Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations

Romain R. Dassonneville, R.F. R. Brøndum, T. T. Druet, Sébastien S. Fritz, François F. Guillaume, B. B. Guldbrandtsen, M.S. M. Lund, Vincent Ducrocq, G. G. Su

J. Dairy Sci. 94:3679–3686
doi:10.3168/jds.2011-4299

Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations

R. Dassonneville,*†[1,2] R. F. Brøndum,‡[1,2] T. Druet,§ S. Fritz,# F. Guillaume,*† B. Guldbrandtsen,‡ M. S. Lund,‡ V. Ducrocq,* and G. Su‡

*INRA, UMR1313 Génétique Animale et Biologie Intégrative (GABI), 78350 Jouy-en-Josas, France
†Institut de l'Elevage, 149 rue de Bercy, 75595 Paris, France
‡Aarhus University, Faculty of Science and Technology, Department of Genetics and Biotechnology, DK-8830 Tjele, Denmark
§Unit of Animal Genomics, GIGA-Research and Department of Animal Production, Faculty of Veterinary Medicine, University of Liège, B-4000 Liège, Belgium
#Union Nationale des Coopératives d'Élevage et d'Insémination Animale (UNCEIA), 149 rue de Bercy, 75595 Paris, France

ABSTRACT

The purpose of this study was to investigate the imputation error and loss of reliability of direct genomic values (DGV) or genomically enhanced breeding values (GEBV) when using genotypes imputed from a 3,000-marker single nucleotide polymorphism (SNP) panel to a 50,000-marker SNP panel. Data consisted of genotypes of 15,966 European Holstein bulls from the combined EuroGenomics reference population. Genotypes with the low-density chip were created by erasing markers from 50,000-marker data. The studies were performed in the Nordic countries (Denmark, Finland, and Sweden) using a BLUP model for prediction of DGV and in France using a genomic marker-assisted selection approach for prediction of GEBV. Imputation in both studies was done using a combination of the DAGPHASE 1.1 and Beagle 2.1.3 software. Traits considered were protein yield, fertility, somatic cell count, and udder depth. Imputation of missing markers and prediction of breeding values were performed using 2 different reference populations in each country: either a national reference population or a combined EuroGenomics reference population. Validation for accuracy of imputation and genomic prediction was done based on national test data. Mean imputation error rates when using national reference animals was 5.5 and 3.9% in the Nordic countries and France, respectively, whereas imputation based on the EuroGenomics reference data set gave mean error rates of 4.0 and 2.1%, respectively. Prediction of GEBV based on genotypes imputed with a national reference data set gave an absolute loss of 0.05 in mean reliability of GEBV in the French study, whereas a loss of 0.03 was obtained for reliability of DGV in the Nordic study. When genotypes were imputed using the EuroGenomics reference, a loss of 0.02 in mean reliability of GEBV was detected in the French study, and a loss of 0.06 was observed for the mean reliability of DGV in the Nordic study. Consequently, the reliability of DGV using the imputed SNP data was 0.38 based on national reference data, and 0.48 based on EuroGenomics reference data in the Nordic validation, and the reliability of GEBV using the imputed SNP data was 0.41 based on national reference data, and 0.44 based on EuroGenomics reference data in the French validation.

Key words: genomic selection, imputation, reliability, reference population

INTRODUCTION

Genomic selection (Meuwissen et al., 2001) is becoming a routine tool for genetic evaluation in dairy cattle breeding. Currently, a SNP panel with 54,000 markers is widely used. A new low-density panel with only 3,000 markers at a lower price, potentially decreasing genotype costs, is now also available (Illumina Inc., San Diego, CA). Using the low-density panel instead of the current one may allow cattle breeders to genotype more bulls and cows.

