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 markers, 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
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:

\[ y_i = \mu + u_i + \sum_{j=1}^{N} z_{ij} a_j \delta_j + e_i, \]
where $y_i$ is the performance value of individual $i$, $\mu$ is an overall mean effect, $u_i$ is the residual polygenic effect of animal $i$, $\mathbf{u} \sim MVN \left( \mathbf{0}, \mathbf{A} \sigma_u^2 \right)$, where $MVN$ is multivariate normal distribution, $\mathbf{A}$ is the additive relationship matrix, and $\sigma_u^2$ is 0.2 times the genetic variance, $N$ is the total number of SNP in the model, $z_{ij}$ is an indicator variable representing the number of copies of one of the alleles at marker $j$ in animal $i$, $a_j$ is the substitution effect of marker $j$, $\delta_i$ is a 0/1 variable indicating whether or not marker $j$ is assumed to have an effect, and $e_i$ is a random error term for animal $i$. The residual polygenic effect was assumed to account for 20% of the total genetic variance, whereas the rest of the genetic variance was attributed to the selected markers. Following the Bayes-C$\pi$ analysis, the $k$ SNP with the largest probability of inclusion in the model were considered to be QTL. These SNP will be called QTL-SNP. This step was done within the framework of a classical validation study, using the same training and validation populations as for the haplotype-based tests (see in detail below). In practice, the first 1,000, 3,000, and 6,000 QTL-SNP were selected for each trait (denoted as 1K, 3K, and 6K, respectively). Due to this selection procedure, for each trait, every smaller set is a subset of the larger set(s). It is expected that these QTL-SNP were in strongest LD with the causative mutations.

The original GS3 software by Legarra et al. (2013) was extended to deal with haplotypes (Croiseau et al., 2014). This haplotype Bayes-C was used for genomic evaluation and for testing the performance of the different haplotype construction methods. Haplotypes were modeled as class variables, with one effect predicted for each haplotype allele. The proportion $\pi$ of haplotypes with no effect was fixed because of practical considerations: the haplotypic Bayes C was very time-consuming due to the increased number of effects to estimate. Fixing $\pi$ allowed us to perform a large number of tests within a reasonable time, without sacrificing accuracy. Moreover, preliminary tests showed that fixing $\pi$ led to validation correlations slightly higher as compared with a scenario where $\pi$ was estimated during the analysis due to poor mixing in the latter case (data not shown). A constant value of $\pi$ (90%) was selected because it gave a number of marker effects to be estimated similar to the number of individuals in the training population. The same model was used for the haplotype-based Bayes-C analyses as for the SNP-based tests, with the SNP effects being replaced by the haplotype effects.

Out of the 2,235 bulls with both phenotype and genotype information, the youngest 20% of individuals were selected as the validation population. Allele effects were estimated using the training population (that is, the oldest 80% of animals) and GEBV were estimated for the individuals in the validation population using only genomic information of that population and the estimated allele effects. Accuracy of the breeding value estimation was measured by the correlation coefficient between GEBV and DYD values of the validation population. The performance of the different haplotype construction methods was evaluated based on this parameter. In addition, the slopes of the regression of DYD on GEBV were calculated and compared.

**Haplotype Selection**

Haplotypes were constructed within each QTL window. The most desirable one was supposed to maximize the number of alleles with an allele frequency higher than a given threshold. As previously mentioned, it is advantageous in genomic prediction to avoid both under- and over-represented alleles.

Once a window of window size ($WS$; the size in number of markers) SNP was defined around each QTL position, every possible haplotype of haplotype size ($HS$; the size in number of markers) SNP was constructed. Three different methods with different criteria were used, and each of these methods resulted in a haplotype within each window. The performances of these haplotypes (methods) in genomic evaluation were compared. These criteria are described in detail below. Considering that the QTL-SNP had the strongest LD within a window with the linked QTL, this SNP was always forced to be part of the final haplotype. The number of haplotypes that can be built within the window is therefore:

$$\binom{WS - 1}{HS - 1} = \frac{(WS - 1)!}{(HS - 1)! \times (WS - HS)!}.$$  

One haplotype was selected from each window to be used in genomic evaluation based on 3 different approaches. These approaches were termed as flanking markers, criterion-A, and criterion-B and their performances were compared. To test the effect of the WS and HS on genomic prediction, windows of size WS = 10, 15, and 20 SNP, as well as haplotypes of size HS = 3, 4, and 5 SNP were constructed. All WS and HS combinations were tested.

