J. Dairy Sci. 100:9656–9666
https://doi.org/10.3168/jds.2017-13288

Predicting bull fertility using genomic data and biological information

Rostam Abdollahi-Arpanahi,*† Gota Morota,‡ and Francisco Peñagaricano*§[1]
*Department of Animal Sciences, University of Florida, Gainesville 32611
†Department of Animal and Poultry Science, University of Tehran, Pakdasht, Iran 3391653755
‡Department of Animal Science, University of Nebraska, Lincoln 68583
§University of Florida Genetics Institute, Gainesville 32611

ABSTRACT

The genomic prediction of unobserved genetic values or future phenotypes for complex traits has revolutionized agriculture and human medicine. Fertility traits are undoubtedly complex traits of great economic importance to the dairy industry. Although genomic prediction for improved cow fertility has received much attention, bull fertility largely has been ignored. The first aim of this study was to investigate the feasibility of genomic prediction of sire conception rate (SCR) in US Holstein dairy cattle. Standard genomic prediction often ignores any available information about functional features of the genome, although it is believed that such information can yield more accurate and more persistent predictions. Hence, the second objective was to incorporate prior biological information into predictive models and evaluate their performance. The analyses included the use of kernel-based models fitting either all single nucleotide polymorphisms (SNP; 55K) or only markers with presumed functional roles, such as SNP linked to Gene Ontology or Medical Subject Heading terms related to male fertility, or SNP significantly associated with SCR. Both single- and multikernel models were evaluated using linear and Gaussian kernels. Predictive ability was evaluated in 5-fold cross-validation. The entire set of SNP exhibited predictive correlations around 0.35. Neither Gene Ontology nor Medical Subject Heading gene sets achieved predictive abilities higher than their counterparts using random sets of SNP. Notably, kernel models fitting significant SNP achieved the best performance with increases in accuracy up to 5% compared with the standard whole-genome approach. Models fitting Gaussian kernels outperformed their counterparts fitting linear kernels irrespective of the set of SNP. Overall, our findings suggest that genomic prediction of bull fertility is feasible in dairy cattle. This provides potential for accurate genome-guided decisions, such as early culling of bull calves with low SCR predictions. In addition, exploiting nonlinear effects through the use of Gaussian kernels together with the incorporation of relevant markers seems to be a promising alternative to the standard approach. The inclusion of gene set results into prediction models deserves further research.

Key words: complex trait prediction, gene set, kernel model, sire conception rate

INTRODUCTION

The prediction of unobserved genetic values or yet-to-be observed phenotypes for complex quantitative traits is relevant not only in animal and plant breeding but also in evolution and personalized medicine. Given that complex traits are controlled by a large number of small-effect genes and by environmental conditions, which in turn can interact in cryptic ways, the accurate prediction of unobserved or future values can be extremely challenging. The recent arrival of high-throughput genotyping and sequencing technologies that allow the assessment of thousands of SNP sites across the entire genome has revolutionized the genetic study of these complex traits. These whole-genome data combined with phenotypic records allow the identification and fine mapping of causal mutations and the development of predictive models. Indeed, high-density SNP data can be effectively used to predict phenotypes or breeding values (Meuwissen et al., 2001). Whole-genome prediction has transformed livestock breeding (Ibáñez-Escriche et al., 2014; Wiggans et al., 2017) and crop breeding (Crossa et al., 2014; Lin et al., 2014) and is gaining ground in human medicine (Vazquez et al., 2012; de los Campos et al., 2013b).

Genomic prediction is largely recognized as a black box tool because it completely ignores any available information about functional features of the genome. For instance, the genomic BLUP (**GBLUP**) method (VanRaden, 2008), considered to be the benchmarking approach for whole-genome prediction, assumes a priori that all SNP have an effect on the trait under study and that all these SNP effects are of similar magnitude (de los Campos et al., 2013a). Similarly, other popular genomic prediction methods, such as Bayes B, Bayes C,

Received June 4, 2017.
Accepted September 13, 2017.
[1] Corresponding author: fpenagaricano@ufl.edu

or even Bayes R, ignore any prior biological knowledge available and assume that all the SNP are equally likely to affect the trait of interest (Gianola, 2013). However, genome-wide association studies have been successful in identifying genomic regions and individual variants associated with numerous complex traits. The incorporation of these genetic findings into predictive models could positively affect both model predictive ability and model robustness.

The use of biological information for prediction of complex traits has recently received some attention. For instance, Zhang and colleagues proposed a weighted GBLUP model in which the genomic relationship matrix is replaced with a trait-specific variance covariance matrix constructed based on either prior publicly available genome-wide association study results (Zhang et al., 2014) or relevant genomic information extracted from the data set at hand (Zhang et al., 2015). Similarly, Tiezzi and Maltecca (2015) evaluated the performance of GBLUP models fitting alternative weighted genomic relationship matrices to account for trait architecture. Furthermore, Kadarmideen (2014) recently proposed the so-called systems GBLUP approach, a GBLUP model that includes 2 genomic relationship matrices, one built with SNP with known biological functions and the other built with SNP with unknown functional roles. Within the Bayesian alphabet, there also have been attempts to use biological knowledge for prediction. For instance, the Bayes RC model, an extension of Bayes R, incorporates biological information by defining classes of SNP likely to be enriched for causal variants (MacLeod et al., 2016). Moreover, there is growing evidence that genetic polymorphisms affecting phenotypic variation are not uniformly distributed across the genome but rather located within or near genes that in turn are connected via molecular pathways or biological processes (Lango Allen et al., 2010). In this sense, Edwards et al. (2016) have extended the GBLUP model by incorporating prior information from gene ontologies. Overall, all these studies have shown that the use of biological information can improve the accuracy of genomic predictions.

