Short communication: The combined use of linkage disequilibrium–based haploblocks and allele frequency–based haplotype selection methods enhances genomic evaluation accuracy in dairy cattle

Dávid Jónás,*†¹ Vincent Ducrocq,* and Pascal Croiseau*

*GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France

†ALLICE, 149 rue de Bercy, 75012 Paris, France

ABSTRACT

The construction and use of haploblocks [adjacent single nucleotide polymorphisms (SNP) in strong linkage disequilibrium] for genomic evaluation is advantageous, because the number of effects to be estimated can be reduced without discarding relevant genomic information. Furthermore, haplotypes (the combination of 2 or more SNP) can increase the probability of capturing the quantitative trait loci effect compared with individual SNP markers. With regards to haplotypes, the allele frequency parameter is also of interest, because as a selection criterion, it allows the number of rare alleles to be reduced, and the effects of those alleles are usually difficult to estimate. We have proposed a simple pipeline that simultaneously incorporates linkage disequilibrium and allele frequency information in genomic evaluation, and here we present the first results obtained with this procedure. We used a population of 2,235 progeny-tested bulls from the Montbéliarde breed for the tests. Phenotype data were available in the form of daughter yield deviations on 5 production traits, and genotype data were available from the 50K SNP chip. We conducted a classical validation study by splitting the population into training (80% oldest animals) and validation (20% youngest animals) sets to emulate a real-life scenario in which the selection candidates had no available phenotype data. We measured all reported parameters for the validation set. Our results proved that the proposed method was indeed advantageous, and that the accuracy of genomic evaluation could be improved. Compared with results from a genomic BLUP analysis, correlations between daughter yield deviations (a proxy for true) and genomic estimated breeding values increased by an average of 2.7 percentage points for the 5 traits. Inflation of the genomic evaluation of the selection candidates was also significantly reduced.

The proposed method outperformed the other SNP and haplotype-based tests we had evaluated in a previous study. The combination of linkage disequilibrium–based haploblocks and allele frequency–based haplotype selection methods is a promising way to improve the efficiency of genomic evaluation. Further work is needed to optimize each step in the proposed analysis pipeline.

Key words: haplotype, haploblock, genomic evaluation

Short Communication

The development of cost-efficient SNP chips, as well as elaborate evaluation methods such as the Bayesian alphabet [A, B, C(- π), D(- π), R by Meuwissen et al., 2001; Habier et al., 2011; Erbe et al., 2012] or the single-step genomic BLUP (Aguilar et al., 2010) has led to the practical implementation of genomic selection in dairy cattle breeding in most developed countries (e.g., Boichard et al., 2012). The majority of currently available methods use biallelic SNP as genetic markers to trace QTL. However, haplotype markers (a combination of 2 or more SNP markers, Hayes et al., 2007; Villumsen et al., 2009; Garrick and Fernando, 2014) can outperform individual SNP markers in genomic evaluation (Croiseau et al., 2015; Jónás et al., 2016). The main advantage of haplotypes lies in their multi-allelic nature: when more alleles can be tracked at a given locus, chances are higher that at least one will be linked to existing QTL. However, allele effects are not always predicted more accurately with haplotypes than with SNP. The accuracy with which allele effects can be estimated is largely influenced by the alleles' frequency, which determines how much phenotypic information can be directly linked to each allele. Rare haplotype alleles are more likely than with SNP, especially if flanking (i.e., neighboring) SNP are combined into a haplotype marker, because of the short genetic distance [i.e., high linkage disequilibrium (**LD**)] between SNP on medium- and high-density SNP chips. It is desirable to maximize the number of haplotype alleles in genomic prediction to increase the probability that at least 1 will be linked to the QTL (if present). However, it is also

Received July 29, 2016.

Accepted November 24, 2016.

¹Corresponding author: david.jonas@jouy.inra.fr

necessary to avoid rare alleles to ensure accurate allele effect estimation and efficient genomic evaluation.

Following these considerations, Jónás et al. (2016) proposed a method of selecting haplotype markers prior to genomic evaluation, based on observed allele frequencies. These preselected haplotypes outperform haplotypes of flanking SNP in genomic evaluation. However, a major drawback of the proposed method was the fact that the approximate location of the QTL had to be determined before genomic evaluation. Here, we present an extension of this work, aimed at removing that initial step by exploiting information about the LD pattern along the genome.

Jónás et al. (2016) proposed 2 criteria for selecting haplotypes, with a small difference between their formulations. In the present study, we will consider only the criterion with the higher performance, Criterion-B (Jónás et al., 2016). This procedure selects the haplotype that leads to the best balance between allele frequencies and number of alleles. In other words, it selects in a given interval the haplotype that carries the largest number of alleles with intermediate frequencies. Indeed, avoiding both rare alleles (because the estimation of their effects is difficult) and very frequent alleles (because overrepresented alleles limit the total number of alleles that can segregate in the population) is advantageous.

