Animal (2011), **5:8**, pp 1162–1169 © The Animal Consortium 2011
doi:10.1017/S1751731111000309

animal

Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection

D. P. Berry[1†] and J. F. Kearney[2]

[1]*Animal & Grassland Research and Innovation Centre, Teagasc, Moorepark, County Cork, Ireland;* [2]*Irish Cattle Breeding Federation, Highfield House, Bandon, County Cork, Ireland*

(Received 16 November 2010; Accepted 23 January 2011; First published online 28 February 2011)

The objective of this study was to quantify the accuracy achievable from imputing genotypes from a commercially available low-density marker panel (2730 single nucleotide polymorphisms (SNPs) following edits) to a commercially available higher density marker panel (51 602 SNPs following edits) in Holstein–Friesian cattle using Beagle, a freely available software package. A population of 764 Holstein–Friesian animals born since 2006 were used as the test group to quantify the accuracy of imputation, all of which had genotypes for the high-density panel; only SNPs on the low-density panel were retained with the remaining SNPs to be imputed. The reference population for imputation consisted of 4732 animals born before 2006 also with genotypes on the higher density marker panel. The concordance between the actual and imputed genotypes in the test group of animals did not vary across chromosomes and was on average 95%; the concordance between actual and imputed alleles was, on average, 97% across all SNPs. Genomic predictions were undertaken across a range of production and functional traits for the 764 test group animals using either their real or imputed genotypes. Little or no mean difference in the genomic predictions was evident when comparing direct genomic values (DGVs) using real or imputed genotypes. The average correlation between the DGVs estimated using the real or imputed genotypes for the 15 traits included in the Irish total merit index was 0.97 (range of 0.92 to 0.99), indicating good concordance between proofs from real or imputed genotypes. Results show that a commercially available high-density marker panel can be imputed from a commercially available lower density marker panel, which will also have a lower cost, thereby facilitating a reduction in the cost of genomic selection. Increased available numbers of genotyped and phenotyped animals also has implications for increasing the accuracy of genomic prediction in the entire population and thus genetic gain using genomic selection.

Keywords: genomic selection, impute, genotype, cattle

Implications

Genomic selection is the method of choice for genetic evaluations of dairy cattle in most countries. However, the cost of acquiring a genotype on a high-density panel is prohibitively expensive for individual farmers, thereby limiting the uptake of this technology on-farm. This study shows that the cost of implementing genomic selection can be reduced by genotyping using a low-density panel of markers and predicting or 'imputing' to a higher density panel. There was little reduction in accuracy of genomic prediction by using the imputed higher density marker panel.

Introduction

Genomic selection (Meuwissen *et al.*, 2001) is increasing in popularity as a method to evaluate the genetic merit of animals (VanRaden, 2008; Harris and Johnson, 2010). Accuracy of prediction of direct genomic values (DGVs) is a function of, amongst others, the quantity of phenotyped and genotyped individuals used to estimate marker effects (Daetwyler *et al.*, 2008). Furthermore, genome-wide association analyses, utilising population-wide linkage disequilibrium among densely positioned genetic markers across the genome, are being increasingly used to attempt to detect regions of the genome associated with performance (Pryce *et al.*, 2010). Large data sets of phenotyped and densely genotyped individuals are also required for successful genome-wide association studies (Purcell *et al.*, 2003), particularly for traits governed by a large number of quantitative trait loci (QTL).

Marker panels are, and will continue to be developed with varying numbers of markers. Using imputation, it may be possible to combine data from different marker panels by imputing missing genotypes across panels with the outcome of a complete set of markers across all individuals

† E-mail: donagh.berry@teagasc.ie

(Nothnagel *et al*., 2009; Druet *et al*., 2010). Furthermore, the current cost of genotyping sufficiently large numbers of individuals for the accurate estimation of single nucleotide polymorphism (SNP) effects within either genomic selection breeding programmes or genome-wide association studies, can be expensive. This is especially true for individual farmers who want to exploit genomic selection on-farm as part of their own breeding programme. Although genomic prediction using reduced marker panels yield high accuracy of prediction relative to high-density panels (Weigel *et al*., 2009) these reduced panels may be trait and breed specific. An alternative is to predict or impute SNPs on a high-density panel, which are not on a lower density panel. However, few studies have investigated the accuracy of imputation from low- to high-density genotype platforms using real life cattle populations (Zhang and Druet, 2010; Weigel *et al*., 2010a) or the impact on the prediction of DGVs (Weigel *et al*., 2010b).

