Effects of reduced panel, reference origin, and genetic relationship on imputation of genotypes in Hereford cattle

Y. Huang, C. Maltecca, J. P. Cassady, L. J. Alexander, W. M. Snelling and M. D. MacNeil

J ANIM SCI 2012, 90:4203-4208.
doi: 10.2527/jas.2011-4728 originally published online August 2, 2012

Effects of reduced panel, reference origin, and genetic relationship on imputation of genotypes in Hereford cattle

Y. Huang,* C. Maltecca,* J. P. Cassady,*[1] L. J. Alexander,† W. M. Snelling,‡ and M. D. MacNeil†[2,3]

*Department of Animal Science, North Carolina State University, Raleigh 27606; †USDA-ARS, Fort Keogh Livestock and Range Research Laboratory, Miles City, MT 59301; ‡USDA-ARS, U. S. Meat Animal Research Center, Clay Center, NE 68933

ABSTRACT: The objective of this study was to investigate alternative methods of designing and using reduced SNP panels for imputing SNP genotypes. Two purebred Hereford populations, an experimental population known as Line 1 Hereford (**L1**, n = 240) and registered Hereford with American Hereford Association (**AHA**, n = 311), were used. Using different reference samples of 62 to 311 animals with 39,497 SNP on 29 autosomes and study samples of 57 or 62 animals for which genotypes were available for ~2,600 SNP (reduced panels), imputations were performed to predict the other ~36,900 loci that had been masked. An imputation package, including LinkPHASE and DAGPHASE, was used for imputation. Four reduced panels differing in minor allele frequency (**MAF**) and marker spacing were evaluated. Reduced panels included every 15th SNP across the genome (**SNP_space**), commercial Illumina Bovine3K Beadchip (**SNP_3K**), SNP with the highest MAF (**SNP_MAF**), and SNP with high MAF that were also evenly spaced across the genome (**SNP_MS**). Imputation accuracy was defined as the correlation of imputed genotypes and real genotypes. Reference samples were either from L1 or AHA. Among animals with genotypes, genetic relationships were estimated based on molecular marker genotypes or pedigree. Reduced panel design, number of animals in the reference sample, reference origin and genetic relationship between animals in the reference, and study samples all affected imputation accuracy ($P < 0.001$). Across genotyping schemes, imputed genotypes from SNP_MS had the greatest accuracy. A 0.1 increase in average pedigree relationship or average molecular relationship between reference and study samples increased imputation accuracy 10 to 20%. Using reference samples from the L1 population resulted in lower imputation accuracy than using reference samples from the admixed population AHA ($P < 0.001$). Increasing the number of animals in the reference panel by 100 individuals increased imputation accuracy by 8% when pedigree relationship was used as a covariate and 6% when molecular relationship was used as a covariate. We concluded that imputation accuracy would be increased through optimization of reduced panel design and genotyping strategy.

Key words: cattle, genomic, SNP

J. Anim. Sci. 2012.90:4203–4208
doi:10.2527/jas2011-4728

INTRODUCTION

For cattle, SNP marker panels differing in density from a few hundred to >750,000 markers are currently available. A cost-effective genotyping strategy may be to develop a reference panel by genotyping key individuals with a high density SNP chip. Animals sampled for a specific study (or candidates for selection) would be genotyped with a reduced set of markers. To make full use of data collected from both panels, marker genotypes that were not directly assayed with the low density panel would be imputed by exploiting linkage disequilibrium (**LD**) and pedigree information (Daetwyler et al., 2011; Druet and Georges, 2010; Hickey et al., 2011). Previously, reduced panels of different densities have been compared for dairy breeds using evenly spaced markers (Weigel et al., 2010; Zhang and Druet, 2010). Genetic relationships among genotyped individuals using panels

[1]Corresponding author: joe_cassady@ncsu.edu.

[2]Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by USDA. The USDA is an equal opportunity provider and employer.

[3]Present address: Delta G, 145 Ice Cave Rd., Miles City, MT 59301

Received Sept. 20, 2011.
Accepted June 19, 2012.

of different densities may affect accuracy of imputation (Druet and Georges, 2010; Huang et al., 2009).

