Genome-Wide Prediction, Association, and Gene Network Analysis of Grain Composition in Sorghum

Sirjan Sapkota[1,2,*], J. Lucas Boatwright[1,2], Richard Boyles[2,3], and Stephen Kresovich[1,2]

[1] *Advanced Plant Technology Program, Clemson University, Clemson, SC 29634*

[2] *Department of Plant and Environmental Sciences, Clemson University, Clemson, SC 29634*

[3] *Pee Dee Research and Education Center, Clemson University, Darlington, SC 29532*

[*] *Correspondence:* ***ssapkot@clemson.edu****; GitHub: sirjansapkota; LinkedIn: sirjan-sapkota-09787049*

Abstract

Starch and protein are two of the most important constituents of grain contributing to most of human and animal caloric needs. The genetic control of starch and protein composition are complex and not completely understood. While genome-wide prediction has been routinely studied for grain yield, little is done in terms of application of predictions for grain composition. Here we present our genotype-phenotype association and prediction study for starch, protein and gross energy using 224,007 SNPs in a sorghum diversity panel with 389 individuals. We did not find any significant difference in predictive ability between various Bayesian models with different priors. On average, the predictive ability was 0.6 for starch, 0.45 for protein, and 0.58 for gross energy. Using a multivariate linear mixed model for starch and protein, we were able to identify significant associations at genomic regions in chromosomes four and eight that were not significant using a univariate model for starch or protein alone. A total of 13 genes within linkage disequilibrium of the associated regions had high confidence (0.7) first interactors using STRING. The gene network analysis of those 13 genes and their first interactors showed significant enrichment for various biochemical pathways including sucrose and starch biosynthesis and nitrogen metabolism. Our results provide new insights into the application of multivariate approaches in statistical learning. Furthermore, the genes and pathways identified could be crucial in understanding genetic mechanisms in source-sink dynamics during grain filling.

Genotype and Phenotype

A total of 389 genetically diverse sorghum accessions, mostly from the sorghum association panel, were characterized with 224,007 SNP markers using genotyping-by-sequencing.

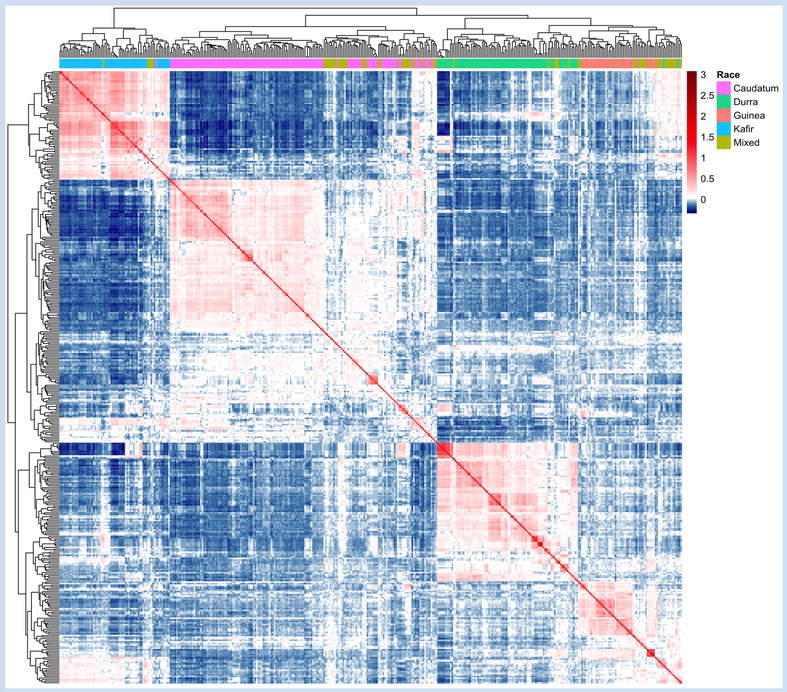

Figure 1. Heatmap of genomic relationship matrix computed using VanRaden(2008) method. Rug plot on top shows the racial classification of each genotype based on population structure analysis using *ADMIXTURE* (K=5).