The objective, therefore, of this study was to evaluate the accuracy of imputing genotypes of animals from the currently available low-density marker panel marketed by Illumina referred to here as the Bovine3K (Illumina, San Diego, CA, USA), to the commonly used (Harris and Johnson, 2010; Pryce *et al*., 2010) BovineSNP50 beadchip also marketed by Illumina (Matukumalli *et al*., 2009) and referred to as the Bovine50K in this study. The impact of imputation on genomic predictions will also be evaluated. The results from this study will be useful in evaluating the potential of reducing the cost of genomic selection by using a lower cost, lower density marker panel, coupled with imputation.

Material and methods

Data

Genotypes on 5489 Holstein–Friesian artificially inseminated (AI) bulls ($n = 4318$) and cows ($n = 1171$) using the Bovine50K marker panel (Illumina; Matukumalli *et al*., 2009) were available. All animals had genotypes called for at least 95% of the 54 001 SNPs on the Bovine50K. No animal had $>0.5\%$ Mendelian inconsistencies with either its parental or offspring genotype(s); 85.6% of the genotyped animals had either a parent or an offspring genotyped. Remaining inconsistent parent–offspring genotypes were set to missing.

Chromosome number and positions of the SNPs on the Bovine50K were obtained from the UMD3.0 bovine genome assembly from the University of Maryland. Of the original 54 001 SNPs on the Bovine50K, all SNPs on the sex chromosomes ($n = 1082$) as well as SNPs of unknown location ($n = 1065$) were discarded; one of a further 24 SNPs that appeared twice on the Bovine50K panel were also discarded. In addtion, a further 228 SNPs that showed $>0.5\%$ Mendelian inconsistencies between parent–progeny pairs were discarded. A total of 51 602 SNPs remained. The number of SNPs per chromosome is summarised in Table 1.

The Bovine3K SNP panel contains 2900 SNPs. Only SNPs on the autosomes were retained ($n = 2731$) and one SNP, of unknown position was also discarded leaving 2730 SNPs. The number of SNPs per chromosome are summarised in Table 1. Across chromosomes, 4.7% Bos taurus autosome (BTA19) to 5.9% (BTA5) of the SNPs on the Bovine50K panel were on the low density Bovine3K panel (Table 1).

Animals were separated into two groups: (i) reference group of animals born before 2006 ($n = 4725$) and (ii) a test group of animals $>50\%$ Holstein, born from 2006 onwards ($n = 764$). Mean (s.d.) Holstein proportion of the animals in the reference group and test group was 78% (30%) and 85% (15%), respectively. Only the 2730 SNPs from the Bovine3K (after editing) were retained in the test group of animals with the remaining SNPs on the Bovine50K in these animals to be imputed.

Imputation

Imputation was undertaken for each chromosome separately using the freely available software Beagle version 3.1.0 (http://faculty.washington.edu/browning/beagle/beagle.html; Browning and Browning, 2007 and 2009). Beagle is written in Java and includes algorithms for haplotype and missing-data inference, single marker and multilocus association analysis, as well as permutation testing. Beagle uses a localised haplotype-cluster model, which defines a Hidden Markov model that can be subsequently used to infer haplotypes or impute missing genotypes for each animal conditional on that animal's genotype. Beagle also provides posterior probabilities for each imputed haplotype, which reflects the degree of uncertainty in the imputed genotype. Beagle can handle phased or unphased genotypes as well as genotypes from unrelated animals, parent–offspring pairs or parent–offspring trios. In this study, direct knowledge of the parental genotypes was exploited in the imputing process. The methodology used in Beagle is described in more detail by Browning and Browning (2007).

Genomic prediction

To quantify the impact of imputation on the estimation of DGVs in the imputed animals, genomic prediction using BLUP (VanRaden, 2008) was undertaken for all traits included in the Irish total merit index, the economic breeding index (EBI; Table 2; Berry *et al*., 2007). Of the 51 602 SNPs remaining, following the editing previously described, a further 7210 were removed because they had a minor allele frequency (MAF) of $\leqslant 0.02$. An additional 2409 SNPs were removed because they had a call rate of $\leqslant 0.95$ and 374 SNPs were removed because they deviated ($P < 0.1 \times 10^{-7}$) from Hardy–Weinberg equilibrium. A total of 41 609 SNPs were therefore included in the genomic prediction. SNP effects were estimated in the reference animals and these SNP effects were applied to either the real genotypes or the imputed genotypes of the test group animals to estimate their DGVs.

