Estimating genomic breed composition of individual animals in cattle: A comparison of SNP panels and statistical methodologies

J. He[1,2], Y. Guo[2,5], J. Xu[2,4], H. Li[2,3], A. Fuller[2], R. G. Tait Jr.[2], X-L. Wu[2,3] & S. Bauck[1]*
[1]College of Animal Science and Technology, Hunan Agricultural University, Changsha, Hunan 410128, China.
nwu@neogen.com (Corresponding Author)
[2]Biostatistics and Bioinformatics, Neogen GeneSeek, Lincoln, NE 68504, USA.
[3]Department of Animal Sciences, University of Wisconsin, Madison, WI 53706, USA.
[4]Department of Statistics, University of Nebraska, Lincoln, NE 68583, USA.
[5]Department of Education, University of Nebraska, Lincoln, NE 68583, USA.

Introduction

Estimation of breed identification or composition of individual animals is useful in a variety of situations, such as breed registration of farm animals, quality control of samples, and verification of breed identifications of breeding animals. SNPs are more accurate for estimating genomic breed composition (**GBC**) of animals than pedigrees because the latter tend to be partially or entirely missing, or incorrectly recorded (VanRaden and Cooper, 2015). Animal populations differ in SNP allele frequencies at many loci as a result of domestication, selection, and genetic drift (Luca et al., 1994). Through SNP genotyping, individuals can be grouped into genetic clusters (breeds) according to their patterns of multiple-loci genotypes (or haplotypes). The objectives of the present study were to 1) select and compare five SNP panels for estimating GBC of individual from 10 cattle breeds; 2) compare two statistical models for calculating GBC; and 3) show the change in genomic prediction accuracies (**GPA**) on nine quantitative traits when using SNP effects estimated on purebred Santa Gertrudis animals versus SNP effects from animals with unverified (mixed) breed composition.

Materials and Methods

Genotype data, reference SNPs, and reference animals

Genotype data. The dataset included a total of 29,609 animals of ten cattle breeds, each genotyped with either the GeneSeek Genomic Profiler low-density (**GGP-LD**) SNP chip (40,660 SNPs) or the GGP Bovine 50K V1 SNP chip (49,463 SNPs). Approximately 53% of the animals were dairy cattle (Holstein and Jersery) and the remaining 47% were beef cattle of eight breeds (namely, Akaushi, Angus, Beefmaster, Red Angus, Brangus, Hereford, Santa Gertrudis, and Wagyu). Mean minor allele frequency (**MAF**) of SNPs varied from 0.188 (Wagyu) to 0.305 (Beefmaster).

Reference SNPs. Five sets of reference SNPs were selected, each having all SNPs in common across five commercial bovine SNP chips, namely: Illumina BovineHD (777K) SNP chip, GGP uHD (150K) SNP chip, GGP HD (80K) SNP chip, GGP Bovine 50K V1 SNP chip, and GGP-LD SNP chip. Prior to SNP selection, there were 15,708 common SNPs among the five bovine SNP chips (denoted as the 16K panel). Then, four subsets of SNPs, each consisting of 1,000, 3,000, 5,000, and 10,000 SNPs (denoted as the 1K, 3K, 5K, and 10K panels), respectively, were selected from these 15,708 SNPs by maximizing average Euclidean distance (AED) of allelic frequencies among the ten breeds, given their respective panel sizes.

Reference animals. Reference animals were selected according to the likelihoods of their genotypes in relation to the 1K SNP genotype frequencies pertaining to that breed. The likelihood that an animal belonged to a specific breed was computed by assuming independent multinomial distributions of its SNP genotypes. To avoid taking logarithm on zero counts of genotypes, SNP genotype frequency was computed based on estimated SNP allele frequencies assuming Hardy-Weinberg disequilibrium and the latter were obtained based on a Bayesian Binomial model. Each reference animal had a value of (-2)log(likelihood) which was smaller than a pre-defined cutoff.

Estimation of GBC

GBC was defined as the proportion of genomic contribution of each breed to an individual animal, which was estimated using two statistical models, namely an admixture model and a linear regression model. For individuals whose ancestors originated in different populations, and those that were admixed, their genetic composition exhibited multiple ancestries associated with multiple different genetic clusters or populations, which was estimated by modeling a population using allele frequencies at multiple loci and each individual's genome as an admixture of alleles from different populations (Pritchard et al.,2000; Tang et al.,2005; Alexander et al.,2009). Alternatively, using a linear regression model, GBC was estimated by regressing discrete random variables corresponding to counts of reference alleles of SNPs across the genome on the allele frequencies of each SNP in the pure breeds involved (Chiang et al., 2010; Kuehn et al., 2011).

Genomic prediction accuracy. GPA on nine quantitative traits were compared between predictions with SNP effects estimated from purebred Santa Gertrudis animals versus those with SNP effects estimated from all the animals with unverified breed composition. The phenotypes included expected progeny difference (**EPD**) of birth weight (**BW**), weaning weight (**WW**), yearling weight (**YW**), maternal weaning weight (**MWW**), scrotal circumference (**SC**), hot carcass weight (**HCW**), marbling score (**MARB**), fat thickness (**FAT**), and ribeye area (**REA**).

