J. Dairy Sci. 99:1–10
http://dx.doi.org/10.3168/jds.2015-10433

Alternative haplotype construction methods for genomic evaluation

Dávid Jónás,*†‡[1] Vincent Ducrocq,* Marie-Noëlle Fouilloux,§ and Pascal Croiseau*
*INRA, UMR1313 Génétique animale et biologie intégrative, 78350 Jouy-en-Josas, France
†AgroParisTech, 16 rue Claude Bernard, 75231 Paris 05, France
‡ALLICE, 149 rue de Bercy, 75012 Paris, France
§Idele, UMR1313 GABI, 78352 Jouy-en-Josas Cedex, France

ABSTRACT

Genomic evaluation methods today use single nucleotide polymorphism (SNP) as genomic markers to trace quantitative trait loci (QTL). Today most genomic prediction procedures use biallelic SNP markers. However, SNP can be combined into short, multiallelic haplotypes that can improve genomic prediction due to higher linkage disequilibrium between the haplotypes and the linked QTL. The aim of this study was to develop a method to identify the haplotypes, which can be expected to be superior in genomic evaluation, as compared with either SNP or other haplotypes of the same size. We first identified the SNP (termed as QTL-SNP) from the bovine 50K SNP chip that had the largest effect on the analyzed trait. It was assumed that these SNP were not the causative mutations and they merely indicated the approximate location of the QTL. Haplotypes of 3, 4, or 5 SNP were selected from short genomic windows surrounding these markers to capture the effect of the QTL. Two methods described in this paper aim at selecting the most optimal haplotype for genomic evaluation. They assumed that if an allele has a high frequency, its allele effect can be accurately predicted. These methods were tested in a classical validation study using a dairy cattle population of 2,235 bulls with genotypes from the bovine 50K SNP chip and daughter yield deviations (DYD) on 5 dairy cattle production traits. Combining the SNP into haplotypes was beneficial with all tested haplotypes, leading to an average increase of 2% in terms of correlations between DYD and genomic breeding value estimates compared with the analysis when the same SNP were used individually. Compared with haplotypes built by merging the QTL-SNP with its flanking SNP, the haplotypes selected with the proposed criteria carried less under- and over-represented alleles: the proportion of alleles with frequencies <1 or >40% decreased, on average, by 17.4 and 43.4%, respectively. The correlations between DYD and genomic breeding value estimates increased by 0.7 to 0.9 percentage points when the haplotypes were selected using any of the proposed methods compared with using the haplotypes built from the QTL-SNP and its flanking markers. We showed that the efficiency of genomic prediction could be improved at no extra costs, only by selecting the proper markers or combinations of markers for genomic prediction. One of the presented approaches was implemented in the new genomic evaluation procedure applied in dairy cattle in France in April 2015.

Key words: single nucleotide polymorphism, haplotype, genomic evaluation, dairy cattle

INTRODUCTION

Virtually all current genomic prediction methods use information from SNP markers (e.g., Meuwissen et al., 2001; Habier et al., 2011), which are abundant all over the genome. However, a major limitation of individual SNP markers as explanatory variables is that each significant causal mutation should be in high linkage disequilibrium (**LD**), with at least 1 SNP to ensure a good prediction. Given the fact that SNP on the commercial SNP chips were selected to have a high minor allele frequency, this requirement is not necessarily fulfilled when the mutated alleles are rare. For example, the development of high-density SNP chips in cattle was expected to overcome this limitation and increase genomic prediction accuracy, but recent studies could show only a limited gain (e.g., Erbe et al., 2012; VanRaden et al., 2013; Ma et al., 2014). Furthermore, the accurate separation and estimation of the effects of closely linked QTL with SNP is not feasible either.

Haplotypes (defined as combinations of 2 or more SNP as in Hayes et al., 2007; Villumsen et al., 2009; Garrick and Fernando, 2014) are multiallelic genomic markers that hold the promise of improving genomic prediction due to higher expected LD between the haplotype and the QTL alleles (e.g., Hayes et al., 2007). Indeed, haplotype information has been used in practical genomic selection in France since 2008, leading to an increased correlation between estimated breeding

Received September 23, 2015.
Accepted February 8, 2016.
[1]Corresponding author: david.jonas@jouy.inra.fr

values and performances as compared with genomic prediction methods based on SNP (Boichard et al., 2012).