We investigated factors affecting accuracy of imputation using two subpopulations of Hereford cattle. Line 1 Hereford (**L1**) cattle at Fort Keogh Livestock and Range Research Laboratory, Miles City, MT, have been maintained as a closed herd under documentable selection for more than 75 yr (MacNeil, 2009). As a consequence, L1 cattle are closely interrelated. Kuehn et al. (2011), in collaboration with the American Herford Association (**AHA**), sampled sires from the American Hereford registry. Specific questions were: 1) do differences in minor allele frequency (**MAF**) for SNP included in the reduced panel affect accuracy of imputation, and 2) does the degree to which average genetic relationships between animals genotyped with the high density and reduced panels affect imputation accuracy?

MATERIALS AND METHODS

Animal Care and Use Committee approval was not obtained for this study because the work was based on previously compiled data.

Data

A sample of 240 animals from L1 that were born from 1953 to 2008 was genotyped for this study. The 240 individuals were divided into 2 groups: calves born in 2008 and ancestral sires. The 2008-born calves (57 females and 62 males) represented 9 paternal half-sib families. The average number of progeny per sire was 12.8 ± 2.6. Individuals in the AHA sample (n = 311) were males from the U.S. Meat Animal Research Center (MARC) 2,000 bull project (Kuehn et al., 2011). Fifty-six of these bulls were previously used in the MARC Germplasm Evaluation program (http://www.ars.usda.gov/Main/docs.htm?docid=6238). The remainder (n = 255) were chosen by AHA from bulls born between 1970 and 2008. A pedigree of 12,356 animals, including 7,300 L1 animals, was extracted from AHA records and the research database at Fort Keogh. For the genotyped animals, mean inbreeding coefficients were 0.29 ± 0.022 and 0.04 ± 0.048 for L1 and AHA, respectively.

Marker genotypes were obtained for the L1 and AHA Hereford populations. All animals were genotyped using the Illumina BovineSNP50 Beadchip (Matukumalli et al., 2009). Chromosome information and physical positions of SNP were mapped to UMD3.0 assembly (Zimin et al., 2009). All animals were pooled for data editing and quality assessment. After removal of SNP which lacked chromosome information and physical position, had a call rate <90%, had a MAF <2%, or were in complete LD with an adjacent SNP, a subset of 39,497 SNP representing 29 autosomes remained. Genotypes were checked for inconsistencies between sire and offspring. After examining the distribution of the number of genotypes that did not follow Mendelian rules on sire offspring pairs, sire offspring pairs with >2,000 inconsistent genotypes were considered as animals with pedigree errors. In this case, a possible sire was sought using genotypes. Pedigree errors of 6 sire offspring pairs were corrected with this method. Less than 50 inconsistent genotypes between sire and offspring pairs were set as missing in the offspring. After editing, <0.1% of genotypes were missing.

Relationships

Average pedigree and molecular relationships were calculated for each study individual relative to the reference sample. To calculate pedigree relationships, the full numerator relationship matrix was constructed using R (http://www.R-project.org) software package GeneticsPed (http://rgenetics.org), following Henderson's method (Henderson, 1976). Pedigree relationships provided estimates of genotypes shared by individuals that were identical by descent. Using only SNP information, a realized molecular relationship matrix was calculated following VanRaden (2008). The molecular relationship matrix described the proportion of genotypes shared by individuals.

Design of Genotyping Schemes

Animals were divided into reference and study samples (Table 1). Reference samples had information on all 39,497 SNP. Study samples included information on a reduced set of 2,600 SNP. Accuracy of imputation was investigated for each reduced panel. Four criteria were used to produce panels denoted as: SNP_space, SNP_3K, SNP_MAF, and SNP_MS (SNP with high MAF that were also evenly spaced across the genome). The SNP_space panel included every 15th SNP across the genome, ordered by physical position. The SNP_3K panel comprised 2,600 markers included on the Illumina Bovine3K Beadchip (http://www.illumina.com/Documents//products/datasheets/datasheet_bovine3k.pdf) that remained after editing. The SNP_MAF panel included those markers with MAF nearest 0.5. Finally, SNP_MS panel included markers that were evenly spaced across the genome and had MAF >0.35.

