Accuracy of genomic prediction using an evenly spaced, low-density single nucleotide polymorphism panel in broiler chickens

C. Wang,* D. Habier,* B. L. Peiris,* A. Wolc,* A. Kranis,† K. A. Watson,† S. Avendano,† D. J. Garrick,* R. L. Fernando,* S. J. Lamont,* and J. C. M. Dekkers*[1]

*Department of Animal Science, Iowa State University, Ames 50011; and †Aviagen Ltd., Newbridge, Midlothian, EH28 8SZ, United Kingdom

ABSTRACT One approach for cost-effective implementation of genomic selection is to genotype training individuals with a high-density (HD) panel and selection candidates with an evenly spaced, low-density (ELD) panel. The purpose of this study was to evaluate the extent to which the ELD approach reduces the accuracy of genomic estimated breeding values (GEBV) in a broiler line, in which 1,091 breeders from 3 generations were used for training and 160 progeny of the third generation for validation. All birds were genotyped with an Illumina Infinium platform HD panel that included 20,541 segregating markers. Two subsets of HD markers, with 377 (ELD-1) or 766 (ELD-2) markers, were selected as ELD panels. The ELD-1 panel was genotyped using KBiosciences KASPar SNP genotyping chemistry, whereas the ELD-2 panel was simulated by adding markers from the HD panel to the ELD-1 panel. The training data set was used for 2 traits: BW at 35 d on both sexes and hen house production (HHP) between wk 28 and 54. Methods Bayes-A, -B, -C and genomic best linear unbiased prediction were used to estimate HD-marker effects. Two scenarios were used: (1) the 160 progeny were ELD-genotyped, and (2) the 160 progeny and their dams (117 birds) were ELD-genotyped. The missing HD genotypes in ELD-genotyped birds were imputed by a Gibbs sampler, capitalizing on linkage within families. In scenario (1), the correlation of GEBV for BW (HHP) of the 160 progeny based on observed HD versus imputed genotypes was greater than 0.94 (0.98) with the ELD-1 panel and greater than 0.97 (0.99) with the ELD-2 panel. In scenario (2), the correlation of GEBV for BW (HHP) was greater than 0.92 (0.96) with the ELD-1 panel and greater than 0.95 (0.98) with the ELD-2 panel. Hence, in a pedigreed population, genomic selection can be implemented by genotyping selection candidates with about 400 ELD markers with less than 6% loss in accuracy. This leads to substantial savings in genotyping costs, with little sacrifice in accuracy.

Key words: genomic selection, broiler, accuracy, low-density panel

2013 Poultry Science 92:1712–1723
http://dx.doi.org/10.3382/ps.2012-02941

INTRODUCTION

High-throughput genotyping techniques make the use of high-density (**HD**)-SNP in animal and plant genetic improvement possible. Meuwissen et al. (2001) were the first to suggest genomic selection as a strategy to employ HD-SNP to predict genomic estimated breeding values (**GEBV**). To implement genomic selection, genotypes for all HD-SNP must be obtained on both training individuals and selection candidates, which may not be cost-effective, especially for breeding programs involving large numbers of selection candidates. To reduce genotyping costs, implementation of low-density (**LD**) marker panels is desirable. A common strategy for creation of LD marker panels is to select a subset of HD markers that is most predictive of breeding values for a particular trait (Habier et al., 2009). A potential problem with this strategy is that the selected subset of HD markers is likely trait-specific, which would still require a large number of markers to be genotyped for breeding programs that select on multiple traits (Habier et al., 2009). For traits that are highly polygenic, it may also be difficult to find a limited number of SNP that provide good predictive ability. As an alternative, Habier et al. (2009) proposed using evenly spaced, low-density (**ELD**)-SNP across the genome, combined with imputation of other SNP genotypes, to estimate GEBV in selection candidates. In this approach, a small subset of HD markers (less than 1,000) is used to genotype selection candidates and genotyping costs can be reduced dramatically. The

Received November 26, 2012.
Accepted March 3, 2013.
[1]Corresponding author: jdekkers@iastate.edu

training population is genotyped with a HD-SNP panel and selection candidates are genotyped with a subset of the HD-SNP, forming an ELD-SNP panel; then, missing HD genotypes in selection candidates are imputed on a within-family basis, using linkage. However, due to imperfect genotype imputation, the accuracy of GEBV prediction is expected to be lower than when selection candidates are genotyped using HD panels. This was illustrated by Habier et al. (2009) using simulation. Habier et al. (2009) assumed that haplotypes of HD markers in the training population are known but they must be inferred in practice, which may introduce errors and reduce the accuracy of imputed genotypes and resulting GEBV. Moreover, in real breeding programs, the number of dams is typically larger than the number of sires. Thus, the genotyping cost could be reduced further if dams are also genotyped with the ELD-marker panel. However, ELD-genotyping of dams will further reduce the accuracy of imputed genotypes and the resulting GEBV of the selection candidates. The extent to which accuracy is reduced for all these situations must be quantified. Huang et al. (2012) analyzed a commercial pig population using different genotyping strategies and found that when offspring and dams were genotyped with 384 SNP and sires with the HD-SNP panel, the accuracy was reduced by about 6.5%, compared with the scenario where all individuals were genotyped with the HD-marker panel. In their study, the ELD marker panels were simulated by extracting markers from the HD-marker panel. In this study, we used real HD- and ELD-SNP panel genotypes in a commercial broiler breeding program for 2 traits to investigate the accuracy and reliability of GEBV with ELD-SNP panels in the scenarios described above.

MATERIALS AND METHODS

Genotype Data

A commercial female broiler line was used in this study. The line was selected for reproductive fitness, live performance (growth, feed efficiency, and breast yield), and welfare in a balanced way. This population has been closed and fully pedigreed for more than 30 generations, the effective population size is managed to achieve less than 1% increase of inbreeding coefficient per generation. From this line, 1,091 birds from 3 consecutive generations, including almost all parents used for breeding in those generations, plus 160 progeny from the third generation were genotyped with a custom-made Illumina Infinium platform SNP panel with 36,455 genome-wide SNP. A total of 20,541 SNP remained for analyses after filtering based on (1) minor allele frequency (**MAF**) per SNP >0.03, (2) missing genotype rate per SNP <0.03, (3) Hardy-Weinberg equilibrium *P*-value per SNP $>10^{-19}$, (4) parent-offspring mismatch rate <0.03, (5) minimum call rate per individual >0.90, and (6) SNP on chromosomes 16 and 25 were not used because of the limited number of SNP on these chromosomes.

Based on the HD-genotype information, an ELD-SNP panel with 384 SNP (ELD-1) was created. The SNP included in the ELD-1 panel were chosen from the HD-SNP panel based on MAF (high MAF desired) and position on the linkage map (evenly spaced). For the latter, linkage maps were developed for each chromosome for HD-SNP with MAF >0.3 using a modified version of Crimap software (Green et al., 1990), following procedures outlined in Groenen et al. (2000). The modified version of Crimap allowed simultaneous analysis of larger numbers of SNP and was kindly provided by X. Liu and M. Grosz of Monsanto Co. (St. Louis, MO). The order of SNP was based on build 2 of the chicken genome (WASHUC2, May 2006), which was the build available at the time the ELD panel was constructed.

