Published December 4, 2014

Predicting breed composition using breed frequencies of 50,000 markers from the US Meat Animal Research Center 2,000 Bull Project[1,2]

L. A. Kuehn,*[3] J. W. Keele,* G. L. Bennett,* T. G. McDaneld,* T. P. L. Smith,* W. M. Snelling,* T. S. Sonstegard,† and R. M. Thallman*

*US Meat Animal Research Center, USDA, ARS, Clay Center, NE 68933; and †Bovine Functional Genomics Laboratory, USDA, ARS, Beltsville, MD 20705

ABSTRACT: Knowledge of breed composition can be useful in multiple aspects of cattle production, and can be critical for analyzing the results of whole genome-wide association studies currently being conducted around the world. We examine the feasibility and accuracy of using genotype data from the most prevalent bovine genome-wide association studies platform, the Illumina BovineSNP50 array (Illumina Inc., San Diego, CA), to estimate breed composition for individual breeds of cattle. First, allele frequencies (of Illumina-defined allele B) of SNP on the array for each of 16 beef cattle breeds were defined by genotyping a large set of more than 2,000 bulls selected in cooperation with the respective breed associations to be representative of their breed. With these breed-specific allele frequencies, the breed compositions of approximately 2,000 two-, three-, and four-way cross (of 8 breeds) cattle produced at the US Meat Animal Research Center were predicted by using a simple multiple regression technique or Mendel (http://www.genetics.ucla.edu/software/mendel) and their genotypes from the Illumina BovineSNP50 array, and were then compared with pedigree-based estimates of breed composition. The accuracy of marker-based breed composition estimates was 89% when using either estimation method for all breeds except Angus and Red Angus (averaged 79%), based on comparing estimates with pedigree-based average breed composition. Accuracy increased to approximately 88% when these 2 breeds were combined into an aggregate Angus group. Additionally, we used a subset of these markers, approximately 3,000 that populate the Illumina Bovine3K (Illumina Inc.), to see whether breed composition could be estimated with similar accuracy when using this reduced panel of SNP makers. When breed composition was estimated using only SNP in common with the Bovine 3K array, accuracy was slightly reduced to 83%. These results suggest that SNP data from these arrays could be used to estimate breed composition in most US beef cattle in situations where pedigree is not known (e.g., multiple-sire natural service matings, non-source-verified animals in feedlots or at slaughter). This approach can aid analyses that depend on knowledge of breed composition, including identification and adjustment of breed-based population stratification, when performing genome-wide association studies on populations with incomplete pedigrees. In addition, SNP-based breed composition estimates may facilitate fitting cow germplasm to the environment, managing cattle in the feedlot, and tracing disease cases back to the geographic region or farm of origin.

Key words: breed composition, cattle, single nucleotide polymorphism

J. Anim. Sci. 2011. 89:1742–1750
doi:10.2527/jas.2010-3530

[1]The authors acknowledge scientists at the Animal Improvement Program Laboratory, USDA, ARS, in Beltsville, MD, and at the Ft. Keogh Livestock and Range Research Laboratory, USDA, ARS, Miles City, MT, for contributing allele frequency data on dairy and Hereford Line 1 cattle. In addition, they acknowledge, at the US Meat Animal Research Center, USDA, ARS, L. Flathman, S. Simcox, and A. Bertles for technical assistance with genotyping and J. Watts for manuscript preparation.

[2]Mention of a trade name, proprietary product, or specific equipment does not constitute a guarantee or warranty by the USDA and does not imply approval to the exclusion of other products that may be suitable.

[3]Corresponding author: Larry.Kuehn@ars.usda.gov

Received September 22, 2010.
Accepted January 19, 2011.

INTRODUCTION

Knowledge of breed composition of cattle would be useful for predicting heterosis (Dickerson, 1973), evaluating adaptability to production environments, and sorting animals into management groups. In addition, estimates of breed composition of crossbred commercial cattle would be useful for mapping loci underlying economically important traits and host resistance to disease. If it were possible to estimate breed composition accurately, unselected control allele frequencies matching a sample of cases (sick animals) could be derived statistically based on allele frequencies from reference samples for each breed potentially contributing to cases. Breed composition is useful for tracing the history of an animal from birth for the purposes of tracking disease transmission and sources of contamination in meat (e.g., DNA forensics, as in Wasser et al., 2008).

Breed identification has been performed in cattle (e.g., Watanabe et al., 2008) and other species (e.g., dogs in Parker et al., 2004) by using microsatellite alleles in specific genomic regions. Methods have been established to predict breed composition or contributions of ancestral populations (e.g., Mendel; Lange et al., 2001). However, predicting breed composition in advanced generations of outcrossed populations is more difficult with limited markers because unique breed alleles are not necessarily passed on to advanced generations.

Here we show that breed composition of cattle can be accurately predicted from allele frequencies estimated from a diverse sample of prominent reference breeds. The allele frequencies of these reference breeds are provided. It is important that the germplasm of the cattle being predicted is represented in the reference breeds. Apparent discrepancies between pedigree-determined breed composition and predictions based on SNP markers are not just a function of errors in prediction; variation attributable to chromosomal sampling around the average (pedigree) also contributes.

MATERIALS AND METHODS

The DNA samples used in this project were either from semen samples provided by national breed associations or from animals raised in conformation with the Guide for the Care and Use of Agricultural Animals in Agricultural Research and Teaching (FASS, 1999), and their care was approved by the US Meat Animal Research Center (**USMARC**) Animal Care and Use Committee.

