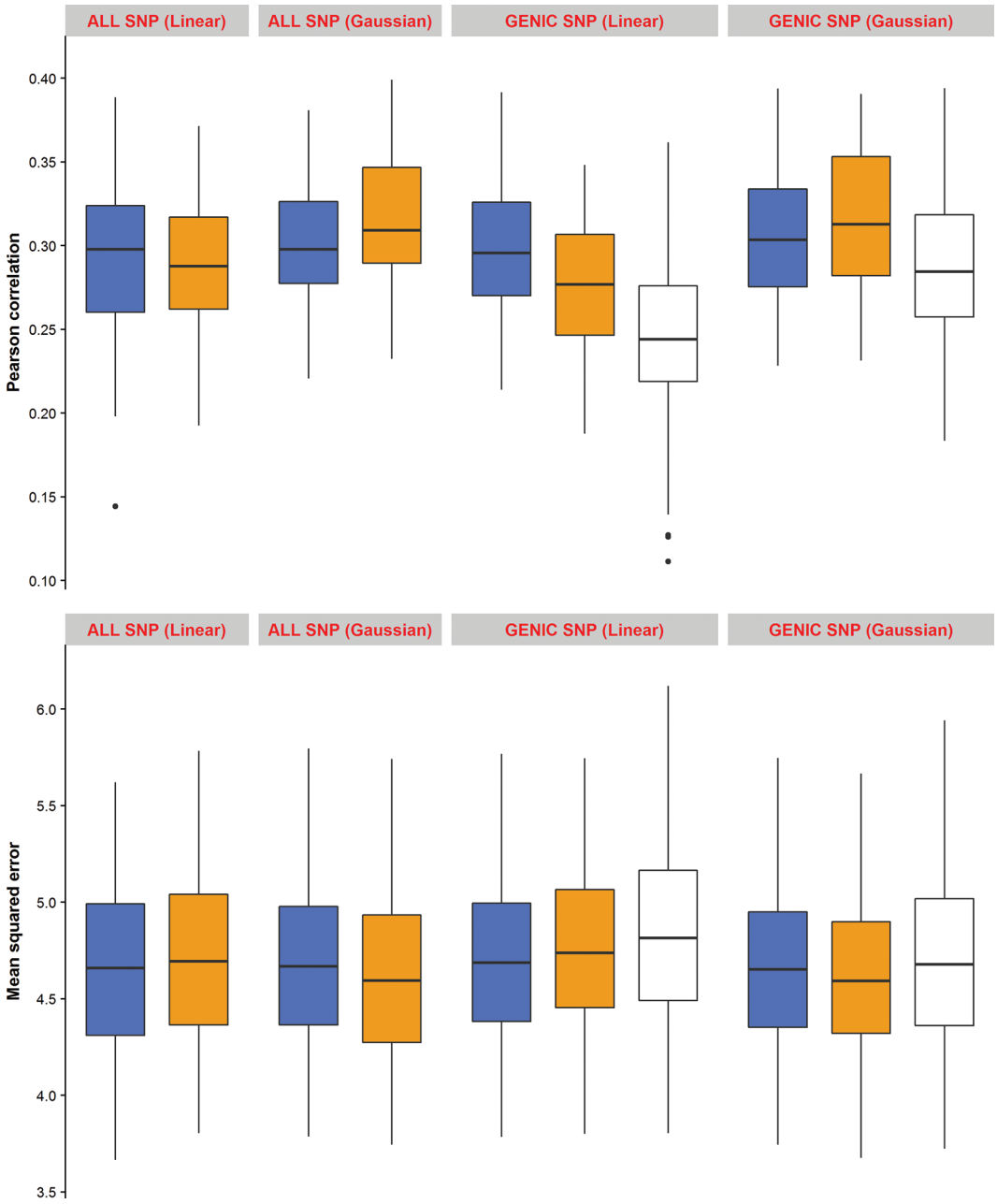

Figure 3. Predictive ability of kernel-based models fitting only Jersey (blue) or Jersey and Holstein records (orange) in the reference population. Models were evaluated using the entire SNP data set (all) or only genic SNP (set of SNP markers located within or near annotated genes). Predictive correlation (top) and mean squared error of prediction (bottom) were calculated for both linear and Gaussian kernel-based models using 5-fold cross-validation with 10 replicates. The performance of a set of SNP with the same size as the genic SNP subset, but randomly sampled from the genome (white), was also evaluated. The bottom and top of the box represent first and third quartiles; the horizontal line denotes the median; the whiskers correspond to 1.5× interquartile distance; and dark dots are outliers.

around 0.57. Notably, calving traits routinely evaluated in US dairy breeds, such as sire calving ease and sire stillbirth rate, have selection accuracies of around 0.55. In addition, novel health traits recently introduced in the US Holstein national evaluation, such as ketosis, lameness, and metritis, have accuracies of around 0.60 for young genomic sires (Parker Gaddis et al., 2014). Taking all together, our findings are promising and revealed that the genomic prediction of US Jersey sire fertility is feasible. This could positively affect the dairy industry, allowing, for example, the culling of newborn bull calves with very low SCR predictions.