Grain of three random plants grown at Clemson Pee Dee Research and Education Center in Florence, SC during 2013, 2014, and 2017 were harvested at physiological maturity, ground and analyzed using near infrared spectroscopy (NIRS). Phenotypic best linear unbiased prediction (BLUPs) for each genotype (G) was calculated by adjusting phenotypic values (y) for random effects of year (Y), genotype-by-year (G × Y), and year-by-replication (Y × R) using a linear mixed model (1) in R package *lme4*.

$$y \sim G + Y + G \times Y + Y \times R + \epsilon \quad (1)$$

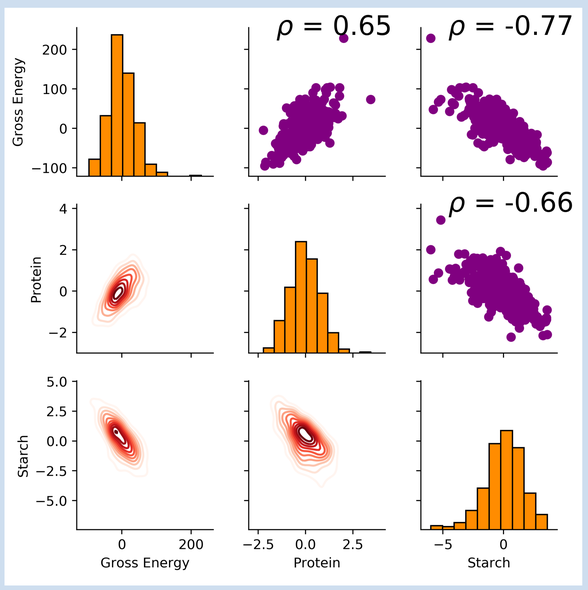

Figure 2. Distribution and correlation of BLUPs.

Genomic Prediction

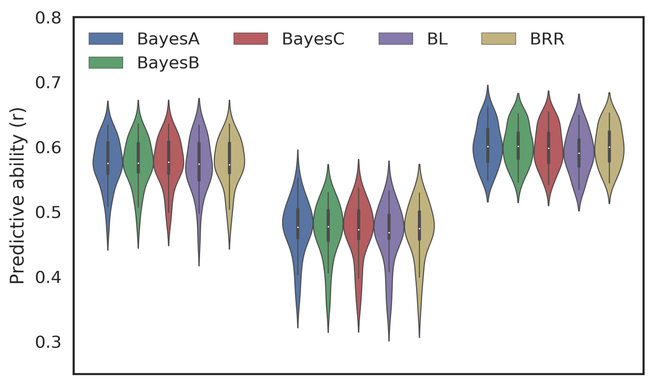

Figure 3. Results from five-fold cross validation of Bayesian whole genome regression using R package *BGLR*. BRR: Bayesian ridge regression, BL: Bayesian lasso.

Gene Network

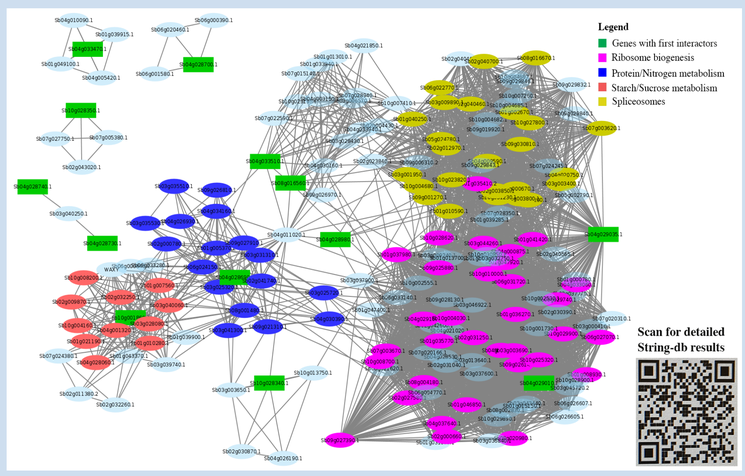

Figure 5. Gene network analysis for genes near significantly associated regions using *STRING*. Among the genes within 20 kb of significant SNPs, we identified 13 (green) with high confidence first interactors (0.7) in STRING.