The SNP for inclusion in the ELD panel were chosen as follows: (1) the number of SNP to be selected on a chromosome i was calculated as $n_i = n_{ELD} \times \frac{l_i}{l_{genome}}$, where n_{ELD} is the number of ELD SNP expected to be included in the ELD-SNP panel, l_i is the genetic length of chromosome i, and l_{genome} is the total genetic length of the genome; (2) for chromosome i, the ideal MAF for selected SNP is 0.5, and the ideal distance between selected SNP is $d = \frac{l_i}{n_i - 1}$. With these notations, the selected set of n_i SNP was derived by iteratively minimizing the function

$$f = \sum_{j=1}^{n_i} (MAF_j - 0.5)^2 + \sum_{j=1}^{n_j - 1} (S_{j+1} - S_j - d)^2,$$

where S_j is the position of jth SNP on the genetic map. To ensure coverage on the telomeres, where accuracy of imputation is expected to be lower as a result in having fewer flanking markers, the first and last SNP on each chromosome were automatically included in the ELD panel.

The 160 progeny were genotyped again with the ELD-1 panel, using the KBiosciences KASPar SNP genotyping chemistry. Seven ELD SNP were removed from analysis because of poor genotyping quality. To investigate the effect of marker density in ELD-marker panels on the accuracy of GEBV, an additional ELD panel (ELD-2) with twice as many SNP was simulated by adding one SNP to each ELD-marker interval. These SNP were chosen from the HD panel by picking the SNP with the highest MAF from the middle third of each interval between 2 flanking ELD SNP; some SNP were also added at the ends of some chromosomes to increase coverage at the telomeres, resulting in 766 SNP on the ELD-2 panel. Genotypes of the 160 progeny for the ELD-2 panel, thus, consisted of a mixture of the actual ELD-1 genotypes and genotypes from the

Table 1. Estimated chromosome map lengths and comparison to the consensus map, number of SNP in high-density (HD) and 2 equally spaced low-density (ELD) SNP panels and the percentage of solved genotype orders for the 1,091 HD-genotyped birds by the rule-based method

Chromosome	Linkage map length (cM)		High-density panel	Low-density panel		Ordered genotypes (%)
	Current	Consensus[1]		ELD-1	ELD-2	
1	403	484	4,152	57	113	99.6
2	287	312	3,365	41	82	99.6
3	262	269	2,336	38	77	99.6
4	183	202	1,754	23	46	99.6
5	146	158	1,285	21	42	99.6
6	93	110	772	13	26	99.6
7	93	113	893	13	27	99.6
8	85	92	612	12	25	99.5
9	85	89	508	12	25	99.4
10	75	90	436	11	22	99.5
11	57	69	401	8	17	99.3
12	68	74	374	10	21	99.3
13	62	59	343	9	19	98.9
14	58	68	293	8	17	98.9
15	51	55	258	7	15	99.5
16	—	—	—	—	—	—
17	50	53	238	7	15	98.6
18	53	52	232	8	16	99.3
19	51	53	249	6	13	98.9
20	51	52	260	7	15	99.4
21	47	53	163	6	11	99.2
22	49	59	69	6	12	97.0
23	53	45	120	7	15	99.2
24	55	48	155	7	14	99.3
25	15	57	—	—	—	—
26	49	47	116	6	13	98.9
27	27	53	91	3	7	98.9
28	50	52	103	6	13	98.1
Z	213	232	963	25	48	99.8
Total	2,771	3,100	20,541	377	766	99.2

[1]Consensus map lengths based on Groenen et al. (2000).

HD panel for the SNP that were added. More information about the distribution of HD- and ELD-SNP across the genome is in Table 1.

Phenotypic Data and Estimation of HD-SNP Effects

Two traits were investigated in this study: BW at 35 d for both males and females and hen house production (**HHP**), which was recorded between wk 28 and 54 of the life of a broiler hen (Wolc et al., 2010). The number of phenotypic records in the training data set was 82,369 and 3,596 for BW and HHP, respectively, including 1,091 and 789 phenotypes on the 1,091 genotyped birds. The remaining phenotypes were on ungenotyped progeny of the 1,091 individuals (excluding the 160 progeny of the last generation), including some sibs of the 160 progeny that were included in the training data set. All phenotypes were preadjusted for the fixed effects of contemporary group and common environment due to dams based on estimates from a multi-trait pedigree-based animal model best linear unbiased prediction (**BLUP**) procedure. For full- or half-sib families that were not genotyped but had at least one genotyped parent, family means were constructed and the family was assigned the average SNP genotypes of parents. For families with only one parent genotyped, the genotype of the other parent was set to twice the population allele frequency of the respective SNP.

The general statistical model used for training analyses was

$$y_i = \mu + \sum_{j=1}^{m} g_{ij}\beta_j + \frac{e_i}{\sqrt{w_i}},$$

where y_i is the preadjusted individual phenotype for genotyped animals and a preadjusted family mean phenotype for ungenotyped individuals that have at least one parent genotyped; μ is the overall mean; m is the number of SNP; β_j is the allele substitution effect for SNP j; g_{ij} is the genotype value of SNP j (coded as allele dosage: 0/1/2 for genotyped individuals or average genotype of the parents for families); e_i is the residual effect, which was assumed to follow a normal distribution with mean zero and variance σ^2; and w_i is the weight for record i to account for heterogenous residual variance of family means. Weights were derived using the method of Garrick et al. (2009), using estimates of heritability obtained from the multi-trait pedigree-based animal model of 0.27 for BW and 0.16 for HHP.

The number of own phenotypes and family means was 1,634 for BW and 1,346 for HHP. Genomic prediction methods Bayes-A, Bayes-B, and Bayes-C with a $\pi = 0.99$ prior probability of a SNP having zero effect and genomic BLUP (**GBLUP**; implemented using Bayes-C with $\pi = 0$; Fernando and Garrick, 2011) were used for training, with 50,000 rounds following 30,000 rounds for burn-in.

Imputation of Missing HD Genotypes

The rapid and simple rule-based method of Habier et al. (2010) was used to infer haplotypes for the HD-genotyped parents of ELD-genotyped individuals. The Gibbs sampler with overlapping blocks, as described in Habier et al. (2009, 2010), was then employed to estimate probabilities of allele segregation indicators at the ELD-SNP for the ELD-genotyped individuals, utilizing the haplotype information at ELD-SNP in the HD-genotyped individuals. Haplotypes of the HD-genotyped individuals and segregation probabilities at ELD-SNP for the ELD-genotyped individuals were then used to estimate genotypes of the missing HD-SNP for the ELD-genotyped individuals (Habier et al., 2010). The imputed HD genotype was computed as the sum of probabilities for the paternal and maternal alleles transmitted to the progeny, such that imputed HD genotypes were on the same allele dosage scale as the observed HD genotypes. With this approach, imputed genotypes are not assigned to 1 of the 3 discrete genotype categories (0, 1, 2) but are given noninteger values between 0 and 2 based on the sum of the 2 allele transmission probabilities. To evaluate the impact of assigning discrete genotypes, imputed genotype values were also rounded to their nearest integer for the ELD-1 panel, which will be referred to as ELD-1R.