We used the dataset described in Jónás et al. (2016), allowing for easy comparison with the results of the earlier study. The dataset included 2,235 progeny-tested bulls from the French Montbéliarde population. Phenotype data in the form of daughter yield deviations (**DYD**) were available on 5 production traits: milk, protein, and fat yield, and protein and fat content. We used genotype data from the Bovine SNP50 BeadChip (50K; Illumina Inc., San Diego, CA). After quality control, we retained 43,801 SNP for genomic evaluation.

We conducted analyses in a validation study with the 20% youngest animals, which formed the validation population (the 80% oldest animals formed the training population). We estimated haplotype allele effects using the training set; then, using these estimated allele effects along with genotype and pedigree information from the validation population, we estimated genomic estimated breeding values (**GEBV**) for all individuals in the validation set. Finally, we calculated correlations between estimated **GEBV** and **DYD**, as well as regression slopes of **DYD** on **GEBV**. We compared these findings with the results from Jónás et al. (2016), which used a genomic BLUP (**GBLUP**) model and the Criterion-B haplotype selection approach, because this approach was found to be optimal. In the latter procedure, we estimated SNP effects using a Bayes-C π analysis, and the SNP with the highest probability of

inclusion in the model were selected (in practice, 1,000, 3,000, or 6,000 SNP were identified). These SNP were assumed not to be the causative mutations themselves, but to merely indicate the approximate location of the QTL that affected the trait of interest. In a 10-SNP-wide window symmetrically surrounding these preselected SNP, all possible combinations of 4 SNP were considered as a different haplotype, and 1 haplotype was selected using Criterion-B to represent the linked QTL. These haplotypes were used to better capture the QTL effects. This procedure will be referred as the preselection method.

Jónás et al. (2016) used a fixed window size. However, it is reasonable to assume that different window boundaries should be used along the genome, adapting to the local LD (e.g., Jeffreys et al., 2005 in humans or Weng et al., 2014 in beef cattle; other drawbacks of fixed window sizes have been outlined by Beissinger et al., 2015). To account for different recombination rates and to remove the need for information about the approximate location of QTL, we built windows of SNP in strong LD along the genome and selected haplotypes to represent these windows. We defined windows as a set of consecutive SNP where the LD between every neighboring SNP exceeded a predefined limit, and we called these windows haploblocks, following Knürr et al. (2013). We used D' as a measure of LD, and set a threshold at 45%, following Cuyabano et al. (2014). We also evaluated a threshold of 90%. After determining the haploblocks, we selected a single haplotype of 4 SNP from all possible haplotypes of 4 SNP to represent each haploblock along the genome. We selected haplotypes within each haploblock using Criterion-B and its optimal parameter values, as identified in Jónás et al. (2016).

This process also allowed for the identification of haplotypes that were expected to contribute most to the reliability of the genomic evaluation before using phenotype data. This was a notable advantage, because significant markers are usually identified in an earlier genomic evaluation after the training population has been split into subpopulations, and this method is clearly suboptimal, particularly for breeds that have only a small reference population. In such situations, the number of animals with both genotype and phenotype data is already low, and their division into further subpopulations is more detrimental.

Another advantage of this procedure was that it allowed the use of the same haplotypes for all traits analyzed. This is because haploblock construction was based on observed LD patterns, but haplotype selection assumed knowledge of allele frequencies only; no information on performance was used to select genetic markers. We could expect the differences between the

Table 1. Characteristics of the haploblocks and the selected haplotypes

Parameter	Value
Total number of markers	43,801
Number of haploblocks	8,393
Number of haplotypes built	7,804
Average number of SNP per haploblock	5.2
Average number of alleles per haplotype	13.3

genetic backgrounds of the traits to be reflected in the different estimated allele effects of the haplotypes.