Bulls Used to Estimate Breed-Specific Allele Frequencies

Semen samples from 2,235 bulls chosen to be representative of their breed by 16 breed associations were used to obtain DNA from which to estimate allele frequencies by breed. A portion of this total (approximately 40 animals from Beefmaster, Hereford, Limousin, and Red Angus) include bulls sampled by The Bovine HapMap Consortium (2009). Bulls were chosen by the US beef cattle breed associations, and semen (DNA source) was provided to the USMARC in 2008 and 2009. Breed associations were fully responsible for the choice of bulls they felt were representative; no constraints on relatedness, progeny numbers, or EPD accuracy were imposed. Representation of each breed increased with its contribution to the national herd, but smaller breeds were more represented in the sample than in the national herd. In addition, Hereford was more highly represented than Angus because of the contribution of Line 1 Hereford bulls from the USDA, ARS Ft. Keogh Livestock and Range Research Laboratory (Miles City, MT). The total numbers of bulls sampled per breed are listed in Table 1. Each bull was genotyped using the Illumina BovineSNP50 array (Matukumalli et al., 2009; Illumina Inc., San Diego, CA).

Table 1. Numbers of bulls sampled per breed in the US Meat Animal Research Center 2,000 Bull Project

Breed	No.
British-derived breeds	
Angus	403
Hereford	491[1]
Red Angus	175
Shorthorn	86
Continental European-derived breeds	
Braunvieh	27
Chianina[2]	47
Charolais	125
Gelbvieh	146
Limousin	141
Maine-Anjou	59
Salers	41
Simmental	254
US-derived breeds	
Beefmaster	65
Brahman	53
Brangus	68
Santa Gertrudis	54

[1]One hundred eighty Hereford bulls from the Line 1 pedigree based at USDA, ARS, Ft. Keogh Livestock and Range Research Laboratory, Miles City, MT.

[2]Most are Chianina × Angus composites (Chiangus), with variable amounts of base breed percentages (most less than 50% Chianina).

This sample of bulls and their genotypic data have been informally termed the USMARC 2,000 Bull Project. In general, the aim of the 2,000 Bull Project was to provide a conduit to transfer research results to the beef cattle industry. In principle, estimates of SNP associations can be combined with genotypes from the 2,000 Bull Project to provide predicted genetic merit for traits not routinely collected by industry, such as feed intake and resistance to disease. An additional benefit of this project was to provide a sample of highly representative bulls from the industry for allelic and haplotypic frequencies. By using the allele frequencies we report herein, in the future it will be possible to quickly estimate the variance contributed by newly dis-

covered QTL to establish the relevance of the finding to the US beef industry.

Genotyping was successful for 52,156 markers using the Illumina BovineSNP50 array. Allele frequencies were calculated using a simple allele counting approach [(number of copies of Illumina allele B)/(2 × number of animals)], ignoring relationships among bulls. Previous phylogenetic analyses indicate that some breeds are more diverse relative to other breeds (Decker et al., 2009; The Bovine HapMap Consortium, 2009). Correlations of breed frequencies among bulls were used to plot the genetic distances between all pairs of breeds, using the Ape package in R (http://cran.r-project.org/web/packages/ape/index.html), to determine cases in which some breeds may be difficult to separate because of greater commonality in allele frequencies. This turned out to be the case for discriminating between Angus and Red Angus (see Results and Discussion section).

Crossbred Test Population

To validate whether breed composition could be accurately predicted using the allelic frequencies derived from the USMARC 2,000 Bull Project, a population with known breed compositions and genotyped for the same markers needed to be identified.

Steers and heifers from Cycle VII of the USMARC Germplasm Evaluation Project had previously been genotyped using the Illumina BovineSNP50 BeadChip (Snelling et al., 2010). Cycle VII was initiated by sampling 149 bulls (22 Angus, 21 Hereford, 21 Red Angus, 22 Charolais, 23 Gelbvieh, 20 Limousin, and 20 Simmental sires) and mating them to Angus, Hereford, or MARC III composite (1/4 Angus, 1/4 Hereford, 1/4 Pinzgauer, 1/4 Red Poll) females to produce F_1 progeny. These F_1 progeny were subsequently mated in multiple-sire mating groups to produce 2-, 3-, and 4-breed cross progeny termed F_1^2 (i.e., crosses of F_1); average breed proportions were therefore multiples of 25%. BovineSNP50 genotypes were obtained on the original purebred bulls, 73 F_1 sires, and 2,014 F_1^2 progeny. Paternity was assigned using BovineSNP50 results; consequently, the expected pedigree breed composition of F_1^2 progeny was established (assuming dams were identified correctly). Indeed, pedigree selection (Mendel; Lange et al., 2001) analysis showed that dams were correctly identified for more than 99% of the F_1^2 cattle (data not shown).

Red Poll and Pinzgauer (component breeds of the MARC III composite) were not part of the USMARC 2,000 Bull Project. However, approximately one-half of the F_1^2 progeny had a MARCIII maternal granddam. Having a source of germplasm that was not in our reference population would potentially bias the resulting estimates of breed composition. Therefore, the frequencies of Red Poll × Pinzgauer alleles were estimated using genotypic data from Cycle VII F_1 steers and heifers (364) with MARC III parents. Allele frequencies were estimated with a generalized linear model with a binomial family and logistic link function using the glm function in R (derived from McCullagh and Nelder, 1989). The analysis was run separately for each SNP. The model was

$$\mathbf{Y} = f(\mathbf{XB}) + \mathbf{e},$$

where $\mathbf{Y}$ is a matrix of allele counts for each breed cross (summed across animals of that cross). Matrix $\mathbf{Y}$ contained a row for each breed cross and a column for each allele. The logistic function is represented by $f(x)$. The matrix $\mathbf{X}$ contains the fraction of breed in each breed cross, $\mathbf{B}$ are the regression coefficients for each allele by breed type, and $\mathbf{e}$ is the residual. Frequency was computed from $\mathbf{B}$ using the logit function (cumulative density function for the logistic distribution).

