Predicting bull fertility using biologically informed genomic models

R. Abdollahi-Arpanahi [1,2], *G. Morota*[3] *& F. Peñagaricano*[1]

[1] *Department of Animal Sciences, University of Florida, USA*
[2] *Department of Animal and Poultry Science, University of Tehran, Iran*
[3] *Department of Animal Science, University of Nebraska, USA*
fpenagaricano@ufl.edu (Corresponding Author)

Summary

The genomic prediction of unobserved genetic values or future phenotypes for complex traits has revolutionized agriculture and human medicine. Fertility traits are complex traits of great economic importance to the dairy industry. While genomic prediction for improved cow fertility has received much attention over the last years, bull fertility has been largely 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 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 SNPs (55K) or only markers with presumed functional roles, such as SNPs linked to Gene Ontology (GO) or Medical Subject Headings (MeSH) terms related to male fertility, or SNPs significantly associated with SCR. Both single- and multi-kernel models were evaluated using linear and Gaussian kernels. Predictive ability was evaluated in 5-fold cross-validation. The entire set of SNPs exhibited predictive correlations around 0.35. Neither GO or MeSH gene sets achieved predictive abilities higher than their counterparts using random sets of SNPs. Notably, kernel models fitting significant SNPs 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 SNPs. Overall, our findings suggest that genomic prediction of bull fertility is feasible in dairy cattle. This provide potential for accurate genome-guided decisions, such as culling bull calves with low SCR predictions. In addition, exploiting nonlinear effects using Gaussian kernels together with the incorporation of relevant variants seems a promising alternative to the standard approach. The inclusion of gene set results into prediction models deserves further research.

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

Introduction

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 GBLUP method, considered as the benchmarking approach for whole-genome prediction, assumes a priori that all SNPs have an effect on the trait under study, and all these SNP effects are of similar magnitude. Similarly, other popular methods, such as Bayes B, Bayes C, or Bayes R also ignore any prior biological knowledge available and assume that all the SNPs are equally likely to affect the trait of interest. However, association studies have been successful in identifying genetic variants, genes and pathways associated with numerous complex traits. The incorporation of this information into predictive models could positively impact both model predictive ability and model robustness.

Fertility traits are very complex traits of great economic importance to the dairy industry. While most studies have focused on cow fertility, dairy bull fertility has been largely ignored. Indeed, 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 Sire Conception Rate (SCR) in US Holstein bulls using high-density SNP data. Recently, we identified biological pathways and gene sets associated with SCR (Han and Peñagaricano, 2016). As such, the second objective of this study was to incorporate biological information into alternative predictive models and evaluate their predictive ability.

Material and methods

Phenotypic and genotypic data

Since 2008, the US dairy industry has access to a national phenotypic evaluation of sire fertility called Sire Conception Rate (SCR). A total of 7,447 Holstein bulls with both SCR records and genome-wide (55k) SNP data were used in this study.

Combining genomic data with biological information

For the first objective, i.e., assess the performance of genomic models for predicting SCR, alternative predictive models using the entire SNP dataset were evaluated. For the second objective, where the goal was to predict bull fertility combining genomic and biological information, different SNP subsets were investigated, such as SNPs within/near genes, SNPs linked to genes in relevant gene sets, and SNPs that were significantly associated with SCR.

Genic SNPs. A given SNP was assigned to a particular annotated gene if it was located within the genomic sequence of the gene or within 15 kb either upstream or downstream the gene.
Gene Set SNPs. Gene sets can be defined as groups of genes that share some properties. Based on our previous work (Han and Peñagaricano, 2016), we evaluated the set of SNPs linked to genes in the GO term Reproduction (GO SNPs) and also the set of SNPs linked to genes associated with a group of MeSH terms (MeSH SNPs) related to sperm biology.
Significant SNPs. The association between each SNP marker and SCR was assessed using a linear model with the SNP allele count as a linear covariate. Those SNP markers with nominal P-value < 0.05 were considered as significant SNPs (TOP SNPs).

The performance of each SNP subset was compared to the performance exhibited by another SNP subset with the same number of markers but randomly selected across the genome.

