J. Dairy Sci. 93:5423–5435
doi:10.3168/jds.2010-3149

Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle

K. A. Weigel,*[1] G. de los Campos,* A. I. Vazquez,* G. J. M. Rosa,*† D. Gianola,*†‡ and C. P. Van Tassell§#
*Department of Dairy Science,
†Department of Biostatistics and Medical Informatics, and
‡Department of Animal Sciences, University of Wisconsin, Madison 53706
§Bovine Functional Genomics Laboratory, and
#Animal Improvement Programs Laboratory, ARS, USDA, Beltsville, MD 20705-2350

ABSTRACT

The objective of the present study was to evaluate the predictive ability of direct genomic values for economically important dairy traits when genotypes at some single nucleotide polymorphism (SNP) loci were imputed rather than measured directly. Genotypic data consisted of 42,552 SNP genotypes for each of 1,762 Jersey sires. Phenotypic data consisted of predicted transmitting abilities (PTA) for milk yield, protein percentage, and daughter pregnancy rate from May 2006 for 1,446 sires in the training set and from April 2009 for 316 sires in the testing set. The SNP effects were estimated using the Bayesian least absolute selection and shrinkage operator (LASSO) method with data of sires in the training set, and direct genomic values (DGV) for sires in the testing set were computed by multiplying these estimates by corresponding genotype dosages for sires in the testing set. The mean correlation across traits between DGV (before progeny testing) and PTA (after progeny testing) for sires in the testing set was 70.6% when all 42,552 SNP genotypes were used. When genotypes for 93.1, 96.6, 98.3, or 99.1% of loci were masked and subsequently imputed in the testing set, mean correlations across traits between DGV and PTA were 68.5, 64.8, 54.8, or 43.5%, respectively. When genotypes were also masked and imputed for a random 50% of sires in the training set, mean correlations across traits between DGV and PTA were 65.7, 63.2, 53.9, or 49.5%, respectively. Results of this study indicate that if a suitable reference population with high-density genotypes is available, a low-density chip comprising 3,000 equally spaced SNP may provide approximately 95% of the predictive ability observed with the BovineSNP50 Beadchip (Illumina Inc., San Diego, CA) in Jersey cattle. However, if fewer than 1,500 SNP are genotyped, the accuracy of DGV may be limited by errors in the imputed genotypes of selection candidates.

Key words: genomic selection, imputation, Jersey, predictive ability

INTRODUCTION

The recent introduction of high-throughput assays for dense genotyping of SNP in cattle (Van Tassell et al., 2008) has stimulated hundreds of research projects and transformed practical breeding programs. Tens of thousands of dairy cattle, mostly progeny-tested bulls in commercial AI programs or young bulls that are candidates for such programs, have been genotyped using the BovineSNP50 BeadChip (Illumina Inc., San Diego, CA) during the past 2 yr. Numerous methods have been proposed for estimating SNP effects and predicting direct genomic values (**DGV**) of selection candidates (e.g., Hayes et al., 2009), some of which are variations of the whole-genome selection model proposed in the ground-breaking publication of Meuwissen et al. (2001). In North America, genomic information has been used in national genetic evaluations for routine calculation of PTA for production, conformation, and fitness of dairy cattle since January 2009 (Wiggans et al., 2009).

Despite the impressive gains in reliability of young selection candidates that have been achieved using current high-density genotyping platforms, such as the BovineSNP50 BeadChip (VanRaden et al., 2009), the commercial price of such assays may limit their application to males and elite females. Inexpensive, low-density genotyping platforms with, for example, 300 to 3,000 SNP, could stimulate widespread commercial use for applications such as preliminary screening of young bulls, selection of replacement heifers, discovery of parentage, identification of optimal mating sires, or development of genome-guided management protocols. In practice, the optimal size of such platforms and the characteristics of the SNP that comprise them will depend on the population and family structure, the

Received February 8, 2010.
Accepted July 5, 2010.
[1]Corresponding author: kweigel@wisc.edu

extent of linkage disequilibrium (**LD**) in the species and breed of interest, the genetic architecture underlying economically important traits, the number and proportion of animals with high-density SNP genotypes and accurate phenotypes (i.e., the training set), and the extent of relationships between these individuals and future candidates for selection (i.e., the testing set).