The phenotypes of the reference animals used in the genomic prediction were, where available, based on deregressed estimated breeding values from multiple-across country evaluation (MACE) evaluations from the August 2010 genetic evaluation run; if MACE evaluations were not available then deregressed breeding values from the August 2010 national evaluation were used. Animals included in the

Table 1 *Number of SNPs per chromosome in the low-density panel (Bovine3K) and high-density panel (Bovine50K) as well as the four measures of imputation accuracy evaluated*

	Number of SNPs					
Chromosome	Bovine3K	Bovine50K	Genotype concordance	Allele concordance	Standardised allele-frequency error	Allelic R^2
1	178	3362	0.96 (0.04)	0.98 (0.02)	0.009 (0.008)	0.94
2	146	2769	0.95 (0.04)	0.98 (0.02)	0.009 (0.008)	0.93
3	129	2491	0.96 (0.04)	0.98 (0.02)	0.009 (0.010)	0.94
4	135	2512	0.96 (0.04)	0.98 (0.02)	0.009 (0.008)	0.94
5	130	2198	0.95 (0.05)	0.97 (0.02)	0.009 (0.010)	0.93
6	126	2540	0.95 (0.04)	0.98 (0.02)	0.009 (0.014)	0.93
7	125	2292	0.95 (0.04)	0.98 (0.02)	0.009 (0.011)	0.93
8	127	2355	0.96 (0.04)	0.98 (0.02)	0.008 (0.007)	0.94
9	115	2036	0.96 (0.04)	0.98 (0.02)	0.008 (0.010)	0.94
10	114	2147	0.95 (0.05)	0.97 (0.03)	0.009 (0.009)	0.93
11	113	2246	0.95 (0.05)	0.97 (0.03)	0.009 (0.008)	0.93
12	101	1720	0.95 (0.04)	0.98 (0.02)	0.009 (0.011)	0.93
13	91	1815	0.95 (0.05)	0.97 (0.02)	0.009 (0.015)	0.93
14	92	1794	0.95 (0.04)	0.97 (0.02)	0.009 (0.009)	0.92
15	91	1679	0.95 (0.04)	0.97 (0.02)	0.009 (0.010)	0.92
16	87	1686	0.95 (0.05)	0.98 (0.02)	0.009 (0.009)	0.93
17	76	1571	0.94 (0.05)	0.97 (0.02)	0.009 (0.007)	0.91
18	71	1332	0.94 (0.04)	0.97 (0.02)	0.010 (0.014)	0.91
19	65	1375	0.93 (0.05)	0.97 (0.03)	0.010 (0.012)	0.90
20	82	1533	0.95 (0.03)	0.98 (0.02)	0.009 (0.008)	0.93
21	79	1444	0.94 (0.06)	0.97 (0.03)	0.010 (0.011)	0.92
22	70	1310	0.95 (0.04)	0.97 (0.02)	0.009 (0.015)	0.93
23	55	1072	0.93 (0.05)	0.97 (0.03)	0.009 (0.007)	0.90
24	71	1285	0.95 (0.04)	0.97 (0.02)	0.009 (0.008)	0.92
25	48	980	0.93 (0.05)	0.96 (0.02)	0.009 (0.009)	0.89
26	56	1099	0.94 (0.05)	0.97 (0.03)	0.009 (0.012)	0.91
27	49	947	0.94 (0.04)	0.97 (0.02)	0.009 (0.008)	0.91
28	53	952	0.93 (0.05)	0.96 (0.02)	0.010 (0.007)	0.90
29	55	1060	0.95 (0.04)	0.97 (0.02)	0.009 (0.010)	0.92

SNPs = single nucleotide polymorphisms.

estimation of SNP effects were born before 2006 and had to have a reliability, less parental contribution, for the respective trait under investigation of $\geqslant$60%. Animals with imputed genotypes were not considered for the estimation of SNP effects. The number of animals used for the estimation of SNP effects are detailed in Table 2 for the different traits; the test-group of animals for which the DGVs were estimated was always 764.

Summary statistics on the accuracy of imputation

Several statistics were calculated to compare the accuracy of imputation in the test group of animals: (i) genotype concordance rate defined as the average proportion of correctly imputed genotypes within SNP or within animal, (ii) allele concordance rate defined as the average proportion of correctly imputed alleles within SNP or within animal; in this instance, a genotype imputed to be heterozygote but was truly homozygote was assumed to have one correct allele imputed, (iii) allelic R^2 defined as the squared correlation between the allele dosage of the most likely imputed genotype and the allele dosage of the real genotype and (iv) the standardised allele-frequency error (Browning and Browning, 2009). The standardised allele-frequency error was defined as:

$$\sqrt{\frac{|p_A - q_A|}{(p_A[1-p_A]/2n)}},$$

where p_A is the real allele frequency in the sample of n animals and q_A is the estimated frequency of the same allele from the most likely imputed genotypes. This statistic standardises the allele frequency error for the allele frequency and sample population size.