Results and discussion

GBC were estimated for 198 Akaushi animals using the five SNP panels based on the admixture model and the linear regression model. The results obtained using the admixture model agreed very well among the five SNP panels. There were 166 animals which were assigned to be 100% Akaushi by all five panels, each animal having GBC of Akaushi breed (GBCA) equal to 1. There were 27 animals with variable GBCA (between 0 and 1) among the five panels, and these animals were considered to be crossbreds of Akaushi cattle. Finally, five animals had 0% GBCA because they were actually Red Angus cattle. The results suggested that a small number (e.g. 1,000) of SNPs can be sufficient to distinguish these ten cattle breeds based on the admixture model. However, the results from the linear regression method showed considerable differences among the five panels, with stable results obtained by 10K SNPs or more (Table 1). The number of purebred animals (GBCA=1) identified by these panels were 57 (1K), 125 (3K), 142 (5K), 150 (10K) and 151 (16K), respectively. Fewer animals were identified with GBCA = 1 based on the linear regression model than were identified based on the admixture model. Purebred Akaushi animals (GBCA=1) identified by the admixture model roughly corresponded to those with GBCA > 0.9 using 5k to 16K panels using the linear

regression model and those with GBCA > 0.8 using 1K to 3K panels based on the linear regression model.

Table 1. Estimated Akaushi genomic breed composition for 198 animals under verification.

GBCA	Admixture model					Linear Regression model				
	1K	3K	5K	10K	16K	1K	3K	5K	10K	16K
=1	166	166	166	166	166	57	125	142	150	151
[0.9, 1.0)	1	1	1	1	0	71	36	24	18	17
[0.8, 0.9)	9	8	9	9	11	41	13	9	7	8
[0.7, 0.8)	4	5	4	4	3	8	4	5	5	4
[0.6, 0.7)	1	0	1	1	1	3	4	1	1	3
[0.5, 0.6)	11	12	11	11	11	6	9	11	11	9
[0.4, 0.5)	0	0	0	0	0	5	1	0	0	0
[0.3, 0.4)	0	0	0	0	0	1	0	0	0	0
[0.2, 0.3)	0	0	0	0	0	0	0	0	0	0
[0.1, 0.2)	1	1	1	1	1	1	1	1	1	1
[0, 0.1)	5	5	5	5	5	5	5	5	5	5

Genomic prediction in Santa Gertrudis cattle

GBC of Santa Gertrudis breed (GBCSG) were estimated for 1,424 test animals based on the admixture model with the 1K SNP panel. The results suggested that 1,138 of them were purebred Santa Gertrudis (GBCSG = 1), and the remaining were crossbred (GBCSG < 1). With SNP effects estimated using the 1,138 purebred Santa Gertrudis animals, GPA on nine traits in purebred Santa Gertrudis cattle were between 0.559 and 0.744, but GPA dropped drastically as the GBCSG decreased. GPA on the nine traits were between 0.52 and 0.67 for crossbred animals with a high level of GBCSG (0.75 ≤ GBCSG <1), and between 0.08 and 0.44 for crossbreds with a GBCSB between 0.50 and 0.75. Two traits even had negative GPA when GBCSG dropped below 0.50. This indicates that SNP effects estimated from purebred Santa Gertrudis animals were not readily applicable to genomic prediction in crossbred animals, and loss in GPA occurred to a varying extent. Nevertheless, by removing crossbred animals from the training set, GPA on the nine traits was increased by between 0.27% (MWW) and 7.31% (YW). These results suggest that estimation of GBC can be useful to obtain increased GPA by eliminating crossbred animals from the training set if genomic prediction is targeted toward the selection of purebred animals.

Conclusions

Five SNP panels were selected, as subsets of common SNPs across five commercial bovine SNP chips. The present results showed that these five SNP panels performed very similarly when estimating GBC for 198 test animals based on the admixture model but the results varied considerably based on the linear regression model. In terms of consistency of results, the admixture model is more robust than the linear model. Of the five SNP panels, the 1K panel is the most cost effective and it is also easy to be included as an obligatory component for future development of bovine chips. Nevertheless, the 10K SNP panel can be more resourceful as an independent LD SNP panel if good imputation accuracy to moderate- or high-density SNP panels is one of the goals. As an extended application for genomic

selection in Santa Gertrudis cattle, SNP effects estimated from purebred animals were more predictive of pure animals *per se* than those estimated from animals with unverified (mixed) breed composition.

List of References

Alexander, D.H., J. Novembre & K. Lange, 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19(9):1655–1664.

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 & J.N. Hirschhorn, 2010. Rapid Assessment of Genetic Ancestry in Populations of Unknown Origin by Genome-Wide Genotyping of Pooled Samples. Plos Genet. 6(3):e1000866.

Luca, M.P. & A. Piazza, 1994. The History and Geography of Human Genes. Princeton, NJ: Princeton University Press.

Pritchard, J.K., M. Stephens & P. Donnelly, 2000. Inference of population structure using multilocus genotype data. Genetics. 155(2):945–59.

Kuehn, L.A., J.W. Keele, G.L. Bennett, T.G. McDaneld, T.P.L. Smith, W.M. Snelling, T.S. Sonstegard, 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(6): 1742-1750.

Tang, H., J. Peng, P. Wang & N.J. Risch, 2005. Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol. 28(4):289–301.

VanRaden, P.M. & T.A. Cooper, 2015. Genomic Evaluations and Breed Composition for Crossbred U.S. Dairy Cattle. Interbull Bulletin. No 49. Orlando, Florida.