Various subsets of L1 and AHA genotypes were used as reference samples to impute additional genotypes in study samples from both L1 and AHA (Table 2). Reference panel sizes ranged from 62 to 311 individuals. The number of animals in the study sample was either 57 or 62. When L1 reference samples were used to impute additional L1 genotypes, 62 paternal half-sibs, 121 ancestral males, or a combination of the 2 were used as reference samples, and the study sample was composed of 57 L1 heifers born in 2008. Another approach included 183 randomly selected

Table 1. Description of SNP[1] on reference and reduced panels, and no. SNP in common on the reduced panels

Marker panel[2]	n	Spacing of SNP, kb ± SD	MAF[3] ± SD	No. common SNP on reduced panels		
				SNP_space	SNP_3K	SNP_MAF
Reference panel	39,497	63.30 ± 66.28	0.264 ± 0.001			
SNP_space	2,634	948.35 ± 346.06	0.271 ± 0.003			
SNP_3K	2,600	949.56 ± 459.12	0.294 ± 0.003	198		
SNP_MAF	2,643	925.43 ± 1,235.64	0.485 ± 0.008	202	211	
SNP_MS	2,602	959.47 ± 440.83	0.425 ± 0.044	201	257	542

[1]All SNP panels included all SNP that remained after quality control.

[2]SNP_space included every 15th SNP; SNP_3K included SNP on the Illumina Bovine3K Beadchip (Illumina, San Diego, CA); SNP_MAF included 2,643 SNP with greatest MAF; and SNP_MS included SNP with MAF >0.35 and were approximately evenly spaced.

[3]MAF = minor allele frequency.

L1 animals as a reference sample to impute genotypes for the remaining 57 L1 animals. When AHA animals were used as reference samples to impute genotypes for the 57 L1 heifers born in 2008, the AHA animals were ordered by either the average pedigree or average molecular relationship to animals in the L1 study sample, and samples of 62, 121, and 183 animals were taken from the extremes. All 311 animals from AHA were also used as a reference sample to impute genotypes for the 57 L1 heifers born in 2008. When the 240 L1 animals were used as reference samples to impute genotypes for the AHA animals, the AHA animals were stratified, by their average molecular relationships to the L1 animals, into 5 strata of 62 animals. Finally, reference panels of 249 randomly sampled AHA animals

Table 2. Summary characterization of genotyping schemes used to investigate accuracy of imputation and relationships between Line 1 (L1) and American Hereford Association (AHA) animals in the reference and study samples, and least squares means of average accuracy of imputation for each scheme that was evaluated

Experiment[1]	Sample acronym[2]	Sample size (n)		Molecular relationship ± SD	Pedigree relationship ± SD	Accuracy of imputation[3]	Cost × US$500
		Reference	Study				
L1 → L1	HS	62	57	0.523 ± 0.008	0.611 ± 0.040	94.0	21.20
	SGS	121		0.506 ± 0.006	0.578 ± 0.039	94.4	35.95
	ALL	183		0.512 ± 0.007	0.589 ± 0.040	96.5	51.45
	RS[4]	183		0.508± 0.016	0.584 ± 0.033	95.5	51.45
AHA → L1	P.hi62	62	57	0.439 ± 0.007	0.251 ± 0.017	83.5	21.20
	P.hi121	121		0.409 ± 0.006	0.174 ± 0.012	84.1	35.95
	P.hi183	183		0.387 ± 0.006	0.125 ± 0.008	83.8	51.45
	P.lo62	62		0.321 ± 0.005	0.000 ± 0.000	63.7	21.20
	P.lo121	121		0.321 ± 0.005	0.001 ± 0.000	66.1	35.95
	P.lo183	183		0.328 ± 0.005	0.009 ± 0.001	71.4	51.45
	M.hi62	62		0.442 ± 0.007	0.245 ± 0.017	83.7	21.20
	M.hi121	121		0.414 ± 0.006	0.170 ± 0.011	84.4	35.95
	M.hi183	183		0.390 ± 0.006	0.124 ± 0.008	84.0	51.45
	M.lo62	62		0.311 ± 0.005	0.002 ± 0.000	62.2	21.20
	M.lo121	121		0.317 ± 0.005	0.002 ± 0.000	65.4	35.95
	M.lo183	183		0.325 ± 0.005	0.012 ± 0.001	70.7	51.45
	AHA_all	311		0.360 ± 0.005	0.074 ± 0.005	83.0	83.45
L1 → AHA	L_AHA.1	240	62	0.442 ± 0.025	0.247 ± 0.113	79.1	66.20
	L_AHA.2	240		0.384 ± 0.013	0.090 ± 0.043	66.9	66.20
	L_AHA.3	240		0.342 ± 0.008	0.034 ± 0.033	57.4	66.20
	L_AHA.4	240		0.323 ± 0.004	0.003 ± 0.007	54.1	66.20
	L_AHA.5	240		0.311 ± 0.005	0.002 ± 0.010	51.8	66.20
AHA → AHA	RS[4]	249	62	0.362 ± 0.009	0.032 ± 0.016	83.8	68.45