The impact of imputation on genomic predictions was quantified by calculating the mean and standard deviation of the bias between DGVs estimated using the real or imputed genotypes in the test group of animal as well as the correlation and regression of the DGVs estimated using the real genotypes on the DGVs estimated using the imputed genotypes in the test group of animal. The imputed genotypes used in the analysis were either the most likely genotypes or the posterior probabilities of each genotype.

Table 2 *Number of animals included in the reference population for the estimation of SNP effects (Reference), s.d. of the DGVs in the test group of animals when using the real genotypes, as well as the summary statistics from comparing the DGVs using the real or imputed genotypes including the mean bias (Bias), the RMSE of the bias and, both the correlation (*r*) and regression (*b*) of the DGVs estimated using the real genotypes on the DGVs estimated using the imputed genotypes*

Trait	Reference[1]	s.d.	Bias (s.e.)	RMSE	*r*	*b* (s.e.)
EBI (€)		54.348	−0.976 (0.433)	11.96	0.98	0.988 (0.008)
Milk yield (kg)	3508	133.340	−3.958 (1.008)	27.87	0.98	0.988 (0.008)
Fat yield (kg)	3508	4.063	−0.119 (0.033)	0.91	0.97	0.998 (0.008)
Protein yield (kg)	3508	3.572	−0.130 (0.027)	0.75	0.98	0.991 (0.008)
Calving interval (days)	1519	2.503	−0.039 (0.018)	0.51	0.98	2.020 (0.008)
Survival (%)	1241	1.179	0.023 (0.010)	0.27	0.97	0.988 (0.008)
Direct calving difficulty (%)	1403	1.165	0.004 (0.013)	0.36	0.95	0.988 (0.012)
Maternal calving difficulty (%)	1112	1.482	−0.007 (0.019)	0.52	0.94	0.994 (0.014)
Gestation length (days)	1089	0.783	−0.021 (0.008)	0.22	0.96	0.980 (0.010)
Perinatal mortality (%)	547	0.544	−0.002 (0.007)	0.20	0.93	0.950 (0.013)
Cow carcass weight (kg)	1042	6.645	−0.048 (0.035)	0.97	0.99	1.012 (0.005)
Progeny carcass weight (kg)	1091	5.407	−0.021 (0.040)	1.10	0.98	0.996 (0.007)
Progeny carcass conformation (scale 1 to 15)	1080	0.194	0.005 (0.002)	0.06	0.95	0.964 (0.011)
Progeny carcass fat (scale 1 to 15)	1040	0.156	0.003 (0.001)	0.02	0.99	1.011 (0.006)
Somatic cell score (log_e units)	3508	0.074	−0.001 (0.001)	0.02	0.97	0.992 (0.009)
Locomotion (scale 1 to 9)	736	6.732	−1.236 (0.115)	3.18	0.92	0.772 (0.012)

SNPs = single nucleotide polymorphisms; DGVs = direct genomic values; RMSE = root mean square error; EBI = economic breeding index.
[1]Number of training animals not included for the EBI as this was the weighted sum of the individual traits and was not estimated directly by using genomic prediction.

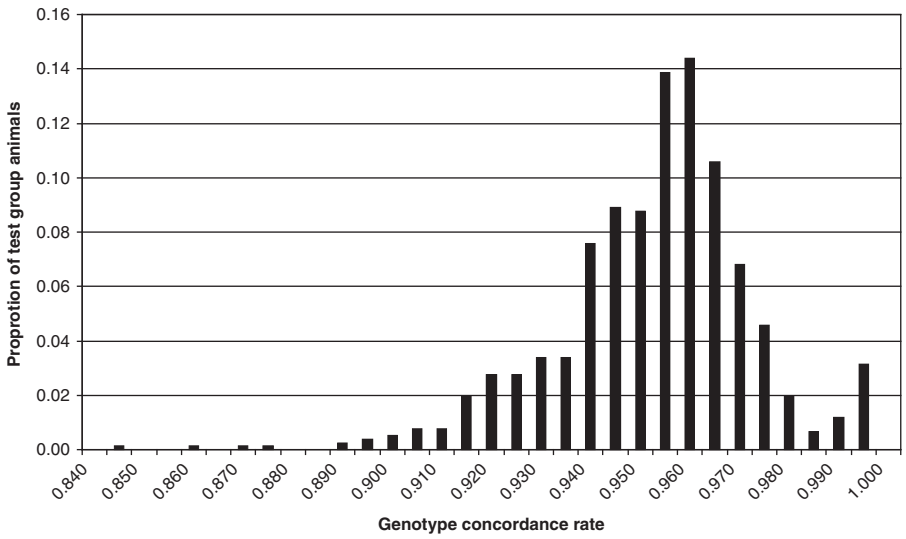

Figure 1 Distribution of individual animal genotype concordance rate between real and imputed genotypes for all animals in the test group.