The accuracy of imputation was quantified based on the correlation and MS error (**MSE**) of imputed versus observed HD-genotype values of the 160 progeny. Correlations and MSE were computed both per individual across SNP and per SNP across individuals. To compute MSE, the genotypes for each HD SNP were divided into 4 categories according to the genotypes of their parents because the latter affect the expected MSE: (1) both parents are heterozygous; (2) the sire is heterozygous and the dam homozygous; (3) the sire is homozygous and the dam heterozygous; and (4) both parents are homozygous. The MSE of imputed genotypes for category k was calculated as

$$\frac{1}{160 \times N_k} \sum_{i=1}^{160} \sum_{j=1}^{N_k} (g_{ij} - \widehat{g}_{ij})^2,$$

where N_k is the number of SNP in category k and g_{ij} and $\widehat{g}_{ij}$ are the observed and imputed genotypes for individual i at SNP j. The MSE by SNP was computed as

$$\frac{1}{160 \times N} \sum_{i=1}^{160} \sum_{j=1}^{N} (g_{ij} - \widehat{g}_{ij})^2,$$

where N is the total number of imputed HD-SNP. The MSE by SNP were then standardized by dividing by the expectation of MSE if average parental genotype values were assigned as the imputed genotypes, which was calculated as

$$\frac{0.5 \times N_{1j} + 0.25 \times N_{2j} + 0.25 \times N_{3j} + 0 \times N_{4j}}{N_{1j} + N_{2j} + N_{3j} + N_{4j}},$$

where N_{kj} $(k = 1, 2, 3, 4)$ are the numbers of genotypes in each category for SNP j. The MSE of imputed genotypes for individual i was calculated as $\frac{1}{N} \sum_{j=1}^{N} (g_{ij} - \widehat{g}_{ij})^2$, where N is the total number of imputed HD-SNP.

Validation Data

The 160 genotyped progeny of the final generation of training population were used for validation data. These birds were the offspring of 117 dams and 68 sires. On average, each sire had 2.35 progeny. To evaluate the impact of ELD genotyping on the accuracy of predicted GEBV of ELD-genotyped progeny, 2 scenarios were created: (1) the 160 progeny were ELD genotyped and (2) the 160 progeny and their dams were ELD genotyped. The ELD genotypes of dams were simulated by keeping HD genotypes only for the ELD SNP. For each scenario, the own phenotypes of the 160 progeny were used as validation data. Based on the estimated HD-SNP effects from a given training analysis method and using observed or imputed HD genotypes, the predicted GEBV of the 160 progeny were calculated as $GEBV_i = \sum_{j=1}^{n_m} g_{ij} \widehat{\beta}_j$, where g_{ij} is the observed or imputed HD genotype value of progeny i at SNP j, $\widehat{\beta}_j$ is the estimated HD-SNP effect at SNP j, and n_m is the number of SNP. Additionally, for comparison, the EBV of the 160 progeny were predicted by traditional pedigree-based BLUP for both traits, in which the 160 animals were assumed to have no phenotypes.

Two criteria were used to evaluate the accuracy of GEBV based on imputed HD genotypes: (1) loss in accuracy of GEBV from using imputed HD genotypes was calculated as $(1 - c) \times 100\%$, where c is the correlation between predicted GEBV obtained when using imputed versus observed HD genotypes for the validation individuals, using the GEBV based on observed HD genotypes as the gold standard; and (2) the impact of using imputed versus observed HD genotypes on the accuracy of the GEBV computed as the correlation of GEBV of the validation individuals with their preadjusted phenotypic values divided by the squared root of the heritability of the trait.

RESULTS

Linkage Maps

Linkage maps were developed using SNP with MAF > 0.3. Map lengths are in Table 1, with lengths from the consensus map (Groenen et al., 2000) as a reference. For most chromosomes, map lengths obtained were smaller than consensus map lengths, possibly as a result of few genotyping errors. Generally, the order of SNP in the linkage map generated from the data of the current study was similar to that obtained from sequence (WASHUC2, May 2006) with a few exceptions, as illustrated in Figure 1 for chromosome 1, where a segment that was at the beginning of the chromosome based on sequence was mapped toward the center of the chromosome in the linkage map. Whereas the original linkage maps were created in reference to the build available at the time (WASHUC2, May 2006), the misplacement of this region is still present in the most recent build (Gallus_gallus-4.0, Nov., 2011). At short distances, the order of SNP obtained from the linkage map was not always consistent with that in the sequence. Generally, for most chromosomes, map position increased almost linearly with physical map position but there were local differences in recombination rates per unit of distance, similar to the observation by Groenen et al. (2009). Also, recombination rate per unit of distance generally increased at the end of the chromosomes, similar to Groenen et al. (2009).

Inference of Haplotypes

In the first iteration of the rule-based method to determine haplotypes, more than 76.6% of SNP genotypes could be ordered for the 1,091 HD-genotyped individuals on the autosomes and 98.0% on chromosome Z. After 6 to 8 iterations, no additional parental origins could be identified and more than 99% of genotypes across the genome were ordered (Table 1). This percentage was greater for the larger chromosomes and chromosome Z. For chromosome Z, females are homozygous and, therefore, SNP are ordered automatically. Chromosome 22 had the smallest percentage of genotypes ordered (97.0%) because it only had 69 HD-SNP, the least of any chromosome. The proportion of ordered genotypes for each chromosome was about 1% less when the dams of the 160 progeny were assumed to be ELD-genotyped and, therefore, not included in the haplotype analysis.

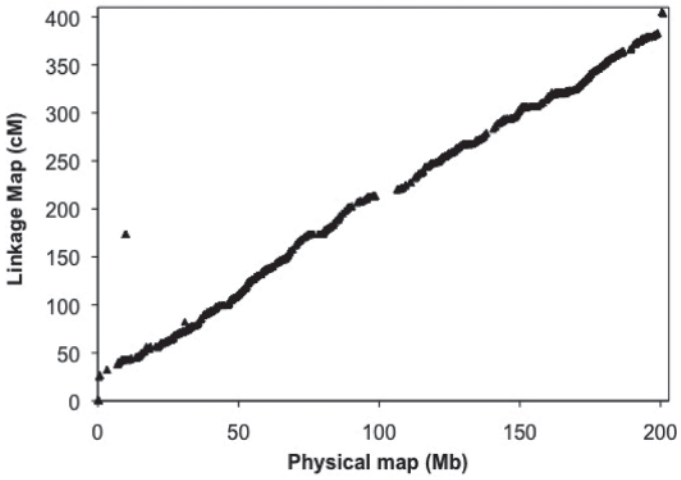

Figure 1. Linkage versus physical map for chromosome 1.

