# Genome-wide association and gene network analysis for starch and protein in sorghum

Sirjan Sapkota[1,2,*], Jon Lucas Boatwright[2], Kathleen Jordan[2], Richard Boyles[1,3], Stephen Kresovich[1,2],

1 Department of Plant and Environmental Sciences, Clemson University, Clemson, SC, USA

2 Advanced Plant Technology Program, Clemson University, Clemson, SC, USA

3 Pee Dee Research and Education Center, Clemson University, Florence, SC, USA

* correspondence: ssapkot@g.clemson.edu

## Abstract

The grains of cereals are ultimate sink for macromolecules such as starch and protein which serve as principal source of nutrition. Dissection of genetic basis of phenotypic variation for starch and protein is important in metabolic engineering of grain. Genetic studies of starch and protein in sorghum, a cereal crop, has involved single traits and metabolic network involved in their regulation are not completely characterized. In this study we used univariate and multivariate (MV) linear mixed models (LMM) to identify associated genomic regions, potential candidate genes and their interactors. Six single nucleotide polymorphism (SNPs) in strong linkage disequilibrium ($r^2$ >0.8) from ∼52 Mb of chromosome 8 were significantly associated with starch content. Five of those SNPs were located within mRNA of a heat shock protein 90 (HSP90), *Sobic.008G111600*, with two of them in the coding sequences of the gene. The HSP90 had a total of 142 high confidence (PPI-score: 0.6) first interactors and the network was enriched for six bichemical pathways including protein processing and export. A SNP, S4_60623675, identified using MV-LMM model was located at 5'UTR of a fatty acid desaturase gene, *Sobic.004G260800*, which interacted with another fatty acid desaturase and several nitrate reductase genes. The two candidates, HSP90 and FAD2, were found to be highly expressed in reproductive tissues. We conclude multivariate analysis of correlated phenotypes can identify biologically important metabolic networks and functional analyses of identified gene candidates can be beneficial in understanding grain filling in sorghum and other cereals.

# Introduction

The seeds of cereals, that represent an important sink for metabolites during grain filling, are principal source of human and animal nutrition. Sorghum [*Sorghum bicolor*. (L.) Moench] is a cereal crop that provides dietary staple for over half a billion people in semi-arid tropics (Mace et al., 2013). While primarily used as animal feed in industrialized economies, the end use products of sorghum grain has diversified to include baking, malting, brewing, and bio-fortification (Zhu, 2014). Understanding the genetic basis of phenotypic variation in grain composition such as starch and protein could provide the basis for metabolic engineering of these macromolecules through selective breeding.

Linkage mapping has been a powerful method to identify quantitative trait loci (QTL) that cosegregate with a given trait but suffers from two fundamental limitations; only allelic diversity that segregates between the parents can be assayed, and the amount of recombination from bi-parental crosses places a limit on the mapping resolution (Korte et al., 2013). In contrast, genome-wide association studies (GWAS) have mapped genetic variants associated to phenotypes to a much higher resolution using whole genome markers in a diverse group of individuals. The cost effectiveness in generating large scale genotypic data has now led to swathe of GWAS in crops and focus has shifted towards computational challenges (Myles et al., 2009). Most application of GWAS has focused on single traits, whereas phenotypes are usually correlated and might be controlled by genetic loci with pleiotropic effects. Meanwhile, studies have shown that joint analysis of correlated phenotypes can exploit the correlation among the phenotypes for detecting additional genetic variants with small effects across multiple traits (Korte et al., 2012; Thoen et al., 2017; Carlson et al., 2019; Rice et al., 2020). Some of the approaches to leverage correlation between traits in association analysis include: use of ratios of directly related traits in univariate GWAS (Gieger et al., 2008), combining test statistics from univariate GWAS of each trait to detect pleiotropic effects (Yang et al., 2010), using dimension reduction technique to derive transformed phenotypes for univariate GWAS (Aschard et al., 2014), and directly modeling multiple traits into a multivariate linear mixed models (Korte et al., 2012; Zhou et al., 2014).