Statistical Analysis

To predict breed composition using the regression approach (Chiang et al., 2010), genotypes of F_1^2 animals were converted to copies of Illumina allele B. Each genotypes of an individual animal ($\mathbf{y}$; copies of allele B divided by 2; 0, 0.5, or 1) were then predicted using the following model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e},$$

where $\mathbf{X}$ is a 52,156 × 17 (16 breeds plus MARC III) matrix of frequencies of allele B for each breed, $\mathbf{b}$ is a vector of regression coefficients representing the percentage contribution of each breed to the animal in $\mathbf{y}$, and $\mathbf{e}$ is a random residual vector. In addition, we estimated breed contributions in the F_1^2 animals using the ethnic admixture option (option 15, model 2) of Mendel (http://www.genetics.ucla.edu/software/mendel; Lange et al., 2001).

Based on preliminary analyses, composite breeds such as Brangus (3/8 Brahman, 5/8 Angus), Beefmaster (1/2 Brahman, 1/4 Hereford, 1/4 Shorthorn), and Santa Gertrudis (3/8 Brahman, 5/8 Shorthorn) would be predicted to have nonzero values in the regression approach in cases where their component breed should be represented. For instance, an animal that was 1/4 Angus may be predicted to be approximately 16% Brangus and approximately 14% Angus. The same animal would then generally be predicted to be approximately −7% Brahman to compensate. As a result, composite breeds were left out of the resource population in this analysis if their ancestral purebred populations were present.

All the bulls used to produce the F_1 parents of the F_1^2 population were part of the USMARC 2,000 Bull Project. To avoid the potential of these grandsires producing results more favorable than expected from a sample of the breeds, these 149 bulls were not included in the breed frequency calculation for this demonstration.

Recently, a reduced marker panel with approximately 3,000 markers has been released (Bovine3K, Illumina

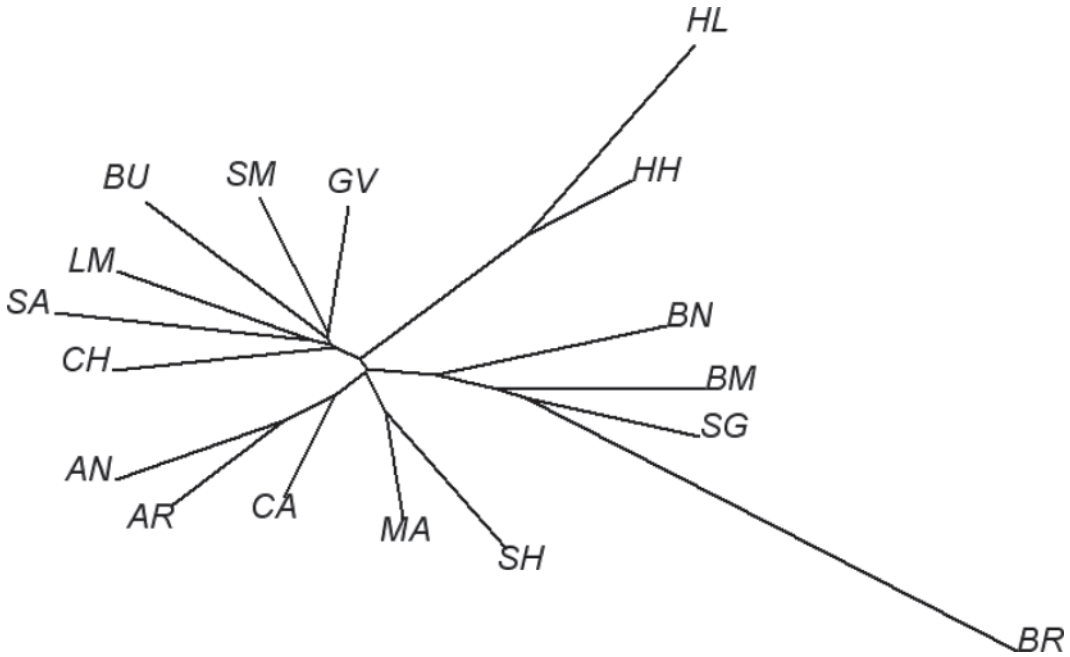

Figure 1. Genetic distance between breeds as estimated by the correlations among frequencies for markers on Illumina BovineSNP50 (Illumina BovineSNP50 BeadChip, Illumina Inc., San Diego, CA). Breeds are Angus (AN), Hereford (HH), Line 1 Hereford (HL), Red Angus (AR), Shorthorn (SH), Braunvieh (BU), Chianina (CA), Charolais (CH), Gelbvieh (GV), Limousin (LM), Maine-Anjou (MA), Salers (SA), Simmental (SM), Beefmaster (BM), Brahman (BR), Brangus (BN), and Santa Gertrudis (SG).

Inc.; http://www.illumina.com/documents/products/datasheets/datasheet_bovine3k.pdf). All the markers on this panel are also on the Illumina BovineSNP50. To compare the accuracy of breed identification in this reduced panel relative to the BovineSNP50, we reduced the marker set to those on the Bovine3K and predicted breed composition of the F_1^2 population using the breed frequencies.

