Short communication: Imputation performances of 3 low-density marker panels in beef and dairy cattle

R. Dassonneville,*† S. Fritz,‡ V. Ducrocq,* and D. Boichard*
*INRA, UMR1313 Génétique Animale et Biologie Intégrative (GABI), 78350 Jouy-en-Josas, France
†Institut de l’Elevage, 149 rue de Bercy, 75595 Paris, France
‡UNCEIA, 149 rue de Bercy, 75595 Paris, France

ABSTRACT

Low-density chips are appealing alternative tools contributing to the reduction of genotyping costs. Imputation enables researchers to predict missing genotypes to recreate the denser coverage of the standard 50K (~50,000) genotype. Two alternative in silico chips were defined in this study that included markers selected to optimize minor allele frequency and spacing. The objective of this study was to compare the imputation accuracy of these custom low-density chips with a commercially available 3K chip. Data consisted of genotypes of 4,037 Holstein bulls, 1,219 Montbéliarde bulls, and 991 Blonde d’Aquitaine bulls. Criteria to select markers to include in low-density marker panels are described. To mimic a low-density genotype, all markers except the markers present on the low-density panel were masked in the validation population. Imputation was performed using the Beagle software. Combining the directed acyclic graph obtained with Beagle with the PHASEBOOK algorithm provides fast and accurate imputation that is suitable for routine genomic evaluations based on imputed genotypes. Overall, 95 to 99% of alleles were correctly imputed depending on the breed and the low-density chip used. The alternative low-density chips gave better results than the commercially available 3K chip. A low-density chip with 6,000 markers is a valuable genotyping tool suitable for both dairy and beef breeds. Such a tool could be used for preselection of young animals or large-scale screening of the female population.

Key words: low-density chip, imputation, genomic selection, single nucleotide polymorphism (SNP) chip

Short Communication

Genomic selection (Meuwissen et al., 2001) is now widely used in dairy cattle breeding to select bulls at an early stage; it requires estimation of effects for SNP covering the whole genome at sufficient density. Thousands of bulls have been genotyped on a SNP panel with 54,000 markers (50K chip), commercially available from Illumina Inc. (San Diego, CA) and developed by Matukumalli et al. (2009). One of the next challenges to take full advantage of genomic selection is to genotype a large proportion of the female population. This would require a substantial reduction in genotyping costs. For this purpose, low-density chips can be considered as an alternative tool. The Golden Gate Bovine 3K chip (GG3K) was developed in 2009 by Illumina Inc. using inexpensive technology (Golden Gate). The drawback of having genotypes for a smaller number of markers can be overcome by applying imputation. Imputation is a statistical method that predicts unobserved genotypes, offering the possibility to infer a dense (e.g., 50K) genotype based on low-density chip data. Reconstructing denser coverage such as the standard BovineSNP50 panel (Illumina Inc.) from low-density chips is now recognized as the best way of using these cheaper tools (Weigel et al., 2010). Therefore, low-density chips should be developed in such a way that imputation accuracy is maximized. Genomic breeding values can then be computed from evaluations based on imputed genotypes (Weigel et al., 2010; Berry and Kearney, 2011; Dassonneville et al., 2011; VanRaden et al., 2011). In addition to genomic evaluation, low-density chips can be used for sexing and parentage assignment or verification. Minor allele frequency (MAF), quality edits, and spacing between markers are assumed to influence imputation performance. To check this hypothesis, 2 custom in silico low-density marker panels were developed optimizing both MAF and spacing. The objective of this study was to compare the imputation accuracy of different chips: the commercially available chip and 2 custom low-density chips differing in marker density (3K and 6K).

The reference population included individuals that were genotyped on the Illumina Bovine 50K chip. The validation population was a subset of the reference population, including the youngest animals that were
considered as selection candidates in this study. For this reason, low-density genotypes were mimicked for these animals. The training population consisted of the remaining individuals from the reference population.

Three breeds were chosen for this study and studied separately: Holstein (dairy), French Montbéliarde (dairy) breed, and Blonde d’Aquitaine (beef). The reference population of the Holstein and Montbéliarde dairy breeds included 4,037 and 1,219 progeny-tested bulls, respectively, distributed across several generations. The validation population for these 2 breeds was defined through a cut-off date such that approximately 25% of the bull’s reference population forming the validation population were born after that date. All of the selection candidates had their sires in the training population and most of the male ancestors were genotyped and included in the training population, whereas no female was genotyped. The reference population of the Blonde d’Aquitaine breed included 961 young bulls and their 30 sires; 237 young bulls were randomly selected to form the validation population. Therefore, for the Blonde d’Aquitaine, the sire was the only densely genotyped ancestor of the validation animals. Table 1 summarizes the number of animals included in training and validation populations for the 3 different breeds.