Several options for selecting a low-density panel have been suggested. One option is to select several markers with large effects for a given trait; another is to use markers evenly spaced across the genome. Previous studies showed that the difference in reliability of the genomic breeding values, when using 3,000 markers with large effect or 3,000 markers evenly spread across the genome, is small (Moser et al., 2010). The option of evenly spaced markers removes the need for trait- and breed-specific low-density SNP panels. The efficiency of a trait-specific marker panel also depends on the linkage disequilibrium (**LD**) between the markers with large effect and the actual QTL. This LD might de-

Received February 23, 2011.
Accepted April 8, 2011.
[1]These authors contributed equally to this work.
[2]Corresponding authors: romain.dassonneville@jouy.inra.fr and rasmusf.brondum@agrsci.dk

crease through generations. The other advantage of evenly spaced markers is the possibility to use statistical methods to impute the missing markers, thus extending the 3,000 markers to 50,000 markers, albeit with some uncertainty. This is also possible with unevenly spread markers, but then the accuracy of imputation is expected to be lower.

It has been reported that a lower marker density leads to lower reliability of genomic prediction (Moser et al., 2010). A feasible strategy is to extend the low-density markers to the current 50,000 markers by imputation. Several methods for imputation of SNP markers, relying on either linkage based on family information (Daetwyler et al., 2010) or LD based on population information (Scheet and Stephens, 2006; Browning and Browning, 2007), have been proposed. It is also possible to combine both types of information (Druet and Georges, 2010). In a study using this combined approach to impute from 3,000 to 50,000 markers, where the 3,000 markers were specially selected for high minor allele frequency, Zhang and Druet (2010) found an allele error rate (i.e., the proportion of incorrectly predicted alleles, of approximately 3%). A study by Weigel et al. (2010) on American Jersey cattle has shown that using 3,000 SNP for candidates imputed to a 50,000-marker SNP panel can provide approximately 95% of the predictive ability achieved using the real 50,000-marker SNP panel.

The accuracy of imputation can be increased by increasing the size of the reference population. EuroGenomics is a collaboration between 4 European AI companies and scientific partners: Deutscher Holstein Verband e.V.-Vereinigte Informationssysteme Tierhaltung w.V. (DHV-VIT; Germany), Union Nationale des Coopératives d'Élevage et d'Insémination Animale-Institut National de la Recherche Agronomique (UNCEIA-INRA; France), Coöperatie Rundvee Verbetering (CRV; the Netherlands, Flanders), and Viking Genetics-Aarhus University (Denmark-Finland-Sweden). The collaboration includes the sharing of reference populations for genomic selection, where each country initially contributed 4,000 genotyped Holstein bulls with progeny-tested breeding values. A previous study showed a significant increase in reliability of genomic breeding values using this combined reference population (Lund et al., 2010). We expect that the accuracy of imputation based on EuroGenomics reference data will be higher than that based on national reference data.

The objective of this study was to investigate the imputation error, when imputing from a 3,000-marker SNP panel to a 50,000-marker SNP panel using a group of reference animals with 50,000-marker information. The 3,000 markers were the same as the Illumina 3,000-marker SNP panel. The imputed SNP markers were used for genomic prediction to assess how the imputation error rate affects the reliability of genomic breeding values and the ranking of the animals. This assessment was carried out in the Nordic countries and France. For both analyses, a validation population consisting of national test animals with 3,000-marker genotype was imputed to the 50,000-marker genotype using a reference population made of either national or EuroGenomics data.

MATERIALS AND METHODS

Data

The combined EuroGenomics reference population contains 15,966 progeny-tested bulls with genotypes from the Illumina Bovine 50,000-marker SNP panel (Matukumalli et al., 2009). Four thousand Dutch bulls were genotyped using a customized CRV 60,000-marker chip, but by double genotyping 972 influential bulls with the Illumina 50,000-marker chip, it was possible to impute markers from the Illumina chip for all Dutch bulls with an imputation error of less than 1% (Druet et al., 2010).

Measurement of imputation error rate and reliability of genomic predictions for Nordic and French bulls were carried out separately, using either national or EuroGenomics reference data. Deregressed proofs (**DRP**) on the scale of the target population, calculated from Interbull 2010–01 multiple-trait across country evaluation (MACE) proofs (www.interbull.org), were used for predicting and validating direct genomic values (**DGV**) and GEBV, if the equivalent daughter contribution was at least 20 (Lund et al., 2010). In the French study, daughter yield deviations (**DYD**) from the October 2009 national evaluation were used as phenotypes for the French bulls. The reference and validation populations were divided according to the bulls' birth dates. The cut-off dates were October 1, 2001 and June 13, 2002 in the Nordic and French case, respectively. Thus, about 25% of national genotyped bulls were taken as a validation set.