Several methods have been used to construct haplotypes for genomic evaluation (Calus et al., 2008, 2009; Boichard et al., 2012; Cuyabano et al., 2014). Allele effect predictability can be defined as the expected prediction accuracy of the effect of haplotype alleles, and it is expected to have a significant effect on the performance of genomic prediction. However, none of the previously mentioned methods take into account any information on this predictability. The construction of haplotypes at a particular SNP position by merging this SNP with the flanking markers is straightforward. However, because of the short distance between the markers, the resulting haplotypes most frequently include a small number of over-represented alleles together with a large number of alleles with low frequencies within the population. An accurate estimation of allele effects for the haplotype alleles that are greatly under-represented is difficult, whereas the abundant information on over-represented alleles does not contribute efficiently to the improvement of genomic estimated breeding value (**GEBV**). The complexity of the statistical model cannot be increased to the range of hundreds of thousands of effects to be estimated, as would happen if all possible nonoverlapping haplotypes of 4 to 5 SNP were considered. Therefore, an efficient haplotype selection procedure is required to identify the haplotypes most suitable for genomic evaluation purposes. In addition, the estimated effects of rare alleles would be generally inaccurate. Hence, the selection of haplotypes with fewer rare alleles would also be beneficial.

For QTL fine mapping, Grapes et al. (2006) showed that it is beneficial to use a selected subset of markers instead of all available markers within a genomic region to build haplotypes, especially when markers are densely distributed. The main objective of the present study was to develop a method to, a priori, construct the most appropriate haplotype for genomic prediction, given a set of SNP previously detected to be in LD with QTL influencing the trait of interest. These SNP will be called QTL-SNP hereafter. Two haplotype selection methods are proposed to select the best haplotype within a window of N SNP around the QTL-SNP based on observed allele frequencies. The goal is to reduce the number of under-represented alleles and to maximize the number of alleles properly represented in the population under study. The predictability of an allele effect also depends on the effect size of the linked QTL (Meuwissen et al., 2001), but this information is not available at the haplotype selection step. The effect on genomic prediction of haplotypes from the 2 haplotype selection methods versus haplotypes built from flanking markers around the QTL-SNP was compared on a real data set.

MATERIALS AND METHODS

General Notation

The term QTL-SNP refers to SNP in strong LD with causative mutations affecting a trait of interest. These SNP were identified using a Bayes-Cπ procedure (see details below). Haplotypes are defined as combinations of N SNP along a chromosome (similar to the definitions of Hayes et al., 2007; Villumsen et al., 2009; Garrick and Fernando, 2014). The term allele refers to the alternative forms of a genetic marker present in a population; considering SNP, 2 alleles are present per marker, whereas haplotypes can be composed of 2^N different alleles, where N is the haplotype size in number of SNP. Flanking SNP of a QTL-SNP are the nearest SNP surrounding the QTL-SNP. Flanking haplotypes are the haplotypes that are built by merging the QTL-SNP and the flanking SNP into a single haplotype. A short genomic segment around the QTL-SNP defined in number of SNP is referred to as a QTL window, or simply as a window.

In this study, the QTL-SNP were considered as markers indicating the approximate positions of the QTL affecting the trait of interest. A short, symmetric genomic window was constructed around each QTL-SNP and these genomic segments were assumed to contain the linked QTL. Our aim was to select a single haplotype of N SNP per window to represent the QTL within that window in genomic prediction. Once haplotypes were selected around each QTL-SNP, all of them were used in genomic prediction to predict breeding values for the individuals in the validation population.

Data and QTL Detection Methods

Performance values in the form of average daughter yield deviations (**DYD**) for 5 dairy cattle production traits (milk quantity, fat content, fat yield, protein content, and protein yield) were available for 2,235 Montbéliarde bulls genotyped with the Bovine SNP50 BeadChip (50K; Illumina Inc., San Diego, CA). Only autosomal chromosomes were used. After quality control, 43,801 SNP were retained from the 50K chip. In a first step, a QTL detection was undertaken using a Bayes-Cπ approach as implemented in the GS3 software by Legarra et al. (2013). The model used in this SNP-based Bayes-C analysis was:

$$y_i = \mu + u_i + \sum_{j=1}^{N} z_{ij} a_j \delta_j + e_i,$$