Only 50K genotype data were used, and the genotypic data were first edited per locus. The study focused on autosomes, and markers mapped on the X chromosome were deleted. Markers with an unknown map position in the Btau 4.0 assembly, departing from Hardy-Weinberg equilibrium, with a call rate <0.85, or with a MAF strictly equal to 0 were removed. Finally, genotype data were checked for Mendelian inconsistencies between parents and offspring, and inconsistent genotypes were set to missing values.

Three different low-density chips were defined. The GG3K chip included 2,900 SNP (Table 2), of which 2,635 were kept after edits and quality control as described above. The custom chips included 2,929 and 6,052 SNP, respectively, covering the 29 autosomes. Markers used for parentage testing were included. It must be mentioned that additional SNP from sex chromosomes should be integrated so that the low density chip can be used to perform sexing of embryos.

To be included in the 2 custom in silico panels, the selection criteria were as follows: (1) markers had to be present on versions 1 and 2 of the Bovine SNP50 (Matukumalli et al., 2009) and on the Bovine HD chip from Illumina; (2) markers had to have a known position on Btau 4.0 (Elsik et al., 2009) and UMD3 (Zimin et al., 2009) assemblies; (3) the marker position had to be consistent between the 2 assemblies; that is, less than 10 Mb apart; (4) call rates needed to exceed 0.98 with no technical problem observed in the sample of genotyped animals at INRA; and (5) markers were checked for Hardy-Weinberg equilibrium (q value >0.01).

From the set of markers meeting these criteria, SNP were chosen to maximize the MAF within chromosome segments. The MAF were available for 8 French dairy and beef breeds (Blonde d’Aquitaine, Brown Swiss, Charolais, Holstein, Limousine, Normande, Montbéliarde, and Maine-Anjou), with 110 to 16,055 samples per breed. For the custom 3K chip, the genome was divided into 1-Mb segments and, within each segment, the SNP with the highest average MAF over the 3 main French dairy breeds (Holstein, Normande, and Montbéliarde) was kept. For the custom 6K chip, in a first step, the SNP set of the custom 3K was retained. In a second step, the SNP with the highest average MAF over the 8 breeds was added within each 1-Mb segment. Finally, a few more SNP were added to cover every 0.5 Mb and to ensure a better coverage of chromosome extremities (4 markers per 1 Mb instead of 2). For these custom low-density chips, MAF was optimized and each 1 Mb (custom 3K chip) or each 0.5 Mb (custom 6K chip) was covered. These rules were simple and did not account for linkage disequilibrium between markers.

The GG3K chip included 2,900 SNP (Table 2), of which 2,635 were kept after edits and quality control as described above. The custom chips included 2,929 and 6,052 SNP, respectively, covering the 29 autosomes. Markers used for parentage testing were included. It must be mentioned that additional SNP from sex chromosomes should be integrated so that the low density chip can be used to perform sexing of embryos.

Table 1. Number of animals in the reference and validation sets for the 3 breeds

<table>
<thead>
<tr>
<th>Breed</th>
<th>Reference</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Montbéliarde</td>
<td>997</td>
<td>222</td>
</tr>
<tr>
<td>Holstein</td>
<td>3,071</td>
<td>966</td>
</tr>
<tr>
<td>Blonde d’Aquitaine</td>
<td>754</td>
<td>237</td>
</tr>
</tbody>
</table>

Table 2. Number of markers included in the 3 low-density chips

<table>
<thead>
<tr>
<th>Low-density chip¹</th>
<th>Markers (no.)</th>
<th>Markers after quality control (no.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Golden Gate Bovine 3K</td>
<td>2,900</td>
<td>2,635</td>
</tr>
<tr>
<td>French custom 3K</td>
<td>2,929</td>
<td>2,929</td>
</tr>
<tr>
<td>French custom 6K</td>
<td>6,144</td>
<td>6,052</td>
</tr>
</tbody>
</table>

¹Golden Gate Bovine 3K chip was from Illumina Inc. (San Diego, CA); French custom 3K and 6K chips were simulated and developed in silico.
The SNP included on the low-density chips were all included on the Bovine 50K chip. To mimic the low-density chip, marker genotypes of validation animals were obtained by erasing markers from the 50K and not present in the low-density chip. Imputation of markers was performed using Beagle 3.2 (Browning and Browning, 2007) with the unphased option. Consequently, pedigree information was not used for imputation. With Beagle, all markers are imputed and the method does not leave any missing markers.