The traits studied were protein yield, SCC, fertility (defined as non-return rate in the Nordic countries and conception rate in France), and udder depth. Heritabilities and number of animals available for the specific traits are shown in Table 1.

Marker data were edited according to procedures used in Nordic countries and in France.

Nordic Marker Editing

The genotypic data was edited both per animal and per locus. At the animal level, the requirements were a

Table 1. Heritabilities (h^2) and number of animals used for protein yield, SCC, fertility, and udder depth (UD) in the Nordic and French studies

	Nordic study				French study			
Trait	h^2	Nordic reference	Euro reference	Nordic validation	h^2	French reference	Euro reference	French validation
Protein yield	0.39	3,038	10,701	899	0.3	3,071	12,078	966
SCC	0.15	3,077	10,800	899	0.15	3,071	12,078	966
Fertility	0.02	3,069	10,712	895	0.02	3,071	12,078	966
UD	0.37	2,958	10,755	900	0.36	3,071	12,078	966

call rate above 95%, except for some old animals that were accepted with call rates of at least 85%. Marker loci were accepted if they had a call rate of at least 95% in a large reference sample. Loci with a minor allele frequency (**MAF**) less than 5% were excluded. Loci without a known map position in the *Bos taurus* (Btau) 4.0 assembly or mapped on the X chromosome were discarded. Animals with an average GenCall score (Illumina Inc., 2005) of less than 0.65 were excluded. Individual marker typings with a GenCall score of less than 0.6 were also discarded.

French Marker Editing

The French genotypic data was first edited per locus. Markers without a known map position in the Btau 4.0 assembly or mapped to the X chromosome were removed. Markers were then filtered for Hardy-Weinberg equilibrium (q-value <0.01). Markers with call rates below 0.85 were removed. Markers with MAF strictly equal to 0 were removed. Genotype data were finally checked for Mendelian inconsistencies between parents and offspring. Inconsistent genotypes were set to missing. Marker-editing procedures differed slightly between France and the Nordic countries (including GenCall score, for example).

While checking for inconsistencies between parents and offspring, Mendelian segregation rules were also applied to determine marker types of ungenotyped ancestors. Inferred marker data was not complete. However, it is important for ancestors with large numbers of progeny. Thus, the French national training population included 3,071 animals with real observed marker types (Table 2) and a total of 3,505 when ancestors with imputed genotypes were included. The corresponding figures for the EuroGenomics population were 12,078 and 13,947 animals, respectively. This might help for further imputation, especially through linkage information.

Simulating Illumina Bovine 3K BeadChip Data

The 2,900 SNP in the Illumina Bovine 3,000-marker (3K) BeadChip are all included in the Bovine 50,000-marker (50K) BeadChip (except for 14 markers located on the Y chromosome). To mimic the low-density chip, marker types of test animals (i.e., animals born after the cut-off date), were obtained by erasing markers from the 50,000-marker type (i.e., in silico chip). As 3,000-marker genotypes are simulated from 50,000-marker data, they do not account for a possibly higher genotyping error rate with the 3K BeadChip. After marker editing as outlined above, 2,285 and 2,635 markers were kept for the Nordic and French data (see Table 2).

Imputation of Missing SNP Markers

Imputation of markers was done using the PHASEBOOK package (Druet and Georges, 2010) in combination with Beagle 2.1.3 (Browning and Browning, 2007). The method was applied as a stepwise procedure using both linkage and LD information. The same procedure as in Zhang and Druet (2010) was applied. First, all markers that can be determined unambiguously using Mendelian segregation rules were phased using the LinkPhase software (part of PHASEBOOK; Druet and Georges, 2010). In the first step, both training and

Table 2. Number of animals and number of markers used

	National		EuroGenomics		No. of markers	
Study	Reference	Validation	Reference	Validation	Reference	Validation
Nordic	3,058	1,086	10,880	1,086	38,545	2,285
French	3,071/3,505[1]	966	12,078/13,947[1]	966	43,582	2,635

[1]Including bulls with partially reconstructed genotypes.

test animals were included. An iterative procedure was then applied, where a directed acyclic graph (DAG), describing the haplotype structure of the genome, was fitted to the partially phased data from the previous step. This was, however, only done for the reference animals. This was done for 10 iterations and then the final DAG, the genotype file, and the output from LinkPhase (partially phased data) were used to reconstruct haplotypes and impute missing markers for both test and training animals using the Viterbi algorithm. With Beagle and PHASEBOOK, all markers are imputed and the method does not leave any missing markers. More details on the imputation procedure can be found in Druet et al. (2010) and Zhang and Druet (2010).

