J. Dairy Sci. 93:2229–2238
doi:10.3168/jds.2009-2849

Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms

K. A. Weigel,*[1] C. P. Van Tassell,†‡ J. R. O'Connell,§ P. M. VanRaden,‡ and G. R. Wiggans‡
*Department of Dairy Science, University of Wisconsin, Madison 53706
†Bovine Functional Genomics Laboratory, and
‡Animal Improvement Programs Laboratory, ARS, USDA, Beltsville, MD 20705-2350
§University of Maryland School of Medicine, Baltimore 21201

ABSTRACT

The availability of dense single nucleotide polymorphism (SNP) genotypes for dairy cattle has created exciting research opportunities and revolutionized practical breeding programs. Broader application of this technology will lead to situations in which genotypes from different low-, medium-, or high-density platforms must be combined. In this case, missing SNP genotypes can be imputed using family- or population-based algorithms. Our objective was to evaluate the accuracy of imputation in Jersey cattle, using reference panels comprising 2,542 animals with 43,385 SNP genotypes and study samples of 604 animals for which genotypes were available for 1, 2, 5, 10, 20, 40, or 80% of loci. Two population-based algorithms, fastPHASE 1.2 (P. Scheet and M. Stevens; University of Washington Tech-Transfer Digital Ventures Program, Seattle, WA) and IMPUTE 2.0 (B. Howie and J. Marchini; Department of Statistics, University of Oxford, UK), were used to impute genotypes on *Bos taurus* autosomes 1, 15, and 28. The mean proportion of genotypes imputed correctly ranged from 0.659 to 0.801 when 1 to 2% of genotypes were available in the study samples, from 0.733 to 0.964 when 5 to 20% of genotypes were available, and from 0.896 to 0.995 when 40 to 80% of genotypes were available. In the absence of pedigrees or genotypes of close relatives, the accuracy of imputation may be modest (generally <0.80) when low-density platforms with fewer than 1,000 SNP are used, but population-based algorithms can provide reasonably good accuracy (0.80 to 0.95) when medium-density platforms of 2,000 to 4,000 SNP are used in conjunction with high-density genotypes (e.g., >40,000 SNP) from a reference population. Accurate imputation of high-density genotypes from inexpensive low- or medium-density platforms could greatly enhance the efficiency of whole-genome selection programs in dairy cattle.

Key words: single nucleotide polymorphism, imputation, cattle

INTRODUCTION

The recent development of high-throughput systems for genotyping SNP in cattle and other food animal species has led to an extraordinary amount of research activity, particularly in areas such as whole-genome selection of livestock and genome-wide association studies for detection of quantitative trait loci (Van Tassell et al., 2008). Tens of thousands of dairy cattle have been genotyped using the BovineSNP50 BeadChip (Illumina Inc., San Diego, CA) or related platforms, and the resulting genomic data have already been incorporated into national genetic evaluation systems for routine prediction of the genetic merit of selection candidates (Wiggans et al., 2009).

To date, most genotyping in North American dairy cattle has been on a common platform, namely the aforementioned BovineSNP50 BeadChip, and gains in reliability due to incorporation of the resulting genomic information into practical breeding programs have been impressive (VanRaden et al., 2009). However, it is likely that other options will become available in the near future, such as ultra-high-density platforms with more than 500,000 SNP and specialized low-density platforms with 300 to 3,000 selected or equally spaced SNP. Recently, Weigel et al. (2009) reported that low-density genotyping of young Holstein bulls for 300 to 2,000 selected SNP provided direct genomic values for Lifetime Net Merit that were moderately correlated (0.43 to 0.57) with the breeding values obtained from progeny testing and highly correlated (0.63 to 0.90) with direct genomic values derived from the BovineSNP50 BeadChip. However, a limitation of low-density genotyping with selected SNP is that such platforms will be breed and trait specific. For this reason, Habier et al. (2009)

Received October 22, 2009.
Accepted January 4, 2010.
[1]Corresponding author: kweigel@wisc.edu

suggested genotyping selection candidates with a low-density platform comprising equally spaced SNP and subsequently imputing genotypes for the remaining SNP using high-density genotypes of their parents. The authors noted that this approach would provide not only low-density platforms that are applicable across traits and breeds, but also platforms that are robust to the number of loci affecting individual traits and the nuances of the statistical methods used for estimating SNP effects. Overall, it seems that imputation of missing genotypes will be necessary for cost-effective, widespread implementation of whole-genome selection, regardless of whether breeding values are derived from models in which SNP effects are estimated directly (e.g., Meuwissen et al., 2001) or from models in which SNP genotypes are used to describe the covariance structure between selection candidates (e.g., Gianola and van Kaam, 2008).

Algorithms for constructing haplotypes and imputing missing genotypes generally fall into 2 categories (Li and Li, 2007). The first category of methods, which are well suited for natural populations, includes population-based algorithms that typically assume that a large group of individuals without known relationships have been genotyped (e.g., Scheet and Stephens, 2006; Howie et al., 2009). The second category of methods, which are well suited for case-control studies, includes family-based algorithms that generally assume that many small, nuclear families exist in which both parents and at least one offspring have been genotyped (e.g., Li and Jiang, 2003; Zhang and Zhao, 2006). The structure of modern dairy cattle breeding populations matches neither of these descriptions, because the populations comprise thousands or millions of individuals related to each other in a complex, multi-generational manner. Furthermore, vast differences in the reproductive capacity of males and females and, hence, in the economic worth of selected candidate parents, ensures that expensive high-density genotyping or DNA sequencing technologies will be used more widely in males. Thus, one cannot assume that detailed genotypic information will be available for both parents of every selection candidate. Last, if low-density genotyping is broadly implemented on commercial dairy farms for routine activities such as selection of replacement heifers or development of genotype-guided mating recommendations, situations in which both parents lack genotypic information may become commonplace.