Efficiency was first measured by the number of correctly imputed alleles (as in Zhang and Druet, 2010). Therefore, the number of errors was counted as 0 when the imputed and observed marker genotypes were identical, 1 if the real genotype was homozygous and the imputed one was heterozygous (or vice versa), and 2 if real and imputed marker genotypes were opposite homozygous. The error rate was calculated as the total number of errors divided by twice the number of imputed (masked) loci.

Several authors (Weigel et al., 2010; Druet et al., 2010; VanRaden et al., 2011) report the same kind of measures (error rate or concordance rate). Hickey et al. (2012) suggest using the correlation between true and imputed genotypes to account for MAF. Indeed, a high concordance rate is expected for low MAF and may overestimate imputation performance. In this study, both measures were reported.

The percentage of alleles correctly imputed is presented in Figure 1 for the 3 breeds and the 3 low-density chips. In the Montbéliarde breed, 97.4% of the masked alleles were correctly imputed from the GG3K chip. This result seems to be sufficiently high to implement genomic selection based on low-density chips; indeed, Weigel et al. (2010) found lower imputation accuracy and were still able to run appropriate genomic evaluations for the Jersey breed. However, imputation accuracy can be improved with other marker panels. The custom 3K chip included more effective markers (markers kept after quality edits) and was optimized for French dairy breeds, including Montbéliarde. For this reason, this optimized 3K chip gave better imputation accuracy (98.0%) than the GG3K chip (97.4%).

With the 6K chip, both marker density and MAF optimization were improved, resulting in higher imputation accuracy (99%).

Holstein was 1 of the 3 breeds involved in the choice of markers for the GG3K chip. For this reason, we might have expected similar imputation results between the two 3K chips. However, the custom 3K chip gave better results (imputation accuracy of 98.1% compared with 97.4%). One possible explanation is the lower number of effective markers on the GG3K chip. Another explanation is the constraints of the Golden Gate technology or chemistry in the choice of markers, which might have limited the possible optimization on MAF and spacing. As expected, results were better with the 6K chip (more than 99% of alleles correctly imputed).

Imputation accuracy was lower in the French Blonde d’Aquitaine beef breed compared with the other 2 breeds, probably because of a smaller and different reference population, with few ancestors genotyped, and...
a larger effective population size. Imputation accuracy was slightly better with the custom 3K chip (95.8%) than with the GG3K chip (95.2%), although MAF of Blonde d’Aquitaine were not taken into account when constructing these 2 chips. This may be related to the better optimization of MAF and spacing or an increased number of efficient markers as reported above. The largest gain for the 6K chip compared with the 3K chips was obtained with the beef breed (97.5% vs. 95–96%). On the one hand, Blonde d’Aquitaine MAF was accounted for in the design of the 6K chip. On the other hand, the advantage of the 6K chip was greater because the performance of the 3K chips was lower in beef breeds than in dairy breeds, leaving a larger margin for improvement.

Using concordance rate as an imputation efficiency criterion may be misleading because it depends on MAF. The lower the MAF, the higher the concordance rate for the same efficiency and this should be accounted for in the interpretation. For example, with an average MAF of 0.2, 80% of the results would be correct after random sampling of the missing alleles. Achieving a 95% concordance rate corresponds to an imputation efficiency of only 75% [(95–80)/(100–80)]. Correlation is an alternative criterion less dependent on MAF. Table 3 presents the correlations between true and imputed genotypes, which ranged from 0.88 to 0.97. Comparison across breeds and across marker panels led to the same conclusions as studying the fraction of alleles correctly imputed. The ranking of the chips was the same and results for Blonde d’Aquitaine were lower than those for dairy breeds.

A previous study by Zhang and Druet (2010) showed that both the number of reference animals and the number of markers in a low-density panel affect the imputation error rate. This error rate is also affected by the relationship between validation and reference animals. Comparing imputation error rates based on different populations is therefore difficult because the relationship between training and validation populations differs, and because the number of reference individuals and the linkage disequilibrium between markers vary.