Allele Imputation Error Rate Calculation

The number of errors was counted as 0 when the imputed and observed marker types were identical, 1 if the real marker type was homozygous and the imputed genotype was heterozygous (or vice versa), and 2 if real and imputed marker types were opposite homozygous. Error counting only considered markers/animals where observed marker types were not missing in the original non-imputed data set. The error rate was calculated as the total number of errors divided by twice the number of imputed loci. This gives the number of falsely predicted alleles, which is an appropriate measure when using an additive prediction model, as in this study. For other purposes, the genotype error rate could be easily found as approximately twice the allele error rate (Zhang and Druet, 2010).

Prediction of Direct Genomic Values in Nordic Countries

Prediction of DGV was performed using a BLUP model at SNP level (VanRaden, 2008). Specifically, the model is given by

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of phenotypic observations, μ is the mean, $\mathbf{u}$ is a vector of SNP effects, $\mathbf{e}$ is the random error vector, and $\mathbf{Z} = \mathbf{M} - \mathbf{P}$ is a design matrix for the random effects. The marker matrix $\mathbf{M}$ is an m by n matrix, where m is the number of animals and n is the number of markers. Entries in the ith row of $\mathbf{M}$ are the genotypes for the ith animal and are represented by -1 if the animal is homozygous aa, 0 if the animal is heterozygous, and 1 if the animal is homozygous AA. The matrix $\mathbf{P}$ has n columns, where the elements in column j are $\mathbf{P}_j = 2p_j - 1$, where p_j is the frequency of allele A at locus j. Subtraction of the allele frequencies standardizes the allele effects to a population mean of 0. Thus, $\mathbf{a} = \mathbf{Zu}$ gives the direct genomic values.

Deregressed proofs were used as phenotypic values in the model. The weighting factor $r^2_{DRP}\Big/\left(1 - r^2_{DRP}\right)$ was used to scale the inversed residual variance of an observation.

The DGV reliability was calculated as the weighted squared correlation between DRP and DGV, divided by the mean reliability of DRP. The weights were given by $r^2_{DRP}\Big/\left(1 - r^2_{DRP}\right)$, standardized to a mean weight of 1.

Prediction of Genomic Breeding Values in France

The French genomic prediction is an extension of the marker-assisted evaluation method by Fernando and Grossman (1989). The model is the following:

$$\mathbf{y} = 1\mu + \mathbf{Zu} + \sum_{i=1}^{nQTL} (h_{i1} + h_{i2}) + \mathbf{e},$$

where $\mathbf{y}$ is the vector of phenotypic observations, μ is the overall mean, $\mathbf{u}$ is a vector of random pedigree-based residual polygenic effects, h_{ij} is the random effect of haplotype j for QTL i, and $\mathbf{e}$ is a vector of residuals, with heterogeneous residual variances inversely proportional to equivalent daughter contributions.

The selection of QTL included in the model was the result of a combination of 2 approaches (Boichard et al., 2010). First, dozens of QTL per trait were detected after QTL fine mapping, as described below. Then, hundreds of haplotypes were chosen using the elastic net algorithm (**EN**).

For QTL mapping, a LD linkage analysis (**LDLA**), combining both within-family linkage information and population-based LD, was used on the EuroGenomics training population (12,078 animals), following the approach described by Druet et al. (2008). Identity by descent probabilities were calculated as in Meuwissen and Goddard (2001). The likelihood ratio test threshold to retain a QTL was arbitrarily set to 6. This resulted in the selection of 80 to 100 QTL, depending on the trait. Fine-mapped QTL were traced by haplotypes of 5 flanking markers.