Statistical Models

The goal was to predict yet-to-be observed SCR phenotypes using genomic data. The predictive ability of either the entire SNP set or the alternative SNP subsets was investigated using Reproducing Kernel Hilbert Spaces (RKHS) regression models (Morota and Gianola, 2014). Kernel-based models are powerful predictive machines, and they allow to integrate, in a very simple way, prior biological information, e.g., functional variants, gene sets, and pathway data. We investigated the performance of both single and multi-kernel (2-kernel) models using either linear or Gaussian kernels. These models were implemented in a Bayesian framework using Gibbs sampling. All the analyses were performed using the R package Bayesian Generalized Linear Regression (BGLR; version 1.0.4) (Pérez and de los Campos, 2014).

Model Comparison

The predictive ability of the different RKHS models was assessed in 5-fold cross-validation with 5 replications comparing observed vs. predicted SCR values in the testing set using both correlation (COR) and the mean-squared error of prediction (MSEP).

Results and discussion

Predicting bull fertility using whole-genome data

Figure 1 displays the predictive performance of the whole-genome (ALL) kernel-based model fitting a single linear kernel. This model is equivalent to the GBLUP model. The average predictive correlation was equal to 0.341. If we divide this predictive correlation by the square root of the trait heritability, we get a predictive accuracy around 0.65. Interestingly, sire calving ease and sire stillbirth rate, 2 calving traits evaluated in US dairy breeds, have selection accuracies (square root of the reliability) of around 0.55. A recent study reported accuracies for novel producer-recorded health traits, such as ketosis and metritis, around 0.60 for young genomic sires (Parker Gaddis et al. (2014). 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 to make accurate genome-guided decisions, such as early culling of predicted subfertile bull calves.

Comparing predictive ability of different SNP classes

Of the 54,807 SNPs evaluated in this study, 25,619 were located within or near annotated genes. A total of 870 and 337 of these genic SNPs pointed to genes within the GO and MeSH terms, respectively. About 35% of these gene-set SNPs were found among the TOP SNPs.

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 SNPs. The gene-set SNP classes, either GO, MeSH, or GO+MeSH, yielded lower predictive performance than the genic SNPs, although in principle these findings are promising if we consider the number of SNPs in each term. However, neither the genic nor the gene-set SNPs outperformed their counterpart with random SNPs. We should conclude that the predictive ability exhibit by the functional SNPs is not driven by their biological roles, but rather by accounting for genomic relationships.

The class TOP SNPs achieved a slightly better predictive ability than the entire SNP dataset, with lower MSEP (4.12 vs 4.16) and higher COR (0.347 vs 0.341) than all the SNPs. This represents an increase in accuracy of about 2%. These TOP SNPs also showed better predictive ability than random SNPs. Note that the significance of each SNP was evaluated in each iteration of the cross-validation in the training data, and only the significant markers used to predict unobserved SCR values in the testing data. Moser et al. (2010) evaluated the predictive ability of different subsets of SNPs in Holstein cattle, and they found that small subsets containing the markers with the largest SNP effects exhibited comparable predictive power than that obtained using all the markers.

Figure 2 shows the comparison in predictive ability between linear and Gaussian kernels. Models fitting Gaussian kernels allow to explore nonlinear relationships between genotypes and phenotypes. Here, irrespective of the set of SNPs under consideration, Gaussian kernel models outperformed their counterparts fitting linear kernels, showing lower MSEP values and higher COR values. The TOP SNPs exhibited again the best predictive power with an increase in predictive correlation of 4.1% compared to whole-genome approach. These findings suggest that considering non-additive effects would benefit bull fertility prediction.

Predictive performance of multi-kernel models

Multi-kernel models fitting Gaussian kernels exhibited better predictive ability than their counterparts fitting linear kernels. Gene-set kernel-based models showed similar predictive ability than single-kernel models fitting all the SNPs, irrespective of kernel under study. The multi-kernel Gaussian model fitting TOP SNPs 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. 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, two traits regulated by few major genes.

Conclusions

Our findings suggest that the genomic prediction of dairy bull fertility is feasible. This could have a positive impact on the dairy industry, allowing the early culling of bull calves with very low SCR predictions. We also evaluated alternative kernel models to incorporate biological information. Indeed, kernel-based analysis provides an elegant framework to perform whole-genome prediction incorporating relevant prior knowledge, e.g., relevant markers or gene networks. While prediction accuracy was improved by using SNPs with the largest effects, neither GO nor MeSH terms outperformed the standard whole-genome approach. The inclusion of gene set results into prediction models deserves further research.