Improving reproductive efficiency is a major goal in dairy cattle. Reproduction is a very complex trait; it involves a large number of events, including gametogenesis, fertilization, implantation, and embryo and fetus development, that should be accomplished in a well-orchestrated manner to achieve a successful pregnancy. Most research in dairy cattle has focused on cow fertility. Indeed, 3 female fertility traits—daughter pregnancy rate, heifer conception rate, and cow conception rate—are routinely evaluated in US dairy cattle. Notably, genomic selection has positively affected the genetic trend of daughter pregnancy rate in US Holsteins, changing from close to zero to large and favorable in a short period of time (García-Ruiz et al., 2016). On the other hand, genetic improvement of dairy bull fertility has been largely ignored. However, some studies have suggested that a significant percentage of reproductive failure is attributable to bull subfertility (DeJarnette et al., 2004); hence, the fertility of service sires should not be overlooked. Since 2008, the US dairy industry has had access to a national phenotypic evaluation of bull fertility called sire conception rate (**SCR**). It should be noted that this evaluation is intended as a phenotypic rather than a genetic evaluation. There is growing evidence that bull fertility is influenced by genetic factors. We recently investigated the genomic architecture underlying SCR in US Holstein bulls (Han and Peñagaricano, 2016). Our analyses included the application of alternative whole-genome association mapping methods and the subsequent use of diverse gene set tools using Gene Ontology (**GO**; Ashburner et al., 2000) and Medical Subject Heading (**MeSH**; Coletti and Bleich, 2001) databases. Interestingly, we identified a set of candidate regions and individual genes strongly associated with SCR; most of the genes were closely related to sperm physiology and male biology. In addition, the gene set analyses reveled a list of significant GO and MeSH terms, including *reproduction*, *fertilization*, *sperm motility*, and *sperm capacitation* (Han and Peñagaricano, 2016).

To our best knowledge, no study to date has explored the possibility of predicting sire fertility using genomic information. Therefore, the first objective of this study was to assess the potential feasibility of genomic prediction of SCR in US Holstein bulls using high-density SNP data. Second, our recent study identified many biological pathways and gene sets associated with SCR. As such, the second objective of this study was to incorporate biological information into alternative predictive models and evaluate their predictive ability.

MATERIALS AND METHODS

Phenotypic and Genotypic Data

Since August 2008, first the Animal Improvement Programs Laboratory of the USDA and now the Council of Dairy Cattle Breeding (**CDCB**) have provided a national phenotypic evaluation of service sire fertility, denoted SCR, to the US dairy industry. Kuhn et al. (2008) and Kuhn and Hutchison (2008) provided a complete explanation of the statistical methodology used for evaluating sire fertility using field data. The term SCR is defined as the expected difference in conception rate of a given sire compared with the mean of all other evaluated sires. Contrary to evaluations for other traits such as production or cow fertility, SCR is designed as a phenotypic rather than a genetic evaluation because the published estimates include not only genetic but also nongenetic effects.

A total of 7,447 Holstein bulls with SCR records were used in this study. The SCR records were obtained from 23 consecutive evaluations provided to the US dairy industry between August 2008 and April 2016. These 23 SCR evaluations are available at the CDCB website (https://www.cdcb.us/). The reliabilities of the SCR records, calculated as a function of the number of breedings, were also available. For bulls with multiple fertility evaluations, the most reliable SCR record (i.e., the SCR record with the most breedings) was used in the analyses. Note that it is recommended to use all the available information when possible. However, in this study, given the complexity of the computational analysis, only 1 record per animal (the most reliable SCR estimate) was used.

Genome-wide SNP data for the 7,447 bulls were provided by the Cooperative Dairy DNA Repository. The SNP markers that mapped to chromosome X had minor allele frequency <5%, calling rate <95%, and Hardy-Weinberg equilibrium P-value $\leq 10^{-6}$ were removed from the data set. After data editing, a total of 54,706 SNP markers were retained for subsequent analysis.

Combining Genomic Data with Biological Information

For the first objective of this study (i.e., assess the performance of genomic models for predicting SCR), alternative predictive models using the entire SNP data set were evaluated. For the second objective, where the goal was to predict bull fertility by combining genomic data with biological information, different subsets of SNP were investigated, such as SNP within or near annotated genes, SNP linked to genes in relevant functional categories (gene sets), and SNP that were significantly associated with SCR.

Genic SNP. The SNP were assigned to genes based on the UMD3.1 bovine genome assembly (Zimin et al., 2009) using the R package *biomaRt* (version 2.26.1; Durinck et al., 2005, 2009). A given SNP was assigned to a particular gene if it was located within the genomic sequence of the gene (between the start of the first exon and the end of the last exon) or within 15 kb either upstream or downstream from the gene. The distance of 15 kb was used to capture regulatory regions and other functional sites that may lie outside (e.g., promoter region) but very close to the gene.

Gene Set SNP. Gene sets or functional terms can be defined as groups of genes that share some properties, typically their involvement in the same biological process or molecular function. Based on Han and Peñagaricano (2016), we evaluated the set of SNP linked to genes in the GO term *reproduction* (GO:0000003; GO SNP) and the set of SNP linked to genes associated with a group of MeSH terms (MeSH SNP) closely related to sperm biology, including *spermatozoa* (D013094), *sperm motility* (D013081), and *sperm capacitation* (D013075).

Significant SNP. The association between each SNP marker and SCR was assessed using a single-marker linear model with the SNP allele count as a linear covariate and the SCR evaluation as a categorical variable (class effect with 23 levels). Those SNP markers with nominal P-value $\leq$0.05 were considered as significant SNP (**TOP SNP**). Note that the major goal was to predict yet-to-be observed phenotypes instead of pinpointing causal mutations; hence, controlling type I error was not a major priority.

The performance of each SNP subset was compared with the performance exhibited by another SNP subset with the same number of markers but randomly selected across the genome. This comparison should give an idea of the benefits of using markers with biological roles beyond simply accounting for population structure (genomic relationships).

Statistical Models

Our goal was to predict yet-to-be observed phenotypes (SCR) using genomic data. The predictive ability of either the entire SNP set or the alternative SNP subsets was investigated using Bayesian reproducing kernel Hilbert spaces (**RKHS**) regression models (Gianola and van Kaam, 2008; Morota and Gianola, 2014). Kernel-based prediction models are very powerful predictive machines, and they allow the integration, in a very simple way, of prior biological information (e.g., functional variants or significant markers). We investigated the performance of both single- and multikernel models using either linear or Gaussian kernels.