The objective of the present study was to evaluate the accuracy that could be achieved when predicting missing SNP genotypes of study samples of Jersey cattle from the high-density genotypes of reference animals of the same breed using publicly available, population-based imputation algorithms that require neither pedigrees nor genotyped parents. As such, the realized accuracy of imputation in the present study should be representative of the lower bound that dairy producers and breeding companies may achieve in practice when imputing genotypes of animals that may lack pedigree information or high-density parental genotypes. Specifically, we sought to evaluate scenarios that may arise in practical breeding programs when selection candidates are genotyped using inexpensive low- or medium-density platforms.

MATERIALS AND METHODS

Genotypes of 3,146 Jersey dairy cattle (2,656 males and 490 females) from 5 countries were provided by the USDA-ARS Animal Improvement Programs Laboratory (Beltsville, MD) and consisted of 43,385 SNP markers distributed across the 29 *Bos taurus* autosomes and the X chromosome. These SNP represent the subset of markers on the BovineSNP50 BeadChip that are used for routine genomic evaluation of US dairy cattle, after removal of SNP with a call rate of <90%, greater than 1% parent-progeny conflicts, complete linkage disequilibrium (**LD**) with an adjacent SNP, or minor allele frequency (**MAF**) of <1% in each of the Holstein, Jersey, and Brown Swiss breeds (Wiggans et al., 2009). Genotypes at each locus were coded as 0 (homozygous for allele B), 1 (heterozygous), 2 (homozygous for allele A), or 5 (missing). After the aforementioned edits, the mean percentage of missing genotypes (averaged across all 43,385 loci) was 0.31%.

Because the objective of this study was to evaluate the accuracy of imputed genotypes in subsets of animals genotyped using low- or medium-density genotyping platforms, the population was divided into study samples and reference panels in the following manner. The reference panels consisted of animals that were genotyped for all 43,385 SNP, whereas in the study samples we randomly masked genotypes for 20, 60, 80, 90, 95, 98, or 99% of the SNP loci. The purpose was to mimic a situation in which animals in the study samples were genotyped using a low- or medium-density platform comprising 80, 40, 20, 10, 5, 2, or 1% of SNP on the high-density platform, respectively. Subsequently, high-density SNP genotypes of animals in the reference panels were used in conjunction with low- or medium-density SNP genotypes of animals in the study samples to infer masked genotypes of the latter.

Reference panels and study samples were constructed in 2 ways, as shown in Table 1. In one scenario, a subset of 2,542 animals that were born between 1953 and 2006 was used as the reference panel, and a subset of 604 animals that were born between 2007 and 2009 was used as the study sample. This approach was analogous

Table 1. Summary of data from the US Jersey cattle population used in the present study

	Random		Future	
Item	Reference panel	Study sample	Reference panel	Study sample
Total animals	2,542	604	2,542	604
Sex				
Male	2,153	503	2,173	483
Female	389	101	369	121
Year of birth				
1953–1958	4	1	5	0
1959–1964	5	0	5	0
1965–1970	16	1	17	0
1971–1976	35	8	43	0
1977–1982	55	8	63	0
1983–1988	175	27	202	0
1989–1994	452	106	558	0
1995–2000	599	145	744	0
2001–2006	733	172	905	0
2007–2009	468	136	0	604
Country of registration				
Australia	18	0	16	2
Canada	94	34	112	16
Denmark	8	2	10	0
New Zealand	2	1	3	0
United States	2,420	567	2,401	586
Ancestors in reference panel				
Sire		405		595
Paternal grandsire		426		561
Maternal grandsire		432		580
Dam		85		127
Paternal granddam		119		199
Maternal granddam		56		81
Offspring in reference panel				
Son(s)		53		0
Paternal grandson(s)		18		0
Maternal grandson(s)		48		0
Daughter(s)		40		0
Paternal granddaughter(s)		10		0
Maternal granddaughter(s)		14		0

to the experimental design used by VanRaden et al. (2009) to evaluate the ability of genomic breeding values to predict the results of progeny testing in the next generation. In the present study, it mimics a situation in which young animals are genotyped with a low- or medium-density assay, and missing SNP are inferred from high-density genotypes of their ancestors. In the second scenario, a random subset of 2,542 animals in the population was used as the reference panel, regardless of year of birth, and the remaining 604 animals were used as the study sample. This approach, which may be unrealistic in practical breeding programs, mimics a situation in which the decision of whether to genotype a particular animal using a high- or low-density platform is made at random. In both study samples, most animals had a genotyped sire, maternal grandsire, and paternal grandsire in the reference panel, as shown in Table 1. However, few animals in either study sample had a genotyped dam, maternal granddam, or paternal granddam in the reference panel. A greater proportion of animals in the future study sample had genotyped ancestors in the reference panel; conversely, some animals in the random study sample had genotyped sons, daughters, grandsons, or granddaughters in the reference panel.

To evaluate the accuracy of imputed genotypes, 3 chromosomes were chosen: BTA1, BTA15, and BTA28. These corresponded to the maximum (2,748), median (1,399), and minimum (802) number of SNP per chromosome, respectively, on the BovineSNP50 BeadChip. After removal of SNP with unknown physical position on the chromosome based on the UMD2 assembly of *B. taurus* (Zimin et al., 2009), a total of 2,693, 1,377, and 795 SNP remained on BTA1, BTA15, and BTA28, respectively, for animals in the reference panels. After randomly masking from 20 to 99% of SNP on these chromosomes for animals in the study samples, the number of available SNP from which to impute masked genotypes ranged from 27 to 2,154 for BTA1, from 14 to 1,102 for BTA15, and from 8 to 636 for BTA28.

For reference purposes, masked genotypes were imputed by random assignment of genotypes of 0, 1, or

2 to masked loci assuming that the probability of a particular genotype was equal to its expectation under Hardy-Weinberg equilibrium with p and q equal to observed allele frequencies at each locus in the reference panels.