List of references

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

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 markers. Genet. Sel. Evol. 42:37.

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.

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.

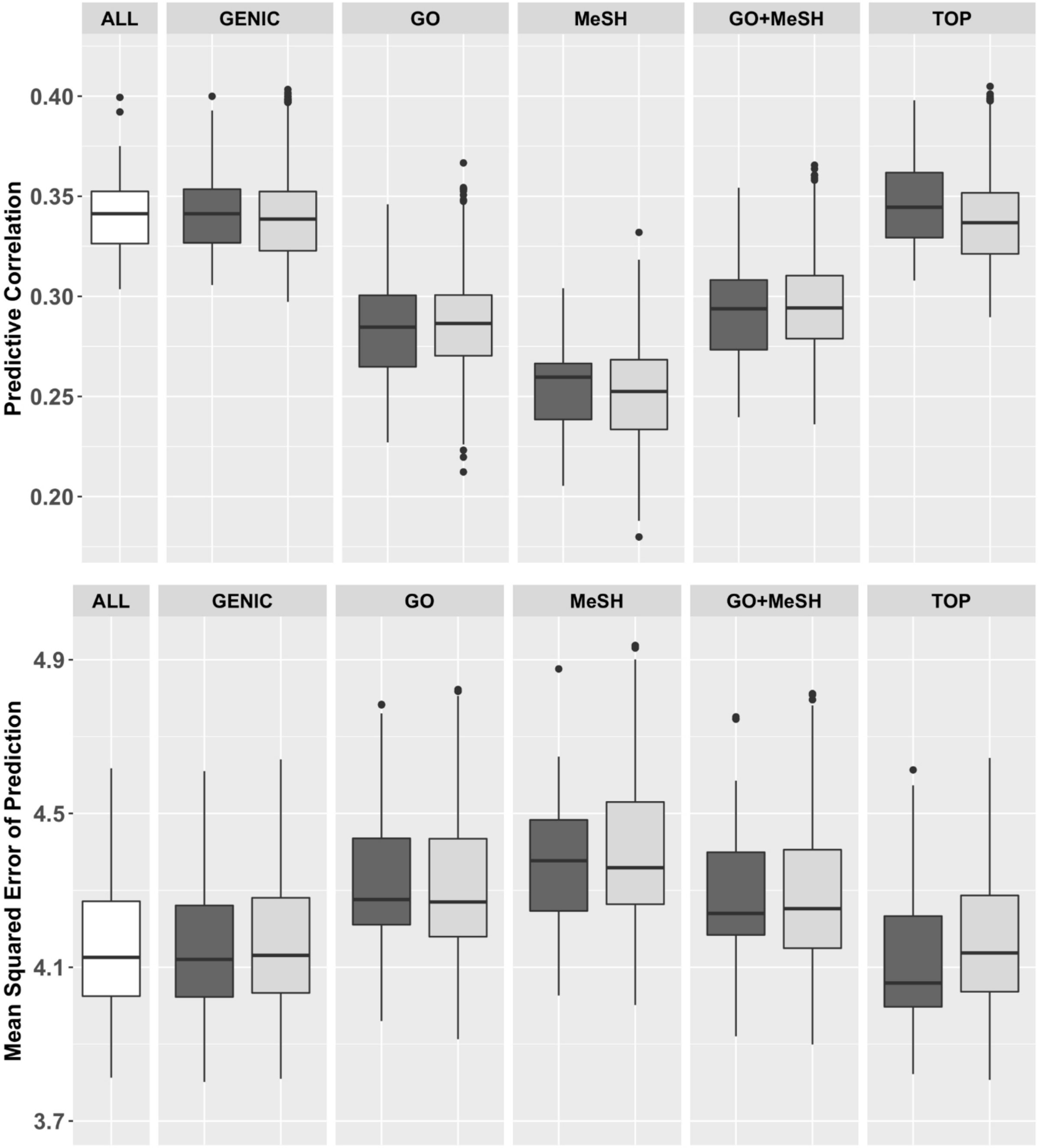

Figure 1. Predictive ability of single-kernel models using different SNPs subsets. Predictive ability was evaluated using predictive correlation (top) and mean squared error of prediction (bottom). Each analysis was performed using either the SNP class of interest (dark grey) or a set of SNPs with the same size but randomly sampled from the genome (light grey).