Results

Across all chromosomes and animals, the mean (s.d. in parenthesis) genotype and allele concordance rate was 0.950 (0.044) and 0.974 (0.023), respectively. However, across chromosomes the mean genotype concordance rate varied from 0.930 to 0.959 (Table 1), whereas the mean allele concordance rate varied from 0.964 to 0.979 (Table 1). The standardised allele-frequency error varied from 0.008 to 0.010 per chromosome, whereas the allelic R^2 values varied from 0.892 to 0.942 (Table 1). There was only a slight reduction in genotype and allele concordance rate of 0.007 and 0.003, respectively when only the 38 980 of the 41 609 SNPs included in the genomic prediction but not on the Bovine3K panel were compared; 101 of the SNPs on the Bovine3K were discarded in the editing of the SNPs for genomic selection based on MAF ($n = 27$, one of which was monomorphic), deviation from Hardy–Weinberg equilibrium ($n = 21$) and poor call rates ($n = 53$). A weak but significant negative correlation existed between average genotype and allele concordance rate per SNP and the respective MAF ($r = -0.05$; $P < 0.001$) suggesting that the accuracy of imputation increased slightly as MAF reduced.

Mean genotype concordance rate per animal varied from 0.843 to 0.994 but was negatively skewed (Figure 1); 98% of animals had a genotype concordance rate of $\geqslant$0.900,

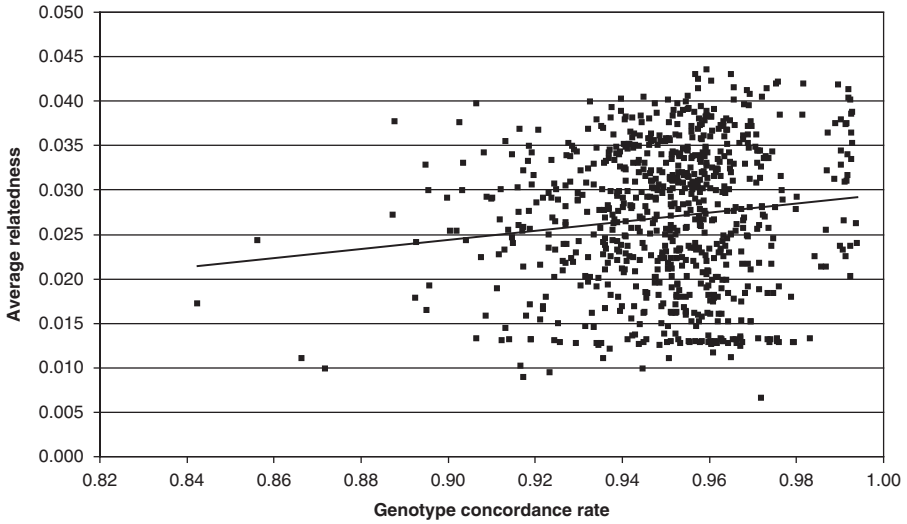

Figure 2 Association ($r = 0.13$) between average genotype concordance rate per animal in the test group and average relatedness to the reference population.

57% had a genotype concordance rate of $\geqslant$0.950 and 3% had a genotype concordance rate of $\geqslant$0.990. Mean allele concordance rate per animal varied from 0.917 to 0.996 and was also negatively skewed; 97.6% had a genotype concordance rate of $\geqslant$0.950 and 4.8% had a genotype concordance rate of $\geqslant$0.990. Allelic R^2 varied from 0.760 to 1.00 per animal; the mean allelic R^2 was 0.93.

The average relatedness between animals in the test group and animals in the reference group was 2.7%; the average relatedness among the test group of animals was 4.2% and among the reference group of animals was 2.3%. The relationship between mean genotype concordance rate per animal and average relatedness of the animal to the reference population is illustrated in Figure 2. A positive correlation of 0.13 ($P < 0.001$) existed between animal mean genotype concordance rate and average relatedness to the reference population estimated using pedigree information; the correlation was 0.44 ($P < 0.001$) when the average relatedness to the reference population was replaced by the maximum relatedness to the reference population. The mean (s.d. in parenthesis) genotype concordance rate per animal, for animals with no parent in the reference population ($n = 98$), only one parent in the reference population ($n = 632$) or both parents in the reference population ($n = 34$) was 0.927 (0.022), 0.952 (0.015) and 0.991 (0.002), respectively; the respective statistics for the allele concordance rate was 0.962 (0.012), 0.975 (0.008) and 0.995 (0.001).