The resulting breed percentages from both the regression and Mendel methods were evaluated for accuracy in the trial by regressing the estimated breed percentage for each of the 7 breeds (excluding MARC III) in the F_1^2 population on their pedigree-based breed fraction. The resulting regression coefficients and the proportion of variance explained (R^2) are reported.

RESULTS AND DISCUSSION

Frequency estimates for each breed in the USMARC 2,000 Bull Project for all 52,156 markers are contained in Supplemental Table 1 (http://jas.fass.org/content/vol89/issue6/). In addition to the breeds from the USMARC 2,000 Bull Project, we report the frequencies of these SNP markers for 3 dairy breeds (Holstein, Jersey, Brown Swiss; provided by the Animal Improvement Program Laboratory, USDA, ARS, Beltsville, MD). These additional breeds increase the utility of these frequencies.

Observed genetic distances between breeds are shown as a phylogenetic tree in Figure 1 and were similar to those reported for specific chromosomes by The Bovine HapMap Consortium (2009) and in the phylogenic analysis performed by Decker et al. (2009). Correlations used to form this tree are reported in Supplemental Table 2 (http://jas.fass.org/content/vol89/issue6/). As expected from these previous analyses, Brahman was the most distant from all the other breeds in the 2,000 Bull Project. Composites containing Brahman were generally intermediate between Brahman and the cluster of *Bos taurus* breeds. Continental European breeds generally were in a cluster, with the exception of Maine-Anjou, which was most similar to Shorthorn. Hereford was surprisingly distant from all the other *B. taurus* breeds; some of this distance is likely to be an ascertainment bias in SNP discovery (which was based in part on the Hereford draft sequence data; Matukumalli et al., 2009). However, there is a possibility that it is truly due to a distant evolutionary relationship. Whether artificial or real, this distance of Hereford from the other breeds likely would increase the accuracy of estimating the percentage of Hereford in crossbred animals when using the BovineSNP50 genotyping array.

Squared correlations between predicted breed composition and pedigree-derived compositions ranged between 77 and 92% for both the Mendel and regression methods (Table 2) when predicting breed composition using all the markers on the BovineSNP50. Estimates of breed composition from SNP based on the regression or Mendel method were highly correlated, at approximately 99% for each breed. The only noticeable difference between the 2 methods was that the regression resulted in some negative estimates of breed percentages.

Both Red Angus and Angus pedigree compositions were predicted least accurately, as measured by both the R^2 and the regression coefficient. We hypothesized that this result was due to the genetic similarity be-

Table 2. Results from regressing the estimated breed percentage using Illumina BovineSNP50[1] and the pedigree breed percentage

	Regression method			Mendel[2] method		
Breed	Intercept	Regression	R^2	Intercept	Regression	R^2
Angus	−0.014 ± 0.003	0.737 ± 0.008	0.798	−0.013 ± 0.003	0.727 ± 0.008	0.811
Red Angus	0.059 ± 0.001	0.883 ± 0.011	0.772	0.060 ± 0.001	0.869 ± 0.010	0.796
Aggregate Angus[3]	−0.012 ± 0.003	0.917 ± 0.007	0.882	−0.008 ± 0.003	0.902 ± 0.007	0.885
Hereford	0.015 ± 0.002	0.981 ± 0.006	0.920	0.007 ± 0.002	0.985 ± 0.006	0.920
Limousin	0.013 ± 0.001	0.925 ± 0.008	0.880	0.014 ± 0.001	0.893 ± 0.007	0.902
Charolais	0.013 ± 0.001	0.873 ± 0.007	0.879	0.012 ± 0.001	0.839 ± 0.006	0.913
Gelbvieh	0.021 ± 0.001	0.922 ± 0.007	0.898	0.021 ± 0.001	0.918 ± 0.006	0.909
Simmental	0.015 ± 0.001	0.882 ± 0.006	0.905	0.016 ± 0.001	0.863 ± 0.006	0.922

[1]Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, CA).

[2]http://www.genetics.ucla.edu/software/mendel; Lange et al. (2001).

[3]The aggregate Angus group results from regression of the estimated Angus + Red Angus percentage on pedigree-derived percentages of Angus + Red Angus.

tween the 2 breeds. Red Angus and Angus were the most closely related pair of breeds in Figure 1, which is not surprising given their recent divergence into separate breed associations (Red Angus was established as a US breed in 1954; http://www.redangus.org). To examine the degree to which each breed was interfering with the estimate of the other, we examined the relationship between residuals (SNP-derived breed percentage minus pedigree-derived breed percentage) for each breed (Figure 2). Our results indicate that it is difficult to distinguish between Angus and Red Angus because of their similarity. The correlation between Angus and Red Angus residuals was −0.61, indicating a high incidence of substituting one related breed for another in the breed prediction process. These results led us to combine frequencies for Angus and Red Angus into an aggregate breed group (Table 2). When combined, the squared correlation between predicted breed composition and pedigree-derived composition of aggregate Angus increased to 88%.

Breed composition was underpredicted overall for each of the 7 breeds, as demonstrated by regression coefficients being less than 1 (Table 2). Figures 3 and 4 provide further demonstration of Angus and Red Angus as an aggregate breed; the mean of the predicted breed composition is less than the 45° line passing through the origin. These figures are representative of other breeds and are shown as examples in that other breeds in the F_1[2] population also display predicted mean breed compositions below the 45° line (data not shown). When comparing Figure 3 with Figure 4, some estimates of aggregate Angus breed percentage were negative under the regression method, whereas the minimum percentage estimate allowed was zero when using Mendel. Otherwise, both methods were strongly correlated. There was minimal overlap (<10%) in predictions among cattle with pedigree differences in breed composition greater than 25% (e.g., aggregate Angus at 0% vs. 25 to 32% in Figure 2). Extreme outliers (e.g., animal predicted to be 50% aggregate Angus with a pedigree of 0%) are likely because of incorrect pedigree (dam) assignment rather than inaccurate prediction.