Single-Kernel Models. This model allows the fitting of one set of SNP per time—either the entire SNP data set or a particular SNP subset. The general single-kernel regression model is

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{K}\boldsymbol{\alpha} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of phenotypic records (SCR values); $\mathbf{b}$ is the vector of fixed effects including a general intercept (μ) and the USDA-CDCB SCR evaluation class effect; $\mathbf{X}$ is the design matrix relating fixed effects to SCR records; $\mathbf{K}$ is an $n \times n$ kernel matrix indexed by the observed SNP genotypes; and $\boldsymbol{\alpha}$ is the vector of

RKHS regression coefficients estimated as the solution that minimizes $l(\boldsymbol{\alpha} \mid \lambda) = (\mathbf{y} - \mathbf{K}\boldsymbol{\alpha})'(\mathbf{y} - \mathbf{K}\boldsymbol{\alpha}) + \lambda\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}$, where λ is a regularization parameter. The 2 random effects $\boldsymbol{\alpha}$ and $\mathbf{e}$ were distributed as $\boldsymbol{\alpha} \sim N(0, \mathbf{K}^{-1}\sigma_g^2)$ and $\mathbf{e} \sim N(0, \mathbf{R}^{-1}\sigma_e^2)$, where σ_g^2 and σ_e^2 are the genetic and residual variances, respectively, and $\mathbf{R}$ is a diagonal matrix with its elements representing reliabilities of the SCR records.

Single-kernel models were fitted using either linear or Gaussian reproducing kernels (**K**). The linear (L) kernel, which is equivalent to the well-known additive genomic relationship matrix formulated by VanRaden (2008), takes the form $\mathbf{K}_{\mathrm{L}} = \mathbf{SS}'/p$, where $\mathbf{S}$ is a matrix of centered and standardized SNP genotypes and p represents the number of SNP. In the case of the Gaussian reproducing kernel, $\mathbf{K}_{\mathrm{G}}$ was evaluated in the (average) squared-Euclidean distance between genotypes as follows:

$$\mathbf{K}(\mathbf{w}_i, \mathbf{w}_{i'}) = \exp\left[-h \times \frac{\sum_{k=1}^{p}(\mathbf{w}_{ik} - \mathbf{w}_{i'k})^2}{p}\right],$$

where $\mathbf{w}_i$ and $\mathbf{w}_{i'}$ are the genotype vector of bulls i and i', h is the bandwidth parameter that was chosen following the Bayesian approach proposed by Pérez-Elizalde et al. (2015), and k is an index for SNP markers.

Multikernel Models. This model allows the simultaneous fitting of multiple sets of SNP. In our case, only 2 subsets were fitted, one with the SNP of interest (given their biological role) and the other with the remaining set of SNP markers. The general multikernel regression model is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{K}_1\boldsymbol{\alpha}_1 + \mathbf{K}_2\boldsymbol{\alpha}_2 + \mathbf{e},$$

where $\mathbf{K}_1$ is the kernel matrix linking SCR records to the SNP of interest and $\mathbf{K}_2$ is other kernel matrix linking SCR records to the remaining SNP. The random genomic and residual effects were assumed to be independent and normally distributed as $\boldsymbol{\alpha}_1 \sim N(0, \mathbf{K}_1{}^{-1}\sigma_{g1}^2)$, $\boldsymbol{\alpha}_2 \sim N(0, \mathbf{K}_2{}^{-1}\sigma_{g2}^2)$, and $\mathbf{e} \sim N(0, R^{-1}\sigma_e^2)$. Multikernel models were also fitted using either linear or Gaussian kernels; these kernels were constructed as described above.

Implementation

Kernel models were implemented in a Bayesian framework using Gibbs sampling. For each model, a Markov chain Monte Carlo was run with a total of 100,000 iterations, with a burn-in of 30,000 and a thinning interval of 5, so that a total of 14,000 samples were used for computing features of the posterior distribution. Most runs lasted less than 24 h. Convergence diagnostics were carried out by visual inspection of trace plots of some parameters, such as variance components. All the analyses were performed using the R package Bayesian Generalized Linear Regression (version 1.0.4; Pérez and de los Campos, 2014).

Model Comparison

The ability of the different RKHS regression models to predict SCR was assessed using 5-fold cross-validation. Briefly, the entire data set was partitioned into 5 sets, with an imposed restriction that all levels of the fixed effects were represented in each of the sets. Then, solutions for all fixed and random effects of the training set (*train*) were estimated and used to predict SCR values in the testing set (*test*). This cross-validation procedure was repeated 5 times, so each analysis resulted in 25 estimates. Additionally, in those analyses with random SNP sets, given that the process of sampling SNP markers across the genome was completely random, the sampling was repeated 20 times, so each analysis resulted in a total of 500 estimates.

In the single-kernel model scenario, the prediction of genetic values in the testing set $(\hat{\boldsymbol{g}}_{test})$ is given by $\hat{\boldsymbol{g}}_{test} = \mathbf{K}_{test,train}\mathbf{K}_{train}^{-1}\hat{\boldsymbol{g}}_{train}$, where $\mathbf{K}_{test,train}$ is a rectangular kernel matrix of genomic relationships between training and testing bulls (a subset of the total $\mathbf{K}$ constructed using all the animals in the data set), $\mathbf{K}_{train}$ is the 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. In the multikernel model scenario, the prediction of genetic values using subset q for animals in the testing set is given by $\hat{\boldsymbol{g}}_{q,test} = \mathbf{K}_{q,test,train}\mathbf{K}_{q,train}^{-1}\hat{\boldsymbol{g}}_{q,train}$, where notations are as in the single-kernel scenario except that $q \in (1,2)$ denotes the kernel matrix of the qth SNP subset. The predicted SCR values were $\hat{\boldsymbol{y}}_{test} = \boldsymbol{X}_{test}\hat{\boldsymbol{b}}_{train} + \hat{\boldsymbol{g}}_{test}$ and $\hat{\boldsymbol{y}}_{test} = \boldsymbol{X}_{test}\hat{\boldsymbol{b}}_{train} + \hat{\boldsymbol{g}}_{1,test} + \hat{\boldsymbol{g}}_{2,test}$ in the single- and multikernel model scenarios, respectively.

