J. Dairy Sci. 97:2444–2454
http://dx.doi.org/10.3168/jds.2013-7143

Strategy for the simulation and analysis of longitudinal phenotypic and genomic data in the context of a temperature × humidity-dependent covariate

T. Yin,* E. C. G. Pimentel,* U. König v. Borstel,† and S. König*[1]
*Department of Animal Breeding, University of Kassel, 37213 Witzenhausen, Germany
†Department of Animal Breeding, Göttingen University, 37075 Göttingen, Germany

ABSTRACT

A simulation study was conducted to evaluate the performance of genomic random regression models for the continuous environmental descriptor temperature-humidity index (THI). Statistically innovative aspects of the study included the combined simulation of both longitudinal phenotypic data representing the same trait in the course of THI and genomic data. The longitudinal trait was simulated (phenotypically expressed) at 5 different values of THI. For a moderate heritability trait, heritabilities were 0.30, 0.35, 0.40, 0.40, and 0.35 for THI of 15, 30, 45, 60 and 75, respectively. In a consecutive run, low heritabilities of 0.05, 0.1, 0.15, 0.15, and 0.10 were simulated, respectively. On the genomic level, simulation combined high and low linkage disequilibrium with 5,000-, 15,000-, and 50,000-SNP chip applications to simulate different scenarios of genomic architecture. With regard to data analyses, 2 strategies were applied to evaluate the accuracy of genomic predictions across THI, with special focus on the extreme ends of the environmental scale. In the first strategy, 100, 80, 50, or 20% of phenotypes at THI 75 were deleted randomly and the remaining data set was used to predict the breeding value at THI 75 for non-phenotyped, but genotyped cows. In the second strategy, 1,600 cows had complete information (genotypes and phenotypes) and 400 cows were genotyped, but with missing phenotypes for all THI. For the first strategy and without phenotypic observations at THI 75, accuracies of genomic predictions were lower than 0.34. When only 20% of cows had phenotypic records at THI 75, accuracies increased (∼0.60). Such a small proportion of phenotyped cows was sufficient to predict reliable genomic breeding values for cows without phenotypes for extreme THI. For the second strategy, also for low linkage disequilibrium combined with a low density 5,000-SNP chip, the average accuracy of genomic predictions was 0.52, which is substantially higher than accuracies based on pedigree relationships. From a practical perspective, genomic random regression models can be used to predict genomic breeding values for scarce phenotypes (e.g., novel traits) traits measured in extreme environments, or traits measured late in life, such as longevity.
Key words: genomic selection, genotype by environment interaction, random regression

INTRODUCTION

Methods for dealing with longitudinal data in genetic evaluations have evolved from the use of repeatability models with permanent environmental effects or multiple-trait models with covariance matrices (Henderson, 1984) to random regression models (**RRM**; Schaeffer and Dekkers, 1994) with covariance functions (Kirkpatrick et al. 1990). The use of RRM for analyzing longitudinal production data are a standard in genetic evaluations for dairy cattle worldwide, because such models provide an overview of genetic parameters and breeding values across the whole lactation trajectory. Additionally, further applications of RRM to describe performances over a range of environments in reaction norm studies have been proposed. Such models are interesting with regard to genotype by environment interactions, where different environments can be defined on a continuous scale. Ravagnolo and Misztal (2000) estimated variance components for milk production traits at different levels of heat stress, defined by a temperature-humidity index (**THI**). In more recent studies, RRM were further elaborated by defining THI as an environmental covariate (Aguilar et al., 2009; Brügemann et al., 2011). The basic idea of RRM applications is to depict the physiological background or genetic mechanisms of traits in a quantitative genetic context, meaning that different genes are switched on or off with, for example, aging of the animal or with environmental changes. Substantial changes of both quantitative genetic parameters and gene expression profiles by inducing heat stress were shown in fertility

Received June 14, 2013.
Accepted December 12, 2013.
[1]Corresponding author: sven.koenig@uni-kassel.de

traits of mice (Cammack et al., 2006, 2009). Genetic studies on heat stress in dairy cattle have an important practical background with regard to semen and livestock exports. For example, the German dairy cattle industry, and especially the dairy cattle breeding organization Masterrind GmbH (Verden, Germany), is strongly involved in exports of livestock and sire semen. Target countries include tropical countries located in Asia, Africa, and South America. In 2012, a total of 26,249 heifers were exported to these countries (DHV, 2013). The tropical and hot climates in the importing countries causes heat stress in the cows, especially when the THI rises above 72 (Bohmanova et al., 2005). In contrast, performance testing within Germany exhibits a shift from population-wide recording schemes toward so-called selected "contract herds" (Schierenbeck et al., 2011). Contract herds are characterized by superior feeding, management, and housing conditions, and by substantially lower THI levels realized by, for example, housing systems with integrated cooling techniques (Brügemann et al., 2012).

