J. Dairy Sci. 98:4107–4116
http://dx.doi.org/10.3168/jds.2014-9005

Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction

R. F. Brøndum,*[1] G. Su,* L. Janss,* G. Sahana,* B. Guldbrandtsen,* D. Boichard,† and M. S. Lund*
*Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Blichers Allé 20, Aarhus University, DK-8830 Tjele, Denmark
†Institut National de la Recherche Agronomique (INRA), UMR 1313 Génétique Animale et Biologie Intégrative, 78350 Jouy-en-Josas, France

ABSTRACT

This study investigated the effect on the reliability of genomic prediction when a small number of significant variants from single marker analysis based on whole genome sequence data were added to the regular 54k single nucleotide polymorphism (SNP) array data. The extra markers were selected with the aim of augmenting the custom low-density Illumina BovineLD SNP chip (San Diego, CA) used in the Nordic countries. The single-marker analysis was done breed-wise on all 16 index traits included in the breeding goals for Nordic Holstein, Danish Jersey, and Nordic Red cattle plus the total merit index itself. Depending on the trait's economic weight, 15, 10, or 5 quantitative trait loci (QTL) were selected per trait per breed and 3 to 5 markers were selected to tag each QTL. After removing duplicate markers (same marker selected for more than one trait or breed) and filtering for high pairwise linkage disequilibrium and assaying performance on the array, a total of 1,623 QTL markers were selected for inclusion on the custom chip. Genomic prediction analyses were performed for Nordic and French Holstein and Nordic Red animals using either a genomic BLUP or a Bayesian variable selection model. When using the genomic BLUP model including the QTL markers in the analysis, reliability was increased by up to 4 percentage points for production traits in Nordic Holstein animals, up to 3 percentage points for Nordic Reds, and up to 5 percentage points for French Holstein. Smaller gains of up to 1 percentage point was observed for mastitis, but only a 0.5 percentage point increase was seen for fertility. When using a Bayesian model accuracies were generally higher with only 54k data compared with the genomic BLUP approach, but increases in reliability were relatively smaller when QTL markers were included. Results from this study indicate that the reliability of genomic prediction can be increased by including markers significant in genome-wide association studies on whole genome sequence data alongside the 54k SNP set.

Key words: custom chip, genomic prediction, quantitative trait loci, *Bos taurus*

INTRODUCTION

The accuracy of genomic prediction is highly dependent on the linkage disequilibrium (**LD**) between the genotyped markers and actual causative variants (de Roos et al., 2008; de Los Campos et al., 2013). In dairy cattle, genomic predictions are usually done using the Illumina Bovine SNP50 (54k) SNP chip (Illumina Inc., San Diego, CA), where the distance between the markers might dictate a low level of LD. Increasing the marker density to 777k or high density (**HD**) has only increased the reliability by around 1 percentage point (VanRaden et al., 2011; Erbe et al., 2012; Su et al., 2012). In a study on simulated whole genome sequence data, it was shown that even at an already high marker density a further inclusion of causative variants led to a higher accuracy of genomic prediction (Meuwissen and Goddard, 2010), but the selection of markers included on the commercial chips is without reference to their association with phenotypes in dairy cattle. With the advent of next-generation sequencing (**NGS**) technologies, whole-genome sequence data has become available at a reasonable price. This makes it possible to sequence substantial numbers of bulls, which can be used as a reference to impute animals already genotyped with 54k or HD chips. The NGS data from the 1,000 Bull Genomes project (Daetwyler et al., 2014) have dramatically increased the number of animals available as a reference for imputation, such that common sequence markers (minor allele frequency >0.1) can now be imputed with an accuracy of more than 0.8 (Brøndum et al., 2014; van Binsbergen et al., 2014). However, with current genomic prediction models and computational resources, the sizes of data sets with imputed sequence are too large for all markers to be included in the model at the same time. Conversely,

Received October 22, 2014.
Accepted March 12, 2015.
[1]Corresponding author: rasmusf.brondum@mbg.au.dk

genome-wide association studies (**GWAS**) based on sequence data have shown higher power to identify putative causative variants and shows stronger signals of association (Daetwyler et al., 2014; Höglund et al., 2014; Sahana et al., 2014a). By using GWAS to identify QTL and including these in genomic prediction models, it might be possible to increase the accuracy of genomic prediction (Van den Berg et al., 2014).

