J. Dairy Sci. 100:1–11
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

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

(de los Campos et al., 2013a). Similarly, other popular genomic prediction methods, such as Bayes B, Bayes C, 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