**Flanking Markers.** The QTL-SNP and its flanking markers were grouped into a haplotype. Haplotype allele frequency was not considered. Flanking markers were always considered symmetrically around the QTL-SNP: the flanking haplotype built from 5 SNP included the QTL-SNP and 2–2 flanking SNP on both sides of the QTL-SNP. When HS was an even number (i.e., an odd number of SNP had to be selected on the 2
sides of the QTL-SNP), a symmetric haplotype of \((HS + 1)\) SNP was created around the QTL-SNP and the marker that was the farthest from the QTL-SNP was excluded from the haplotype. The same principle was used when asymmetric windows had to be constructed around the QTL-SNP.

**Criterion-A.** A threshold level denoted as allele frequency threshold (AFT) was used to determine which alleles are considered predictable (i.e., which allele effects can be predicted with satisfactory accuracy). The following AFT values were tested: 1, 3, 5, and 8%.

With criterion-A, a 2-step approach was implemented. First, for each haplotype \(i\) within a specific window, the number of predictable alleles (i.e., with a frequency higher than AFT) was determined. Then for the haplotypes carrying the maximum number \((N_{\text{max}})\) of predictable alleles within the window, a score \(SD_{hi}\) was calculated as the squared deviation of observed allele frequencies from the ideally balanced allele frequency, where the latter was equal to \(1/N_{\text{max}}\). The score can be written as

\[
SD_{hi} = \sum_{k=1}^{N_i} \left( OF_{i,k} - \frac{1}{N_i} \right)^2,
\]

where \(h_i\) is haplotype \(i\), \(N_i (=N_{\text{max}})\) is the number of predictable alleles of haplotype \(i\), and \(OF_{i,k}\) is the observed frequency of allele \(k\) of haplotype \(i\). Retaining the haplotype with the lowest squared deviation score guarantees that the observed allele frequencies are as balanced as possible.

**Criterion-B.** A drawback of criterion-A is that the allele frequencies can still be unbalanced to a high degree, because haplotypes with more predictable alleles are always preferred over haplotypes with fewer predictable alleles. This is true even if, for example, many alleles of a certain haplotype have a frequency that barely exceeds the threshold level, whereas a small number of alleles are greatly over-represented in the population. Criterion-B consists of 2 parts, from which the first part is a modified version of the SD score calculated for criterion-A. The difference is that \(1/N_i\) is replaced by \(1/2^{HS}\) to ensure that this part is, assuming similar variations in the allele frequencies, smaller for haplotypes with a higher number of predictable alleles. This is guaranteed because the observed frequencies of the predictable alleles will on average get closer to \(1/2^{HS}\) as their number is increasing. The second part is a weighted number of predictable alleles. It ensures that out of haplotypes that carry the same number of alleles, the haplotype(s) that include more predictable alleles have a lower score. A parameter that we call maximum deviation (MD) was introduced in the computation of the weight (see Supplemental Materials for details; http://dx.doi.org/10.3168/jds.2015-10433). It is defined as the average acceptable deviation of \((n - 1)\) alleles from the ideal frequency \(1/2^{HS}\), expressed as a proportion of the ideal frequency. The \(n\)th allele must have a frequency equal to or larger than AFT. The MD parameter can be interpreted as follows: the smaller its value is, the less the allele frequencies are allowed to deviate from their mean. For example, if MD is set to a relatively strict value of 10%, haplotypes with fewer predictable alleles are favored when their allele frequencies are more balanced against haplotypes with more predictable alleles, but with a larger variation among the frequencies of those alleles.

In practice, criterion-B is calculated as

\[
\text{Criterion-B}_{hi} = \sum_{k=1}^{N_i} \left( OF_{i,k} - \frac{1}{2^{HS}} \right)^2 - w \times N_i,
\]

where \(w\) is the weighing factor of the number of predictable alleles. The second term of criterion-B is negative to be consistent with the first term, which is optimal when it takes the smallest value.

Table 1 illustrates the difference between criterion-A and -B. Criterion-A would prefer the second haplotype over the fourth despite of its highly unbalanced allele frequencies. This preference is reversed with criterion-B, assuming appropriate AFT and MD values.

An analysis using only the QTL-SNP as genomic markers was conducted to obtain a basis for comparisons. This analysis was conducted on all sets of QTL-SNP (1K, 3K, and 6K) and the optimal number of QTL-SNP was selected for each trait. The benefit of haplotypes versus SNP was judged by analyzing the same SNP selected by each method in a Bayes C model utilizing them as single-SNP information. The observed correlations between DYD and GEBV from these analyses were compared with those obtained with their haplotype counterparts. A genomic BLUP analysis with all retained SNP markers was also performed to complete the tests.