We estimated haplotype allele effects using a haplotypic Bayes-C π approach (Croiseau et al., 2014). The model included an overall mean effect and a residual polygenic effect, in addition to the haplotype marker effects (Jónás et al., 2016). It can be written as

$$y_i = \mu + u_i + \sum_{j=1}^N \mathbf{z}_{ij} \mathbf{a}_j \delta_j + e_i,$$

where y_i is the performance value (DYD) of an individual i ; μ is an overall mean effect; u_i is the residual polygenic effect of animal i [$u \sim MVN(0, \mathbf{A}\sigma_u^2)$]; where MVN refers to multivariate normal distribution, $\mathbf{A}$ is the pedigree relationship matrix, and σ_u^2 is the residual polygenic variance; N is the total number of haplotypes in the model; $\mathbf{z}_{ij}$ is a vector of dimension $1 \times k_j$ (where k_j is the number of alleles at haplotype j) indicating the number of each haplotype allele copies animal i carries at haplotype j for every allele of that haplotype (i.e., the vector sum of $\mathbf{z}_{ij}$ is 2); $\mathbf{a}_j$ is a vector of substitution effects for haplotype j (of dimension $k_j \times 1$); δ_j is a 0/1 variable indicating whether marker j is present or absent; and e_i is a random error term for animal i . The proportion of genetic variance attributed to the residual polygenic effect was allowed to vary.

We tested 2 threshold values of the D' parameter: 45 and 90%. The 45% threshold was found to be optimal

in Cuyabano et al. (2014), and our tests confirmed those results (data not shown). Therefore, only results with a D' threshold of 45% will be presented here.

Table 1 provides a short summary of the characteristics of the haploblocks (selected with a D' threshold of 45%) and the selected haplotypes. The 43,801 SNP were divided into 8,393 haploblocks with an average of 5.2 SNP per haploblock. This number of SNP per haploblock was relatively small due to the long distance between the markers on the 50K SNP chip panel (average ~57,300 bp, exceeding 100,000 bp only in 11.5% of cases). Sometimes haploblocks were shorter than the desired haplotype size (4 SNP). In such cases, we built haplotypes using all SNP from the haploblock and added the closest flanking SNP to extend the haplotypes to 4 SNP. When short haploblocks were adjacent, it was likely that we had built the same haplotypes for each of them, so we kept only 1 for analysis. For this reason, we observed fewer haplotypes than haploblocks (Table 1). The average number of alleles per haplotype was higher using the haploblocks (13.3; Table 1) than with fixed window sizes (12.4; Jónás et al., 2016), despite the fact that haplotypes were selected from narrower windows.

Table 2 presents the results for GBLUP, as well as the results of the preselection method (Jónás et al., 2016), together with the new results obtained using haploblock information. The correlation coefficients between DYD and GEBV and regression slopes of DYD on GEBV are presented. The preselection method column corresponds to the second-last row of Supplementary Table S5 in Jónás et al. (2016), displaying the best results obtained in that study.

With the haploblock method, the proportion of variance attributed to the residual polygenic effect converged to 5.7% (average of the 5 traits). The rest of the genetic variance was explained by the haplotypes. Results obtained with the combination of LD-based haploblocks and allele frequency-based haplotype selection outperformed traditional GBLUP analysis by

Table 2. Correlation coefficients between DYD and GEBV and regression slopes of DYD on GEBV using GBLUP, the preselection method, or haploblock information¹

Trait name	GBLUP ²		Preselection method ²		Haploblock information ³	
	Correlation	Slope	Correlation	Slope	Correlation	Slope
Milk quantity	0.490	0.810	0.496	0.789	0.504	0.910
Fat yield	0.551	0.850	0.562	0.806	0.564	0.943
Protein yield	0.478	0.738	0.476	0.697	0.493	0.803
Fat content	0.570	0.785	0.594	0.865	0.637	0.933
Protein content	0.584	0.987	0.609	0.971	0.613	1.071
Average ⁴	0.535	0.166	0.547	0.174	0.562	0.096

¹DYD = daughter yield deviation; GBLUP = genomic BLUP; GEBV = genomic EBV.

²Results were taken from Jónás et al. (2016).

³Results obtained using haploblock information with a D' threshold of 45%.

⁴For regression slopes, average absolute deviations from one are shown.

2.7 percentage points in correlation coefficients, and the best preselection method by 1.5 percentage points. We observed the greatest improvements for fat content (4.3 percentage points in correlation, compared with the other 2 methods) and protein yield (1.7 percentage points in correlation). Although the observed increases were very limited for certain traits, a significant Wilcoxon signed-rank test ($P = 0.03$) between the haploblock and preselection results showed that a significant global increase was observed when haploblock information was taken into account. The large improvement with these traits was most likely because when regions are preselected based on a prior Bayes-C π analysis, multiple SNP are linked to the same major genes (such as diacylglycerol O-acyltransferase 1, or *DGAT1*) and as a consequence, SNP that were linked to other QTL were missed. In contrast, these SNP are necessarily retained when all markers from all regions are kept in the haploblock-based analysis, leading to higher selection accuracies.

Regression slopes of DYD on GEBV were also substantially improved with the haploblock-based method. On average, deviation of the regression slopes from their optimal value (i.e., 1) was 0.078 smaller than with the preselection or GBLUP method. We also implemented a test using all consecutive haplotypes of 4 SNP along the genome, resulting in inferior correlations and regression slopes compared with haploblock-based analyses (data not shown).