Then, an EN procedure was run on the French training population (3,071 animals), following the approach described by Croiseau et al. (2010). The selected SNP were grouped into haplotypes of 3 to 5 SNP. The 2 sets of haplotypes were included in the model. For computational reasons, the number of markers detected by the EN procedure included in the model was limited so that the total number of QTL was at maximum 700.

Table 3. Number (n) of QTL selected for the French prediction model using either linkage disequilibrium linkage analysis (LDLA) or an elastic net (EN) procedure and percentage of allocated genetic variance (% var) for protein yield, SCC, fertility, and udder depth (UD)

	LDLA		EN		Total	
Trait	n	% var	n	% var	n	% var
Protein yield	100	24	593	30	693	54
SCC	80	27	362	30	442	57
Fertility	80	21	392	30	472	51
UD	80	27	482	30	562	57

The genetic variance attached to each QTL detected through LDLA mapping was proportional to the variance estimated in the single QTL analysis. The variance explained by each haplotype selected by EN was assumed to be constant and their sum over all EN haplotypes was set to 30% of genetic variance. As shown in Table 3, 442 to 693 QTL were included in the model, explaining 51 to 57% of genetic variance.

RESULTS AND DISCUSSION

Accuracy of Imputation

Imputation in the Nordic population showed a mean error rate of 5.5% when using only Nordic animals as the reference set (Table 4). The extension of this reference set with the EuroGenomics animals gave an error rate of 4.0%. The same pattern was found in the French population, where a French reference set gave a mean imputation error rate of 3.9%, whereas increasing the reference set with EuroGenomics animals decreased it to 2.1%. The lower error rate in the French study is likely due to the inclusion of more markers for the study on the 3K BeadChip (2,635 vs. 2,285) because of different marker-editing rules (such as selection on MAF), giving a denser genome coverage and a higher homozygosity. A previous study by Zhang and Druet (2010) showed that both the number of reference animals and the number of markers in the low-density panel affect the imputation error rate. This error rate is also affected by the relationship between validation and reference animals.

In the Nordic study, it was observed that the mean error rate depended on whether or not the animals had their sire in the reference data, confirming that a closer relationship to the reference population decreases the imputation error rate in the test population. All of the animals in the French validation population had their sire in the reference population, and an additional step based on Mendelian segregation rules was carried out to partially reconstruct genotypes of ungenotyped ancestors. The results indicate that if low-density genotyping and imputation are widely used in the future, the imputation accuracy might decrease unless all breeding bulls are genotyped with the 50,000-marker panel.

The results in the present study are consistent with the error rates obtained by Zhang and Druet (2010) using the same method for imputation (i.e., between 2.1 and 4%). Their reference population was smaller (500 to 2,000 animals) but their 3,000-marker panel was optimized according to MAF for their population; thus, all markers on the 3,000-marker panel were available, whereas some were excluded during the quality control in the present study. Comparing imputation error rates based on different studies is, however, difficult because the relationship between training and validation populations differs, and because the number of reference individuals and the number of markers vary.

Table 4. Imputation allele error rates (%) for Nordic and French test animals using a national reference population or the EuroGenomics reference population

	National		EuroGenomics	
Test population	N	Error rate	N	Error rate
Nordic				
All	1,086	5.5	1,086	4.0
Sire in ref.[1]	795	4.5	1,039	3.8
Sire not in ref.	291	8.3	47	7.0
Sire and maternal grandsire in ref.	650	4.3	953	3.8
French				
Sire in ref.	966	3.9	966	2.1

[1]Ref. = reference population.

Table 5. Reliabilities of direct genomic values for Nordic candidates with full (50,000; 50K) or imputed (3,000; 3K imp) marker data for protein yield, SCC, nonreturn rate (NRR), and udder depth (UD) using either Nordic reference population (Nor-ref) or EuroGenomics reference population (EU-ref)

Trait	N	Nor-ref 50K	Nor-ref 3K imp	EU-ref 50K	EU-ref 3K imp
Protein yield	899	0.41	0.32	0.56	0.51
SCC	899	0.41	0.39	0.55	0.49
NRR	895	0.44	0.42	0.49	0.45
UD	900	0.40	0.36	0.55	0.49
Average		0.41	0.38	0.54	0.48

Reliability of Genomic Prediction

Prediction of DGV based on either true or imputed genotypes in the Nordic data (Table 5) showed that using the Nordic reference population, the observed marker types had a mean reliability of DGV over the 4 traits of 0.41, whereas the imputed marker types led to a mean reliability of 0.38. Using the EuroGenomics data as the reference population for prediction of DGV resulted in a mean reliability of 0.54 with the observed marker types, whereas using imputed genotypes resulted in a mean reliability of 0.48.