Association studies for grain quality has been reported in several cereal crops such as maize (Wilson et al., 2004; Cook et al., 2012), rice (Zhao et al., 2011; Wang et al., 2017), sorghum (Sukumaran et al., 2012; Rhodes et al., 2017), and wheat (Reif et al., 2011; Gaire et al., 2019). In diverse sorghum accessions, starch and protein show continuous variation ranging from 60 to 72% and 8 to 18% of total grain, respectively (Rhodes et al., 2017). Previous association analyses for starch and protein

content have identified significantly associated genomic regions in sorghum (Sukumaran et al., 2012; Rhodes et al., 2017; Boyles et al., 2017). Starch and protein content represent majority of grain composition and are likely to be controlled by genetic loci with pleiotropic effects. Such genetic loci could have gone undetected in single trait GWAS, and multivariate GWAS might be able to identify such associations. Furthermore, the path from genetic association to biology is not always straightforward because an association between a genetic variant and a trait may not be informative with respect to the target gene (Gallagher et al., 2018). In this study, we implemented univariate and multivariate linear mixed models for starch and protein content using sorghum association panel, and identified candidate genes within significantly associated loci. We also performed candidate gene network analysis using protein-protein interaction and studied expression profile of candidate genes and their interactors.

# Materials and Methods

## Plant material

A panel of approximately 400 diverse sorghum accessions was planted in randomized complete block design with two replications in 2013, 2014, and 2017 field seasons at the Clemson University Pee Dee Research and Education Center in Florence, SC. This diversity panel, with over 80% of the accessions from the original United States sorghum association panel (SAP) developed by Casa et al. (2008), will be referred to as SAP. The details on experimental field design and agronomic practices have been described in details in Boyles et al. (2016) and Sapkota et al. (2020). Succinctly, the experiments were planted in a two row plots each 6.1 m long, separated by row spacing of 0.762 m with an approximate density of 130,000 plants $ha^{-1}$. Fields were irrigated only when signs of drought stress was seen across the field. Primary panicle of three plants selected from each plot was harvested at physiological maturity. The plants from beginning and end of the row were excluded to account for border effect. Panicles were air dried to a constant moisture (10-12%) and threshed. A 25g of cleaned and homogenized subsample of grain ground to 1-mm particle size with a CT 193 Cyclotec Sample Mill (FOSS North America) was used in near infrared spectroscopy (NIRS) for compositional analysis.

## Phenotypic data

A DA 7250$^{TM}$ NIR analyzer (Perten Instruments) was used for compositional analysis. The predicted phenotypic values were obtained from the calibrated curves for spectral measurements of ground grain samples. The calibration curve was built using wet chemistry values from a subset of samples. The wet chemistry was performed by Dairyland Laboratories, Inc. (Arcadia, WI) and the Quality Assurance Laboratory at Murphy-Brown, LLC (Warsaw, NC). The details on the prediction curves and wet chemistry can be found in Boyles et al. (2017).

The phenotypic values were fitted into a linear mixed model analysis using lme4 package in R (Bates et al., 2015; R Core Team, 2019). The following mixed model equation was fit:

$$y_{ijk} \sim G_i + Y_j + G_i \times Y_j + Y_j \times R_k + \epsilon_{ijk} \tag{1}$$

where $y_{ijk}$ represents the phenotypic value for the combination of genotype $i$, year $j$, and replication $k$; $G_i$, $Y_j$, $G_i \times Y_j$, and $Y_j \times R_k$ are random effects of genotype, year, genotype-by-year, and replication-by-year, respectively; and $\epsilon_{ijk}$ is the random effect of residuals, with $N(0, \sigma^2_\epsilon)$. Best linear unbiased predictors (BLUPs) for the traits were calculated as the random effects of genotypes in the model. Variance components for genotype (G), environment/year (Y), and genotype $\times$ environment interactions were used to calculate the broad sense heritability:

$$H^2 = \frac{\sigma^2_G}{\sigma^2_G + \frac{\sigma^2_{G\times Y}}{Y} + \frac{\sigma^2_\epsilon}{YR}} \tag{2}$$

## Genotypic data

The population was genetically characterized using genotyping-by-sequencing (Morris et al., 2013; Boyles et al., 2016). Sequenced reads were aligned to the BTx623 v3.1 reference assembly (phytozome) using Burrows-Wheeler aligner (Heng Li et al., 2010). TASSEL 5.0 pipeline was used for SNP calling, imputation and filtering (Glaubitz et al., 2014). The missing genotypes were imputed using the TASSEL plugin FILLIN (Swarts et al., 2014). Following imputation SNPs with minor allele frequency (MAF) <0.01, and sites missing in more than 30% genotypes in diversity panel were filtered. Genotypes with more than 10% of SNP sites missing were filtered. A total of 389 genotypes with 224,007 SNPs were used in the study. The SNP genotype file was converted into plink (Purcell et al., 2007) binary ped and bed format for association and linkage disequilibrium (LD) analysis.