Subsequently, masked genotypes were imputed using the population-based haplotype clustering algorithm of Scheet and Stephens (2006), which was implemented via the fastPHASE version 1.2 software (University of Washington TechTransfer Digital Ventures Program, Seattle, WA). This method assumes that, over short regions of a chromosome, haplotypes of individuals within the population will tend to cluster into groups. Because of recombination, groups of haplotypes that are similar in one region of a chromosome will differ in another region, according to the rate of decay in LD. Therefore, this method allows membership in haplotype clusters to change as one moves along the chromosome, using a hidden Markov model (**HMM**) to describe each observed haplotype as a mosaic of a small number of common haplotypes. This approach is more flexible than block-based methods that divide the genome into regions of high LD and then allow changes in cluster membership to occur only at the boundaries of these blocks, such as the method of Kimmel and Shamir (2005), which underlies the GERBIL software, or the method of Greenspan and Geiger (2004), which underlies the HaploBlock software. Furthermore, this method offers greater computational feasibility in large data sets than many other flexible models, such as the product of approximate conditional likelihoods approach of Li and Stephens (2003), which underlies the PHASE software, because computing time of the Scheet and Stephens (2006) algorithm increases linearly with the number of genotyped individuals. In the present study, the number of haplotype clusters was fixed at values of 8 (the default value), 16, or 32. Although the software allows estimation of the optimal number of haplotype clusters, it is very computationally demanding for large data sets. In addition, consideration of more than 32 haplotype clusters was not feasible in the present study, because computing time increases quadratically with the number of haplotype clusters.

In addition, masked genotypes were imputed using the HMM-based algorithm of Howie et al. (2009), which was implemented via the IMPUTE version 2.0 software (Department of Statistics, University of Oxford, UK). This algorithm is very flexible, in the sense that it can accommodate multiple reference panels of phased or unphased genotypes that may be composed of different sets of SNP from assays of varying density. Furthermore, it is computationally feasible for large data sets because of separation of the phasing and imputation steps. Specifically, rather than simultaneously estimating missing genotypes and integrating over the unknown phase of SNP that are present in both the reference panels and the study sample (as is the case for most other HMM-based algorithms), this algorithm estimates haplotypes at SNP that are present in both populations and then imputes genotypes in the study sample, assuming that these haplotype guesses are correct. Uncertainty about phasing is taken into account by iterating these steps in a Markov chain Monte Carlo (**MCMC**) framework. Thus, unlike many competing algorithms for which phasing accuracy does not depend on the size of the study sample (e.g.,Scheet and Stephens, 2006; Marchini et al., 2007), IMPUTE 2.0 gains accuracy by using information from both the reference panels and study sample during the phasing step. The computational feasibility of IMPUTE 2.0 is enhanced by using only a subset of haplotypes at each MCMC iteration to build the conditional distribution of haplotypes of observed SNP for an animal in the study sample, given the animal's genotype, the haplotypes of other animals in the study sample, and the haplotypes of animals in the reference panel. Rather than sampling these "conditioning states" randomly (e.g., Li et al., 2006), this algorithm selects sets of haplotypes that are closest to the animal in question, based on the Hamming distance (i.e., the minimum number of substitutions required to change one haplotype into the other) between the current-guess haplotype for this animal and for other animals in the population. For computational reasons, the number of conditioning states used in the present study was set to 40, although Howie et al. (2009) notes that slight increases in accuracy might be achieved by increasing this parameter, albeit with a quadratic increase in computing time.

With large data sets, the computational feasibility of programs such as fastPHASE 1.2 and IMPUTE 2.0 can be enhanced by breaking chromosomes into pieces of more manageable size. In the present study, BTA28 was analyzed in its entirety using both fastPHASE 1.2 and IMPUTE 2.0, because in some scenarios (e.g., with 98% or 99% of SNP masked in the study samples), the number of remaining SNP on BTA28 that were available for imputation was extremely limited. In analyses with fastPHASE 1.2, BTA1 and BTA15 were also analyzed in their entirety, whereas in analyses with IMPUTE 2.0, computational feasibility was enhanced by breaking BTA1 and BTA15 into 4 and 2 pieces of equal size, respectively.

RESULTS AND DISCUSSION

The proportion of masked SNP genotypes that were imputed correctly was computed for every animal in each of the random and future study samples for sub-

Table 2. Mean (minimum, maximum) proportion of masked SNP genotypes on chromosome 1 (2,693 total SNP) that were imputed correctly in random and future study samples comprising US Jersey cattle, using reference panels comprising 2,542 animals of the same breed[1,2]