The predictive ability of the alternative RKHS models was assessed comparing observed SCR values $(\boldsymbol{y})$ with predicted SCR values $(\hat{\boldsymbol{y}}_{test})$ in the testing set using the Pearson product–moment correlation coefficient (**COR**) and the mean squared error of prediction (**MSEP**), defined as $\mathrm{MSEP} = n^{-1}\sum_{f=1}^{5}\sum(\boldsymbol{y} - \hat{\boldsymbol{y}}_{test})^2$. Although COR is a very intuitive way of measuring predictive ability, MSEP is in fact a preferred metric because it considers both prediction bias and variability, whereas the predictive correlation provides only a measure of association (Abdollahi-Arpanahi et al., 2016).

RESULTS AND DISCUSSION

Both female and male fertility are very important traits for the dairy industry. Fertility traits are arguably good examples of complex phenotypes—traits that are influenced simultaneously by a large number of small-effect genes and environmental factors. Therefore, the genetic dissection of these traits is in general very challenging for pinpointing causal mutations or predicting future values. Dairy cow fertility has received much attention in the last decades, whereas service sire fertility has been largely ignored. This study was specially conducted to assess genomic prediction of sire conception rate, the US national phenotypic evaluation of dairy bull fertility. We first evaluated the feasibility of whole-genome prediction of SCR. Second, we investigated the predictive performance of alternative biologically informed genomic models.

Predicting Sire Conception Rate Using Whole-Genome Data

Figure 1 displays the predictive performance of the whole-genome kernel-based model fitting a single linear kernel. Note that this particular single-kernel model is mathematically equivalent to the GBLUP model. The average predictive correlation was equal to 0.341; this value is the average of 25 independent predictive correlations (i.e., 5-fold cross-validation repeated 5 times). By definition, the mean of the SCR values per evaluation is zero; hence, the categorical variable evaluation has a negligible effect in prediction. Therefore, the predictive correlation of 0.341 is a good estimate of the correlation between observed phenotypic values and predicted breeding values. The corresponding prediction accuracy, defined as the correlation between the true and the predicted breeding values, often is obtained by dividing the predictive correlation by the square root of the trait heritability (e.g., Ober et al., 2012). Here, if we divide the estimated predictive correlation by the square root of SCR heritability ($h^2 \approx 0.30$; Supplemental Table S1, https://doi.org/10.3168/jds.2017-13288), we get a predictive accuracy equal to 0.63. Interestingly, sire calving ease and sire stillbirth rate, 2 calving traits routinely evaluated in US dairy breeds, have selection accuracies (square root of the reliability) of around 0.55 (Wiggans et al., 2007). Recently, Parker Gaddis et al. (2014) reported accuracies for novel producer-recorded health traits, such as ketosis, lameness, and metritis, of around 0.60 for young genomic sires. Overall, our findings are promising and suggest that genomic prediction of service sire fertility is feasible. This study could be the foundation for the development of genomic tools that help the dairy industry make accurate genome-guided decisions, such as early culling of predicted subfertile bull calves.

Comparing Predictive Ability of Different SNP Classes

Table 1 shows the distribution of the SNP among different functional categories. Of the 54,807 SNP evaluated in this study, a total of 25,619 were located within or surrounding annotated genes. In addition, a total of 870 and 337 of these genic SNP pointed to genes within the GO and MeSH terms, respectively. About 35% of these SNP linked to GO and MeSH terms were found among the TOP SNP—that is, the set of markers associated (nominal P-value ≤ 0.05) with sire conception rate.

The predictive performance of single-kernel models fitting linear kernels with different subsets of SNP markers is shown in Figure 1. The genic SNP class achieved similar predictive ability in terms of both COR and MSEP than all the SNP. In addition, the gene set SNP classes (GO, MeSH, or the combination of GO and MeSH) yielded lower predictive performance than the genic SNP, although in principle these findings are promising if we consider the number of SNP in each of the terms. The predictive power of each of these SNP classes was compared with that exhibited by SNP randomly sampled across the genome. In this sense, neither the genic SNP nor the gene set SNP outperformed their counterpart with random SNP sets. We should conclude, therefore, that the predictive ability exhibited by the functional classes of SNP is not driven by their biological roles but rather by accounting for genomic relationships. Evaluating the predictive ability of different genomic regions in broiler chickens, both Morota et al. (2014) and Abdollahi-Arpanahi et al. (2016) concluded that SNP within coding regions yielded similar predictions compared with SNP in intergenic regions regardless of the trait under study.

Of particular interest, the class TOP SNP, consisting of markers that showed a significant association with SCR, achieved a slightly better predictive ability than the entire SNP data set. Indeed, this SNP class showed lower mean squared error of prediction (4.12 vs. 4.16) and higher predictive correlation (0.347 vs. 0.341) than all the SNP. This represents an increase in accuracy of about 2%. These TOP SNP also showed better predictive power than random SNP. It should be emphasized that the significance of each SNP was systematically evaluated in each iteration of the cross-validation in the training data, and only the markers