### Genome wide association analysis

Genome-wide association between SNPs and phenotypes were computed using a univariate or multivariate linear mixed model (LMM) fit with GEMMA v0.94 (Zhou et al., 2012; Zhou et al., 2014). Eigenvalues (-d) and eigenvectors (-u) from a genomic relationship matrix calculated using (VanRaden, 2008) was used to account for relatedness between individuals. P-values of each marker association tests were computed using Wald's statistics (-lmm 1). The SNPs with minor allele frequency less than 5% were filtered out during association analysis. Significance of marker association was determined using bonferroni threshold ($\frac{\alpha}{p}$), where $\alpha = 0.05$ and p = total number of markers.

### Gene network and expression analysis

Candidate genes in LD with significantly associated SNPs were identified using annotations for BTx623 v3.1.1 (phytozome). Python codes were used to isolate associated candidate genes from annotation file and to convert gene names from *Sobic* to *Sb* gene format. Once converted, candidate genes from associated region were used to identify their high confidence (0.6) first interactors (neighbors) using sorghum protein interaction data from STRING v11.0 (www.string-db.org). Gene expression results from Olson et al. (2014) was used to examine the gene expression pattern of genes and interactors for various tissue types.

## Results

### Phenotypic analysis

We fit a linear mixed model to account for random effects due to environment. The genotypic effects acccounted for about 30% and 45% of total variance in protein and starch, respectively (Supplementary Table S1). The environmental variable (Year) didn't have any effect on starch, whereas year effects amounted to 17% of total variance for protein. Genotype $\times$ environment effect was slightly higher for starch (14%) than for protein (10%). The broad sense heritability was high for both protein (0.75) and starch (0.8). Both protein and starch were normally distributed, and were strongly negatively correlated to each other (Fig 1).

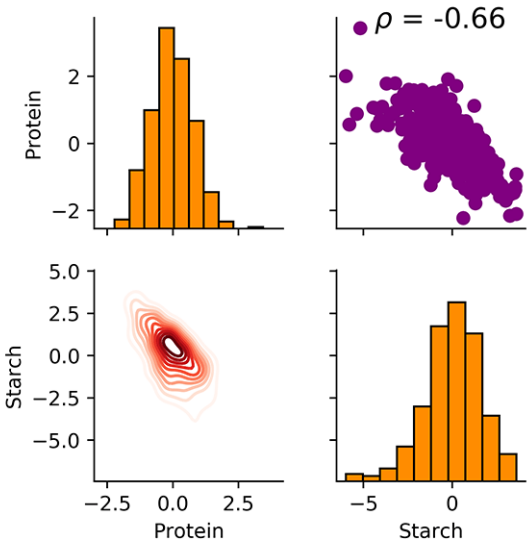


**Figure 1.** Distribution of the adjusted phenotypic mean (BLUPs). Right and left of the diagonal shows scatterplot and density plot, respectively, and the diagonal shows histogram. $\rho$=pearson correlation coefficient.

## Association mapping

We filtered the SNPs with minor allele frequency less than 5% to avoid false positives. There were no SNPs that were significantly associated with protein content (Fig 2a). Starch, on the other hand, had four SNPs (S8_51720767, S8_51721062, S8_51721065, and S8_51726098) in chromosome 8 that were above the significance threshold (Fig 2b). Since starch and protein were strongly correlated, we fit a multivariate (MV) LMM to identify any other associated regions. We found two SNPs, S4_60623675 and S4_63400335, in chromosome 4 that showed significant association for MV model (Fig 2c). The SNPs on chromosome 8 that were significant for starch were also significant for the MV-LMM. Additionally, two more SNPs (S8_51715166 and S8_51719704) on chromosome 8 nearby the other associated SNPs were significantly associated in the MV analysis. All six significant SNPs in chromosome 8 were in strong linkage disequilibrium (LD) with each other (Fig 3). We also fit a univariate LMM for starch with protein as a covariate (say, *StaCovPrt*) to compare with results from MV-LMM for starch and protein. All six chromosome 8 SNPs and the chromosome 4, SNP S4_60623675, significant for MV-LMM were also significantly associated in the *StaCovPrt* model (Fig 2d). Additionally, three SNPs near 64 Mb on chromosome 4 (S4_64019577, S4_64019590 and S4_64019619) were also found to be significantly associated for the *StaCovPrt*, while the chromosome 4 SNP (S4_63400335) from MV-LMM didn't show significant association in *StaCovPrt*. The SNPs in chromosome 4 didn't show strong LD with the neighboring SNPs (Supplementary Table S2).