Accuracy of Imputed Genotypes

The average MSE of imputed genotypes was substantially lower than expected based on random assignment for all ELD genotyping scenarios (Table 2). The average MSE by SNP was greatest when both parents were heterozygous and lowest when both were homozygous. Some SNP, however, had greater MSE than expected in each category (Figure 2). Overall, the number of SNP with MSE greater than expected was small. The mean MSE was approximately halved when the size of the LD panel was doubled (ELD-1 vs. ELD-2 in Table 2). Availability of only ELD genotypes on dams increased the average MSE by 0.01 to 0.03, depending on the parental genotypes and the ELD marker panel. When standardizing the MSE for each category by its expectation and averaging across categories by SNP, the mean standardized MSE was 0.21 for ELD-1 when dams were HD-genotyped (Table 3). This is relative to a value of 1 when genotypes would be assigned at random based on parental genotypes. Thus, on average, imputation captured almost 80% of the Mendelian sampling of SNP alleles from parents to progeny. When the ELD panel was doubled, this increased to almost 90%. These percentages decreased to almost 70% for ELD-1 and to 80% for ELD-2 when dams were also ELD-genotyped. The above results are for SNP on autosomes; SNP on the Z chromosome were imputed with greater accuracy.

The correlations of imputed and observed HD genotypes were computed both by SNP and by individual. Average correlations for the different scenarios are in Table 3 and distributions of these correlations by SNP and by individual are in Figures 3 and 4. The mean correlation by SNP was 0.94 with ELD-1 but increased to 0.97 with ELD-2. Note that some SNP had correlations less than 0.80, likely because of genotyping errors, misplacement, or both. The mean correlation by SNP dropped when dams were ELD-genotyped but was still above 0.91 even for the worst-case scenario considered. Average correlations between imputed and observed HD genotypes by individual were higher than average correlations by SNP (Table 3). All individuals had correlations greater than 0.95 for ELD-1 and dams HD-genotyped and most individuals had correlations greater than 0.97, although some SNP were imputed less accurately (Figure 4). The lowest correlation by individual dropped to 0.93 when dams were also ELD-genotyped.

Table 2. Average MS errors of imputed versus observed high-density (HD) genotypes on autosomes by SNP for different categories in the 160 progeny, with imputation based on the evenly spaced, low-density (ELD) panels

		DamHD[1]		DamLD[2]		
Parental genotype		ELD-1	ELD-2	ELD-1	ELD-2	Expectation
Sire	Dam					
Het[3]	Het	0.09	0.05	0.12	0.06	0.50
Het	Hom[4]	0.06	0.03	0.07	0.04	0.25
Hom	Het	0.05	0.03	0.07	0.04	0.25
Hom	Hom	0.00	0.00	0.01	0.01	0.00

[1]Dams of the 160 progeny were genotyped with the HD panel.

[2]Dams of the 160 progeny were genotyped with the low-density (LD) panel.

[3]The parent's genotype is heterozygous.

[4]The parent's genotype is homozygous.

Accuracy of GEBV Prediction in Validation Data

Using the GEBV obtained from observed HD genotypes on validation individuals as the gold standard, the percentage loss of accuracy of GEBV with imputed HD genotypes for the ELD-1 panel and dams HD-genotyped was less than 6% for BW and less than 3% for HHP (Table 4). The percentage loss differed slightly between methods used for training. These differences are likely the result of the variability that was observed in the accuracy of imputation by SNP and differential

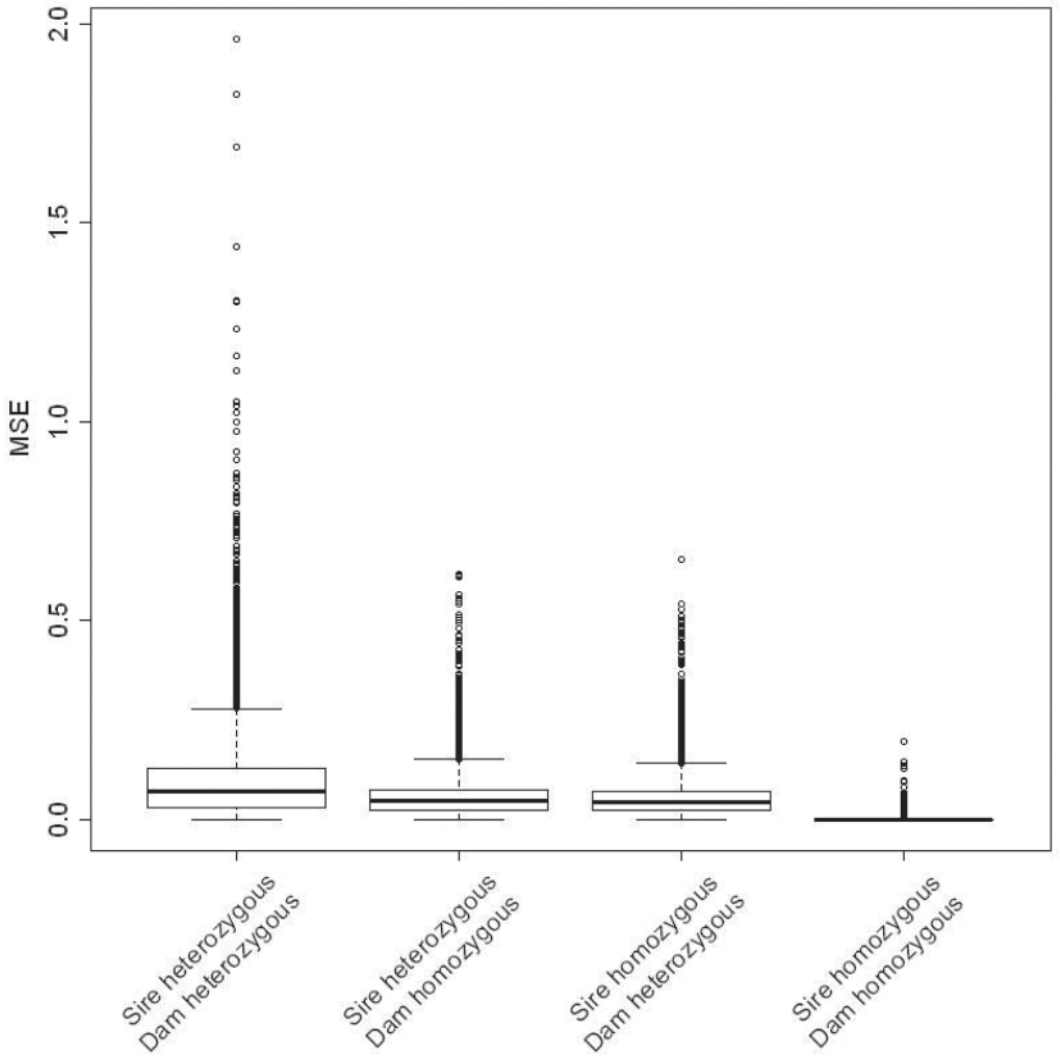

Figure 2. Boxplot of MS errors (MSE) of imputed versus observed high-density (HD) genotypes on autosomes by SNP, categorized by genotypes (homozygous/heterozygous) of the parents for 160 progeny genotyped with the evenly spaced, low-density (ELD)-1 marker panel.