In a recent study, Weigel et al. (2009) found that genotyping young Holstein bulls for 300 to 2,000 highly selected SNP could provide DGV for lifetime net merit with correlations of 0.43 to 0.57 with future PTA from progeny testing, compared with a correlation of 0.61 when all markers on the BovineSNP50 BeadChip were used. Furthermore, Vazquez et al. (2009) reported that low-density genotyping platforms with 500 to 1,000 selected SNP, where selection was based on the largest estimated SNP effects for individual production or fitness traits, could provide correlations of 0.55 to 0.65 with subsequent PTA from progeny testing in Holstein cattle. Furthermore, Vazquez et al. (2009) noted that a low-density platform composed of SNP with the largest effects for lifetime net merit in Holsteins provided correlations of 0.40 to 0.55 with progeny-test PTA for specific production and fitness traits, but correlations tended to be higher for production than for fitness. These challenges, coupled with technological advances that continue to drive down the per locus cost of genotyping, suggest that one should consider genotyping a slightly larger set of equally spaced SNP that would facilitate imputation of high-density SNP genotypes using haplotypes from a reference population that has already been genotyped on a high-density platform.

Numerous algorithms and public domain software packages have been developed for construction of haplotypes and imputation of genotypes in humans (e.g., Scheet and Stephens, 2006; Kong et al., 2008; Howie et al., 2009). In most applications involving humans, the objective is to combine data sets containing subjects (usually cases and controls) who have been genotyped with different high-density platforms for the purpose of carrying out a genome-wide association study for one or more disease traits (e.g., Halperin and Stephan, 2009; Hao et al., 2009). Combining data from competing high-density SNP arrays has not yet been a major issue in cattle, because the majority of genotyping has involved the Illumina BovineSNP50 BeadChip. However, it is likely that multiple high-density (e.g., tens of thousands of SNP) and ultra-high-density (e.g., hundreds of thousands of SNP) genotyping platforms will soon be commercially available, and combining data from different platforms may become relevant in cattle as well.

In food animal species, an alternative use of haplotyping and imputation could involve genotyping a large number of potential selection candidates using an inexpensive, low-density genotyping platform containing equally spaced SNP, and subsequently using this information in conjunction with high-density SNP genotypes (and perhaps pedigrees) of animals in a reference panel to impute missing high-density genotypes in the selection candidates. In a simulation study, Habier et al. (2009) noted that low-density genotyping of selection candidates in the current generation with equally spaced SNP, coupled with high-density genotyping of their parents and grandparents, could lead to rapid and cost-effective genetic progress in commercial breeding programs, particularly if the newly selected parents in each generation were re-genotyped on the high-density platform. Recently, Weigel et al. (2010) masked varying proportions of BovineSNP50 genotypes on 3 chromosomes in Jersey cattle and showed that publicly available, population-based algorithms could impute more than 90% of masked genotypes correctly when as few as 2,000 to 3,000 loci were unmasked in the study sample. However, the effect of imputation errors on the predictive ability of resulting DGV has not yet been documented.

The objective of the present study was to evaluate the predictive ability of DGV for economically important dairy traits when a large proportion of SNP genotypes have been imputed rather than measured directly. Specifically, we sought to determine the reduction in accuracy of DGV for milk yield, milk composition, and female fertility that should be expected if the majority of young selection candidates, and perhaps also a significant proportion of their ancestors, were genotyped using low-density panels consisting of varying numbers of equally spaced SNP.

MATERIALS AND METHODS

Genotypes of 1,762 Jersey sires were provided by the USDA-ARS Animal Improvement Programs Laboratory (Beltsville, MD) and consisted of 42,552 SNP markers distributed across the 29 *Bos taurus* autosomes and the X chromosome. These represented the subset of SNP on the BovineSNP50 BeadChip that are used for routine genomic evaluation of dairy cattle in the United States (Wiggans et al., 2009) and were obtained after removal of SNP with (1) a call rate of less than 90%; (2) greater than 1% parent-progeny conflicts; (3) complete LD with an adjacent SNP; (4) minor allele frequency of less than 1% in each of the Holstein, Jersey, and Brown Swiss breeds, or (5) unknown physical position on the chromosome based on the UMD2 assembly of *B. taurus* (Zimin et al., 2009). Genotypes at each locus were coded as 0 (homozygous for allele B), 1 (heterozygous), 2 (homozygous for allele A), or missing.