One major concern regarding the use of the Beagle software is the computational time. In a large population from the Eurogenomics reference population (Lund et al., 2011), with 12,068 animals in the training population and 3,987 animals in the validation population, computing time with Beagle alone was from 40 to 80 h per chromosome. Alternative methods using long-range phasing and pedigree information (Hickey et al., 2011; VanRaden et al., 2011) are known to be much faster and have been compared with each other and with Beagle by Johnston et al. (2011). However, we propose an alternative that benefits from the accuracy of Beagle and the fast algorithm of the PHASEBOOK package. This package was developed by Druet and Georges (2010), takes advantage of pedigree relationships, and uses this family information in addition to population-based linkage disequilibrium through the DAgphase software. Indeed, it is possible to run Beagle first from scratch, to obtain and store the directed acyclic graph (DAG). This demanding analysis needs to be run only once. Then, the DAG obtained from Beagle can be used within DAgphase. This solution is as accurate as Beagle (in terms of error rate; results not shown), and much less time consuming (1 h per chromosome compared with the 40 to 80 h of Beagle on the same data set including 16,055 individuals). This 2-step approach is quite efficient. The second step, corresponding to only one iteration of DAgphase, is fast and can be run as often as necessary (i.e., monthly or weekly for genomic evaluations). An additional advantage of this combined approach over other methods (Hickey et al., 2011; Sargolzaei et al., 2011) is that all masked markers are called.

The DAgphase software was used by Zhang and Druet (2010) to impute from 3,000 to 50,000 markers, where the 3,000 markers were especially selected for high MAF. They found an allele error rate (i.e., the proportion of incorrectly predicted alleles) of around 3% with a reference population of 4,734 Holstein bulls. Our hypothesis was that better optimization of MAF and spacing for the choice of markers to include in low-density chips may lead to a gain in imputation accuracy. Comparison of imputation results of the GG3K

Table 3. Correlation between true and imputed genotypes for the 3 breeds

<table>
<thead>
<tr>
<th>Low-density chip</th>
<th>Montbéliarde</th>
<th>Holstein</th>
<th>Blonde d’Aquitaine</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoldenGate Bovine3K</td>
<td>0.94</td>
<td>0.93</td>
<td>0.88</td>
</tr>
<tr>
<td>French custom 3K</td>
<td>0.95</td>
<td>0.94</td>
<td>0.89</td>
</tr>
<tr>
<td>French custom 6K</td>
<td>0.97</td>
<td>0.96</td>
<td>0.92</td>
</tr>
</tbody>
</table>

1Golden Gate Bovine 3K chip was from Illumina Inc. (San Diego, CA); French custom 3K and 6K chips were simulated and developed in silico.
and the 3K custom in silico chip (same marker density), for both Holstein and Montbéliarde breeds confirmed this hypothesis. The relatively lower performance of the GG3K chip may be due to the constraints related to Golden Gate chemistry and the smaller effective number of markers kept after quality control (2,635 markers instead of 2,900, leading to a decrease of 10% in marker density).

Imputation using Beagle software was efficient to reconstruct a dense (50K) genotype from low-density chip data. Accuracy, measured by the allelic concordance rate, ranged between 95 and 99%. The highest values were obtained with the highest density 6K chip and the Holstein population (characterized by a large reference population). Using the DAG obtained from Beagle in the PHASEBOOK algorithm allowed us to take advantage of family information and speed up the imputation process with no loss in imputation accuracy.

Low-density chips are appealing alternative tools that reduce genotyping costs, allowing the genotyping of more animals. They can be used for preselection of young animals and is an interesting option for large-scale genomic selection of females.

The existing GG3K chip yielded satisfactory results. However, other choices of markers are possible for low-density chips to optimize MAF and spacing for various breeds so that imputation is more accurate. The 6K chip appears to be the method of choice and provides high imputation efficiency, even for a beef breed with a small reference population such as the French Blonde d’Aquitaine breed. Consequently, a low-density chip with around 6,000 markers is an appealing genotyping tool that is suitable for both dairy and beef breeds. This option was chosen by Illumina to produce the new LD chip in collaboration with an international consortium (Boichard et al., 2012).

ACKNOWLEDGMENTS

Agence Nationale de la Recherche (Paris, France) and APISGENE (Paris, France) are gratefully acknowledged for funding the AMASGEN project. Blonde d’Aquitaine data were obtained from the Qualvigene project.

REFERENCES