Table 3. Average MS errors (MSE) and correlations between imputed and observed high-density (HD) genotypes on autosomes by individual and by SNP for 160 progeny, using evenly spaced, low-density (ELD) panels, depending on whether the dams are HD or low-density (LD) genotyped

	DamHD[1]		DamLD[2]	
Item	ELD-1	ELD-2	ELD-1	ELD-2
Standardized MSE by SNP	0.21	0.11	0.32	0.17
MSE by individual	0.03	0.02	0.05	0.03
Correlation by SNP	0.94	0.97	0.91	0.95
Correlation by individual	0.97	0.99	0.96	0.98

[1]Dams of the 160 progeny were genotyped with the HD panel.

[2]Dams of the 160 progeny were genotyped with the low-density (LD) panel.

weighting of SNP when computing GEBV, depending on the estimate of the SNP effect. Because SNP effects were estimated from HD-genotyped individuals, estimates are expected to be unrelated to the accuracy of imputation of each SNP. Thus, the differences observed between methods are expected to be random rather than systematic, i.e., Bayes-B likely showed greater loss for BW than GBLUP because the SNP that received larger estimated effects with Bayes-B on average had lower imputation accuracy, but this relationship is expected to occur by chance. Percent loss in accuracy was lower for HHP than for BW.

Doubling the size of the ELD panel more than halved the percent loss in accuracy of GEBV (Table 4). The

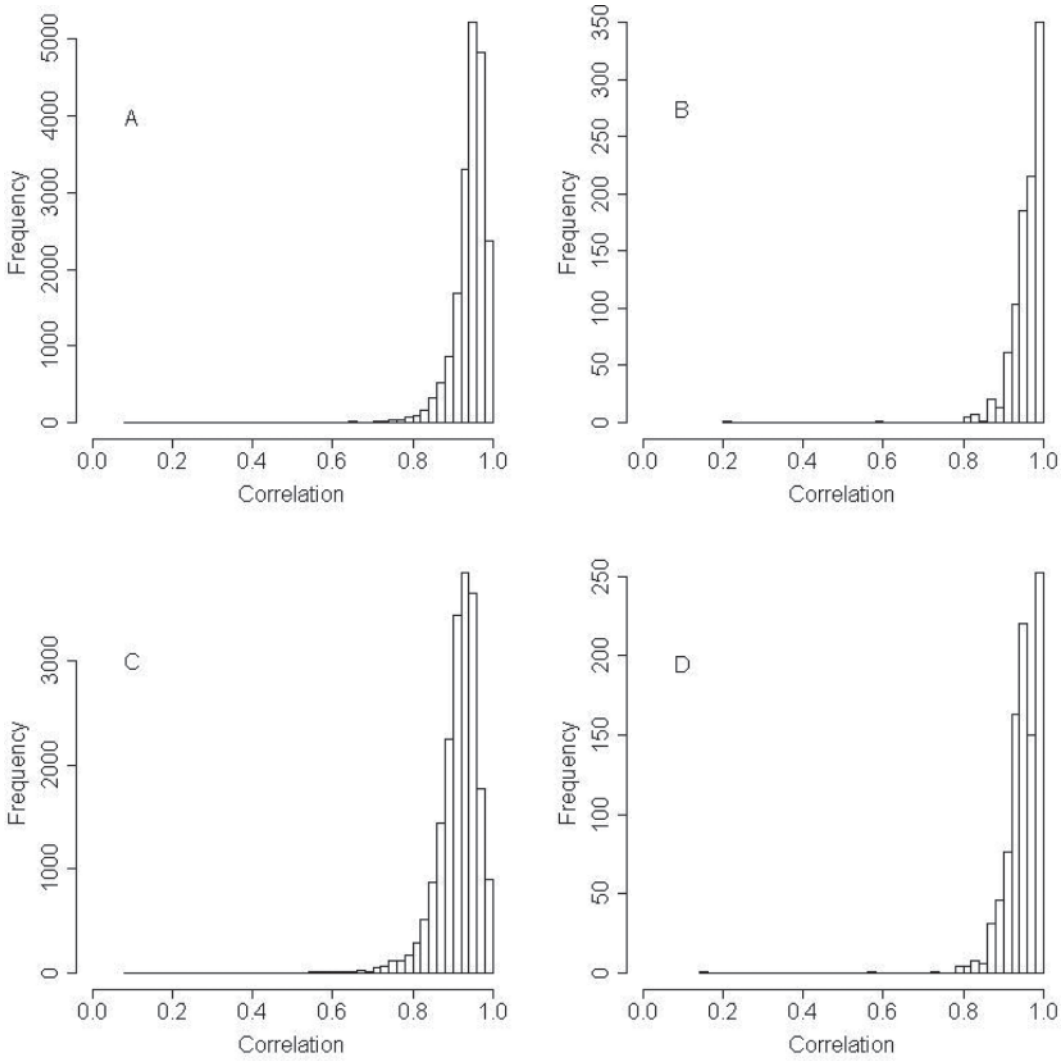

Figure 3. Histogram of correlations of observed and estimated high-density (HD) genotypes by SNP for the 160 progeny genotyped with evenly spaced, low-density (ELD) panels. (A) Autosomes with ELD-1 panel; (B) chromosome Z with ELD-1 panel; (C) autosomes with ELD-2 panel; (D) chromosome Z with ELD-2 panel.

Table 4. Percentage (%) loss in accuracy[1] of genomic estimated breeding values (GEBV) in the progeny validation generation, based on imputed versus observed high-density (HD) SNP genotypes

		DamHD[2]			DamLD[3]		
Trait	Method	ELD-1	ELD-2	ELD-1R	ELD-1	ELD-2	ELD-1R
BW	Bayes-A	3.6	1.6	4.6	5.2	2.7	6.8
	Bayes-B	5.9	3.0	8.3	7.9	4.5	10.0
	Bayes-C	5.3	2.6	7.4	7.3	4.2	9.2
	GBLUP[4]	3.4	1.5	4.5	5.2	2.6	6.7
HHP[5]	Bayes-A	2.3	1.0	2.7	3.6	1.7	4.1
	Bayes-B	2.0	1.0	2.7	3.0	1.6	3.9
	Bayes-C	1.7	0.8	2.2	2.7	1.4	3.3
	GBLUP	1.6	0.6	1.9	2.6	1.3	3.0

[1]The loss was calculated as $(1 - c) \times 100\%$, where c is the correlation between predicted GEBV from imputed versus observed HD genotypes for the 160 progeny.