Using information on LD patterns along the genome in combination with allele frequency information to build haplotypes specifically for genomic evaluation is a promising way to improve the accuracy of genomic evaluation. A very interesting feature of the proposed method is that the same haplotypes can be used to analyze all traits of interest. Further significant improvements can be expected following refinement of the proposed process. For example, Beissinger et al. (2015) have developed a smoothing spline technique to better identify window boundaries. Application of this method could lead to better haploblock definition, which in turn could further improve selection efficiency. Another interesting aspect of the proposed method is that it allows for the use of genotype data from the selection candidates (or that of the validation population in an experimental setup) in combination with genotype data from the training population to build haplotypes for genomic evaluation (because no phenotype data were used for haplotype construction).

ACKNOWLEDGMENTS

The CARTOFINE and AMASGEN projects are funded by the Agence Nationale de la Recherche (ANR-

10-GENM-0014, Paris, France) and APISGENE (Paris, France).

REFERENCES

- Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* 93:743–752.
- Beissinger, T. M., G. J. M. Rosa, S. M. Kaeppler, D. Gianola, and N. de Leon. 2015. Defining window-boundaries for genomic analyses using smoothing spline techniques. *Genet. Sel. Evol.* 47:30.
- Boichard, D., F. Guillaume, A. Baur, P. Croiseau, M. N. Rossignol, M. Y. Boscher, T. Druet, L. Genestout, J. J. Colleau, L. Journaux, V. Ducrocq, and S. Fritz. 2012. Genomic selection in French dairy cattle. *Anim. Prod. Sci.* 52:115–120.
- Croiseau, P., A. Baur, D. Jónás, C. Hozé, J. Promp, D. Boichard, S. Fritz, and V. Ducrocq. 2015. Comparison of different marker-assisted BLUP models for a new French genomic evaluation. Page 248 in Abstracts of the 66th Annual Meeting of the European Federation of Animal Science, Warsaw, Poland. European Federation of Animal Science, Rome, Italy. Accessed Jan. 23, 2017. https://asas.org/docs/default-source/wcgalp-posters/708_paper_10043_manuscript_1181_0bFD602C6D9AD3.pdf?sfvrsn=2.
- Croiseau, P., M. N. Fouilloux, D. Jónás, S. Fritz, A. Baur, V. Ducrocq, F. Phocas, and D. Boichard. 2014. Extension to haplotypes of genomic evaluation algorithms. Abstract 708 in Proc. 10th World Congress on Genetics Applied to Livestock Production, Vancouver, Canada.
- Cuyabano, B. C. D., G. Su, and M. S. Lund. 2014. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. *BMC Genomics* 15:1171.
- Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* 95:4114–4129.
- Garrick, D. J., and R. Fernando. 2014. Genomic prediction and genome-wide association studies in beef and dairy cattle. Pages 474–501 in *The Genetics of Cattle*. 2nd ed. D. J. Garrick and A. Ruvinsky, ed. CABI, Wallingford, UK.
- Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* 12:186 <https://doi.org/10.1186/1471-2105-12-186>.
- Hayes, B. J., A. J. Chamberlain, H. McPartlan, I. Macleod, L. Sethuraman, and M. E. Goddard. 2007. Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. *Genet. Res.* 89:215–220. <https://doi.org/10.1017/S0016672307008865>.
- Jeffreys, A. J., R. Neumann, M. Panayi, S. Myers, and P. Donnelly. 2005. Human recombination hot spots hidden in regions of strong marker association. *Nat. Genet.* 37:601–606.
- Jónás, D., V. Ducrocq, M.-N. Fouilloux, and P. Croiseau. 2016. Alternative haplotype construction methods for genomic evaluation. *J. Dairy Sci.* 99:4537–4546.
- Knürr, T., I. Strandén, M. Koivula, G. P. Aamand, and E. A. Mäntysaari. 2013. Haplotype-assisted genomic evaluations in Nordic red dairy cattle. Page 454 in Abstracts of the 64th Annual Meeting of the European Federation of Animal Science, Nantes, France. European Federation of Animal Science, Rome, Italy.
- Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829.
- Villumsen, T. M., L. Janss, and M. S. Lund. 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. *J. Anim. Breed. Genet.* 126:3–13. <https://doi.org/10.1111/j.1439-0388.2008.00747.x>.
- Weng, Z. Q., M. Saatchi, R. D. Schnabel, J. F. Taylor, and D. J. Garrick. 2014. Recombination locations and rates in beef cattle assessed from parent-offspring pairs. *Genet. Sel. Evol.* 46:34.