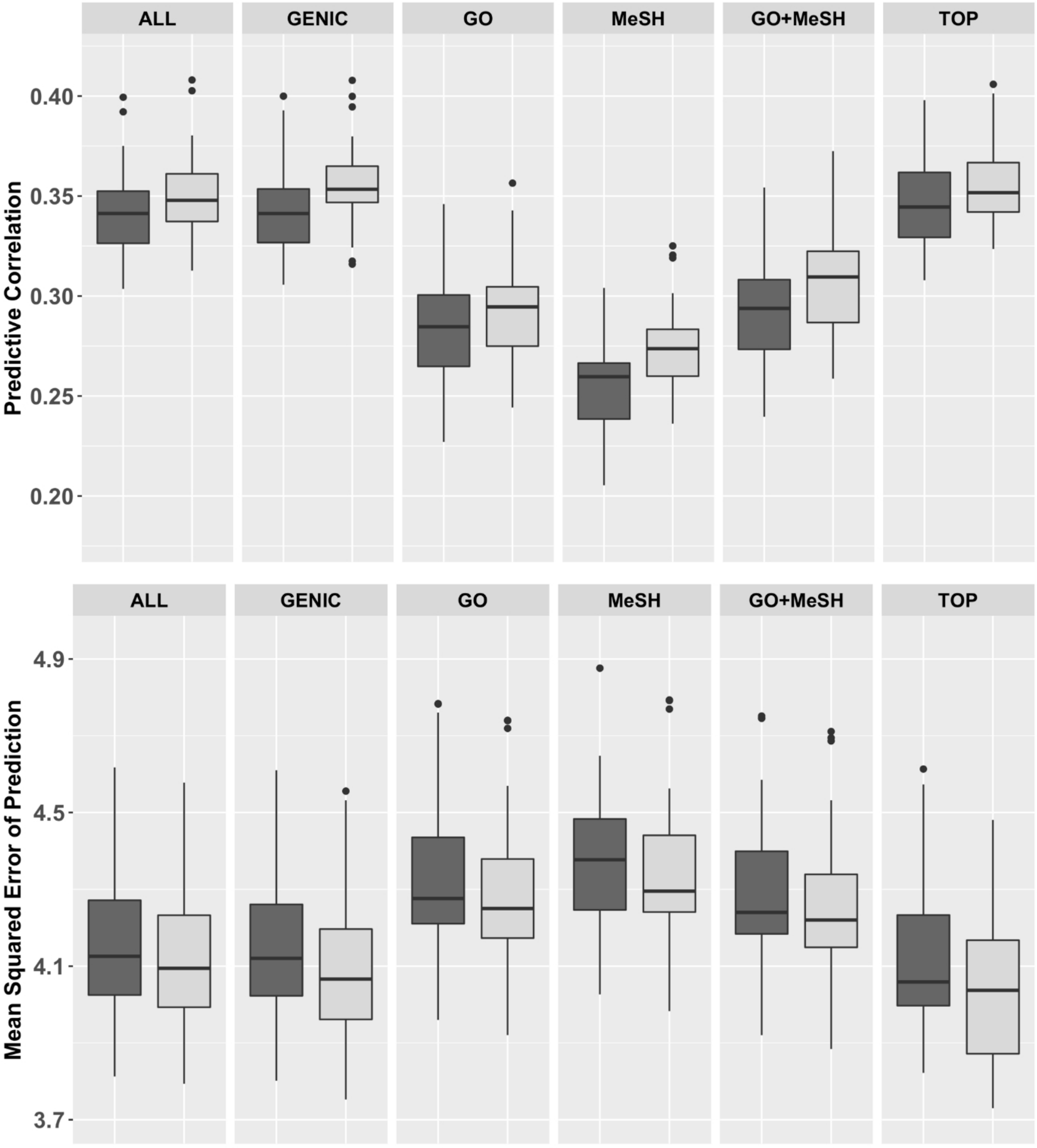

Figure 2. Predictive ability of single-kernel models using different SNPs subsets. Predictive ability was assessed using predictive correlation (top) and mean squared error of prediction (bottom). Each analysis was performed using a linear kernel (dark grey) or a Gaussian kernel (light grey).