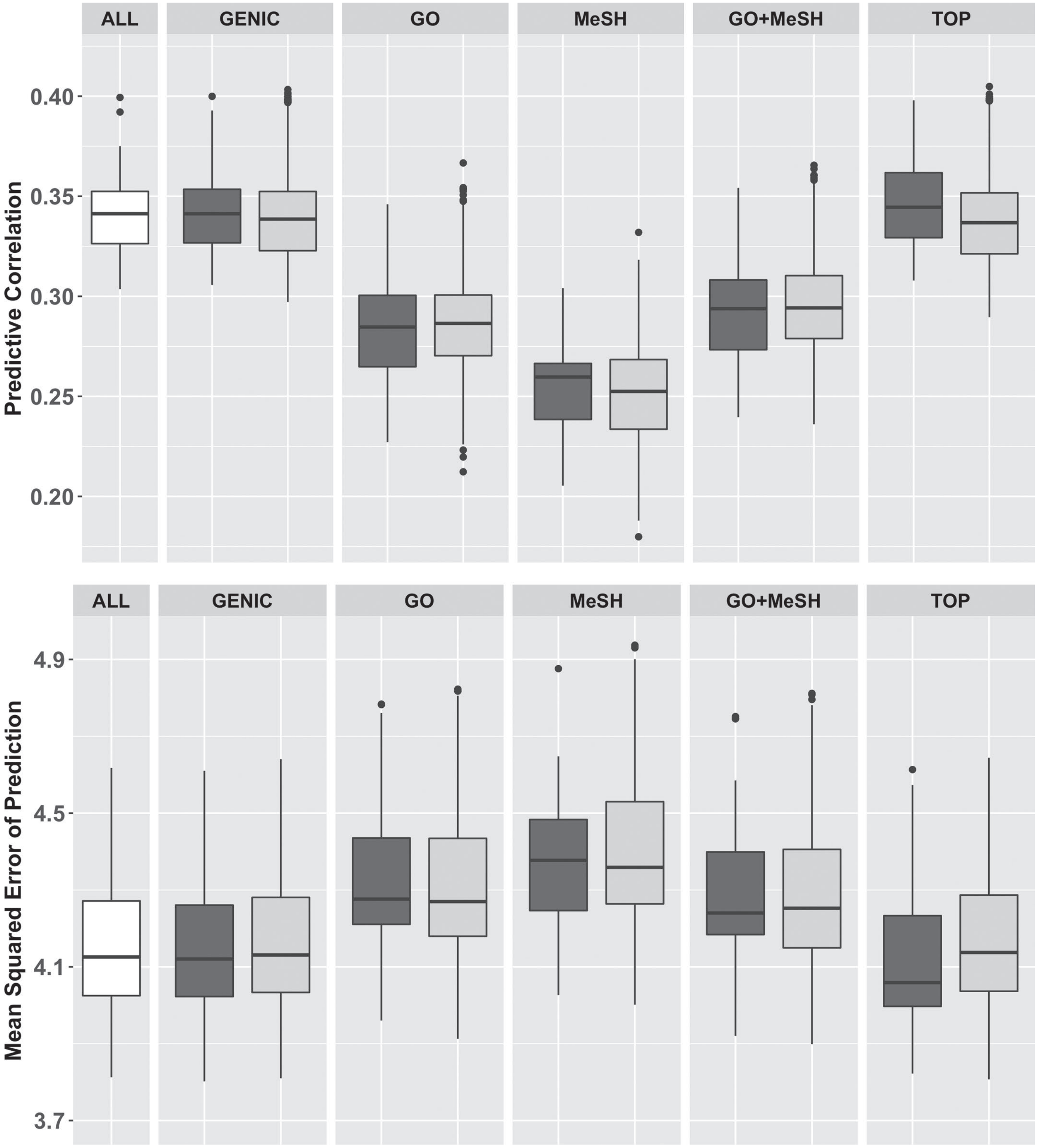

Figure 1. Predictive ability of single-kernel models using different SNP subsets. Predictive ability was evaluated using predictive correlation (top) and mean squared error of prediction (bottom). Each analysis was performed using either the SNP class of interest (dark gray) or a set of SNP with the same size but randomly sampled from the genome (light gray). GO = Gene Ontology; MeSH = Medical Subject Headings; TOP = SNP markers with a nominal P-value ≤ 0.05. The bottom and top of the box represent first and third quartiles; the vertical line denotes the median; the whiskers correspond to 1.5 × interquartile distance; and dark dots are outliers.

Table 1. Distribution of SNP markers per functional category[1]

Item	All SNP	Genic SNP	GO SNP	MeSH SNP	GO + MeSH SNP	TOP SNP
All SNP	54,807	25,619	870	337	1,142	18,659
Genic SNP		25,619	870	337	1,142	8,952
GO SNP			870	65	870	319
MeSH SNP				337	337	114
GO + MeSH SNP					1,142	408
TOP SNP						18,659

[1]GO = Gene Ontology; MeSH = Medical Subject Headings; TOP = SNP markers with a nominal P-value $\leq$ 0.05.

with P-value $\leq$0.05 were used to predict unobserved SCR values in the testing data. Weigel et al. (2009) and Moser et al. (2010) evaluated the predictive ability of different subsets of SNP in Holstein cattle, and they found that small subsets containing the markers with the largest SNP effects exhibited predictive power comparable with that obtained using all the markers. Similarly, Abdollahi-Arpanahi et al. (2014) reported in chicken that subsets containing the SNP with the largest effects provided similar or even higher predictive correlations than using all the 350K markers that they had available.

Figure 2 shows the comparison in predictive performance between linear and Gaussian kernels. The use of nonadditive kernels such as Gaussian allows the exploration of nonlinear relationships between genotypes and phenotypes. Here, irrespective of the set of SNP under consideration, Gaussian kernel models outperformed their counterparts fitting linear kernels, showing systematically lower MSEP values and higher COR values. The class TOP SNP exhibited again the best predictive performance with an increase in predictive correlation of 4.1% compared with a standard whole-genome approach. These findings suggest that considering nonadditive effects would benefit the prediction of bull fertility.

Predictive Performance of Multikernel Models

Figure 3 displays the predictive performance of the alternative multikernel models fitting either linear or Gaussian kernels. Kernel-based models fitting Gaussian kernels exhibited again better predictive ability than their counterparts fitting linear kernels. The gene set kernel-based models (namely GO, MeSH, or combining simultaneously GO and MeSH information) showed predictive ability similar to that of single-kernel models fitting all the SNP irrespective of the type of kernel under consideration. In other words, the use of gene set informed models did not improve predictive ability. On the contrary, Edwards et al. (2016) recently reported that the use of prior GO information improved the prediction of different quantitative traits (startle response, starvation resistance, and chill coma recovery) in *Drosophila melanogaster*. Note that these authors worked with a population of unrelated individuals using whole-genome sequence data.