Because of the decreased cost of genotyping using the Illumina Bovine3K BeadChip relative to the BovineSNP50 BeadChip, the set of markers on this chip were evaluated for their ability to predict breed composition relative to the BovineSNP50 BeadChip. Predicted breed compositions with the regression approach using only SNP restricted to the Illumina Bovine3K BeadChip were slightly less accurate than the 50K array; R^2 averaged 83% for 3K across breed relative to 89% for the 50K chip. Results using Mendel were similar to those of the regression although slightly greater, with an R^2 of 84% averaged across breeds. This R^2 was almost as great as the 89% observed for the 50K chip. The relatively small difference in accuracy between 3K and 50K (with 16× more markers) reflects the linkage disequilibrium (**LD**) among SNP. If the LD were less in the 50K chip and the 3K chip did not saturate the genomic information, we would expect to see a larger advantage in accuracy of the 50K chip relative to the 3K chip. In both cases, Red Angus and Angus were combined as aggregate Angus. Because of the decreased R^2 when using the Bovine3K markers, we attempted to increase the accuracy of prediction by including markers on the X chromosome in the prediction. In the case of this reduced panel, including markers in the X chromosome improved accuracy slightly for females (R^2 was 0.86 including X and was 0.85 if X was not included) but had no effect on male accuracy (R^2 was 0.84 including X and excluding X).

The accuracy values (R^2) were based on estimating breed composition with SNP data relative to pedigree-based composition. In evaluating the ability to determine breed composition from SNP genotype data, we assumed that pedigree records represent the "true" breed composition. Although this assumption is true for F_1 crosses, it is not necessarily the case for F_2 and 4-way cross cattle in our crossbred population. Because of crossover events and the chromosomal assortment

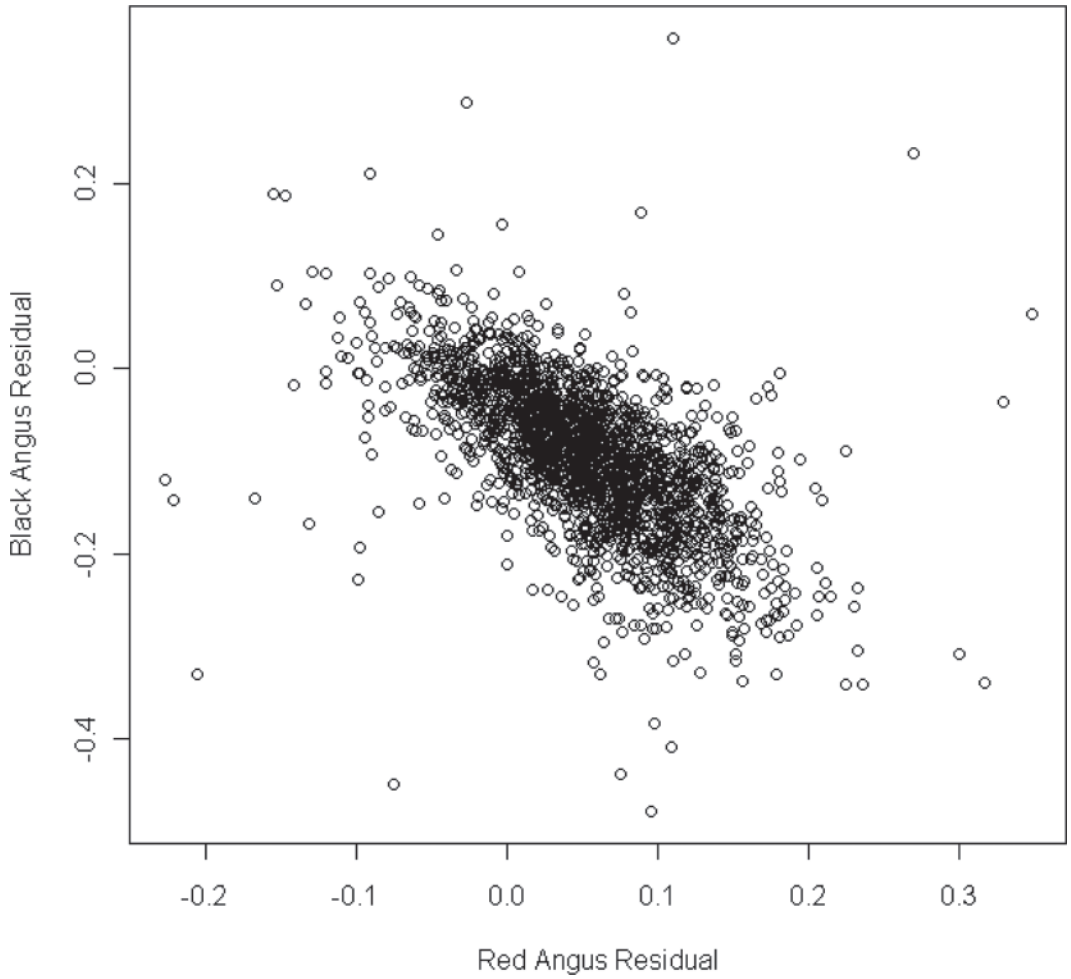

Figure 2. Residuals resulting from regressions of estimates of Angus breed percentages on pedigree Angus breed composition relative to residuals from regressions of estimates of Red Angus breed percentages on pedigree Red Angus breed percentages.

during gametogenesis in the F_1 parents of the subsequent generation, there will be variation in breed composition not captured by strict pedigree-based breed composition. This variable sampling explains part of the loss in R^2 from unity. To evaluate the contribution of chromosomal sampling to variation in breed composition, we simulated crossover intervals as exponential deviates with a mean of 1 morgan. Average breed composition based on pedigree accounted for 96% of the variation in individual breed composition when using simulation with chromosome lengths from the bovine genome assembly, build 4.0 (The Bovine Genome Sequencing and Analysis Consortium et al., 2009). Given the average R^2 of 89%, approximately 7% of the variation in predicted breed composition is the result of errors in estimation and pedigree errors.