Sample	Proportion of SNP genotyped in study samples	Random imputation	fastPHASE 1.2 (8 haplotype clusters)	fastPHASE 1.2 (16 haplotype clusters)	fastPHASE 1.2 (32 haplotype clusters)	IMPUTE 2.0 (40 conditioning states)
Random study sample	0.01	0.545 (0.445, 0.685)	0.703 (0.601, 0.750)	0.704 (0.618, 0.758)	0.705 (0.614, 0.759)	**0.720** (0.597, 0.869)
	0.02	0.545 (0.446, 0.687)	0.717 (0.602, 0.767)	0.720 (0.629, 0.762)	0.723 (0.629, 0.770)	**0.765** (0.632, 0.916)
	0.05	0.543 (0.443, 0.684)	0.775 (0.623, 0.853)	0.776 (0.659, 0.838)	0.784 (0.685, 0.856)	**0.892** (0.712, 0.996)
	0.10	0.545 (0.442, 0.685)	0.863 (0.659, 0.945)	0.882 (0.742, 0.955)	0.894 (0.775, 0.966)	**0.942** (0.732, 0.998)
	0.20	0.544 (0.447, 0.685)	0.922 (0.689, 0.989)	0.948 (0.803, 0.994)	**0.964** (0.827, 0.997)	0.951 (0.775, 0.999)
	0.40	0.542 (0.447, 0.682)	0.959 (0.751, 0.999)	0.981 (0.845, 1.000)	**0.989** (0.870, 1.000)	0.948 (0.747, 0.998)
	0.80	0.534 (0.417, 0.685)	0.975 (0.797, 1.000)	0.990 (0.866, 1.000)	**0.994** (0.890, 1.000)	0.946 (0.737, 1.000)
Future study sample	0.01	0.542 (0.469, 0.660)	0.699 (0.598, 0.763)	0.697 (0.597, 0.744)	0.702 (0.612, 0.756)	**0.712** (0.568, 0.841)
	0.02	0.543 (0.470, 0.660)	0.718 (0.608, 0.776)	0.716 (0.617, 0.770)	0.722 (0.622, 0.785)	**0.754** (0.621, 0.900)
	0.05	0.541 (0.470, 0.659)	0.768 (0.615, 0.834)	0.777 (0.642, 0.838)	0.780 (0.670, 0.855)	**0.880** (0.735, 0.993)
	0.10	0.542 (0.468, 0.659)	0.851 (0.648, 0.950)	0.878 (0.712, 0.958)	0.887 (0.763, 0.963)	**0.935** (0.802, 0.994)
	0.20	0.541 (0.468, 0.659)	0.913 (0.681, 0.989)	0.942 (0.773, 0.995)	**0.960** (0.861, 0.998)	0.944 (0.794, 0.998)
	0.40	0.540 (0.459, 0.655)	0.953 (0.749, 0.998)	0.978 (0.868, 0.999)	**0.989** (0.893, 0.999)	0.941 (0.753, 0.998)
	0.80	0.533 (0.436, 0.674)	0.969 (0.776, 1.000)	0.989 (0.892, 1.000)	**0.995** (0.879, 1.000)	0.937 (0.736, 0.998)

[1]fastPHASE 1.2 (P. Scheet and M. Stevens; University of Washington TechTransfer Digital Ventures Program, Seattle, WA) and IMPUTE 2.0 (B. Howie and J. Marchini; Department of Statistics, University of Oxford, UK) are population-based algorithms.

[2]Bold font denotes the greatest imputation accuracy for a given masking rate.

sequent computation of the mean proportion of SNP that were imputed correctly, as well as the minimum and maximum proportion of SNP that were imputed correctly for an individual animal.

As shown in Tables 2, 3, and 4, random assignment of masked genotypes assuming Hardy-Weinberg equilibrium with allele frequencies estimated from data of the reference panels led to proportions of correct genotypes that ranged from 0.51 to 0.56. Imputation of genotypes using fastPHASE 1.2 with 8 haplotype clusters led to proportions of correct genotypes of approximately 0.70 for BTA1 and BTA15 and 0.66 for BTA28 when only 1% of genotypes were available for animals in the study samples. Corresponding values were 0.72 and 0.68, respectively, when 2% of genotypes were available, and 0.77 and 0.73, respectively, when 5% of genotypes were available. Presumably, the slightly lower accuracy for BTA28, which had only 795 SNP in the reference panels, was because of the small number of SNP that remained for imputation when the proportion discarded was large. Significant variation existed between animals in the proportion of genotypes that were imputed correctly, with minimum (maximum) proportions ranging from 0.55 to 0.60 (0.73 to 0.76) when 1% of genotypes were available to 0.58 to 0.63 (0.83 to 0.86) when 5% of genotypes were available.

The mean and maximum proportions of genotypes that were imputed correctly by fastPHASE 1.2 increased rapidly as the percentage of available SNP in the study samples increased to 10% or more of the total SNP on a given chromosome. Mean proportions ranged from 0.79 to 0.85 when 10% of genotypes were available and from 0.86 to 0.92 when 20% of genotypes were available, whereas the maximums ranged from 0.95 to 0.98 when 10% of genotypes were available and 0.98 to 0.99 when 20% of genotypes were available. The remaining scenarios, with 40 or 80% of genotypes available for animals in the study samples, are not likely to arise in practical situations involving a low- or medium-density SNP assay but might arise when data from 2 competing high-density assays of different size are combined (e.g., a 50,000-SNP assay and a 30,000-SNP assay). In these cases, the mean proportion of genotypes that were imputed correctly ranged from 0.96 to 0.99, whereas minimums ranged from 0.64 to 0.75 and maximums approached unity.

Table 3. Mean (minimum, maximum) proportion of masked SNP genotypes on chromosome 15 (1,377 total SNP) that were imputed correctly in random and future study samples comprising 604 US Jersey cattle, using reference panels comprising 2,542 animals of the same breed[1,2]

