Abstract #15

Predicting breed composition in an Angus-Brahman crossbred population using genomic data

M. Gobena, J. Leal, H. Hamblen, M. Elzo, and R. G. Mateescu*

Department of Animal Sciences, University of Florida, Gainesville, FL

Pedigree information has been used conventionally to estimate breed composition. However, these estimates begin to lose accuracy beyond F1 generations due to Mendelian sampling during crossing over. It has been suggested that breed composition or ancestry information derived from genomic data can be more accurate than data based on pedigree information. It can also be especially useful when pedigree information is missing or incomplete. The goal of this study was to examine the feasibility and accuracy of using genotype data to estimate breed composition and subsequently identify the minimum number of markers needed to determine breed composition with adequate accuracy. A total of 782 cattle consisting of purebred Angus, purebred Brahman and crossbreeds across the spectrum were used. DNA was extracted from blood or semen samples and genotyped with the Illumina Bovine 250k SNP chip (GGP F250). After filtering for minor allele frequency (>0.1) and call rate (>0.1), 96671 SNP remained and were used in subsequent analysis. Breed membership and level of admixture were estimated using Bayesian model based clustering implemented by the software STRUCTURE. Principal component analysis was performed in order to find major axes of variation that efficiently capture population structure. The first principal component (PC1) had strong correlation with pedigree-based breed composition (0.96) and level of admixture inferred by Bayesian model based clustering (>0.99). Genome wide association was performed with PC1 as a response variable, and SNP were ranked based on the strength of their association with PC1. Based on ranking, 20 subsets of SNP were selected starting from the top 5 and increasing the number by 5 up to 100. For each subset of SNP, a 5-fold cross validation was used to assess the accuracy of prediction. The dataset was randomly divided into 5 groups and a linear regression model was trained on 4 groups and tested on the 5^{th} group. Accuracy of prediction was measured by correlating the resulting breed composition with the composition known from pedigree. The results show that few SNP can be used to predict breed proportion, and accuracies > 0.9 can be reached with 25 SNP. This indicates a small number of DNA markers can be used to accurately predict breed proportion which would reduce genotyping cost. The methods described here can also be generalized to infer level of admixture in animals with potential ancestry from more than two breeds.