CONCLUSIONS

Our study reports promising results regarding the potential prediction of service sire fertility in US Jersey cattle using genomic data. Indeed, prediction accuracies are similar than those reported for some traits currently evaluated in US dairy breeds. The use of markers with relevant roles, such as SNP with large effects or SNP within or near annotated genes, yielded comparable or even better predictive abilities than the standard whole-genome approach. In general, predictive models fitting Gaussian kernels outperformed their counterparts fitting linear kernels irrespective of the set of SNP. The use of a multi-breed reference population including US Holstein records exhibited the best prediction performance when SNP were fit using a Gaussian kernel. Future research may investigate the effect on predictive power of increasing the training population size, adding phenotypes and genotypes of Jersey bulls from other countries. Overall, this study has the potential to help the dairy industry to improve conception rates in US Jersey herds, through accurate genome-guided decisions on service sire fertility.

ACKNOWLEDGMENTS

This study was funded by the American Jersey Cattle Club Research Foundation. The authors thank the Cooperative Dairy DNA Repository (Columbia, MO) and the Council on Dairy Cattle Breeding (Bowie, MD) for providing the genotypic data.

REFERENCES

Abdollahi-Arpanahi, R., G. Morota, and F. Peñagaricano. 2017. Predicting bull fertility using genomic data and biological information. J. Dairy Sci. 100:9656–9666.

Abdollahi-Arpanahi, R., G. Morota, B. D. Valente, A. Kranis, G. J. M. Rosa, and D. Gianola. 2016. Differential contribution of genomic regions to marked genetic variation and prediction of quantitative traits in broiler chickens. Genet. Sel. Evol. 48:10.

Breiman, L., and P. Spector. 1992. Submodel selection and evaluation in regression. The X-random case. Int. Stat. Rev. 60:291–319.

Churchill, G. A., and R. W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. Genetics 138:963–971.

De los Campos, G., D. Gianola, G. J. M. Rosa, K. A. Weigel, and J. Crossa. 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. (Camb.) 92:295–308.

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.

Dechow, C. D., W. S. Liu, J. S. Idun, and B. Maness. 2018. Short communication: Two dominant paternal lineages for North American Jersey artificial insemination sires. J. Dairy Sci. 101:2281–2284.

Do, D. N., L. L. G. Janss, J. Jensen, and H. N. Kadarmideen. 2015. SNP annotation-based whole genomic prediction and selection: An application to feed efficiency and its component traits in pigs. J. Anim. Sci. 93:2056–2063.

Edwards, S. M., I. F. Sørensen, P. Sarup, T. F. C. Mackay, and P. Sørensen. 2016. Genomic prediction for quantitative traits is improved by mapping variants to gene ontology categories in *Drosophila melanogaster*. Genetics 203:1871–1883.

García-Ruiz, A., J. B. Cole, P. M. VanRaden, G. R. Wiggans, F. J. Ruiz-López, and C. P. Van Tassell. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc. Natl. Acad. Sci. USA 113:E3995–E4004.

Gianola, D., R. L. Fernando, and A. Stella. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761–1776.

Gianola, D., and J. B. C. H. M. van Kaam. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289–2303.

Han, Y., and F. Peñagaricano. 2016. Unravelling the genomic architecture of bull fertility in Holstein cattle. BMC Genet. 17:143.

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

Koufariotis, L., Y.-P. P. Chen, S. Bolormaa, and B. J. Hayes. 2014. Regulatory and coding genome regions are enriched for trait associated variants in dairy and beef cattle. BMC Genomics 15:436.

Kropp, J., F. Peñagaricano, S. M. Salih, and H. Khatib. 2014. Invited review: Genetic contributions underlying the development of preimplantation bovine embryos. J. Dairy Sci. 97:1187–1201.

Kuhn, M. T., and J. Hutchison. 2008. Prediction of dairy bull fertility from field data: Use of multiple services and identification and utilization of factors affecting bull fertility. J. Dairy Sci. 91:2481–2492.