The multikernel Gaussian model fitting TOP SNP delivered again the highest predictive ability, in this case with an increase in accuracy of 4.7% (0.357 vs. 0.341) compared with the standard genomic approach. Using a similar idea, de los Campos et al. (2013b) reported that a GBLUP model with the **G** matrix weighted with rescaled P-values showed better performance than the standard GBLUP on predicting complex traits in humans. Similarly, Tiezzi and Maltecca (2015) showed that informing the **G** matrix with estimated marker effects resulted in increased predictive performance in dairy cattle, especially for fat percentage and protein percentage—2 traits regulated by few major genes.

Gene Sets as a Source of Relevant Biological Information

Gene set enrichment analysis has proven to be a great complement of genome-wide association analysis (Gambra et al., 2013; Abdalla et al., 2016). Among available gene set databases, GO is probably the most popular, whereas MeSH is a relatively new tool that is gaining ground in livestock genomics (Morota et al., 2015, 2016). We had hypothesized that the use of gene set information could improve prediction. However, neither of the gene set SNP classes outperformed the standard whole-genome approach. Gene sets have been primarily developed using data from model organisms, such as mice and flies, so it is possible that some of the genes included in these terms are irrelevant for bull fertility. Indeed, in our study only 35% of the SNP linked to GO and MeSH terms were significantly associated with SCR (Table 1). It is likely that a better understanding of the biology underlying bull fertility specifically, plus an advance in the annotation of the bovine genome, can provide new opportunities for predicting SCR using gene set information. Moreover, there is growing evidence that a large number of variants that explain the variation in complex phenotypes reside in regulatory re-

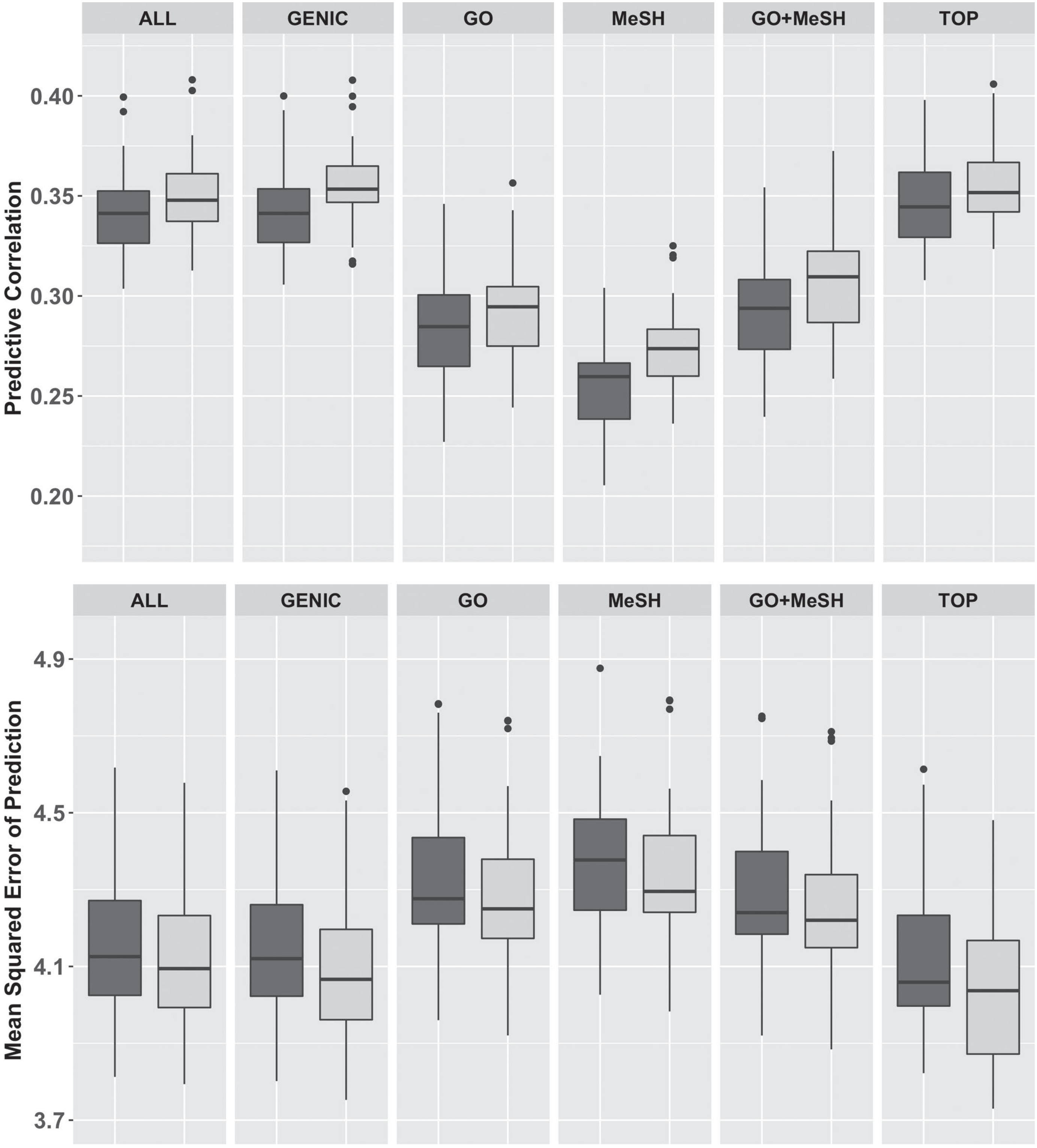

Figure 2. Predictive ability of single-kernel models using different SNP subsets. Predictive ability was evaluated using predictive correlation (top) and mean squared error of prediction (bottom). Each analysis was performed using a linear kernel (dark gray) or a Gaussian kernel (light gray). GO = Gene Ontology; MeSH = Medical Subject Headings; TOP = SNP markers with a nominal *P*-value ≤ 0.05. The bottom and top of the box represent first and third quartiles; the vertical line denotes the median; the whiskers correspond to 1.5 × interquartile distance; and dark dots are outliers.