**RESULTS AND DISCUSSION**

Table 2 shows the number of haplotypes that can be built for several different WS and HS values. The windows have a reasonably small number of combinations. Haplotype selection was performed on a single processor and running time was less than 1 min for windows of 10 SNP, haplotypes of 4 SNP and 3,000 QTL-SNP, where the total number of evaluated haplotypes was 252,000.
**Distribution of Allele Frequencies**

The number of alleles with very low allele frequencies (<1%) decreased with criterion-A and -B compared with the flanking markers approach. With flanking markers and 6K QTL-SNP in the model, 2,660 alleles (i.e., 3.6% of the alleles in the population had frequency >40%) were termed as over-represented alleles; almost half of the flanking haplotypes included one such allele. The proportion of over-represented alleles with the haplotypes selected by either criterion-A or criterion-B was approximately half of this value: 2.1 and 1.56%, respectively. In case of haplotypes of 4 and 5 SNP, criterion-B tended to select haplotypes with slightly fewer rare and over-represented alleles than criterion-A.

Figure 1 shows the distribution of alleles present in the population according to their allele frequency for HS = 4, WS = 10 SNP, and 6,000 QTL-SNP. The use of criterion-A and -B led to a higher proportion of haplotype alleles in the 5 to 30% frequency range, but also to a lower proportion of over-represented alleles. These trends were observed whatever the haplotype size. The difference between the haplotypes built from the flanking markers and from the selected markers decreased when the haplotype size increased (data not shown).

Table 1 shows the average number of alleles per haplotype for different haplotype selection methods, haplotype sizes, and number of QTL-SNP. As expected, with the increase of the haplotype size, the number of segregating alleles increased rapidly. However, it was close to its theoretical maximum value (2^{HS}; i.e., 8, 16, or 32 for HS = 3, 4, or 5) only when HS = 3. This is not surprising, given the relatively dense SNP chips available and the corresponding high LD.

Interestingly, the average number of segregating alleles per haplotype was decreasing as the number of QTL was increasing from 1,000 to 6,000 (Table 3). One interpretation is that QTL with smaller effects (i.e., those QTL-SNP added when moving from 1,000 to 6,000 QTL in the model) are segregating in less polymorphic regions of the genome compared with QTL with larger effects. The reduced number of haplotype alleles might also slightly affect the prediction accuracy, as the probability of having at least 1 allele in strong LD with the QTL is reduced. This trend was apparent with all marker construction methods; however, the magnitude of the decrease is larger with criterion-A and -B than it is with the flanking marker haplotypes.

The number of rare and over-represented alleles was lower with criterion-B. The frequencies of these alleles were also more favorable with criterion-B than with criterion-A; rare alleles had a higher average frequency with criterion-B, whereas the average frequency of the over-represented alleles decreased when compared with criterion-A (data not shown). All of these are beneficial features for genomic prediction, which can be attributed to the changes made in criterion-B. These are the additional constraint on the allele frequency equilibrium and the replacement of 1/N by 1/2^{HS} in the equation of the SD. The total number of segregat-

---

**Table 1.** Allele frequencies for 4 haplotypes; the selection order with both criterion-A and -B is also shown

<table>
<thead>
<tr>
<th>Allele frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

1 As the first 2 haplotypes have 6 predictable alleles (assuming a threshold of allele frequency threshold = 5%), these haplotypes are not considered in the second step of criterion-A.

**Table 2.** Number of possible haplotypes with different window and haplotype sizes

<table>
<thead>
<tr>
<th>Window size</th>
<th>Without forcing the QTL-SNP</th>
<th>With forcing the QTL-SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HS = 3</td>
<td>HS = 4</td>
</tr>
<tr>
<td>10</td>
<td>120</td>
<td>210</td>
</tr>
<tr>
<td>15</td>
<td>455</td>
<td>1,365</td>
</tr>
<tr>
<td>20</td>
<td>1,140</td>
<td>4,845</td>
</tr>
</tbody>
</table>

1 All possible haplotypes within the window are considered, whether they include the QTL-SNP or not.
2 Within a window, only haplotypes that include the QTL-SNP are considered. Good candidate QTL-SNP are required.
3 HS = haplotype size.
ing alleles with criterion-B did not change as the AFT threshold increased, in contrast with criterion-A (see Supplemental Table S1; http://dx.doi.org/10.3168/jds.2015-10433). The number of alleles with very low (<1%) allele frequencies tended to increase with increasing AFT, whereas the number of the moderately frequent alleles (1–10%) systematically decreased (data not shown).