In this study, we observed close agreement between the Mendel and regression methods ($R^2 = 0.99$). However, there are situations in which one or the other might be preferred. Mendel may be preferable because estimates of breed composition are always nonnegative and in the parameter space. On the other hand, there are scenarios in which negative regressions could be useful. For example, the unknown animal might actually be a cross between breed A and breed B, and the only available reference breed populations are breed A, breed B × C, and breed C. In this case, we would expect to estimate regression coefficients of 0.5 for breed A, 1.0 for breed B × C, and −0.5 for breed C, which would be an accurate representation of breed composition because the negative coefficient adjusts breed C out of the B × C cross contribution.

EigenStrat (http://genepath.med.harvard.edu/~reich/Software.htm; Price et al., 2006) and Structure (http://pritch.bsd.uchicago.edu/structure.html; Pritchard et al., 2000) are 2 other tools that can be used to characterize an unknown population structure by using individual animal genotypes. An advantage of the regression and Mendel methods is that we were able to attribute fractions of the breed composition of an individual animal to reference breeds summarized according to their allele frequencies. In our experience, EigenStrat correctly assigned animals to the correct cluster as long as the animals with the same breed composition (e.g., 1/2 Hereford, 1/2 Charolais) existed in the individual reference populations (data not shown). When animals with the same breed composition did not exist among the reference populations, even though the

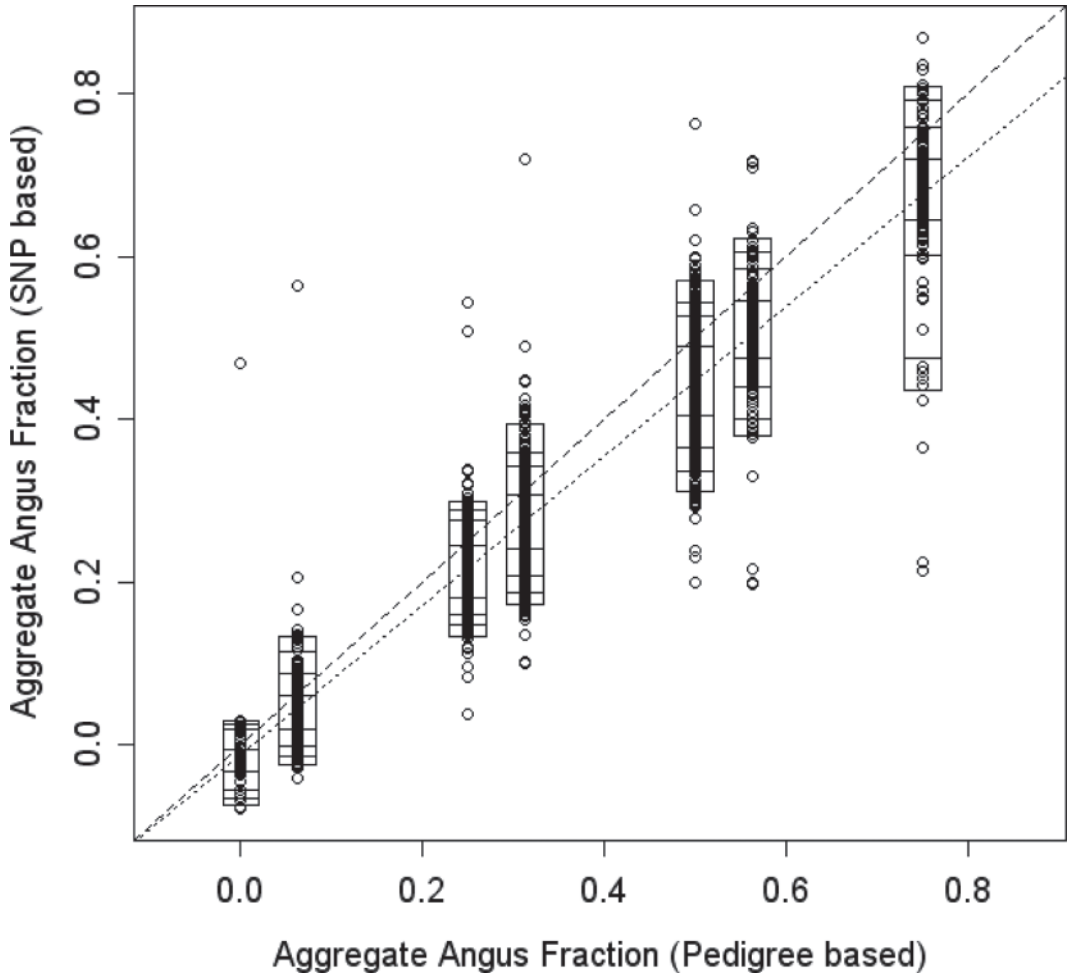

Figure 3. Graphical representation of SNP (BovineSNP50 BeadChip, Illumina Inc., San Diego, CA) estimated percentage of Angus or Red Angus (aggregate Angus), using the regression method, in the F_1^2 population relative to the pedigree-derived percentage of aggregate Angus. Box plots around pedigree-based fractional distributions represent percentiles of 2.5, 5, 10, 25, 75, 90, 95, and 97.5. The finely dashed line is the regression equation from predicting the SNP-derived aggregate Angus percentage using pedigree aggregate Angus, and the second line is a 45° line through the origin.

purebred ancestral breeds were present, the unknown animal did not necessarily cluster in a position of principal component space that permitted interpretation. We did not evaluate Structure because EigenStrat has similar capabilities while requiring substantially fewer computer resources (Price et al., 2006).