[2]Dams of the 160 progeny were genotyped with the HD panel. ELD = evenly spaced, low density.

[3]Dams of the 160 progeny were genotyped with the low-density (LD) panel.

[4]GBLUP = genomic BLUP.

[5]HHP = hen house production.

ELD-genotyping of dams increased losses by a factor of 1.1 to 2.0 for BW and by 0.5 to 2.1 for HHP. The regression coefficient of GEBV from imputed genotypes on GEBV from observed HD genotypes was around 1.00 for all scenarios (results not shown), suggesting that imputation did not introduce any biases. Table 4

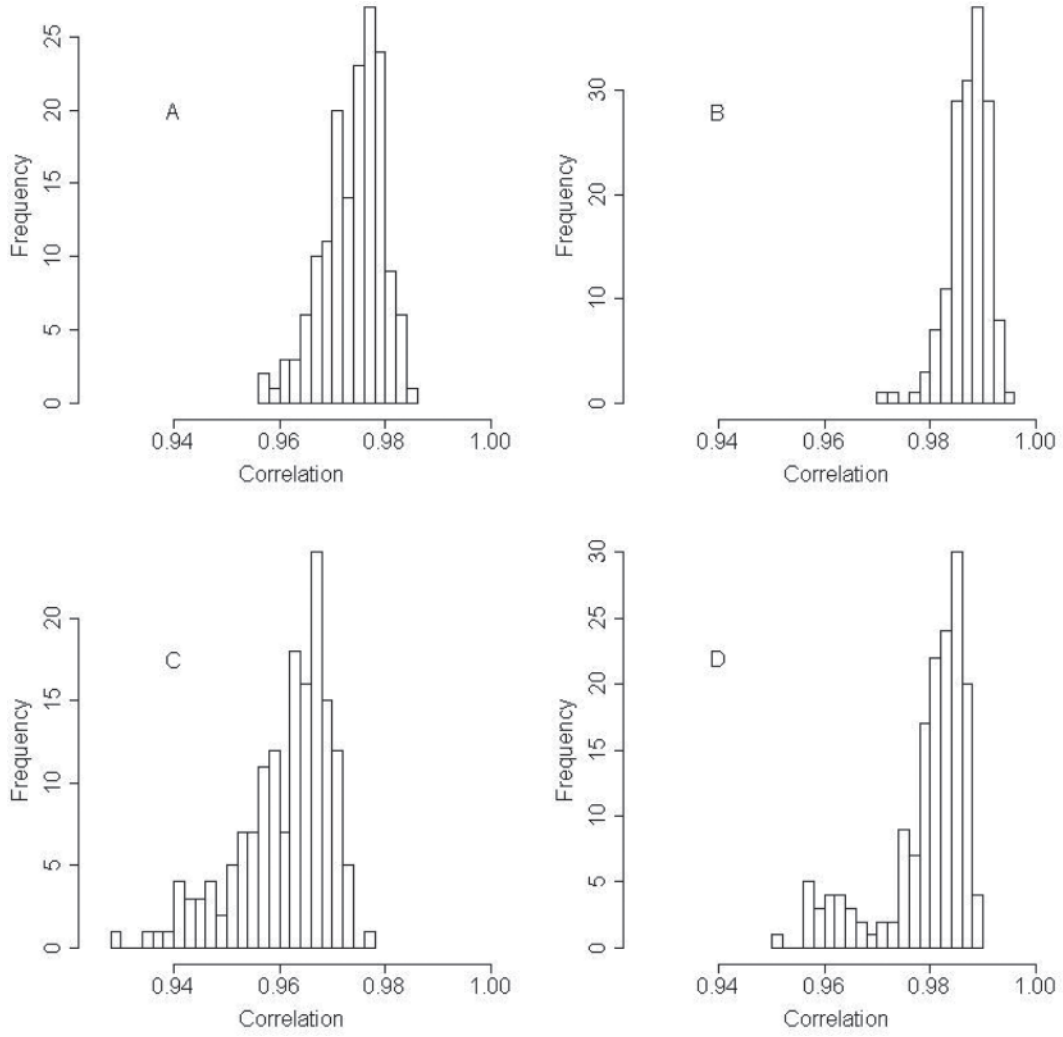

Figure 4. Histogram of correlations of observed and estimated high-density (HD) genotypes by individual for the 160 progeny. The 160 progeny were genotyped with evenly spaced, low-density (ELD)-1 (A) and ELD-2 (B); the 160 progeny and their mothers were genotyped with ELD-1 (C) and ELD-2 (D).

Table 5. Accuracy[1] of genomic estimated breeding values (GEBV) in the progeny validation generation based on observed or imputed genotypes

			DamHD[3]			DamLD[4]		
Trait	Method	ProgHD[2]	ELD-1	ELD-2	ELD-1R	ELD-1	ELD-2	ELD-1R
BW	PBLUP[5]	0.42	0.42	0.42	0.42	0.42	0.42	0.42
	Bayes-A	0.67	0.62	0.62	0.64	0.67	0.63	0.68
	Bayes-B	0.77	0.66	0.68	0.64	0.66	0.67	0.67
	Bayes-C	0.76	0.66	0.68	0.64	0.67	0.67	0.68
	GBLUP[5]	0.66	0.62	0.62	0.64	0.67	0.63	0.68
HHP[6]	PBLUP	0.43	0.43	0.43	0.43	0.43	0.43	0.43
	Bayes-A	0.72	0.72	0.71	0.69	0.75	0.73	0.75
	Bayes-B	0.79	0.77	0.76	0.72	0.78	0.78	0.76
	Bayes-C	0.74	0.72	0.71	0.69	0.74	0.73	0.73
	GBLUP	0.70	0.70	0.69	0.68	0.71	0.70	0.71

[1]Accuracy computed as the correlation between GEBV and pre-adjusted phenotype divided by the square root of heritability, based on 160 progeny with own phenotype for BW and 68 progeny with phenotype for hen house production (HHP).

[2]The progeny were genotyped with high-density (HD) panel.

[3]Dams of the 160 progeny were genotyped with the HD panel. ELD = evenly spaced, low density.

[4]Dams of the 160 progeny were genotyped with the low-density (LD) panel.

[5]PBLUP = pedigree-based BLUP; GBLUP = genomic BLUP.

also showed results for imputed genotypes from panel ELD-1 when genotypes were called as homozygous versus heterozygous based on rounding. This increased losses by 1.0 to 2.4 percentage points for BW and by 0.3 to 0.9 percentage points for HHP.

