J. Dairy Sci. 101:1–14
https://doi.org/10.3168/jds.2017-13587

Use of meta-analyses and joint analyses to select variants in whole genome sequences for genomic evaluation: An application in milk production of French dairy cattle breeds

M. Teissier,*[1] M. P. Sanchez,† M. Boussaha,† A. Barbat,† C. Hoze,†‡ C. Robert-Granie,* and P. Croiseau†
*GenPhySE, Université de Toulouse, INRA, INPT, ENVT, 31326 Castanet-Tolosan, France
†GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France
‡Allice, 75012 Paris, France

ABSTRACT

As a result of the 1000 Bull Genome Project, it has become possible to impute millions of variants, with many of these potentially causative for traits of interest, for thousands of animals that have been genotyped with medium-density chips. This enormous source of data opens up very interesting possibilities for the inclusion of these variants in genomic evaluations. However, for computational reasons, it is not possible to include all variants in genomic evaluation procedures. One potential approach could be to select the most relevant variants based on the results of genome-wide association studies (GWAS); however, the identification of causative mutations is still difficult with this method, partly because of weak imputation accuracy for rare variants. To address this problem, this study assesses the ability of different approaches based on multi-breed GWAS (joint and meta-analyses) to identify single-nucleotide polymorphisms (SNP) for use in genomic evaluation in the 3 main French dairy cattle breeds. A total of 6,262 Holstein bulls, 2,434 Montbéliarde bulls, and 2,175 Normande bulls with daughter yield deviations for 5 milk production traits were imputed for 27 million variants. Within-breed and joint (including all 3 breeds) GWAS were performed and 3 models of meta-analysis were tested: fixed effect, random effect, and Z-score. Comparison of the results of within- and multi-breed GWAS showed that most of the quantitative trait loci identified using within-breed approaches were also found with multi-breed methods. However, the most significant variants identified in each region differed depending on the method used. To determine which approach highlighted the most predictive SNP for each trait, we used a marker-assisted best unbiased linear prediction model to evaluate lists of SNP generated by the different GWAS methods; each list contained between 25 and 2,000 candidate variants per trait, which were identified using a single within- or multi-breed GWAS approach. Among all the multi-breed methods tested in this study, variant selection based on meta-analysis (fixed effect) resulted in the most-accurate genomic evaluation (+1 to +3 points compared with other multi-breed approaches). However, the accuracies of genomic evaluation were always better when variants were selected using the results of within-breed GWAS. As has generally been found in studies of quantitative trait loci, these results suggest that part of the genetic variance of milk production traits is breed specific in Holstein, Montbéliarde, and Normande cattle.
Key words: multi-breed genomic evaluation, meta-analysis, sequence, quantitative trait locus detection

INTRODUCTION

Around the world, the majority of routinely used procedures for genomic evaluation are based on chips containing tens of thousands of SNP. Several methodologies have been developed for genomic evaluation, of which the most commonly used are genomic BLUP (Meuwissen et al., 2001), which assumes that all SNP have a small effect on the trait, and BayesC (Kizilkaya et al., 2010) and BayesCπ (Habier et al., 2011), which assume a proportion (π) of SNP have zero effect. Those methodologies do not take into account prior biological knowledge of traits, and estimate the effects of causative mutations only through SNP in linkage disequilibrium (**LD**) with them. Therefore, one way to improve the accuracy of genomic evaluations is through the identification and localization of causative mutations, which can then be directly included in the evaluation model. Unfortunately, extending genomic evaluation to the whole genome is not realistic due to the computational challenges involved, and selecting a reduced panel of SNP to include in the genomic evaluation is a very difficult challenge (VanRaden et al., 2017).

Received July 28, 2017.
Accepted December 18, 2017.
[1]Corresponding author: marc.teissier@inra.fr

Genome-wide association studies (**GWAS**) are widely used to study the genetic architecture of complex traits (Shi et al., 2012; Karlsson et al., 2013). They aim to screen the genome and detect associations between SNP and a disease or a QTL. However, the regions detected generally have large confidence intervals and contain many candidate genes, which makes it challenging to identify the causative mutation itself. As an example, despite the fact that a large number of regions have been associated with traits of economic importance in dairy cattle at the 50k or HD (800k) SNP genomic densities (Pryce et al., 2010; Meredith et al., 2012), very few causative mutations have thus far been validated.