The allele frequency estimates reported here (Supplemental Table 1; http://jas.fass.org/content/vol89/issue6/), combined with the results of genome-wide association studies, will be useful for predicting heterosis (Abasht and Lamont, 2007; Kusterer et al., 2007) at small numbers of loci in LD with SNP. For the case of whole-genome heterozygosity, we report expected heterozygosity (averaged for all SNP) for all breeds and F_1 crosses (Supplemental Table 2; http://jas.fass.org/content/vol89/issue6/). These estimates predict specific heterosis (hybrid advantage) for each breed cross when heterosis is proportional to heterozygosity (Dickerson, 1973). We do not provide estimates of retained heterozygosity for advanced-generation composites formed from these breeds because of the huge number of possible breed combinations.

Feedlot operators need to allocate cattle to pen groups based on their expected level of performance; breed composition is likely one of the strongest genetic indicators of performance. Likewise, choosing an appropriate match between cow germplasm and the environment (e.g., arid, extensive, intensive, tropic) would be facilitated through knowledge of breed composition.

The estimates of allele frequencies by breed are useful for matching cases and controls in disease association studies (Homer et al., 2008). Allele frequencies for individual breeds can be used to identify and correct for population stratification in disease case control studies.

Breed composition is useful for tracing the herd of origin of an animal for the purposes of tracking disease transmission and sources of contamination in meat (e.g., DNA forensics as in Wasser et al., 2008). Tracing back disease cases (e.g., foot-and-mouth disease, bovine spongiform encephalopathy) to their geographical or farm origin would be facilitated through knowledge of breed composition. For instance, if a sample were identified as a positive case after slaughter, and the source of the animal was unknown but restricted to a