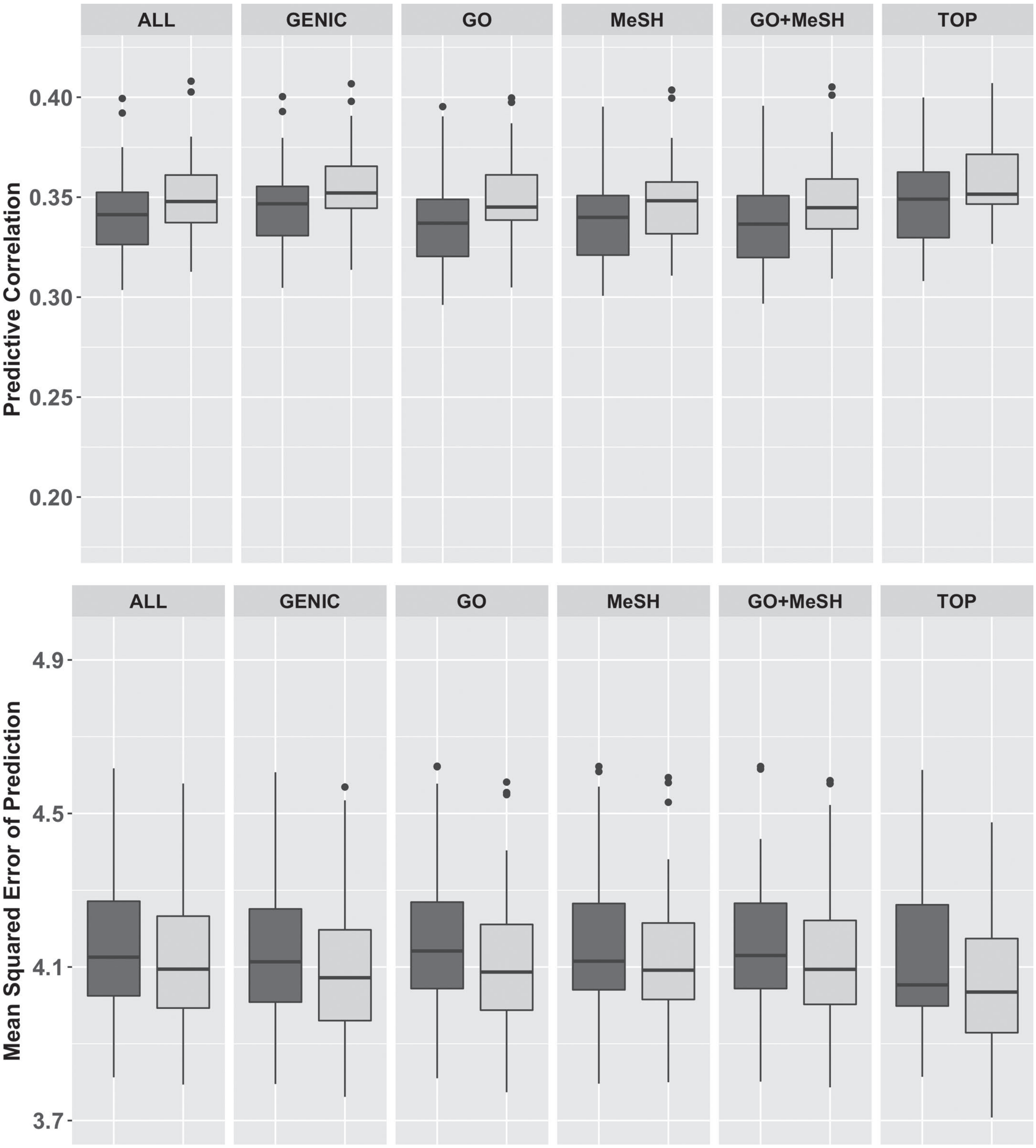

Figure 3. Predictive ability of multikernel models using different SNP subsets. Predictive ability was evaluated using predictive correlation (top) and mean squared error of prediction (bottom). Each analysis was performed using both linear kernels (dark gray) and Gaussian kernels (light gray). GO = Gene Ontology; MeSH = Medical Subject Headings; TOP = SNP markers with a nominal P-value ≤ 0.05. The bottom and top of the box represent first and third quartiles; the vertical line denotes the median; the whiskers correspond to 1.5 × interquartile distance; and dark dots are outliers.

gions that alter gene expression (Cookson et al., 2009). Some of these regulatory regions, such as enhancers, are located far from the genes. Therefore, although the gene might be part of the analysis, the relevant variant would probably not be included in the gene set SNP class. Finally, linkage disequilibrium interferes with the use of biological information in prediction because irrelevant regions (regions without any biological role) capture part of the information encoded in relevant regions, causing both regions to exhibit similar predictive abilities. The use of very high density SNP data or even whole-genome sequence data could alleviate some of these issues.

CONCLUSIONS

Our findings suggest that the genomic prediction of dairy bull fertility is feasible. This could have a positive effect on the dairy industry, allowing, for example, the early culling of bull calves with very low SCR predictions. We also evaluated alternative kernel models to incorporate biological information and thus try to improve prediction. Indeed, kernel-based analysis provides a very flexible and elegant framework for performing whole-genome prediction incorporating relevant prior knowledge (e.g., markers with large effects, major genes, or gene networks). Although prediction accuracy was improved by using SNP with the largest effects, neither GO nor MeSH terms outperformed the standard whole-genome approach. The potential inclusion of gene set results in prediction models deserves further research.

ACKNOWLEDGMENTS

The authors thank the Cooperative Dairy DNA Repository (Beltsville, MD) for providing the genotypic data. This research was supported by the Florida Agricultural Experiment Station (Gainesville) and the Department of Animal Sciences, University of Florida (Gainesville).