Table 5 showed the observed accuracy of GEBV based on the correlation between GEBV and phenotypes for the validation individuals, divided by the square root of heritability. Pedigree-based EBV were also included for comparison but were, obviously, not affected by SNP genotyping. Observed accuracies of GEBV using HD-genotyping were substantially higher than accuracies of EBV based on pedigree BLUP for both traits. Differences between methods were limited but slightly higher for Bayes-B and -C than for Bayes-A and GBLUP. The ELD-genotyping of validation individuals with ELD-1 reduced observed accuracies by 5 to 11% for BW and by less than 5% for HHP. Doubling the size of the ELD panel or also ELD genotyping dams had limited effects on observed accuracies. Rounding imputed genotypes slightly reduced observed accuracies when dams were HD genotyped but slightly increased accuracies when dams were ELD genotyped.

Accuracies of the parental average GEBV (Table 6), computed for each validation individual as the average of the GEBV of the sire and dam, were much lower than accuracies of GEBV based on own genotypes of validation individuals.

For BW, the accuracy of GEBV was 1 to 2% lower when the dams of validation individuals were ELD-genotyped than when they were HD-genotyped (Table 5); for HHP, the accuracy with dams ELD-genotyped was 2 to 4% lower than when dams were HD-genotyped. Use of the different ELD genotypes had little effect on the accuracy for BW, but for HHP, the accuracy from panel ELD-1R was slightly higher than accuracies from

Table 6. Accuracy[1] of parental average genomic estimated breeding values in the progeny validation generation based on observed or imputed genotypes

			DamLD[3]		
Trait	Method	ProgHD[2]	ELD-1	ELD-2	ELD-1R
BW	PBLUP[4]	0.42	0.42	0.42	0.42
	Bayes-A	0.48	0.47	0.47	0.47
	Bayes-B	0.50	0.48	0.48	0.48
	Bayes-C	0.49	0.48	0.48	0.48
	GBLUP[4]	0.49	0.47	0.47	0.47
HHP	PBLUP	0.42	0.42	0.42	0.42
	Bayes-A	0.53	0.51	0.49	0.52
	Bayes-B	0.57	0.54	0.55	0.56
	Bayes-C	0.55	0.52	0.52	0.54
	GBLUP	0.54	0.51	0.50	0.52

[1]Accuracy computed as the correlation between GEBV and pre-adjusted phenotype divided by the square root of heritability, based on 160 progeny with own phenotype for BW and 68 progeny with phenotype for hen house production (HHP).

[2]The progeny were genotyped with the high-density (HD) panel.

[3]Dams of the 160 progeny were genotyped with the low-density (LD) panel. ELD = evenly spaced, low density.

[4]PBLUP = pedigree-based BLUP; GBLUP = genomic BLUP.

the other 2 panels and the accuracy from panel ELD-1 was slightly higher than the accuracy from panel ELD-2.

DISCUSSION

Genotype imputation is increasingly used in studies and applications that use DNA marker genotype data for prediction of breeding values. The purpose of genotype imputation varies from filling in occasional missing genotypes that result from genotyping errors or cases where some genotypes cannot be called, to large-scale imputation to allow combining of data on individuals that were genotyped with different sets of markers, in particular when selection candidates are imputed using a much smaller number of markers.

Information used for imputation can consist of 2 sources: (1) linkage disequilibrium between markers that exists across the population and (2) co-segregation of linked markers from parents to progeny. The use of linkage disequilibrium requires the presence of extensive linkage disequilibrium of the markers that are genotyped with the markers that are not genotyped on the individuals that are to be imputed, such that haplotypes can be identified and used to infer the missing genotypes. This requires reasonably high density of genotyped markers in the to-be-imputed individual such that sufficient linkage disequilibrium exists. This is the information that is used in programs such as phase and Fastphase (Scheet and Stephens, 2006) and the nonpedigree option of Beagle (Browning and Browning, 2007).

Imputation by co-segregation can be accomplished even with very low-density genotyping on the to-be-imputed individuals because it relies on the inheritance of nonrecombined segments of chromosomes from parents to progeny, which can be accomplished with a SNP every 10 cM (Habier et al., 2009). By using linkage or co-segregation information, the missing HD genotypes can be imputed with high accuracy with ELD marker panels and loss of accuracy is limited, as shown in the current study and by Habier et al. (2009). Imputation by co-segregation does require knowledge of pedigree and of the phase of SNP genotypes in the parents. For maximum accuracy of imputation, parents should be genotyped using the HD panel but genotyping parents or even grandparents with the ELD panel and more distant ancestors with the HD panel can be used for imputation also, but with additional losses in accuracy (Habier et al., 2009). Because it requires lower density, imputation by co-segregation is a more attractive alternative to imputation based on linkage disequilibrium and, depending on the cost of low-density SNP genotyping, allows genomic selection to be implemented in species with large numbers of selection candidates and low value individuals, such as pigs and poultry.

Several programs exist that can be use for genotype imputation, including AlphaImpute (Hickey et al., 2011), Beagle (Browning and Browning, 2007, 2009), and a Gibbs sampler (Habier et al., 2009, 2010). Wolc et al. (2011) compared different imputation methods and found that the Gibbs sampler gives slightly higher accuracy with ELD panels of the size used in the present study, so it was adopted for our data.

The Gibbs sampler for imputation (Habier et al., 2009, 2010) requires the genotypes of HD-genotyped ancestors to be phased. Methods used for imputation using linkage disequilibrium can also be used for phasing HD genotypes on unrelated individuals. However, when several generations of individuals are HD-genotyped, as was the case here, simple rule-based methods based on Mendelian segregation and co-segregation can be used for phasing (Habier et al., 2010). This method was implemented here and found to be a very efficient approach for haplotype inference and was able to solve the phase of most genotypes (Table 1). The accuracy of haplotype inference cannot be estimated in the current study because the true haplotypes are unknown in real data. However, the proportion of loci that were homozygous for HD-genotyped individuals was high (0.45 $\pm$ 0.21) and for these loci, the ordered genotypes can be determined easily, which supplies information for the parents or offspring of the individuals. Moreover, the hypothesis that no recombination occurs in smaller chromosome segments has been verified in haplotype inference in many other studies (Zhang et al., 2005; Kong et al., 2008; Meuwissen and Goddard, 2010). With the high proportion of homozygous loci, the simple rule-based method assuming zero recombination in small chromosome segments can yield high accuracy for haplotype inference. Incorrect phasing would reduce the accuracy of imputation. Because the loss of accuracy from imputation was small and of a similar magnitude as that observed in the simulation studies of Habier et al. (2009), in which the true phase was used, incorrect phasing is expected to have contributed little to the loss in accuracy from imputation in this study.

In this study, most SNP were imputed with high accuracy but for some SNP the imputed genotypes were not consistent with observed HD genotypes (Figure 2). For example, the correlation of imputed versus observed HD genotypes for one SNP was only 0.08. For this SNP, 38 of the 160 imputed genotypes were different from observed. For 8 progeny, the observed HD genotype at this SNP was not consistent with the genotypes of their parents, likely due to genotyping errors. The other 30 erroneous genotypes were likely caused by the genotypes of one or both parents being heterozygous, which can lead to imputation errors.

