J. Dairy Sci. 99:1–14
http://dx.doi.org/10.3168/jds.2016-11073

Comparing power and precision of within-breed and multibreed genome-wide association studies of production traits using whole-genome sequence data for 5 French and Danish dairy cattle breeds

Irene van den Berg,*†[1] Didier Boichard,† and Mogens Sandø Lund*
*Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark,
†GABI, INRA, AgroParisTech, Université Paris Saclay, 78350 Jouy-en-Josas, France

ABSTRACT

The objective of this study was to compare mapping precision and power of within-breed and multibreed genome-wide association studies (GWAS) and to compare the results obtained by the multibreed GWAS with 3 meta-analysis methods. The multibreed GWAS was expected to improve mapping precision compared with a within-breed GWAS because linkage disequilibrium is conserved over shorter distances across breeds than within breeds. The multibreed GWAS was also expected to increase detection power for quantitative trait loci (QTL) segregating across breeds. GWAS were performed for production traits in dairy cattle, using imputed full genome sequences of 16,031 bulls, originating from 6 French and Danish dairy cattle populations. Our results show that a multibreed GWAS can be a valuable tool for the detection and fine mapping of quantitative trait loci. The number of QTL detected with the multibreed GWAS was larger than the number detected by the within-breed GWAS, indicating an increase in power, especially when the 2 Holstein populations were combined. The largest number of QTL was detected when all populations were combined. The analysis combining all breeds was, however, dominated by Holstein, and QTL segregating in other breeds but not in Holstein were sometimes overshadowed by larger QTL segregating in Holstein. Therefore, the GWAS combining all breeds except Holstein was useful to detect such peaks. Combining all breeds except Holstein resulted in smaller QTL intervals on average, but this outcome was not the case when the Holstein populations were included in the analysis. Although no decrease in the average QTL size was observed, mapping precision did improve for several QTL. Out of 3 different multibreed meta-analysis methods, the weighted z-scores model resulted in the most similar results to the full multibreed GWAS and can be useful as an alternative to a full multibreed GWAS. Differences between the multibreed GWAS and the meta-analyses were larger when different breeds were combined than when the 2 Holstein populations were combined.
Key words: genome-wide association studies (GWAS), multibreed, meta-analysis, whole genome sequence

INTRODUCTION

Genome-wide association studies (**GWAS**) are used to find associations between traits and polymorphisms. With the increasing number of sequences available, more causative mutations will be among the variants and can be detected directly in the GWAS. However, markers in high linkage disequilibrium (**LD**) with the causative mutations can show an equally high or even higher association than the true causative mutations. Genome-wide association studies often result in large regions associated with the same QTL because of long-range LD observed in dairy cattle breeds (de Roos et al., 2008). Across different breeds, LD is only conserved over short distances. Therefore, multibreed mapping could help improve the precision of GWAS. Furthermore, with the large number of variants studied in a sequence-based GWAS, a high detection threshold is necessary to avoid too many false positives. Quantitative traits can be influenced by many causative mutations with small individual effects, and unless large data sets are used, these effects could be too small to pass the thresholds. Assuming a mutation is indeed shared between different breeds, combining data of multiple breeds increases the sample size and thereby the detection power.

In addition to aiding the identification of causative mutations, improved GWAS precision could also help in selecting variants that are subsequently used for genomic prediction. Genomic prediction is widely used in dairy cattle, with high accuracies within breed, although accuracies of across-breed predictions are much lower (Hayes et al., 2009; Erbe et al., 2012; Lund et al.,

Received February 22, 2016.
Accepted June 23, 2016.
[1]Corresponding author: irene.vandenberg@unimelb.edu.au

2014). Unless only markers in full LD with the causative mutations are used, a loss in prediction reliability occurs (de los Campos et al., 2013). This decrease is larger across breeds than within a breed (van den Berg et al., 2014). Using variants in high LD with causative mutations could therefore increase prediction accuracy. This increase could be especially beneficial for multibreed prediction, if variants in high LD with causative mutations shared across breeds are used. A multibreed GWAS could be used to detect such variants.