Sample	Proportion of SNP genotyped in study samples	Random imputation	fastPHASE 1.2 (8 haplotype clusters)	fastPHASE 1.2 (16 haplotype clusters)	fastPHASE 1.2 (32 haplotype clusters)	IMPUTE 2.0 (40 conditioning states)
Random study sample	0.01	0.539 (0.462, 0.684)	0.703 (0.592, 0.761)	0.700 (0.611, 0.760)	0.701 (0.596, 0.759)	**0.743** (0.549, 0.961)
	0.02	0.539 (0.464, 0.684)	0.725 (0.604, 0.784)	0.723 (0.620, 0.778)	0.725 (0.624, 0.790)	**0.801** (0.607, 1.000)
	0.05	0.539 (0.461, 0.684)	0.769 (0.629, 0.862)	0.782 (0.677, 0.856)	0.790 (0.693, 0.866)	**0.905** (0.687, 1.000)
	0.10	0.540 (0.463, 0.682)	0.857 (0.641, 0.948)	0.873 (0.742, 0.958)	0.887 (0.737, 0.959)	**0.938** (0.740, 1.000)
	0.20	0.540 (0.462, 0.682)	0.920 (0.708, 0.986)	0.949 (0.797, 0.998)	**0.960** (0.794, 0.998)	0.944 (0.742, 0.999)
	0.40	0.545 (0.456, 0.695)	0.953 (0.753, 1.000)	0.979 (0.852, 1.000)	**0.987** (0.851, 1.000)	0.945 (0.742, 1.000)
	0.80	0.556 (0.447, 0.681)	0.969 (0.777, 1.000)	0.987 (0.874, 1.000)	**0.993** (0.882, 1.000)	0.943 (0.706, 1.000)
Future study sample	0.01	0.539 (0.434, 0.662)	0.695 (0.557, 0.756)	0.693 (0.573, 0.749)	0.701 (0.574, 0.766)	**0.730** (0.533, 0.925)
	0.02	0.539 (0.436, 0.661)	0.712 (0.580, 0.773)	0.719 (0.576, 0.784)	0.726 (0.596, 0.797)	**0.780** (0.568, 0.981)
	0.05	0.538 (0.433, 0.661)	0.758 (0.581, 0.839)	0.768 (0.591, 0.875)	0.780 (0.644, 0.856)	**0.890** (0.682, 0.999)
	0.10	0.540 (0.438, 0.663)	0.839 (0.596, 0.946)	0.862 (0.655, 0.952)	0.874 (0.732, 0.960)	**0.924** (0.762, 0.998)
	0.20	0.539 (0.432, 0.659)	0.898 (0.634, 0.990)	0.929 (0.738, 0.994)	**0.951** (0.841, 0.993)	0.932 (0.778, 1.000)
	0.40	0.543 (0.427, 0.658)	0.938 (0.684, 0.998)	0.967 (0.780, 1.000)	**0.984** (0.890, 1.000)	0.935 (0.772, 1.000)
	0.80	0.556 (0.404, 0.684)	0.956 (0.692, 1.000)	0.981 (0.875, 1.000)	**0.992** (0.946, 1.000)	0.930 (0.663, 1.000)

[1]fastPHASE 1.2 (P. Scheet and M. Stevens; University of Washington TechTransfer Digital Ventures Program, Seattle, WA) and IMPUTE 2.0 (B. Howie and J. Marchini; Department of Statistics, University of Oxford, UK) are population-based algorithms.

[2]Bold font denotes the greatest imputation accuracy for a given masking rate.

Increasing the number of clusters tended to increase the accuracy of imputation with fastPHASE 1.2, as shown in Tables 2, 3, and 4. This seems to indicate that 8 clusters could not adequately accommodate the size and complexity of the Jersey population considered herein. With 16 haplotype clusters, the mean proportion of correct genotypes changed negligibly for scenarios in which only 1 or 2% of genotypes were available for animals in the study samples. However, as the proportion of available genotypes increased to 5 or 10%, significant increases in the accuracy of imputation became apparent, particularly for the mean and minimum values. This pattern continued, and became more pronounced, as the proportion of available genotypes for animals in the study samples increased up to 20, 40, or 80%.

The aforementioned gains in accuracy with an increase in the number of haplotype clusters continued when 32 clusters were modeled, as shown in Tables 2, 3, and 4. In this case, the mean proportion of genotypes imputed correctly ranged from 0.66 to 0.73 when only 1% or 2% of genotypes were available for animals in the study samples. This proportion increased to 0.75 to 0.89 when 5 to 10% of genotypes were available, as would be the case for a medium-density panel comprising 2,000 to 4,000 SNP. As the percentage of available genotypes increased to 20, 40, or 80%, the mean proportion of correct genotypes increased from 0.90 to >0.99. This indicates that allowing a large number of haplotype clusters is desirable, but the corresponding gains in accuracy come with a considerable computational cost.

Accuracy of imputed genotypes for animals in the study samples using IMPUTE 2.0 is also shown in Tables 2, 3, and 4. The IMPUTE 2.0 algorithm provides genotype probabilities for each locus and allows the user to specify the threshold at which a genotype should be called. In the present study, we called all genotypes, regardless of the magnitude of the probability for the most likely genotype, such that imputation accuracy could be compared with other methods. In practice, it may be advantageous to refrain from calling genotypes, and to instead use the corresponding genotype probabilities directly when building the coefficient matrix for subsequent prediction of genomic breeding values. When the proportion of masked genotypes was large, such as 98 or 99%, IMPUTE 2.0 was slightly more accurate, with gains in accuracy relative to fast-

Table 4. Mean (minimum, maximum) proportion of masked SNP genotypes on chromosome 28 (795 total SNP) that were imputed correctly in random and future study samples comprising 604 US Jersey cattle, using reference panels comprising 2,542 animals of the same breed[1,2]