REFERENCES

Abdalla, E. A., F. Penagaricano, T. M. Byrem, K. A. Weigel, and G. J. Rosa. 2016. Genome-wide association mapping and pathway analysis of leukosis incidence in a US Holstein cattle population. Anim. Genet. 47:395–407.

Abdollahi-Arpanahi, R., G. Morota, B. D. Valente, A. Kranis, G. J. 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.

Abdollahi-Arpanahi, R., A. Nejati-Javaremi, A. Pakdel, M. Moradi-Shahrbabak, G. Morota, B. D. Valente, A. Kranis, G. J. Rosa, and D. Gianola. 2014. Effect of allele frequencies, effect sizes and number of markers on prediction of quantitative traits in chickens. J. Anim. Breed. Genet. 131:123–133.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene ontology: Tool for the unification of biology. Nat. Genet. 25:25–29.

Coletti, M. H., and H. L. Bleich. 2001. Medical subject headings used to search the biomedical literature. J. Am. Med. Inform. Assoc. 8:317–323.

Cookson, W., L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop. 2009. Mapping complex disease traits with global gene expression. Nat. Rev. Genet. 10:184–194.

Crossa, J., P. Perez, J. Hickey, J. Burgueno, L. Ornella, J. Ceron-Rojas, X. Zhang, S. Dreisigacker, R. Babu, Y. Li, D. Bonnett, and K. Mathews. 2014. Genomic prediction in cimmyt maize and wheat breeding programs. Heredity 112:48–60.

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. Calus. 2013a. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–345.

de los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis, and D. Sorensen. 2013b. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. 9:e1003608.

DeJarnette, J. M., C. E. Marshall, R. W. Lenz, D. R. Monke, W. H. Ayars, and C. G. Sattler. 2004. Sustaining the fertility of artificially inseminated dairy cattle: The role of the artificial insemination industry. J. Dairy Sci. 87(Suppl.):E93–E104.

Durinck, S., Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber. 2005. Biomart and bioconductor: A powerful link between biological databases and microarray data analysis. Bioinformatics 21:3439–3440.

Durinck, S., P. T. Spellman, E. Birney, and W. Huber. 2009. Mapping identifiers for the integration of genomic datasets with the R/bioconductor package biomart. Nat. Protoc. 4:1184–1191.

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

Gambra, R., F. Peñagaricano, J. Kropp, K. Khateeb, K. A. Weigel, J. Lucey, and H. Khatib. 2013. Genomic architecture of bovine κ-casein and β-lactoglobulin. J. Dairy Sci. 96:5333–5343.

García-Ruiz, A., J. B. Cole, P. M. VanRaden, G. R. Wiggans, F. J. Ruiz-Lopez, 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. 2013. Priors in whole-genome regression: The Bayesian alphabet returns. Genetics 194:573–596.

Gianola, D., and J. B. 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.

Ibáñez-Escriche, N., S. Forni, J. L. Noguera, and L. Varona. 2014. Genomic information in pig breeding: Science meets industry needs. Livest. Sci. 166:94–100.

Kadarmideen, H. N. 2014. Genomics to systems biology in animal and veterinary sciences: Progress, lessons and opportunities. Livest. Sci. 166:232–248.

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

Lango Allen, H., K. Estrada, G. Lettre, S. I. Berndt, and M. N. Weedon. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467:832–838.

Lin, Z., B. J. Hayes, and H. D. Daetwyler. 2014. Genomic selection in crops, trees and forages: A review. Crop Pasture Sci. 65:1177–1191.

MacLeod, I. M., P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, K. E. Kemper, A. J. Chamberlain, C. Schrooten, B. J. Hayes, and M. E. Goddard. 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics 17:144.

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.

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., T. M. Beissinger, and F. Penagaricano. 2016. Mesh-informed enrichment analysis and mesh-guided semantic similarity among functional terms and gene products in chicken. G3 (Bethesda) 6:2447–2453.

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

Morota, G., F. Peñagaricano, J. L. Petersen, D. C. Ciobanu, K. Tsuyuzaki, and I. Nikaido. 2015. An application of mesh enrichment analysis in livestock. Anim. Genet. 46:381–387.

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.

Ober, U., J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu, R. A. Gibbs, C. Stricker, D. Gianola, M. Schlather, T. F. C. Mackay, and H. Simianer. 2012. Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. PLoS Genet. 8:e1002685.

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.

Pérez, P., and G. de los Campos. 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483–495.

Pérez-Elizalde, S., J. Cuevas, P. Pérez-Rodríguez, and J. Crossa. 2015. Selection of the bandwidth parameter in a Bayesian kernel regression model for genomic-enabled prediction. J. Agric. Biol. Environ. Stat. 20:512–532.

Tiezzi, F., and C. Maltecca. 2015. Accounting for trait architecture in genomic predictions of US Holstein cattle using a weighted realized relationship matrix. Genet. Sel. Evol. 47:24.

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

Vazquez, A. I., G. de los Campos, Y. C. Klimentidis, G. J. M. Rosa, D. Gianola, N. Yi, and D. B. Allison. 2012. A comprehensive genetic approach for improving prediction of skin cancer risk in humans. Genetics 192:1493–1502.

Weigel, K. A., G. de los Campos, O. Gonzalez-Recio, H. Naya, X. L. Wu, N. Long, G. J. 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.

Wiggans, G. R., J. B. Cole, S. M. Hubbard, and T. S. Sonstegard. 2017. Genomic selection in dairy cattle: The USDA experience. Annu. Rev. Anim. Biosci. 5:309–327.

Zhang, Z., M. Erbe, J. He, U. Ober, N. Gao, H. Zhang, H. Simianer, and J. Li. 2015. Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix. G3 (Bethesda) 5:615–627.

Zhang, Z., U. Ober, M. Erbe, H. Zhang, N. Gao, J. He, J. Li, and H. Simianer. 2014. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PLoS One 9:e93017.

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.