For the prediction of GEBV based on either observed or imputed marker types for the French validation data (Table 6), with a French national reference population and observed marker types, a mean reliability across 4 traits of 0.46 was obtained. The corresponding value for imputed marker types was 0.40. Using the EuroGenomics data as a training population, the mean reliability of GEBV of young animals was 0.48 and 0.46 for observed and imputed marker types, respectively.

Lund et al. (2010) reported that reliabilities of genomic prediction using the EuroGenomics reference data were considerably higher than those using national data, because of the increased size of the reference data. The French validation in this study, however, showed a small difference between reliabilities of GEBV predicted from the national and the EuroGenomics reference data. The small difference can be explained by the way the QTL were chosen for the prediction model. For both, the prediction model based on national reference data and the prediction model based on EuroGenomics reference data, the QTL selected using the LDLA procedure were based on EuroGenomics data, and the QTL selected using the EN procedure were based on national data. On one hand, the genomic prediction based on French data gained from LDLA based on the whole EuroGenomics population. On the other hand, genomic predictions based on EuroGenomics data were probably suboptimal because the EN procedure used only French data. The only way to properly measure the effect of increasing the reference population on genomic reliability based on real genotypes would have been to do 2 LDLA QTL mappings (as in Lund et al., 2010), but the main focus of this study was on imputation. The haplotype effects were, however, estimated on either EuroGenomics data or national data, leading to a gain in reliability when increasing the reference population.

The patterns of differences between reliabilities of genomic predictions using observed 50,000-marker types and the imputed marker types were not consistent between the Nordic and French validations. The difference was smaller when using national reference data than when using EuroGenomics reference data in the Nordic evaluation, whereas an opposite pattern was observed in the French validation. The reasons for the inconsistent pattern were not clear. A possible reason was that the markers with high imputation error rate might give different contribution to genomic prediction when using different reference data sets. For example, MAF for the loci with high imputation error rate might

Table 6. Reliabilities of genomically enhanced breeding values for French candidates with full (50,000; 50K) or imputed (3,000; 3K imp) marker data for protein yield, SCC, conception rate (CR), and udder depth (UD) using either French reference population (FR-ref) or EuroGenomics reference population (EU-ref)

Trait	N	FR-ref 50K	FR-ref 3K imp	EU-ref 50K	EU-ref 3K imp
Protein yield	966	0.40	0.32	0.37	0.36
SCC	966	0.55	0.52	0.58	0.57
CR	966	0.44	0.41	0.47	0.44
UD	966	0.45	0.40	0.51	0.48
Average		0.46	0.41	0.48	0.46

Table 7. Correlations between direct genomic values or genomically enhanced breeding values predicted using observed or imputed marker data for Nordic and French candidates for protein yield, SCC, fertility, and udder depth (UD) using either Nordic reference population (Nor-ref), French reference population (FR-ref), or EuroGenomics reference population (EU-ref)

	Nordic		French	
Trait	EU-ref	NOR-ref	EU-ref	FR-ref
Protein yield	0.94	0.92	0.97	0.94
SCC	0.93	0.92	0.95	0.94
Fertility	0.96	0.95	0.96	0.94
UD	0.93	0.93	0.94	0.91

be smaller (less informative) in one set of reference data, whereas they might be larger (more informative) in another set of reference data.

Correlations between DGV/GEBV based on imputed or observed marker types were high (Table 7). The correlation ranged from 0.92 to 0.95 using national reference data and from 0.93 to 0.96 using EuroGenomics data in the Nordic validation. Similarly, the correlations ranged from 0.91 to 0.94 using national reference data and from 0.94 to 0.97 using EuroGenomics data in the French validation. These results indicate no serious re-ranking of animals when using imputed data.