Sample	Proportion of SNP genotyped in study samples	Random imputation	fastPHASE 1.2 (8 haplotype clusters)	fastPHASE 1.2 (16 haplotype clusters)	fastPHASE 1.2 (32 haplotype clusters)	IMPUTE 2.0 (40 conditioning states)
Random study sample	0.01	0.513 (0.427, 0.685)	0.666 (0.585, 0.741)	0.664 (0.589, 0.729)	0.664 (0.593, 0.729)	**0.689** (0.536, 0.914)
	0.02	0.511 (0.425, 0.685)	0.680 (0.584, 0.754)	0.678 (0.596, 0.757)	0.678 (0.591, 0.751)	**0.727** (0.517, 0.999)
	0.05	0.512 (0.426, 0.685)	0.738 (0.613, 0.849)	0.755 (0.625, 0.858)	0.767 (0.634, 0.874)	**0.886** (0.617, 1.000)
	0.10	0.511 (0.427, 0.681)	0.807 (0.622, 0.922)	0.827 (0.649, 0.953)	0.848 (0.647, 0.956)	**0.917** (0.609, 1.000)
	0.20	0.512 (0.424, 0.677)	0.875 (0.651, 0.983)	0.916 (0.637, 0.992)	**0.942** (0.672, 0.998)	0.921 (0.614, 1.000)
	0.40	0.515 (0.406, 0.690)	0.928 (0.683, 1.000)	0.968 (0.686, 1.000)	**0.981** (0.709, 1.000)	0.926 (0.612, 1.000)
	0.80	0.518 (0.390, 0.681)	0.955 (0.710, 1.000)	0.985 (0.716, 1.000)	**0.993** (0.753, 1.000)	0.921 (0.525, 1.000)
Future study sample	0.01	0.509 (0.435, 0.694)	0.663 (0.549, 0.734)	0.659 (0.573, 0.735)	0.662 (0.573, 0.733)	**0.679** (0.532, 0.825)
	0.02	0.508 (0.432, 0.693)	0.677 (0.550, 0.743)	0.674 (0.560, 0.745)	0.677 (0.584, 0.757)	**0.720** (0.501, 0.997)
	0.05	0.509 (0.433, 0.695)	0.733 (0.584, 0.843)	0.748 (0.590, 0.851)	0.754 (0.630, 0.866)	**0.870** (0.602, 1.000)
	0.10	0.507 (0.426, 0.693)	0.788 (0.604, 0.925)	0.811 (0.598, 0.946)	0.823 (0.651, 0.955)	**0.902** (0.681, 1.000)
	0.20	0.509 (0.430, 0.691)	0.862 (0.613, 0.986)	0.901 (0.613, 0.994)	**0.929** (0.745, 0.997)	0.904 (0.632, 1.000)
	0.40	0.511 (0.415, 0.690)	0.907 (0.635, 1.000)	0.956 (0.762, 1.000)	**0.978** (0.873, 1.000)	0.910 (0.654, 1.000)
	0.80	0.511 (0.395, 0.671)	0.944 (0.710, 1.000)	0.980 (0.846, 1.000)	**0.991** (0.932, 1.000)	0.896 (0.463, 1.000)

[1]fastPHASE 1.2 (P. Scheet and M. Stevens; University of Washington TechTransfer Digital Ventures Program, Seattle, WA) and IMPUTE 2.0 (B. Howie and J. Marchini; Department of Statistics, University of Oxford, UK) are population-based algorithms.

[2]Bold font denotes the greatest imputation accuracy for a given masking rate.

PHASE 1.2 ranging from approximately 0.02 to 0.07. However, IMPUTE 2.0 was significantly more accurate for scenarios in which 90 or 95% of genotypes were masked in the study sample. For example, when 5% of genotypes were available for animals in the study samples, accuracy of imputation ranged from 0.87 to 0.91 for IMPUTE 2.0 and from 0.75 to 0.77 for fastPHASE 1.2 with 32 haplotype clusters. When 10% of genotypes were available, accuracy ranged from 0.90 to 0.94 for IMPUTE 2.0 and from 0.82 to 0.85 for fastPHASE 1.2 with 32 haplotype clusters. However, when 20, 40, or 80% of SNP genotypes were available for animals in the study samples, fastPHASE 1.2 was more accurate, because accuracy of imputation peaked at approximately 0.90 to 0.95 for IMPUTE 2.0. Perhaps the reason why accuracy of IMPUTE 2.0 did not approach unity as the proportion of masked genotypes decreased, as was the case for fastPHASE 1.2, was because of the aforementioned approximation in this algorithm that involves sampling of conditioning states rather than integration over all possible haplotype configurations. Another possibility is that remaining genotype errors (after removal of SNP with <90% call rate and SNP with >1% parent-progeny conflicts), as well as errors in the map order, inhibit the ability of such algorithms to achieve high imputation accuracy for some animals.

Differences in the accuracy of imputation between the random and future study samples were small, although accuracy was slightly higher for the random study samples. Likewise, differences between chromosomes were also small, with slightly greater accuracy for the larger chromosomes (BTA1 and BTA15) than for BTA28. Because this study utilized population-based methods that ignored pedigree information, the accuracy of imputation for animals in the study samples that had genotyped ancestors or offspring in the reference panels was similar to that of animals that lacked such information. Methods that utilize family data, such as the long-range phasing approach of Kong et al. (2008), would most likely provide greater accuracy than the approaches implemented herein for animals with genotyped parents.

Figure 1 shows frequency distributions of the proportion of SNP genotypes on BTA15 that were imputed correctly for each animal in the future study samples using fastPHASE 1.2 with 32 haplotype clusters or