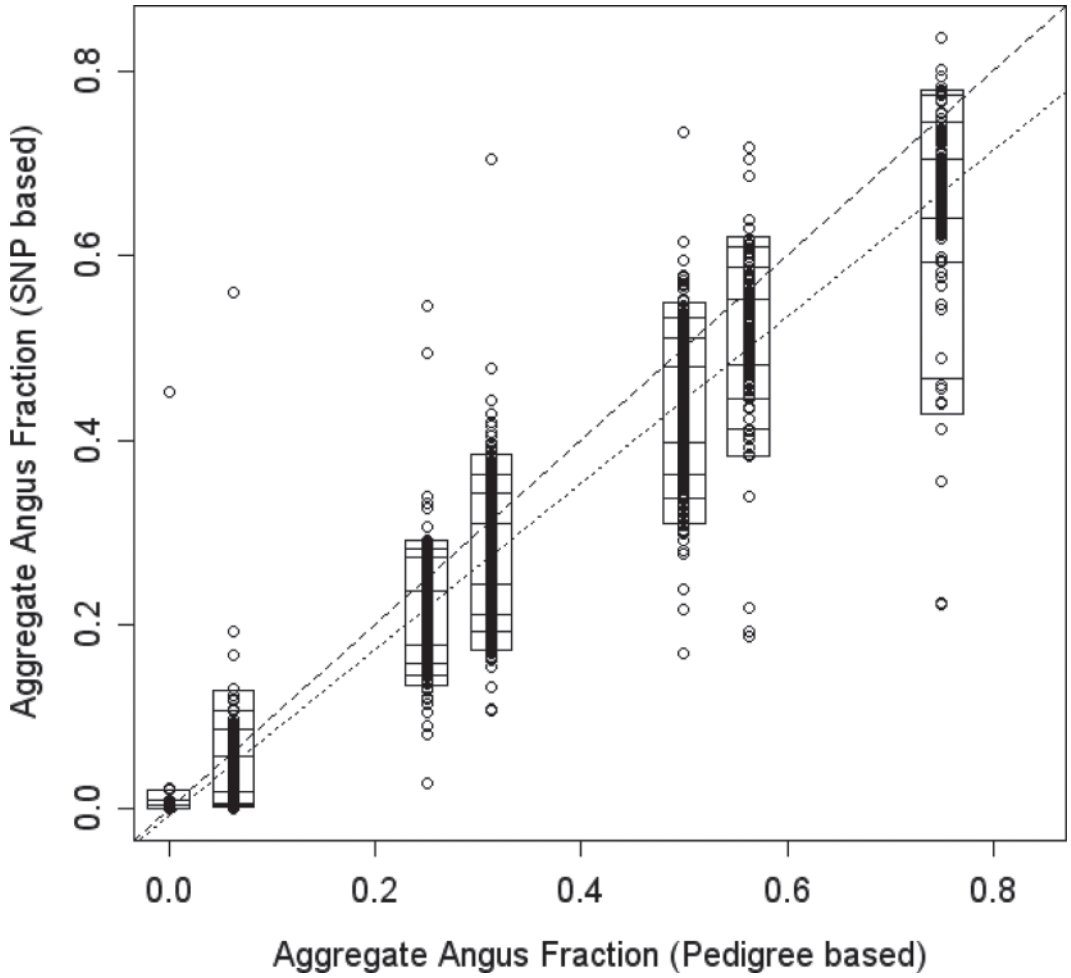

Figure 4. Graphical representation of SNP (BovineSNP50 BeadChip, Illumina Inc., San Diego, CA) estimated percentage of Angus or Red Angus (aggregate Angus), using Mendel (http://www.genetics.ucla.edu/software/mendel; Lange et al., 2001), in the F_1^2 population relative to the pedigree-derived percentage of aggregate Angus. Box plots around pedigree-based fractional distributions represent percentiles of 2.5, 5, 10, 25, 75, 90, 95, and 97.5. The finely dashed line is the regression equation from predicting the SNP-derived aggregate Angus percentage using pedigree aggregate Angus, and the second line is a 45° line through the origin.

candidate set of farms, the breed composition would reduce the number of candidate farms that would need to be screened for the disease (in candidate farms that differ in breed composition), saving money required to sample more animals and resulting in less collateral damage to disease-free farms that are suspected because of proximity.

These results indicate that breed frequencies predicted from a high-density SNP panel can be used to predict breed composition of crossbred animals. The breed frequency estimates were not all-encompassing relative to beef cattle breeds. More reference breeds should be added to this resource to facilitate accurate estimation of breed composition.

LITERATURE CITED

Abasht, B., and S. J. Lamont. 2007. Genome-wide association analysis reveals cryptic alleles as an important factor in heterosis for fatness in chicken F_2 population. Anim. Genet. 38:491–498.

The Bovine Genome Sequencing and Analysis Consortium, C. G. Elsik, R. L. Tellam, and K. C. Worley. 2009. The genome sequence of taurine cattle: A window to ruminant biology and evolution. Science 324:522–528.

The Bovine HapMap Consortium. 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science 324:528–532.

Chiang, C. W. K., Z. K. Z. Gajdos, J. M. Korn, F. G. Kuruvilla, J. L. Butler, R. Hackett, C. Guiducci, T. T. Nguyen, R. Wilks, T. Forrester, C. A. Haiman, K. D. Henderson, L. Le Marchand, B. E. Henderson, M. R. Palmert, C. A. McKenzie, H. N. Lyon, R. S. Cooper, X. Zhu, and J. N. Hirschhorn. 2010. Rapid assessment of genetic ancestry in populations of unknown origin by genome-wide genotyping of pooled samples. PLoS Genet. 6:e1000866.

Decker, J. E., J. C. Pires, G. C. Conant, S. D. McKay, M. P. Heaton, K. Chen, A. Cooper, J. Vilkki, C. M. Seabury, A. R. Caetano, G. S. Johnson, R. A. Brenneman, O. Hanotte, L. S. Eggert, P. Wiener, J. J. Kim, K. S. Kim, T. S. Sonstegard, C. P. Van Tassell, H. L. Neibergs, J. C. McEwan, R. Brauning, L. L. Coutinho, M. E. Babar, G. A. Wilson, M. C. McClure, M. M. Rolf, J. Kim, R. D. Schnabel, and J. F. Taylor. 2009. Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. Proc. Natl. Acad. Sci. USA 106:18644–18649.

Dickerson, G. E. 1973. Inbreeding and heterosis in animals. Pages 54–77 in Proc. Anim. Breeding Genet. Symp. in Honor of Dr. J. L. Lush. Am. Soc. Anim. Sci. and Am. Dairy Sci. Assoc., Champaign, IL.

FASS. 1999. Guide for the Care and Use of Agricultural Animals in Agricultural Research and Teaching. 1st rev. ed. Fed. Anim. Sci. Soc., Savoy, IL.

Homer, N., S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 4:e1000167.

Kusterer, B., H.-P. Piepho, H. F. Utz, C. C. Schön, J. Muminovic, R. C. Meyer, T. Altmann, and A. E. Melchinger. 2007. Heterosis for biomass-related traits in *Arabidopsis* investigated by quantitative trait loci analysis of the triple testcross design with recombinant inbred lines. Genetics 177:1839–1850.

Lange, K., R. Cantor, S. Horvath, M. Perola, C. Sabatti, J. Sinsheimer, and E. Sobel. 2001. Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. Am. J. Hum. Genet. 69(Suppl.):504.

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

McCullagh, P. and J. A. Nelder. 1989. Generalized Linear Models. 2nd ed. Chapman and Hall, London, UK.

Parker, H. G., L. V. Kim, N. B. Sutter, S. Carlson, T. D. Lorentzen, T. B. Malek, G. S. Johnson, H. B. DeFrance, E. A. Ostrander, and L. Kruglyak. 2004. Genetic structure of the purebred domestic dog. Science 304:1160–1164.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904–909.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Snelling, W. M., M. F. Allan, J. W. Keele, L. A. Kuehn, T. McDaneld, T. P. L. Smith, T. S. Sonstegard, R. M. Thallman, and G. L. Bennett. 2010. Genome-wide association study of growth in crossbred beef cattle. J. Anim. Sci. 88:837–848.

Wasser, S. K., W. J. Clark, E. Drori, E. S. Kisamo, C. Mailand, B. Mutayoba, and M. Stephens. 2008. Combating the illegal trade in African elephant ivory with DNA forensics. Conserv. Biol. 22:1065–1071.

Watanabe, T., T. Hirano, Y. Takano, A. Takano, Y. Mizoguchi, Y. Sugimoto, and A. Takasuga. 2008. Linkage disequilibrium structures in cattle and their application to breed identification testing. Anim. Genet. 39:374–382.