The mean error rate of genotype imputation for validation individuals, defined as the mean of the absolute difference of observed and imputed genotypes (Zhang and Druet, 2010), was 0.32 for panel ELD-1, which included only 377 markers. This compares to imputation error rates of 0.26 obtained in Jersey cattle (Weigel et al., 2010) and 0.20 in Holstein dairy cattle (Zhang and

Druet, 2010). The error rate was 0.15 for panel ELD-2 of 766 markers, compared with 0.20 for Jersey (Weigel et al., 2010) and 0.13 for Holstein (Zhang and Druet, 2010). Differences in error rates with those of Weigel et al. (2010) and Zhang and Druet (2010) may due to several reasons: (1) the number of HD markers in dairy cattle was almost twice that of our study; (2) the accuracy of marker orders based on sequence may be different between the 2 species; and (3) family structures were different. Although the imputed genotypes were not very good for some SNP, the correlations of imputed genotypes versus observed HD genotypes by individual were very high, which ensures limited loss in accuracy of predicted GEBV from use of the ELD approach.

The use of imputation by co-segregation from low-density SNP panels (less than 400) in genomic selection was evaluated by simulation by Habier et al. (2009). The purpose of the current study was to validate these results using real data from a broiler breeding program. An additional feature of this study was that the LD genotypes were not simulated by dropping data on SNP from HD-genotyped individuals, but individuals were genotyped for the ELD-1 panel using a platform specific to genotyping limited numbers of SNP on a cost-effective basis. This adds SNP-specific quality differences that tend to exist between genotyping platforms.

The percentage loss in accuracy of GEBV in validation when based on imputed versus observed HD genotypes, was not consistent with the reduction in accuracy of GEBV when evaluated based on the correlation between GEBV and adjusted phenotype. This may be the result of random sampling because of the small validation data set that was used. For example, opposite to expectation, ELD- versus HD-genotyping of dams resulted in an increase in accuracy of GEBV based on the correlation between GEBV and phenotype for HHP.

Increasing the density of the ELD-marker panel can reduce the loss in accuracy of predicted GEBV but increases the genotyping cost. In breeding practice, a balance must be found between genotyping cost and genetic gain (Huang et al., 2012).

Accuracies of the parental average GEBV were much lower than accuracies of GEBV based on own genotypes of validation individuals, which means that individual genotypes captured more information than parental average for estimation of GEBV, the difference being due to the Mendelian sampling term.

Genomic selection using ELD-SNP panels is desirable because of the lower genotyping cost for selection candidates, especially for breeding programs in which the number of selection candidates is large (e.g., in poultry and pigs). Use of an ELD-SNP panel is more attractive than using trait-specific LD marker panels because of its neutrality across traits and populations and its independence of the number of QTL and methods used to estimate marker effects in training data (Habier et al., 2009). In addition, ELD-genotyped individuals can be used for training after the HD genotypes have been imputed. Thus, in pedigreed populations, genomic selection can be implemented by genotyping the selection candidates for as few as 400 SNP across the genome, with less than 6% loss in accuracy. Use of an ELD panel may reduce genotyping costs by an order of magnitude or more, thus facilitating the application of genomic selection in populations with large numbers of selection candidates.

ACKNOWLEDGMENTS

The authors thank the funding support from Aviagen Ltd. (Newbridge, Midlothian, EH28 8SZ, United Kingdom) and Sustainable Animal Breeding (SABRE) Demonstration Funds of EU (European Union) Framework 7 (Biosciences KTN, The Roslin Institute, Easter Bush, UK).

REFERENCES

Browning, B. L., and S. R. Browning. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84:210–223.

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.

Fernando, R. L., and D. J. Garrick. 2011. GenSel—User manual for a portfolio of genomic selection related analyses. Accessed Sep. 2011. http://taurus.ansci.iastate.edu/.

Garrick, D. J., J. F. Taylor, and R. L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:55.

Green, P., K. Falls, and S. Crooks. 1990. Documentation for CRI-MAP, version 2.4. Washington School of Medicine, St. Louis, MO.

Groenen, M. A. M., H. H. Cheng, N. Bumstead, B. F. Benkel, W. E. Briles, T. Burke, D. W. Burt, L. B. Crittenden, J. Dodgson, J. Hillel, S. Lamont, A. P. de Leon, M. Soller, H. Takahashi, and A. Vignal. 2000. A consensus linkage map of the chicken genome. Genome Res. 10:137–147.

Groenen, M. A. M., P. Wahlberg, M. Foglio, H. H. Cheng, H. Megens, R. P. M. A. Crooijmans, F. Besnier, M. Lathrop, W. M. Muir, G. K. Wong, I. Gut, and L. Andersson. 2009. A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res. 19:510–519.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2009. Genomic selection using low-density marker panels. Genetics 182:343–353.

Habier, D., R. L. Fernando, and D. J. Garrick. 2010. A combined strategy to infer high-density SNP haplotypes in large pedigrees. Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany, CD–ROM communication 0915. http://www.kongressband.de/wcgalp2010/.

Hickey, J. M., B. P. Kinghorn, B. Tier, J. F. Wilson, N. Dunstan, and J. H. van der Werf. 2011. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet. Sel. Evol. 43:12.

Huang, Y., J. M. Hickey, M. A. Cleveland, and C. Maltecca. 2012. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genet. Sel. Evol. 44:25.

Kong, A., G. Masson, M. L. Frigge, A. Gylfason, P. Zusmanovich, G. Thorleifsson, P. I. Olason, A. Ingason, S. Steinberg, T. Rafnar, P. Sulem, M. Mouy, F. Jonsson, U. Thorsteinsdottir, D. F. Gudbjartsson, H. Stefansson, and K. Stefansson. 2008. Detection of sharing by descent, long-range phasing and haplotype imputation. Nat. Genet. 40:1068–1075.

Meuwissen, T. H. E., and M. E. Goddard. 2010. The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data. Genetics 185:1441–1449.

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.

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.

Weigel, K. A., C. P. Van Tassell, J. R. O'Connell, P. M. VanRaden, and G. R. Wiggans. 2010. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. J. Dairy Sci. 93:2229–2238.

Wolc, A., J. M. Hickey, M. Sargolzaei, J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, D. Habier, R. L. Fernando, D. J. Garrick, C. Wang, and J. C. M. Dekkers. 2011. Comparison of the accuracy of genotype imputation using different methods. Page 76 in Proc. 7th Eur. Symp. Poult. Genet., Peebles, UK. http://www.roslin.ed.ac.uk/7espg/assets/7espg-edited-proceedings.pdf. (Abstr.)

Wolc, A., I. M. S. White, W. G. Hill, and V. E. Olori. 2010. Inheritance of hatchability in broiler chickens and its relationship to egg quality traits. Poult. Sci. 89:2334–2340.

Zhang, K., F. Sun, and H. Zhao. 2005. HAPLORE: A program for haplotype reconstruction in general pedigrees without recombination. Bioinformatics 21:90–103.

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