Classically, RRM applications (also when studying genotype by environment interactions) are based on longitudinal phenotypic records combined with genetic relationships from pedigree data. Nowadays, the availability of high-throughput genotyping technologies with decreasing costs encourages dairy cattle farmers worldwide to genotype an increasing percentage of cows, heifers, and female calves. Especially for novel traits, reliable, conventional EBV of bulls do not exist. Hence, basing genomic selection on calibration groups of cows might be a promising alternative (Buch et al., 2012). Examples include health traits (Pintus et al., 2013), and traits reflecting energy balance (Verbyla et al., 2010). Furthermore, Misztal et al. (2010) suggested the inclusion of genomic information to improve the accuracy of genetic evaluations of young animals for heat tolerance. Availability of cow genotypes combined with longitudinal phenotypic data enable the application of genomic RRM (**gRRM**) to estimate genomic breeding values (**GBV**) for scarcely recorded traits, or for environmental descriptors that are not or poorly represented in a data set. In this latter context, Suchocki and Szyda (2011) estimated SNP effects over time by applying a mixed model with orthogonal polynomials and genotyped animals for longitudinal growth data. An alternative might be the direct estimation of GBV using gRRM and BLUP. Simulations are a powerful tool to evaluate a broad variety of statistical procedures based on longitudinal phenotypic and genomic data and, in consequence, to study the effects of various scenarios on selection and mating schemes. Accuracies of genomic predictions strongly depend on technical parameters (e.g., size of calibration group and pattern of SNP chips), the quantity and quality of phenotypic data, quantitative genetic parameter estimates, and the genomic architecture of the trait. To our knowledge, no simulation package exists that simultaneously addresses those aspects for longitudinal data structures and directly provides true breeding values (**TBV**), GBV, and phenotypes in the course of a continuous environmental descriptor.

Consequently, the objectives of the present study were to (1) develop a framework for the simulation of longitudinal phenotypic data combined with high-throughput genotypes, (2) evaluate the performance of a gRRM in the context of reaction norms, and (3) investigate the accuracy of genomic predictions for cows that are poorly or not at all represented in the group of cows with records for environmental descriptors. For illustration and based on experiences from previous studies, the environmental descriptor THI was chosen, but applications to further problems will be discussed. The study was performed by varying the assumptions related to the genomic architecture of traits.

MATERIALS AND METHODS

Simulation of Populations

Populations were simulated using the software QMSim (Sargolzaei and Schenkel, 2009). With QMSim, the simulation process is divided in 2 stages. First, a historical population is simulated for several generations to generate a desired level of linkage disequilibrium (**LD**). In a second step, using animals from the last historical generation as founders, further recent populations are simulated for a desired number of generations. Within this second simulation step, population parameters can be varied to generate the appropriate population structure. In the present study, 2 different types of historical populations were simulated to create scenarios with either low or high LD. For both low-LD and high-LD scenarios, 10 recent generations were simulated based on the parameters as specified in Table 1.

Additionally, Table 1 summarizes the parameters of the simulated genome. The simulated genome consisted of 30 chromosomes of 100 cM each. The number of QTL per chromosome was set to 10 and QTL positions on the chromosome were randomly assigned. Effects of QTL alleles were drawn from a gamma distribution with a shape parameter 0.4. The number of QTL alleles at each locus was randomly assigned and was 2, 3, or 4. To achieve resemblance with different densities of SNP chips, 3 scenarios with respect to the number of markers on the genome were simulated. The simulation of 167, 500, and 1,667 biallelic markers per chromosome depicts applications with 5,000 (**5K**)-, 15,000 (**15K**)-,