As causative mutations are (generally) not included on SNP chips, GWAS highlight SNP in LD with a causative mutation, rather than the mutation itself. Instead, more accurate GWAS results can be obtained through the use of whole-genome sequences (**WGS**), because this approach enables the inclusion of millions of variants, including causative mutations. Toward this end, the 1000 Bull Genome Consortium aims to produce a large data set of sequenced animals (Daetwyler et al., 2014). For example, Run 4 contained WGS data of 1,147 bulls from 36 different breeds. This population is large enough to enable the imputation of WGS of all animals for which genotypes (e.g., 50k SNP) are available, which means that GWAS can be performed at the sequence level for all animals having both genotypes and phenotypes. However, the identification of causative mutations is still very difficult in a within-breed analysis, for 2 reasons: (1) the resolution is limited by the high level of LD between SNP in dairy cattle breeds, which leads to positive signals of association over large regions, and (2) high error rates of imputation have been observed for rare variants, which has led to false signals of association. Instead, an analysis that includes data from different breeds, each with its own pattern of LD, should address these shortcomings in 2 ways: first, it increases the population size with respect to a single-breed analysis, and second, it refines the locations of QTL shared across breeds (Raven et al., 2014). In addition, a false positive signal of association is unlikely to be present in multiple separate breeds. Furthermore, in a multi-breed (**MB**) approach, the long-range LD is expected to be lower than in a within-breed approach, and consequently, it should be possible to identify causative mutations with more accuracy.

A MB analysis can be performed 2 different ways: a MB GWAS on a joint data set, or a meta-analysis of results from multiple within-breed GWAS. Joint analyses are expected to yield the best resolution, as they minimize the effects of the long-range LD present within dairy cattle breeds. However, MB GWAS are time consuming, and meta-analyses are a faster alternative. To this end, several methods of meta-analysis have been developed (e.g., fixed effect, random effect, or Z-score; Evangelou and Ioannidis, 2013) and have been used in dairy cattle (Buitenhuis et al., 2016; van den Berg et al., 2016).

In this study, we compared the ability of within-breed GWAS and MB analyses to detect QTL for milk production and milk composition in French Holstein (**HO**), Montbéliarde (**MO**), and Normande (**NO**) cattle. In the absence of functional analysis, it was not possible to know if a given method was able to highlight a causal mutation. However, by using genomic evaluation, we were able to measure the ability of a given approach to predict a phenotype. To this end, the lists of most significant QTL generated by the different methods were analyzed using marker-assisted best linear unbiased prediction (**MABLUP**).

MATERIALS AND METHODS

Samples and Genotypes

Reference populations for the association studies and genomic evaluations consisted of genotyped or sequenced bulls from the 3 main French dairy cattle breeds: HO, MO, and NO. The study population included 6,262 HO, 2,434 MO, and 2,175 NO bulls. For genomic evaluations, reference populations were split into 2 groups: (1) a training set containing 5,107 HO, 1,948 MO, and 1,740 NO bulls, in which performance, pedigrees, and genotypes were recorded and used to establish prediction equations, and (2) a validation set comprising younger bulls (1,155 HO, 486 MO, and 435 NO) from which only pedigrees and genotypes were known. To use the lists of QTL from GWAS in genomic evaluation, animals from the validation population were excluded from GWAS.

Five routinely collected production traits were analyzed: milk production (**MLK**), fat yield (**FY**), protein yield (**PY**), fat content (**FC**), and protein content (**PC**). For all traits, the phenotypes used in analyses were the daughter yield deviations (**DYD**) of each bull (VanRaden and Wiggans, 1991), defined as the average value of daughters' performance, adjusted for fixed and nongenetic random effects and for the additive genetic value of their respective dams. Each DYD was weighted by the effective daughter contribution (VanRaden and Wiggans, 1991). To limit the influence of bulls with higher numbers of daughters, weights were bound to a maximal number of daughters corresponding to a reliability of 0.9.

Bulls were genotyped at different densities. Key ancestors (i.e., bulls responsible for a considerable part