Impact of imputation on genomic prediction

The number of animals included in the reference population for the estimation of SNP effects as well as the summary statistics comparing the DGVs of the 764 test group estimates using real or imputed genotypes is detailed in Table 2. Slightly better concordance, in the population as a whole, was observed when the posterior genotype probabilities were used instead of the most likely genotype for the test group animals, especially in animals with no parent in the reference population; on average the correlation with the DGVs estimated using the real genotypes increased by 0.004 when the posterior genotype probabilities were used instead of the most likely genotypes and therefore it is the results using the former that are presented in Table 2.

The differences among traits in number of animals included in the reference population for the estimation of SNP effects is a function of the heritability of the traits, which is reflected in the number of sires reaching the reliability threshold imposed, as well as whether or not international genetic evaluations are available for the trait under investigation.

DGVs estimated using the imputed genotypes were overestimated ($P < 0.05$) compared with when the real genotypes were used for all three yield traits, calving interval, gestation length and locomotion; underestimation ($P < 0.05$) was evident for survival as well as both carcass conformation and carcass fat score. The Irish total merit index, the EBI (Berry *et al.*, 2007), was also overestimated ($P < 0.05$; Table 2). The differences between the DGVs predicted using the real or imputed genotypes were normally distributed for each trait. The s.d. of the difference between the DGVs predicted using either the real or imputed genotypes varied from 0.13 (progeny carcass fat score) to 0.47 (locomotion) of the s.d. of the DGVs of the 746 test group animals when estimated using their real genotypes.

The correlation between the DGVs of the test group animals when their real or imputed genotypes were used in the genomic prediction varied from 0.92 (locomotion) to 0.99 (cow carcass weight and progeny carcass fat). The correlation between the EBI estimated from the weighted sum of the individual DGVs predicted using the real or imputed genotypes was 0.97. With the exception of locomotion, the regression of the DGVs of the test group of animals estimated using the real genotypes on the DGVs estimated using the imputed genotypes were close to unity and only differed ($P < 0.05$) from unity for perinatal mortality, progeny carcass conformation and fat score, and locomotion.

Discussion

The motivation for this study was to quantify the accuracy of imputation from a commercially available low-density SNP marker panel to higher density SNP marker panel with the objective that, if successful, reducing the cost of genomic predictions. The results clearly show that, on average, high accuracy of imputation can be achieved, especially where the pedigree of an animal has been genotyped on the Bovine 50K and is included in the reference group. It should be noted, however, that the accuracy of imputation is likely to be biased downwards slightly as genotyping errors in the 'real' genotypes to which the imputed genotypes were compared, may exist, although the impact is likely to be small. Furthermore, incorrect genome positioning of some SNPs may also affect the accuracy of haplotyping and therefore imputation.

Imputation accuracy

The accuracy of imputing the unobserved, approximately 95% of the SNPs on the higher density panel was, on average, high although variation among animals did exist corroborating the large variation in imputation accuracy per animal also reported by Weigel *et al*. (2010a). Furthermore, the high allelic R^2 indicates little loss in statistical power for association studies by using the most likely imputed genotype compared with, if the real genotype was used.

Several studies have evaluated the accuracy of imputation in human populations (Browning and Browning, 2007; Hao *et al*., 2009), some of which also compared different software packages available for imputation (Howie *et al*., 2009; Nothnagel *et al*., 2009). However, few studies have attempted to quantify the accuracy of imputation from lower to higher density panel in real life cattle populations (Druet and Georges, 2010; Weigel *et al*., 2010a; Zhang and Druet, 2010) or between different marker panels in cattle (Druet *et al*., 2010). Studies using simulated cattle populations have also been undertaken (Habier *et al*., 2009). Allele concordance rate was evaluated in this study as well as genotype concordance rate because genomic selection in Ireland (Berry *et al*., 2009) currently uses SNP effects based on the number of copies of a given allele.