Although the proposed methods favor haplotypes with intermediate allele frequencies, rare alleles are inevitable. For example, with haplotypes of 4 SNP and AFT of 8%, the proportion of alleles with frequency less than 8% was 63 to 64% with the haplotypes selected by criterion-A or -B instead of ~69% with the flanking markers.

**Correlations Between DYD and GEBV Values**

Genomic prediction of a set of dairy cattle production traits was implemented to investigate the performance of the haplotypes selected by the different methods.

**AFT Tests.** The optimal AFT for the studied population with both criterion-A and -B was 8% (see Supplemental Table S2; http://dx.doi.org/10.3168/jds.2015-10433). The effect of the choice of AFT on correlations decreased when the number of QTL increased (data not shown). This may be related to the fact that the smaller QTL were segregating in less polymorphic parts of the genome, where fewer but more frequent alleles were segregating. The AFT parameter had only a minor effect on the prediction accuracy; it also had a smaller effect on the results of criterion-B than on those of criterion-A (Supplemental Table S2). The AFT was fixed to 8% for the rest of the analysis.

**MD Tests.** Several values were tested for the MD parameter of criterion-B, which were chosen to cover

![Figure 1](image.png)

**Figure 1.** Overall distribution of haplotype allele frequencies according to the haplotype construction approach (haplotype size: 4 SNP; window size: 10 SNP; 6,000 QTL-SNP). The 0 to 10% region is also depicted with more detailed scale on the x-axis.
the whole range between 0 and 1. No large differences were observed in correlations with regard to this parameter (see Supplemental Table S3; http://dx.doi.org/10.3168/jds.2015-10433). As the MD value had only a marginal effect on the results, its value was fixed to 10% (i.e., more strongly favoring more balanced allele frequencies over a higher number of predictable alleles).

**Comparison of the Haplotype Construction Methods.** Table 4 shows the correlations between DYD and GEBV in the validation population obtained with the analysis using either only the QTL-SNP as genomic markers or the haplotypes built from the flanking markers. Hereafter, all correlations and differences in correlations are reported in percentage points. Flanking markers outperformed the analyses, which solely used the QTL-SNP in all scenarios. The observed gain ranged between 0.8 and 2.9%, and it was larger with longer haplotypes and with a higher number of QTL-SNP in the model. The optimal number of QTL-SNP was 6,000 for most of the traits. The average gain observed for the 5 traits was 2.1 to 2.9%, increasing with haplotype size. Similar results were found with criterion-A and -B, except that haplotype size 5 did not result in higher correlations than haplotypes of 4 SNP (see Supplemental Table S4; http://dx.doi.org/10.3168/jds.2015-10433).

Figure 2 shows the obtained correlations between DYD and GEBV values of the validation population with the different haplotype sizes and haplotype selection methods after selecting the optimal number of QTL-SNP for each trait. The solid lines represent the analyses using the selected SNP as haplotypes and the dashed lines correspond to the analyses using the same SNP as individual SNP information sources in genomic prediction. Average correlations of the 5 production traits are shown (for the individual results, see Supplemental Table S5; http://dx.doi.org/10.3168/jds.2015-10433). Merging the SNP into haplotypes was beneficial in all cases, leading to an increase of 1.4% in correlation when the obtained gain was averaged across the 3 haplotype construction methods. This increase in correlation was 2% when only the highest correlation for each trait was considered from those observed with 1K, 3K, and 6K haplotypes in the model. This gain was positively correlated with the increase of number of haplotypes in the model, showing an increase of 0.7, 1.6, and 1.9% with 1,000, 3,000, and 6,000 QTL modeled, respectively. No large differences were observed between the haplotype selection methods in this aspect. With the presented criteria in general, haplotypes of 5 SNP performed worse than the shorter haplotypes; on average for the 5 production traits, no additional gain was observed with criterion-A and HS = 5, compared with its flanking haplotypes counterpart (see Supplemental Table S5). The poor performance of haplotypes of 5 SNP might be a result of over-parameterization of the model. The average gains with criterion-A compared with the flanking marker haplotypes were 1.3 and 0.6% with haplotypes of 3 and 4 SNP, respectively. Haplotypes selected by criterion-B outperformed those selected by criterion-A by 0.3% on average. The observed gain compared with the flanking haplotypes with both criterion-A and criterion-B was decreasing as the haplotype size increased. This can be attributed to the diminishing differences in the number of alleles between the haplotype construction methods with increasing haplotype size (data not shown). Finally, the average correlation of the 5 production traits with genomic BLUP was 0.535; the correlations between DYD and GEBV were 1.1% higher with haplotypes built with criterion-A or -B than with a standard genomic BLUP analysis.