[1]L1 → L1 denotes genotypes of L1 reference animals used to impute genotypes of L1 study animals; AHA → L1 denotes genotypes of AHA reference animals used to impute genotypes of L1 study animals; L1 → AHA and AHA → AHA following the same naming convention.

[2]For L1 → L1, HS = paternal half-sibs; SGS = ancestral sires; ALL = HS and SGS combined; and RS = random samples. For AHA → L1, P denotes use of pedigree relationship; M denotes use of molecular relationship; hi denotes most related; lo denotes least related; and numerical suffix denotes sample size. For L1 → AHA, L_AHA with numerical suffix identifies the sample. For AHA → AHA, RS = random samples.

[3]Least squares means of average imputation accuracy across reduced panel designs. Standard errors of these least squares means are all 0.14.

[4]Average of 10 replicates with 5 technical repeats per replicate.

were used to impute genotypes for the remaining 62 AHA animals. Genotyping schemes that used random sampling were replicated 10 times with a new random sample of animals selected each time. These sampling strategies enabled the examination of changes in imputation accuracy due to differences in relationship, size, and structure of the reference population.

Cost of Genotyping

Genotyping costs differ depending, in part, on the number of SNP. To reflect current costs on an Illumina BovineSNP50 Beadchip and a reduced panel, relative costs were set to US$125 and US$50, respectively. Both SNP_MAF and SNP_MS were customized reduced panels for the Hereford breed. It was recognized that custom panels may cost more than an available commercial panel. However, the cost of all reduced panels was assumed to be the same in this study.

Imputation

Imputation was performed with the LinkPHASE and DAGPHASE package (Druet and Georges, 2010), following the procedure of Zhang and Druet (Zhang and Druet, 2010). Using all SNP on the reference individuals and reduced set of SNP on study individuals, haplotypes were constructed partially based on linkage and Mendelian segregation rules. Imputation was then performed iteratively to construct complete haplotypes based on linkage and LD. For individuals that had genotyped parent(s), linkage information was also used. Population LD was used for individuals that did not have genotyped parent(s).

The LinkPHASE and DAGPHASE package (Druet and Georges, 2010) forms a directed acyclic graph to conduct localized haplotype phasing and uses a hidden Markov model to sample haplotypes conditional on the individual's genotype. Scale and shift parameters control complexity of the model. These 2 parameters were set to 1 and 0.05 to fit the sample sizes and marker density in this study. Twenty iterations were performed for each imputation process. Genotypes on each chromosome were imputed independently. The physical map was converted to a genetic map assuming that 1 Mb equaled 1 cM. Five imputation technical repeats were used for each genotyping scheme and reduced panel.