Weigel *et al*. (2010a) using a reference panel of 43 385 SNPs on 2542 Jersey cattle evaluated two freely available software packages for imputation (i.e. fastPHASE and IMPUTE). The test population in the study of Weigel *et al*. (2010a), which is comparable with how the test group of animals were generated in this study, was the 'future study sample' and consisted of 604 Jersey cattle. When only 5% (i.e. similar to the 5.3% average in this study) of the 43 385 SNP genotypes in the reference population of Weigel *et al*. (2010a) were known in the test population, the genotype concordance rate varied from 0.77 to 0.88 for BTA1 depending on the approach used; similar results were reported for BTA15 and BTA28, which were the only other two chromosomes investigated in that study. In this study, the accuracy of imputation, even after the removal of SNPs with low MAF and poor call rate as undertaken by Weigel *et al*. (2010a), was higher.

Zhang and Druet (2010) using a panel of 3000 SNPs chosen based on a combination of equal distribution across the genome and high MAF, reported an allelic imputation error rate (identical to one minus the allele concordance rate reported in this study) of approximately 0.03 for 2734 Dutch Holstein bulls using a reference population of 1000 Dutch Holstein bulls. This is similar to the allele concordance rate reported in this study.

A contributing factor to the increased accuracy of imputation in this study over the accuracy reported by Weigel *et al*. (2010a) could be the larger size of the reference population in this study (4725 *v.* 2542), which has been shown to improve the accuracy of imputation, for some imputing algorithms at least, in both cattle (Zhang and Druet, 2010; Weigel *et al*., 2010b) and human populations (Browning and Browning, 2009). However, Zhang and Druet (2010) reported little benefit in accuracy of imputation, based on the algorithms they used, for reference population sizes $>$1000 individuals. Another reason for the difference in imputation accuracy compared with that of Weigel *et al*. (2010a) was that Weigel *et al*. (2010a) included a random selection of SNPs in the lower density panel, whereas SNPs on the Bovine3K were selected so as to achieve an even distribution of polymorphic SNPs across the bovine genome, much like the approach adopted by Zhang and Druet (2010). Other reasons for the differences between studies include differences in software (i.e. algorithms) used as well as other possible factors including population structure differences such as linkage disequilibrium within the breed/population under investigation as well as the relatedness between the reference and test group population, which is known to influence imputation accuracy (Figure 2; Zhang and Druet, 2010).

No description of the average relatedness of the test group evaluated by Weigel *et al*. (2010a) with the reference population was presented, although 595 sires and 580 maternal grandsires of the 604 animals in the test group were genotyped for the higher density marker panel and included in the reference population. Nonetheless, Weigel *et al*. (2010a) reported no difference in imputation accuracy in animals that had progeny or pedigree genotyped compared with having no such information. Druet *et al*. (2010) and Zhang and Druet (2010) using both linkage disequilibrium and linkage analysis in different, although overlapping data sets, also clearly showed an increased accuracy of imputation when the animals to be imputed were more closely related to the reference population. Therefore, to maximise confidence in the imputation of genotypes, high-density genotypes of the pedigree of the animal should be included in the reference population.

Effect on genomic prediction

Higher accuracies of genomic predictions were obtained when the posterior probability of an imputed genotype was used rather than the most likely genotype, which is not

unexpected as the use of the posterior probability accounts for the uncertainty of the imputation. Although several studies in human genetics have evaluated the impact of imputation on the power of genome-wide association studies (Marchini *et al*., 2007; Servin and Stephens, 2007; Hao *et al*., 2009), we are aware of only one recent study (Weigel *et al*., 2010b), which has quantified the impact of imputation of genotypes on genomic predictions using real data in cattle. In the present study, the accuracy of genomic predictions from imputed genotypes were compared with DGVs derived from real genotypes rather than comparing with predicted transmitting abilities from traditional genetic evaluations as a large proportion of the animals in the test group were cows or young test sire that had no phenotypic information and therefore were not of high reliability based on traditional genetic evaluations.

Weigel *et al*. (2010b) using a reference population of 1446 Jersey AI sires for milk yield, protein percentage and daughter pregnancy rate, genotyped on the Bovine50K panel reported that the predictive ability of DGVs using 2942 SNPs (i.e. similar to this study) coupled with imputation, was approximately 97% of that of when using the real genotypes; SNP effects in that study were estimated using a Bayesian approach. This conclusion is similar to the mean correlation (0.97) in the present study between DGVs estimated using the Bovine3K panel and the DGVs predicted using the Bovine50K panel. The lower correlations between DGVs predicted using real or imputed genotypes evident for some traits in this study appear to be related to the number of animals included in the reference population for the estimation of SNP effects. The exceptions were direct and maternal calving difficulty, which had relatively weak correlations between the DGVs estimated using real or imputed genotypes (0.94 to 0.95) despite relatively large training populations sizes (1112 to 1403 animals) for the estimation of SNP effects. One possible contributing factor could be the contribution of large QTLs to these phenotypes and the accuracy of imputation may be poor in these genomic regions. Cole *et al*. (2009) reported a region of BTA18 in Holstein cattle to have a large association with genetic merit for direct and maternal calving ease using SNPs from the Bovine50K panel.