**WS Tests.** The effect of window size used for haplotype construction on genomic prediction results was also investigated. Windows of 10, 15, and 20 SNP were constructed and haplotypes were selected from these windows for genomic prediction, using a value of 8% for AFT and 10% for MD. Table 5 shows the results obtained with the different window sizes for both criterion-A and criterion-B and for the 3 tested haplotype sizes. It was expected that wider windows would result in lower correlation due to a decreasing LD between QTL and haplotypes. This was indeed observed for

<table>
<thead>
<tr>
<th>Number of QTL-SNP</th>
<th>Flanking markers</th>
<th>Flanking markers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>QTL-SNP</td>
<td>HS = 3</td>
</tr>
<tr>
<td>1K</td>
<td>0.480</td>
<td>0.491</td>
</tr>
<tr>
<td>3K</td>
<td>0.499</td>
<td>0.523</td>
</tr>
<tr>
<td>6K</td>
<td>0.512</td>
<td>0.534</td>
</tr>
<tr>
<td>Optimal</td>
<td>0.512</td>
<td>0.534</td>
</tr>
</tbody>
</table>

\(^1\) HS = haplotype size.

\(^2\) For each trait separately, the number of QTL-SNP/haplotypes is the one leading to the highest correlation.
the correlations obtained with haplotypes constructed using criterion-B. However, the results obtained with criterion-A showed a small increase in correlations with the increase of window sizes. The apparent inconsistency in the results with respect to the effect of window size might be a result of different LD patterns around the different QTL-SNP in the model, for which the same window size was applied in our study. This might have resulted in windows that overlap with recombination sites or hotspots, greatly reducing the LD between the selected haplotypes and the linked QTL. Undoubtedly, the frequency of such windows increases with the increase of the window size. Therefore, in practical applications, it might be beneficial to take into account additional information for the definition of the windows, such as recombination hotspots or the LD pattern of the SNP along the genome. However, the testing of the effect of this information was outside the scope of our study.

Obviously, it is desirable to adjust parameter values for the model to the studied population. For example, population size has a major effect on the optimal AFT value; in larger populations, lower AFT values can be used. However, the presented criteria (especially criterion-B) appear to be robust to the choice of parameter values within the tested limits. With criterion-B, an increased risk of over-parameterization was noted with haplotypes of 5 SNP (compared with the flanking haplotype situation) due to the higher number of segregating alleles per haplotype (11.5% larger, on average).

**Slope of Regression**

The average slope of regression of DYD on GEBV with haplotypes of 4 SNP over the 5 traits were 0.80, 0.80, and 0.83 with the flanking, criterion-A, and criterion-B haplotypes, respectively. When the same markers were used as single-SNP information, the slopes of regression were in the same order, 0.71, 0.73, and 0.75, respectively. The regression slope was 0.83 with the genomic BLUP model. In all cases, these values are relatively far from the desirable value of 1. Higher values were obtained when the fraction of the total genetic variance allocated to the residual polygenic effect was increased (data not shown); however, optimization of this slope was outside the scope of this paper.

---

**Table 5.** Correlations between the daughter-yield deviations and genetic EBV values for the tested window sizes<sup>1</sup>

<table>
<thead>
<tr>
<th>Haplotype selection method</th>
<th>Window size</th>
<th>Haplotype size</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion-A</td>
<td>10</td>
<td>0.537</td>
<td>0.541</td>
<td>0.547</td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>0.542</td>
<td>0.550</td>
<td>0.543</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.538</td>
<td>0.548</td>
<td>0.547</td>
<td></td>
</tr>
<tr>
<td>Criterion-B</td>
<td>10</td>
<td>0.548</td>
<td>0.540</td>
<td>0.546</td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>0.541</td>
<td>0.546</td>
<td>0.549</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.540</td>
<td>0.546</td>
<td>0.545</td>
<td></td>
</tr>
</tbody>
</table>