Joining data of individual GWAS is not always possible, however. Alternatively, rather than using the full data, results of individual GWAS can be combined using meta-analysis (Begum et al., 2012; Evangelou and Ioannidis, 2013), as is commonly done in human genetics. Using both simulated and real data, Lin and Zeng (2010) found equal efficiency with a meta-analysis as when data from individual studies were used.

The objective of this study was to compare different strategies to perform a multibreed analysis using sequence data from 6 French and Danish dairy cattle populations. Results obtained by within-breed GWAS were compared with those of a multibreed GWAS. Furthermore, several meta-analysis methods were compared with a multibreed GWAS. Specifically, we tested the following hypotheses: (1) the power of detecting QTL is larger for a multibreed GWAS than a within-breed GWAS, (2) the QTL detection is more precise for a multibreed GWAS than a within-breed GWAS, and (3) a multibreed GWAS can be approximated by a meta-analysis of within-breed GWAS.

MATERIALS AND METHODS

Data

Imputed sequences of 4,993 Danish Holstein, 984 Jersey, 768 Danish Red, 5,626 French Holstein, 1,935 Montbéliarde, and 1,725 Normande bulls were used. The majority of sequences were obtained by imputation of bulls genotyped with the 50K and HD SNP chips (Illumina Inc., San Diego, CA). First, bulls genotyped with the 50K chip were imputed to the HD chip. For the French data (Hozé et al., 2013), this step was performed using Beagle 3.0.0 (Browning and Browning, 2007), while for the Danish breeds IMPUTE2 was used (Howie et al., 2012). Subsequent imputation to whole-genome sequence was done for all breeds, using IMPUTE2. Imputation was done within each country, imputing Danish Holstein, Jersey, and Danish Red together, and French Holstein, Montbéliarde, and Normande together. The reference used for imputation to sequences of the Danish bulls consisted of the bulls in run 4 of the 1,000 Bull Genomes project (Daetwyler et al., 2014), while for the imputation of the French bulls, a combined French and Danish reference set was used. The latter consisted of 122 Holstein, 27 Jersey, 28 Montbéliarde, 23 Normande, and 45 Danish Red bulls. More details on the Danish imputation can be found in Höglund et al. (2015). For all bulls, sires were known and deregressed proofs (**DRP**) for milk, fat, and protein yields were used in the analysis.

All genotypic and phenotypic data were obtained from preexisting routine genetic evaluation data for the dairy cattle populations and required no ethical approval.

Principal Component Analysis

To study genomic relationships between breeds, a genomic relationship was constructed using SNP from the 50K chip for 500 randomly selected individuals of each breed. For these individuals, genomic relationship matrix $\mathbf{G}$ was constructed following VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum p_j(1-p_j)},$$

where $\mathbf{Z}$ was a standardized genotype matrix, and p_j was the allele frequency of the alternative allele for locus j. Subsequently, $\mathbf{G}$ was used for a principal component analysis using the prcomp() command in R (R Development Core Team, 2015).

GWAS Within Breed

Variants with a within-breed minor allele frequency below 0.005 or an IMPUTE2 info score below 0.60 were discarded. These exclusions resulted in 24,550,115 SNP and small indels (insertion-deletions). The following model was run for each of these polymorphisms within each breed:

$$y_{ij} = \mu + S_j + \beta g_{ij} + e_{ij},$$

where y_{ij} is the DRP of milk yield, fat yield, or protein yield for individual i with sire j, S_j the random effect of sire j, β the effect of the polymorphism, g_{ij} the allele dose (ranging from 0 to 2) of individual i with sire j, and e_{ij} a random residual.

Multibreed GWAS

Three multibreed GWAS were run, combining French and Danish Holstein (**HOL**), combining Jersey, Dan-