Kuhn, M. T., J. Hutchison, and H. Norman. 2008. Modeling nuisance variables for prediction of service sire fertility. J. Dairy Sci. 91:2823–2835.

Legarra, A., C. Robert-Granié, E. Manfredi, and J.-M. Elsen. 2008. Performance of genomic selection in mice. Genetics 180:611–618.

Li, X., C. Zhu, C.-T. Yeh, W. Wu, E. M. Takacs, K. A. Petsch, F. Tian, G. Bai, E. S. Buckler, G. J. Muehlbauer, M. C. P. Timmermans, M. J. Scanlon, P. S. Schnable, and J. Yu. 2012. Genic and nongenic contributions to natural variation of quantitative traits in maize. Genome Res. 22:2436–2444.

Meuwissen, T. H., 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.

Morota, G., R. Abdollahi-Arpanahi, A. Kranis, and D. Gianola. 2014. Genome-enabled prediction of quantitative traits in chickens using genomic annotation. BMC Genomics 15:109.

Morota, G., and D. Gianola. 2014. Kernel-based whole-genome prediction of complex traits: A review. Front. Genet. 5:363.

Moser, G., M. S. Khatkar, B. J. Hayes, and H. W. Raadsma. 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet. Sel. Evol. 42:37.

Ni, G., D. Cavero, A. Fangmann, M. Erbe, and H. Simianer. 2017. Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture. Genet. Sel. Evol. 49:8.

Nicolini, P., R. Amorín, Y. Han, and F. Peñagaricano. 2018. Whole-genome scan reveals significant non-additive effects for sire conception rate in Holstein cattle. BMC Genet. 19:14.

Norman, H. D., J. R. Wright, S. M. Hubbard, R. H. Miller, and J. L. Hutchison. 2009. Reproductive status of Holstein and Jersey cows in the United States. J. Dairy Sci. 92:3517–3528.

Ortega, M. S., J. G. N. Moraes, D. J. Patterson, M. F. Smith, S. K. Behura, S. Poock, and T. E. Spencer. 2018. Influences of sire

conception rate on pregnancy establishment in dairy cattle. Biol. Reprod. In press.
Parker Gaddis, K. L., J. B. Cole, J. S. Clay, and C. Maltecca. 2014. Genomic selection for producer-recorded health event data in US dairy cattle. J. Dairy Sci. 97:3190–3199.
Peñagaricano, F., K. A. Weigel, and H. Khatib. 2012. Genome-wide association study identifies candidate markers for bull fertility in Holstein dairy cattle. Anim. Genet. 43:65–71.
Peñagaricano, F., K. A. Weigel, G. J. M. Rosa, and H. Khatib. 2013. Inferring quantitative trait pathways associated with bull fertility from a genome-wide association study. Front. Genet. 3:307.
Pérez, P., and G. de los Campos. 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483.
Rezende, F. M., G. O. Dietsch, and F. Peñagaricano. 2018. Genetic dissection of bull fertility in US Jersey dairy cattle. Anim. Genet. 49:393–402.
VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.
Vazquez, A. I., G. J. M. Rosa, K. A. Weigel, G. de los Campos, D. Gianola, and D. B. Allison. 2010. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J. Dairy Sci. 93:5942–5949.
Weigel, K. A., G. de los Campos, O. González-Recio, H. Naya, X. L. Wu, N. Long, G. J. M. Rosa, and D. Gianola. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J. Dairy Sci. 92:5248–5257.
Weller, J. I., E. Ezra, and M. Ron. 2017. Invited review: A perspective on the future of genomic selection in dairy cattle. J. Dairy Sci. 100:8633–8644.
Weller, J. I., G. Glick, A. Shirak, E. Ezra, E. Seroussi, M. Shemesh, Y. Zeron, and M. Ron. 2014. Predictive ability of selected subsets of single nucleotide polymorphisms (SNPs) in a moderately sized dairy cattle population. Animal 8:208–216.
Yang, J., T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso, J. M. Cunningham, M. de Andrade, B. Feenstra, E. Feingold, M. G. Hayes, W. G. Hill, M. T. Landi, A. Alonso, G. Lettre, P. Lin, H. Ling, W. Lowe, R. A. Mathias, M. Melbye, E. Pugh, M. C. Cornelis, B. S. Weir, M. E. Goddard, and P. M. Visscher. 2011. Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. 43:519–525.
Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marcais, M. Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg. 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. Genome Biol. 10:R42.