Statistical Analysis

Accuracy of imputation was calculated as the correlation of imputed and masked genotypes. For a few genotypes (<0.1%) that were missing before being masked, real genotypes were considered the same as imputed genotypes. The Proc GLM method (SAS Inst. Inc., Cary, NC) was used to test the significance of different factors affecting accuracy of imputation. An initial analysis evaluated fixed effects of genotyping scheme (n = 23, Table 2) and panel design (n = 4, SNP_space, SNP_3K, SNP_MAF, and SNP_MS). Subsequently, panel design (n = 4), origin of reference samples (n = 2; AHA and L1), reference sample size, and genetic relationship were tested as separate fixed effects. Size of reference sample and genetic relationship were treated as linear covariates. Separate analyses were performed using pedigree and molecular relationships. In total, 820 imputation runs were evaluated. Multiple comparisons were performed using Tukey's range test.

RESULTS AND DISCUSSION

The model including genotyping scheme and panel design explained essentially all (99.8%) of the variation in accuracy of imputation observed in this study. When effect of genotyping scheme, conditional on panel design, was further partitioned to study main effects of origin of reference samples, number of reference samples, and pedigree-based (molecular-based) relationship, 95.8% (97.1%) of the variation in accuracy of imputation was explained. Thus, we concluded that interactions among these sources of variation were not important for the accuracy of imputation. All of the modeled effects were significant ($P < 0.01$).

Imputation accuracy for each genotyping scheme was reported as least squares means across the 4 reduced panels (Table 2). The genotyping schemes in which L1 genotypes were used to impute other L1 genotypes had the greatest accuracy (94.0 to 96.5%) among the schemes evaluated. When the AHA individuals were randomly selected and used to impute the genotypes for the remaining AHA individuals, relatively high accuracy was achieved.

On average, imputation accuracies were 69.1% and 72.5%, using L1 as reference samples with pedigree and molecular relationships, respectively. Corresponding imputation accuracies were 92.6% and 89.9%, when using the more genetically diverse AHA as reference samples (Table 3). The imputation method adopted here involved a process of sampling haplotypes from the reference sample. When reference samples came from L1, a smaller number of haplotypes were found. If these haplotypes did not represent the haplotypes in study samples, then imputation accuracy decreased. According to records from the AHA herd book, AHA was an admixed population that has been influenced by Line 1 Hereford. In 1984, 57% of the proven bulls recorded in AHA's Sire Summary were of predominantly Line 1 ancestry (Dickenson, 1984). However, a substantial portion of the AHA sample also has little, if any, pedigree relationship to L1 (Table). Thus, AHA was a sufficient reference population for imputation of L1 genotypes. However, L1 was a less adequate reference population for imputation of genotypes of AHA animals.

Table 3. Accuracy of imputation expressed as least squares means of correlation (× 100%) ± SE of imputed and masked genotypes for reference panel origin and reduced panel design, using molecular, pedigree, and genotyping scheme models

	Model		
Effect	Genotyping scheme[1]	Pedigree[2]	Molecular[3]
Reference origin[4]			
L1	-	69.07 ± 0.19[a]	72.46 ± 0.14[a]
AHA	-	92.58 ± 0.16[b]	89.92 ± 0.12[b]
Reduced panel[5]			
SNP_space	76.61 ± 0.05 [a]	80.94 ± 0.18 [a]	81.31 ± 0.15[a]
SNP_3K	76.44 ± 0.05[b]	80.77 ± 0.18 [a]	81.14 ± 0.15[a]
SNP_MAF	74.82 ± 0.05[c]	79.15 ± 0.18[b]	79.52 ± 0.15[b]
SNP_MS	78.10 ± 0.05[d]	82.43 ± 0.18[c]	82.80 ± 0.15[c]

[1]Genotyping scheme model denotes linear model with fixed effects of genotyping scheme (n = 23, Table 2) and panel design (n = 4, SNP_space, SNP_3K, SNP_MAF, and SNP_MS).

[2]Pedigree Model denotes linear model with fixed effects of panel design (n = 4), origin of reference samples (n = 2; AHA and L1), reference sample size, and pedigree relationship. Size of reference sample and genetic relationship were treated as linear covariates.

[3]Molecular model denotes linear model with fixed effects of panel design (n = 4), origin of reference samples (n = 2; AHA and L1), reference sample size, and molecular relationship. Size of reference sample and genetic relationship were treated as linear covariates.