Weigel *et al*. (2010b) did not report any other statistics comparing the DGVs estimated using high- and low-density marker panels but the results from this study suggest little mean difference in DGVs generated from the two methods and that a unit change in the DGVs estimated using the real genotypes was associated with a near unity change in DGV estimated using the imputed genotypes. Furthermore, where bias in the mean did exist it was neither a systematic underestimation nor overestimation of the DGVs, which has important ramifications for the ranking of animals using different genotype platforms. The bias in the total merit index, although statistically significant, is biologically insignificant.

In conclusion, the cost of genomic selection can be considerably reduced by genotyping on the Bovine3K and, through the use of imputation algorithms, obtain *in silico* genotypes on the Bovine50K. However, key to accurate imputation is knowledge of the high-density genotypes of the pedigree of the animal being imputed. Lowering the costs of genomic selection will increase its uptake, especially by farmers. This will, over time, increase the size of the reference population for the estimation of SNP effects, thereby increasing the accuracy of prediction and subsequently genetic gain.

Acknowledgements

Useful discussions with Brian Browning on the implementation of the Beagle software are gratefully acknowledged. Genotypes made available through sharing with international partners as well as funding from national (RSF-06-0353; RSF-06-0428) and European Union Framework 7 funding (http://www.robustmilk.eu) are also gratefully acknowledged.

References

Berry DP, Kearney JF and Harris BL 2009. Genomic selection in Ireland. Proceedings of the Interbull International Workshop – Genomic Information in Genetic Evaluations, Uppsala, Sweden, 26–29 January 2009. Interbull Bulletin 39, 29–34.

Berry DP, Shalloo L, Cromie AR, Olori V, Veerkamp RF, Dillon P, Amer PR, Evans RD, Kearney JF and Wickham B 2007. The economic breeding index: a generation on. Technical report to the Irish Cattle Breeding Federation. Irish Cattle Breeding Federation Society Ltd, Bandon, Co Cork, Ireland.

Browning SR and Browning BL 2007. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. American Journal of Human Genetics 81, 1084–1097.

Browning BL and Browning SR 2009. A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. American Journal of Human Genetics 84, 210–223.

Cole JB, VanRaden PM, O'Connell JR, Van Tassell CP, Sonstegard TS, Schnabel RD, Taylor JF and Wiggans GR 2009. Distribution and location of genetic effects for dairy traits. Journal of Dairy Science 92, 2931–2946.

Daetwyler HD, Villanueva B and Woolliams JA 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3, e3395.

Druet T and Georges M 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, Schrooten C and de Roos APW 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. Journal of Dairy Science 93, 5443–5454.

Habier D, Fernando RL and Dekkers JCM 2009. Genomic selection using low-density marker panels. Genetics 182, 343–353.

Hao K, Chudin E, McElwee J and Schadt EE 2009. Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. BMC Genetics 10, 27.

Harris BL and Johnson DL 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. Journal of Dairy Science 93, 1243–1252.

Howie BN, Donnelly P and Marchini J 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics 5, e1000529.

Marchini J, Howie B, Myers S, McVean G and Donnelly P 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics 39, 906–913.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TPL, Sonstegard TS and Van Tassell CP 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4, e5350.

Meuwissen THE, Hayes BJ and Goddard ME 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

Nothnagel M, Ellinghaus D, Schreiber S, Krawczak M and Franke A 2009. A comprehensive evaluation of SNP genotype imputation. Human Genetics 125, 163–171.

Pryce JE, Bolormaa S, Chamberlain AJ, Bowman PJ, Savin K, Goddard ME and Hayes BJ 2010. A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. Journal of Dairy Science 93, 3331–3345.

Purcell S, Cherny SS and Sham PC 2003. Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics 19, 149–150.

Servin B and Stephens M 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genetics 3, e114.

VanRaden PM 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91, 4414–4423.

Weigel KA, van Tassell CP, O'Connell JR, VanRaden PM and Wiggans GR 2010a. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. Journal of Dairy Science 93, 2229–2238.

Weigel KA, de los Campos G, Vazquez AI, Rosa GJM, Gianola D and Van Tassell CP 2010b. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. Journal of Dairy Science 93, 5423–5435.

Weigel KA, de los Campos G, González-Recio O, Naya H, Wu XL, Long N, Rosa GJM and Gianola D 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. Journal of Dairy Science 92, 5248–5257.

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