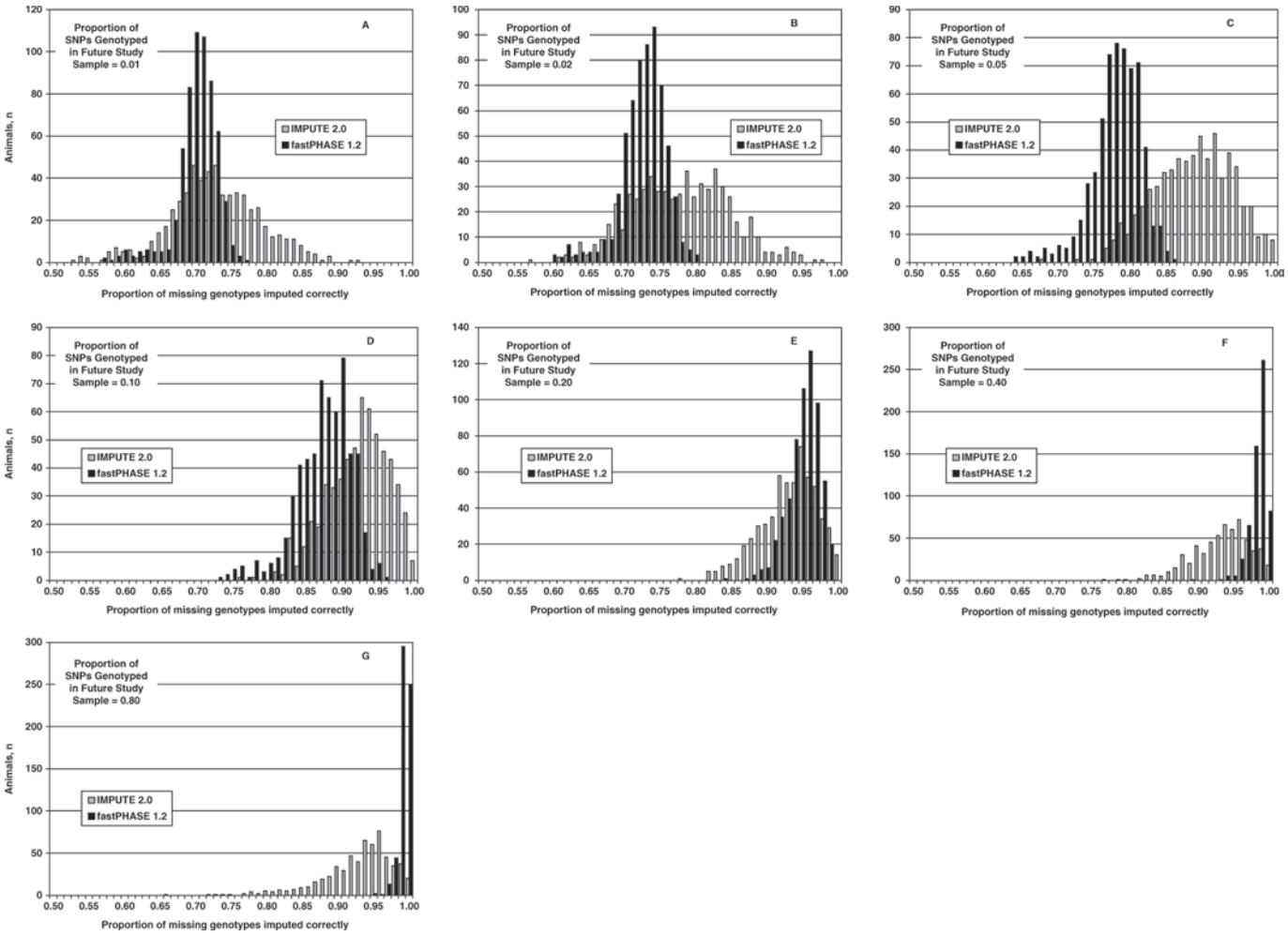

Figure 1. Frequency distribution of the proportion of masked SNP genotypes on chromosome 15 (1,377 total SNP) that were imputed correctly in future study samples comprising 604 US Jersey cattle, using reference panels comprising 2,542 animals of the same breed with the IMPUTE 2.0 (40 conditioning states; B. Howie and J. Marchini; Department of Statistics, University of Oxford, UK) or fastPHASE 1.2 (32 haplotype clusters; P. Scheet and M. Stevens; University of Washington TechTransfer Digital Ventures Program, Seattle, WA) population-based algorithms.

IMPUTE 2.0 with 40 conditioning states. In general, the standard deviation of the proportion of genotypes imputed correctly was greater for IMPUTE 2.0 than for fastPHASE 1.2, and this difference was most noticeable when the proportion of SNP genotyped in the study sample was <0.05 or >0.20. As noted previously, fastPHASE 1.2 provided greater predictive ability when the proportion of SNP genotyped in the study sample was 0.20 or greater, and the degree of superiority increased as the proportion of masked genotypes decreased. Conversely, IMPUTE 2.0 offered greater predictive ability when the proportion of SNP genotyped in the study samples was 0.10 or less, and the superiority of this algorithm was most obvious when the aforementioned proportion was equal to 0.05.

It should be noted that the accuracy of genotype imputation depends critically on the level of LD within a population. As such, the results presented herein should be considered specific to the US Jersey cattle population. Previous studies have indicated that LD may be higher in Jerseys than many other common breeds of dairy cattle. For example, Villa-Angulo et al. (2009) reported that average r^2 across 101 high-density regions of BTA6, BTA14, and BTA25 that were 100 kb in length, with an average of 19.6 SNP per region, was 0.380 in Jersey cattle, compared with 0.377, 0.333, 0.323, and 0.324, respectively, in the Brown Swiss, Guernsey, Holstein, and Norwegian Red breeds.

Several additional analyses were carried out in which we attempted to use information regarding popula-

tion structure to enhance the accuracy of imputation using the fastPHASE 1.2 algorithm. First, animals were divided into subpopulations based on country of registration. Second, animals were grouped into 5 subpopulations based on allele frequencies at 8,677 evenly spaced loci (i.e., by choosing every fifth SNP from a list ordered by physical position), using the model-based clustering algorithm implemented in the Structure 2.3.1 software (Pritchard et al., 2000). Third, animals were grouped into 35 subpopulations according to sire family (34 families with at least 25 members and 1 subpopulation for the remainder). Results of these preliminary analyses are not shown because none led to an improvement in the accuracy of genotype imputation. Last, additional analyses with IMPUTE 2.0 with a greater number of MCMC iterations for integration over the space of possible phase reconstructions for observed data genotypes (50 rather than the default value of 30) failed to provide any increase in the accuracy of imputed genotypes for animals in the study samples.