[4]L1 = Line 1 Hereford and AHA = American Hereford Association animals.

[5]Panel designs: SNP_space included every 15th SNP; SNP_3K included SNP on the Illumina Bovine3K Beadchip (San Diego, CA); SNP_MAF included 2,643 SNP with greatest MAF; and SNP_MS included SNP with MAF >0.35 and were approximately evenly spaced. MAF = minor allele frequency.

[a–d]Least squares means of correlation of imputed and real genotype (%) within a column and within either reference origin or reduced panel with different superscripts differ ($P < 0.05$).

The size of AHA reference panels used to impute L1 genotypes ranged from 62 to 311 animals. Regression of imputation accuracy on size of the reference panel conditional on pedigree relationship was 0.08 ± 0.002. Regression of imputation accuracy on size of the reference panel conditional on molecular relationship was 0.06 ± 0.001. Thus, if relationship between the reference and study samples could be held constant, adding 100 animals to the reference panel would be expected to increase the accuracy of imputation by 6 to 8%. When using L1 as a reference population to impute genotypes for L1, the marginal utility of adding reference animals was <50% of that observed above.

Average pedigree and molecular relationships between reference and study sample were obtained for each of the investigated genotyping schemes (Table 1). Using an AHA reference panel to impute L1 genotypes conditional on the size of the reference panel, regressions of imputation accuracy on pedigree and molecular relationships were 9.2% ± 0.13% and 17.9% ± 0.15% per 0.1 increase in relationship, respectively. When the entire set of L1 genotypes was used to impute AHA genotypes, regressions of imputation accuracy on pedigree and molecular relationships were 10.8% ± 0.22% and 21% ± 0.081% per 0.1 increase in relationship, respectively. These findings confirm those of Zhang and Druet (2010) who found that imputation error decreased as the relationship between reference and study samples of Dutch Holstein cattle increased. This study examined alternative methods for identification of animals based on pedigree and molecular relationships to be used in reference or study panels for imputation. These 2 relationships were numerically different and measured on different scales. However, they were highly correlated (r = 0.93). Pedigree relationship uses the numerator relationship matrix to estimate the probability of identity by descent among animals and ranges from 0 to 1. Molecular relationship used the realized relationship matrix of the genotyped individuals to measure identity by state over all markers and ranges from –1 to 1. In this study, the molecular relationship provided greater resolution for predicting animals that have high or low imputation error than did pedigree relationship.

Average pedigree (molecular) relationships and maximum pedigree (molecular) relationships were also estimated for study animals. Maximum relationships indicated the highest relationship with one animal in the reference group. In cross-population imputation, the correlations of maximum pedigree (molecular) relationships with average pedigree (molecular) relationships were 0.88 to 1.00 (0.31 to 0.85). In this study, only average relationships were used to select subsets of animals in imputation, based on the assumption that using average relationships would result in good imputation accuracy for the overall study sample, which was of greater interest here. In L1 and AHA populations, a large correlation was found on these 2 relationships when they were estimated based on pedigree information. However, molecular relationship indicated the average and maximum relationship with the reference group can have large discrepancy. Further study is needed to determine the tradeoffs of using average or maximum relationships for optimal imputation accuracy.

Reduced Panels

The 4 reduced panels differed in the spacing and MAF of SNP, and shared 198 to 542 SNP (Table 1). Comparisons of least squares means of imputation accuracy across all genotyping schemes were shown in Table 3. A similar pattern of changes in accuracy of imputation among reduced panels was found when either molecular or pedigree relationships were used.