CONCLUSIONS

Imputation of the commercially available low-density bovine 3K BeadChip to the bovine 50K BeadChip gave allele error rates between 2.1 and 5.5%. The accuracies of imputation were higher when using the EuroGenomics reference data sets than when using national reference data sets. Imputation was more accurate when the sire of the candidate was genotyped on the 50,000-marker panel. Using the imputed markers for candidates, the mean reliability of DGV was 0.38 based on based on national reference data, and 0.48 based on EuroGenomics reference data in the Nordic validation, and the reliability of GEBV using the imputed SNP data was 0.41 based on national reference data, and 0.44 based on EuroGenomics reference data in the French validation. Therefore, a 3K SNP BeadChip imputed to 50K could be a feasible alternative for pre-selection of young animals. One may also consider 3,000-marker genotyping as an attractive tool for a large pre-screening of the female population.

ACKNOWLEDGMENTS

Data for this study were provided by the EuroGenomics consortium. This study was partially supported by Apis-Gene (Paris, France) and the French Agence Nationale de la Recherche (ANR) project Approches Méthodologiques et Applications de la Sélection Génomique (AMASGEN). Tom Druet is a research associate from the Fonds de la Recherche Scientifique (F.R.S.–FNRS, Brussels, Belgium).

REFERENCES

Boichard, D., F. Guillaume, A. Baur, P. Croiseau, M. N. Rossignol, M. Y. Boscher, T. Druet, L. Genestout, A. Eggen, L. Journaux, V. Ducrocq, and S. Fritz. 2010. Genomic selection in French dairy cattle. Manuscript 716 in Proc. World Congress of Genetics Applied to Livestock Production, Leipzig, Germany. www.wcgalp2010.org.

Browning, S. R., and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81:1084–1097.

Croiseau, P., C. Colombani, A. Legarra, F. Guillaume, S. Fritz, A. Baur, R. Dassonneville, C. Patry, C. Robert-Granié, and V. Ducrocq. 2010. Improving genomic evaluation strategies in dairy cattle through SNP pre-selection. Manuscript 360 in Proc. World Congress of Genetics Applied to Livestock Production, Leipzig, Germany. www.wcgalp2010.org.

Daetwyler, H. D., G. R. Wiggans, B. J. Hayes, J. A. Woolliams, and M. E. Goddard. 2010. Imputation of missing genotypes from sparse to high density using long-range phasing. Manuscript 539 in Proc. World Congress of Genetics Applied to Livestock Production, Leipzig, Germany. www.wcgalp2010.org.

Druet, T., S. Fritz, M. Boussaha, S. Ben-Jemaa, F. Guillaume, D. Derbala, D. Zelenika, D. Lechner, C. Charon, D. Boichard, I. G. Gut, A. Eggen, and M. Gautier. 2008. Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map. Genetics 178:2227–2235.

Druet, T., and M. Georges. 2010. A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184:789–798.

Druet, T., C. Schrooten, and A. P. W. de Roos. 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. J. Dairy Sci. 93:5443–5454.

Fernando, R., and M. Grossman. 1989. Marked assisted selection using best linear unbiased prediction. Genet. Sel. Evol. 21:467–477.

Illumina Inc. 2005. Illumina GenCall data analysis software. GenCall software algorithms for clustering, calling, and scoring genotypes. Illumina. Pub. No. 370–2004–009.

Lund, M. S., A. P. W. de Roos, A. G. de Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, R. Rents, C. Schrooten, M. Seefried, and G. Su. 2010. Improving genomic prediction by EuroGenomics collaboration. Manuscript number 880 in Proc. World Congress of Genetics Applied to Livestock Production, Leipzig, Germany. www.wcgalp2010.org.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith,

T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4:e5350.

Meuwissen, T. H. E., and M. E. Goddard. 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genet. Sel. Evol. 33:605–634.

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.

Moser, G., M. S. Khatkar, B. J. Hayes, and H. W. Raadsma. 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet. Sel. Evol. 42:37.

Scheet, P., and M. Stephens. 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78:629–644.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.

Weigel, K. A., G. de los Campos, A. I. Vazquez, G. J. M. Rosa, D. Gianola, and C. P. Van Tassell. 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. J. Dairy Sci. 93:5423–5435.

Zhang, Z., and T. Druet. 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. J. Dairy Sci. 93:5487–5494.