CONCLUSIONS

Incorporation of low- or medium-density SNP genotypes into routine genomic evaluations, with the remaining genotypes imputed via algorithms such as those utilized herein, could greatly enhance the efficiency of breed improvement programs. In summary, it appears that publicly available, population-based imputation algorithms can provide predicted SNP genotypes for Jersey cattle with mean accuracy of 0.80 to 0.95 when a subset of animals is genotyped with a medium-density panel comprising 2,000 to 4,000 SNP. With 1,000 or fewer SNP, the proportion of missing genotypes that are assigned correctly by such algorithms may be less than 0.80 for many animals, whereas with more than 8,000 SNP accuracy of imputation will exceed 0.95, on average, and will approach unity for some individuals. Improvements in the accuracy of genotype imputation might be achieved by selecting SNP that are most informative (i.e., those with high MAF) for low- or medium-density assays, compared with random selection of SNP in the present study. Furthermore, the impact of genotype errors, as well as errors in map order, should be considered in future studies. Finally, the computational feasibility of imputation algorithms such as those utilized herein would be greatly enhanced by incorporating reference panels consisting of phased haplotypes, rather than unphased diploid genotypes as in the present study. Furthermore, "tag" SNP could be used to track phased haplotypes, as well as quantitative trait loci that are in LD with SNP in these haplotypes, from one generation to the next. The estimated accuracies presented herein should be considered conservative because pedigree information was not considered, and the availability of high-density genotypes of the sires and dams of animals in the study samples was not a prerequisite. As such, future studies should evaluate the gains in imputation accuracy that can be achieved by incorporating pedigree information, when it is available. In practice, many animals will have high-density genotypes for both parents, and estimation of the probabilities of descent of marker alleles may be possible for other animals based on the genotypes of grandparents and other close relatives. One may wish to use a 2-step approach in which animals with genotyped parents (or other close ancestors) are processed first using a family-based method, and animals that lack such information are processed subsequently using a population-based algorithm. Such an approach could be considered as a form of "boosting" (e.g., Freund and Schapire, 1996), in which 2 or more complementary models, each of which treats a significant percentage of the data optimally, are implemented jointly to solve an estimation or classification problem. The resulting genotype probabilities could then be combined with high-density genotypes of other animals in the population for estimation of SNP effects and prediction of genomic breeding values. Future studies should evaluate the effect of imputation errors on the accuracy of predicted genomic breeding values, on average, as well as for specific subsets of animals that are more or less closely related to animals in the reference panel.

ACKNOWLEDGMENTS

This project was supported by National Research Initiative competitive grant no. 2009-35205-05099 from the USDA National Institute for Food and Agriculture Animal Genome Program. The fastPHASE 1.2 software, which was written by Paul Scheet (Department of Epidemiology, University of Texas M. D. Anderson Cancer Center, Houston) and Matthew Stephens (Department of Human Genetics, University of Chicago, IL), was provided by the University of Washington TechTransfer Digital Ventures Program (Seattle). The IMPUTE 2.0 software was provided by Bryan Howie and Jonathan Marchini, Department of Statistics, University of Oxford (United Kingdom). Kent Weigel acknowledges financial support from the National Association of Animal Breeders (Columbia, MO).

REFERENCES

Freund, Y., and R. E. Schapire. 1996. Experiments with a new boosting algorithm. Pages 148–156 in Proceedings of the Thirteenth International Conference on Machine Learning. L. Saitta, ed. Morgan Kaufmann, San Francisco, CA.

Gianola, D., and J. B. C. H. M. van Kaam. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289–2303.

Greenspan, G., and D. Geiger. 2004. Model-based inference of haplotype block variation. J. Comput. Biol. 11:493–504.
Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2009. Genomic selection using low-density SNP. Genetics 182:343–353.
Howie, B. N., P. Donnelly, and J. Marchini. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5:e1000529.
Kimmel, G., and R. Shamir. 2005. A block-free hidden Markov model for genotypes and its application to disease association. J. Comput. Biol. 12:1243–1260.
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.
Li, J., and T. Jiang. 2003. Efficient inference of haplotypes from genotypes on a pedigree. J. Bioinform. Comput. Biol. 1:41–69.
Li, N., and M. Stephens. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165:2213–2233.
Li, X., and J. Li. 2007. Comparison of haplotyping methods using families and unrelated individuals on simulated rheumatoid arthritis data. BMC Proc. 1(Suppl. 1):S55.
Li, Y., J. Ding, and G. R. Abecasis. 2006. Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. Am. J. Hum. Genet. 79:S2290.
Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39:906–913.
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.
Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.
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.
Van Tassell, C. P., T. P. L. Smith, L. K. Matukumalli, J. F. Taylor, R. D. Schnabel, C. T. Lawley, C. D. Haudenschild, S. S. Moore, W. C. Warren, and T. S. Sonstegard. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods 5:247–252.
VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, and F. Schenkel. 2009. Reliability of genomic predictions for North American dairy bulls. J. Dairy Sci. 92:16–24.
Villa-Angulo, R., L. K. Matukumalli, C. A. Gill, J. Choi, C. P. Van Tassell, and J. J. Grefenstette. 2009. High-resolution haplotype block structure in the cattle genome. BMC Genet. 10:19.
Weigel, K. A., G. de los Campos, O. González-Recio, H. Naya, X. L. Wu, N. Long, G. J. M. Rosa, and D. Gianola. 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J. Dairy Sci. 92:5248–5257.
Wiggans, G. R., P. M. VanRaden, L. R. Bacheller, F. A. Ross Jr, T. S. Sonstegard, G. te Meerman, and C. P. Van Tassell. 2009. Transition of genomic evaluation from a research project to a production system. J. Anim. Sci. 87(E-Suppl. 2):313–314. (Abstr.)
Zhang, K., and H. Zhao. 2006. A comparison of several methods for haplotype frequency estimation and haplotype reconstruction for tightly linked markers for general pedigrees. Genet. Epidemiol. 30:423–437.
Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassel, T. S. Sonstegard, G. Marçais, M. Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg. 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. Genome Biol. 10:R42.