<sup>1</sup>Average correlations over the 5 traits are shown (allele frequency threshold: 8%; maximum deviation: 10%). The optimal number of QTL-SNP was selected, as described in the manuscript. Allele frequency threshold = only alleles with a frequency higher than this threshold are assumed to be sufficiently predictable; maximum deviation = controls the acceptable level of variation among allele frequencies.
Statistical Analysis

The average differences between the correlations of criterion-B and those of the flanking markers (short horizontal lines), as well as the calculated lower confidence bounds of the tests (triangles), are shown on Figure 3. Criterion-B led to a small increase in correlation in almost all of the cases (see also Supplemental Table S5; http://dx.doi.org/10.3168/jds.2015-10433). The significance of the observed increase in correlation between DYD and GEBV was tested using Fisher's Z-transform, as implemented in the “cocor” R package (Diedenhofen and Musch, 2015) based on the work of Zou (2007). As the results of criterion-B were slightly better than those with criterion-A, these were compared with the flanking haplotypes. To test whether haplotypes selected with criterion-B outperform flanking haplotypes, a one-tailed test with $\alpha = 0.05$ and the null hypothesis that the 2 correlations are equal was performed. Out of the 15 correlations (5 traits $\times$ 3 haplotype sizes; the correlation coefficients are present in Supplemental Table S5), 3 were found to be significantly better with criterion-B than with the flanking haplotypes.

A Wilcoxon signed-rank test was performed to assess whether criterion-B, compared with the flanking markers, led globally to increased correlations. The Wilcoxon signed-rank test was chosen because normality could not be assumed due to the low sample size ($n = 15$) and because the available data were paired; for every HS or trait combination, a correlation coefficient was available in both the flanking marker and criterion-B cases.

$$gain_{z,t} = \left( \frac{p_{z,hap,t}}{p_{QTL-SNP,t}} \right) - 1,$$

where $z$ refers to one of the haplotype selection scenarios (flanking marker, criterion-A, or criterion-B), $p_{z,hap,t}$ is the observed correlation coefficient with the haplotype-based analysis using scenario $z$ for trait $t$, $p_{QTL-SNP,t}$ is the observed correlation coefficient with the analysis using only the QTL-SNP as genetic markers for trait $t$, and $gain_{z,t}$ is the observed relative gain in correlation between the 2. The Wilcoxon signed-rank test was performed using $\alpha = 0.05$ (one-tailed test). The test results ($W = 111$ and $P = 0.001$) indicate that the haplotypes selected by criterion-B significantly increased the correlations between DYD and GEBV compared with the flanking haplotypes. The test with criterion-A was also significant ($W = 76$, $P = 0.02$).

Final Remarks

The alleles that are considered predictable based solely on their allele frequencies and those that are actually well predicted in genomic selection are not equivalent because the predictability of an allele also depends on the effect size of the linked QTL. Therefore, whereas alleles carried by a sufficiently large number of individuals in the population are always predictable, effects of rare alleles can be also accurately predicted if those alleles are in strong LD with large QTL. Hence, the efficiency of haplotype selection procedures can be further improved in the future, once objective measures of QTL effect sizes will be available.

At present, interest is increased in using haplotypes as genomic markers in genomic evaluation procedures. The efficiency of the methods presented in our study might be further improved by, for example, identifying window boundaries in a more precise way [for examples, see Cuyabano et al. (2014) and Beissinger et al. (2015)].

Criterion-B is part of the new genomic evaluation procedure, which was implemented for the 4 dairy cattle breeds (Holstein, Montbéliarde, Normande, and Brown Swiss) in France in April 2015 (Croiseau et al., 2015).

CONCLUSIONS

Two methods to improve haplotype allele predictability based on observed allele frequencies were presented and compared with haplotypes created from the
flanking markers. The obtained results indicate that an a priori selection of haplotypes from a small genomic region around each QTL-SNP can improve the correlations between DYD and GEBV at no extra costs. In addition, the proposed methods are data-independent and require neither large computing power nor excessive running time. The inclusion of additional constraints on the allele frequency equilibrium in the haplotype selection procedure was beneficial, further increasing the correlations between DYD and GEBV by 0.3% on average over 5 production traits.

ACKNOWLEDGMENTS

The GEMBAL (Génomique Multi-race des Bovins Allaitants et Laitiers) project is funded by the Agence Nationale de la Recherche (ANR-10-GENM-0014), Agisgene, Races de France, and INRA “AIP Bioresources.”

REFERENCES