Association Mapping

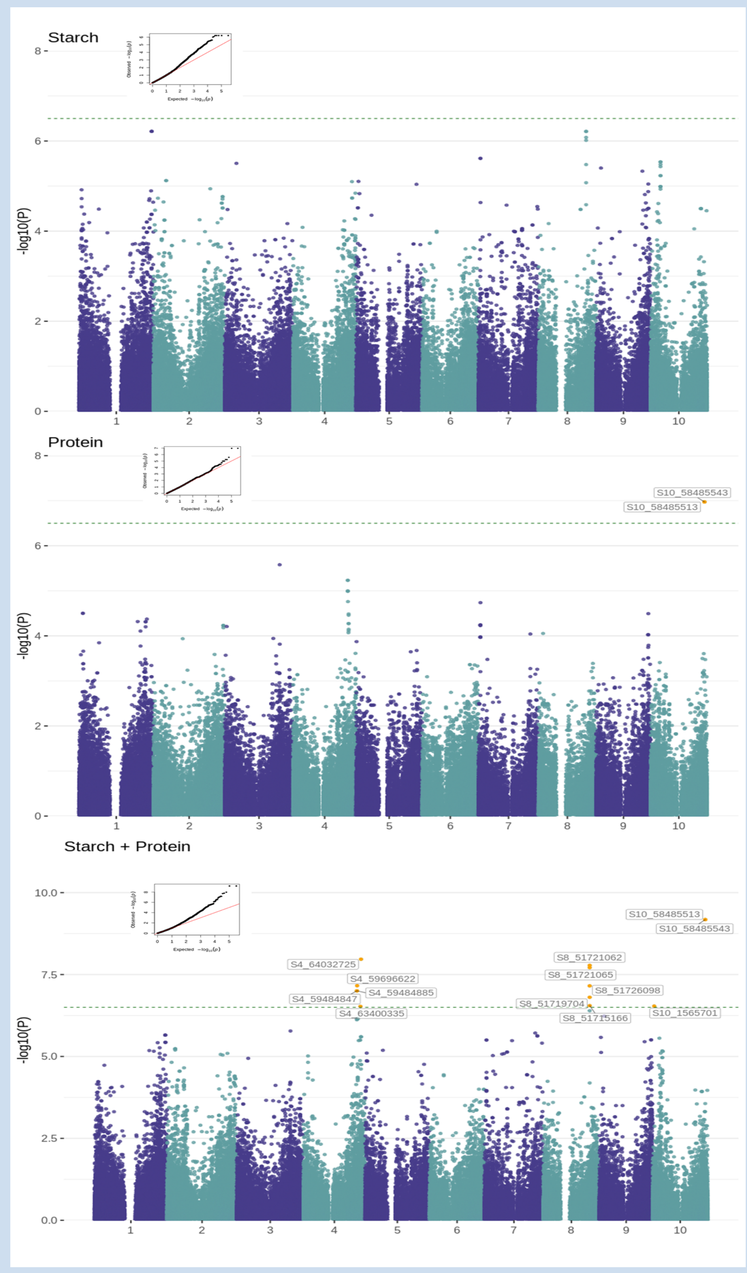

Figure 4. Manhattan plots for univariate (top 2) and multivariate (bottom) mixed model association analysis of starch and protein using software *GEMMA*. Bonferroni significance threshold is shown by green dashed line.

Summary

- Protein and gross energy were significantly positively correlated to each other, but were significantly negatively correlated to starch.
- Starch had higher predictive ability (r=0.6) than protein (r=0.45), but there were no significant differences in predictive ability between various prediction models.
- Use of multivariate (MV) mixed model for highly correlated starch and protein showed increased statistical power over univariate approach. Significant associations were identified in chromosomes 4 and 8 using the MV approach.
- Several metabolic pathways, including protein metabolism, were found enriched in a KEGG enrichment analysis using high confidence first interactors (0.7) of the genes within 20 kb of significant SNPs.
- The genes identified within the associated regions could be potential hub genes involved in key regulatory and functional pathways for sugar/starch metabolism and grain filling.

Acknowledgments

Wade Stackhouse Graduate Fellowship Program