Except for SNP_space and SNP_3K, the reduced panels had significantly different accuracy of imputation ($P < 0.01$; Table 3). The commercial SNP_3K panel was designed based on approximately equal spacing and high MAF of the selected SNP across different breeds of beef and dairy cattle. The SNP_MAF reduced panel had the

least imputation accuracy. However, SNP with the greatest MAF were not evenly spaced across the genome, as indicated by the SE of the average interval between SNP being 3.6 times greater for SNP_MAF than SNP_space (Table 1). The SNP_MS reduced panel had the greatest accuracy in imputation (Table 3). Despite using a similar method in selecting a subset of SNP, SNP_MS outperformed SNP_3K by 1.66% imputation accuracy across all genotyping schemes. An arbitrary threshold on MAF was used in SNP_MS so that the average MAF was slightly less than SNP_MAF and the SE of the interval between adjacent SNP was similar to that of SNP_3K. Previous studies indicated that reduced panels of evenly spaced SNP have low rates of error in imputation (Weigel et al., 2010; Zhang and Druet, 2010). Thus, it was concluded that reduced SNP panels should contain evenly spaced SNP and that there would be some marginal benefit from a custom-designed panel with high MAF for the population of interest.

Comparisons of Genotyping Cost of Different Genotyping Schemes

Because study samples had similar size, increases in genotyping costs were mainly due to increase of reference panel sizes (Table 2). It was possible to design a cost-effective genotyping scheme by taking into account genetic relationships and reference population origins. For example, with similar imputation accuracy, total cost for genotyping the 62 most related individuals (**P.hi62**) was estimated at 25% of the cost for genotyping all 311 animals (**AHA_all**) with the high density panel. Therefore, P.hi62 will be a more cost-effective design than AHA_all in the imputation study shown here.. Similarly, with the same genotyping cost, imputation accuracy decreased 25% from Sample 1, which had the greatest average molecular relationship between L1 and AHA (i.e., L_AHA.1), to Sample 5, which had the least average molecular relationship between L1 and AHA (i.e., L_AHA.5). Thus, L_AHA.1 was a superior genotyping design, compared with L_AHA.5.

Conclusions

The low density SNP panel used as a frame of reference in this study has been replaced in the marketplace by an augmented panel with approximately twice the number of SNP. The increased number of markers would be expected to improve accuracy of imputation. However, general conclusions reached here regarding even spacing of SNP, minor allele frequency, size of reference population, and genetic relationship between reference and study populations are believed to remain valid. Previous studies of accuracy of imputation in dairy cattle (Weigel et al., 2010; Zhang and Druet, 2010) have found results similar to those observed here using samples from L1 to impute genotypes for other L1 animals, albeit with substantially larger reference populations. It seems apparent that to obtain accurate imputation, the reference panel should contain some individuals with close relationships to those whose genotypes are being imputed. In experimental populations composed entirely of related animals, the number of animals genotyped for the reference panel may be fairly small. However, reference panels used across an entire breed need to be sufficiently large so as to contain individuals representing the vast majority of haplotypes that characterize the breed. For experimental populations, use of a generic reduced panel composed of evenly spaced markers may not greatly compromise accuracy of imputation relative to a custom-designed panel.

LITERATURE CITED

Daetwyler, H. D., G. R. Wiggans, B. J. Hayes, J. A. Woolliams, and M. E. Goddard. 2011. Imputation of missing genotypes from sparse to high density using long-range phasing. Genetics 189:317–327.

Dickenson, H. H. 1984. The influence of Line 1 in the Hereford breed. Fort Keogh Livestock and Range Research Station Field Day Report. USDA-ARS, Miles City, MT.

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

Henderson, C. R. 1976. Simple method for computing inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–83.

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, L., Y. Li, A. B. Singleton, J. A. Hardy, G. Abecasis, N. A. Rosenberg, and P. Scheet. 2009. Genotype-imputation accuracy across worldwide human populations. Am. J. Hum. Genet. 84:235–250.

Kuehn, L. A., J. W. Keele, G. L. Bennett, T.G. McDaneld, T. P. L. Smith, W. M. Snelling, T. S. Sonstegard, and R. M. Thallman. 2011. Predicting breed composition using breed frequencies of 50,000 markers from the US Meat Animal Research Center 2,000 Bull Project. J. Anim. Sci. 89:1742–1750.

MacNeil, M. D. 2009. Invited review: Research contributions from seventy-five years of breeding Line 1 Hereford cattle at Miles City, Montana. J. Anim. Sci. 87:2489–2501.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One 4:e5350.

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

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.

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

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marcais, 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.