In Nordic countries, cows are genotyped on a large scale using the Illumina BovineLD chip (7k). This chip has been designed for imputation performance such that 54k markers can be imputed with high accuracy (Boichard et al., 2012). This chip can be augmented by additional markers, for example, better imputation by increasing the overlap with the 54k chip or with markers that are expected to increase the accuracy of genomic prediction. These markers might then be imputed with higher accuracy than from next generation sequence data into existing 54k data. In the current study, we investigated the effect on the accuracy of genomic prediction when markers strongly associated with QTL in GWAS with whole-genome sequence data are included in the prediction models.

MATERIALS AND METHODS

GWAS

Data. For GWAS, phenotypic data were available as EBV for the 16 index traits included in the Nordic total merit index as well as the index itself; all traits are listed in Table 1 and Table 2. Data were available for the 3 major Nordic dairy breeds and comprised approximately (varies slightly for different traits) 5,000 Nordic Holstein bulls, 1,100 Danish Jersey bulls, and 4,500 Nordic Red bulls. All bulls were imputed to full-sequence data using a 2-step approach, where 54k data were initially imputed to HD data and subsequently to the full sequence level. For more detail on this imputation, see Höglund et al. (2014).

Association Analysis. Genome-wide association analysis was carried out separately for the 3 breeds. The analysis was done using a linear mixed model approach where sires were fitted as a random variable that only considered sire-son relationships (for details see Sahana et al., 2014a). The sequence variants were filtered based on their imputation accuracy ($AR^2 \geq 0.95$, where AR^2 is the square of the estimated correlation between the imputed genotype with highest posterior probability with the true genotype of the marker; Browning and Browning, 2009). Larger values of AR^2 indicate more accurate genotype imputation. The total number of markers [SNP or INDELS (insertion-deletions)] analyzed was ~10 million on 29 autosomes.

Table 1. Distribution of traits in categories based on the economic value of the trait[1]

Category I	Category II	Category III
Mastitis	Longevity	NTM[2]
Fertility	Other diseases	Growth
Legs and Feet	Birth	Udder conformation
Milk yield	Calving	Body-conformation
Fat yield		Milking-speed
Protein yield		Yield
		Temperament

[1]The number of QTL selected per trait was 15, 10, or 5 for category I, II, and III, respectively.

[2]Nordic total merit index.

Marker Selection. The 17 indices were loosely classified into 3 categories based on their relative economic importance (see Table 1). The QTL were ranked within breed based on the strength of the association signal and visual assessment of the QTL peak. The numbers of QTL selected were at most 15, 10, and 5 for the 3 categories for each breed (only if there was a sufficient number of genome-wide significant QTL segregating within the breed). Three to 5 associated markers were initially selected to cover each QTL. This was done manually for each QTL. The selection criteria were *P*-values, functional annotation (such as missense variants), and representation for multiple peaks within a QTL region and distance between markers, so if 2 consecutive markers showed similar association strength only the one with highest $\log_{10}(P\text{-value})$ was included. All markers selected for a breed for all the traits, as described previously, were collected and the duplications were removed (if the same marker was selected for 2 traits). The markers were then pruned within

Table 2. Number of QTL per trait and breed covered by ≥1 SNP in the linkage disequilibrium chip

Indices	Holstein	Nordic Red	Jersey
Birth	10	15	10
Body confirmation	6	8	5
Calving	13	12	15
Fat yield	15	14	14
Fertility	19	15	15
Growth	8	9	5
Legs and feet	17	18	10
Longevity	8	11	14
Mastitis	16	17	25
Milk yield	13	18	16
Milking speed	5	7	5
NTM index[1]	6	8	9
Other diseases	23	21	13
Protein yield	22	10	14
Temperament	6	7	0
Udder conformation	8	8	8
Yield index	16	8	13

[1]Nordic total merit index.