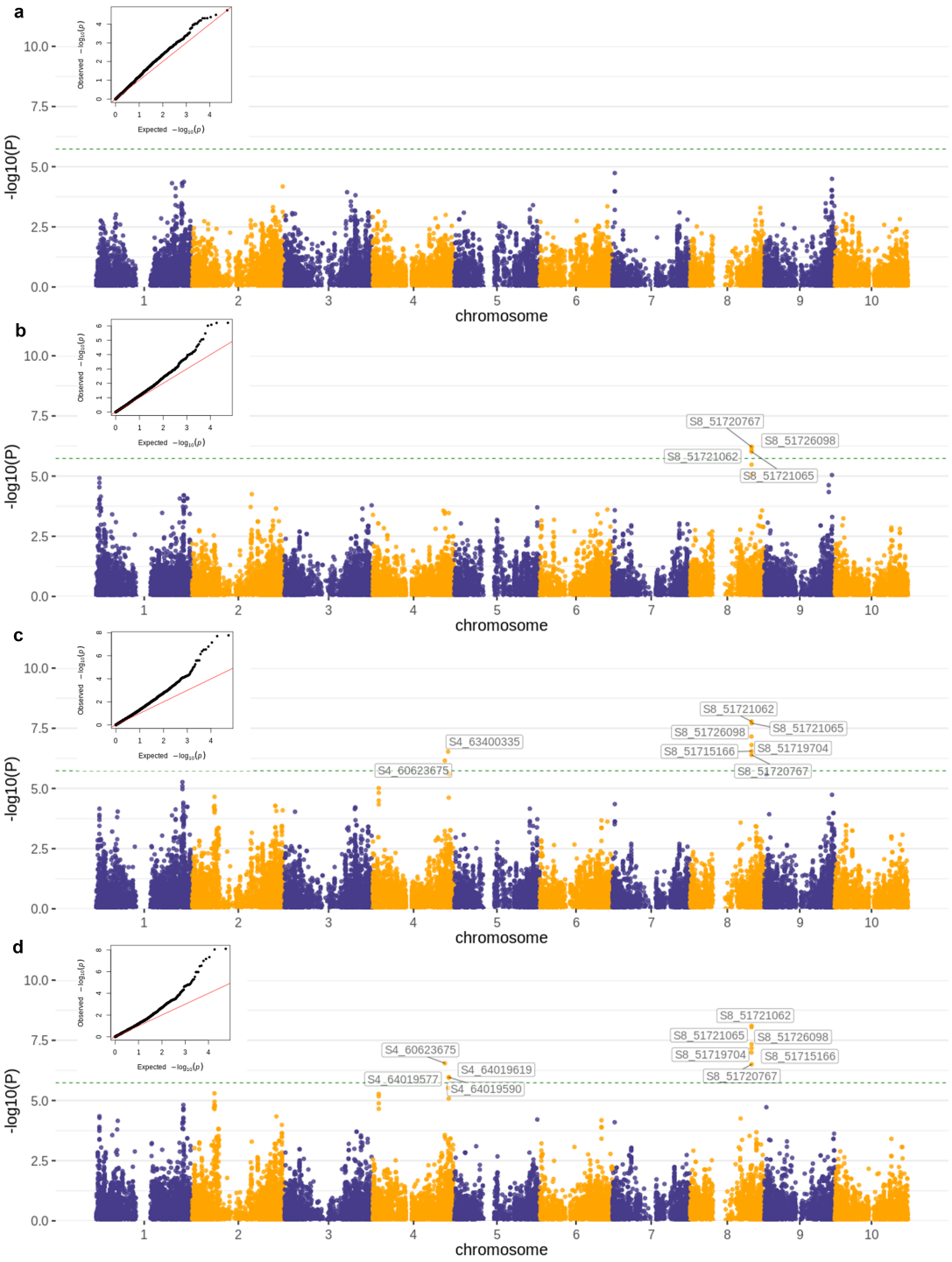


**Figure 2.** Manhattan plot showing genome-wide association using linear mixed model (LMM). Subfigures show univariate LMM for **a**. protein and **b**. starch, and multivariate LMM for **c**. starch and protein. Subfigure **d**. shows univariate LMM for starch with protein as covariate. Horizontal dashed green line represents Bonferroni-corrected significance threshold for $\alpha$=0.05. Quantile-quantile plots for association analysis are presented as inset at top-left of subfigures. Significantly associated SNPs are annotated in the plots.

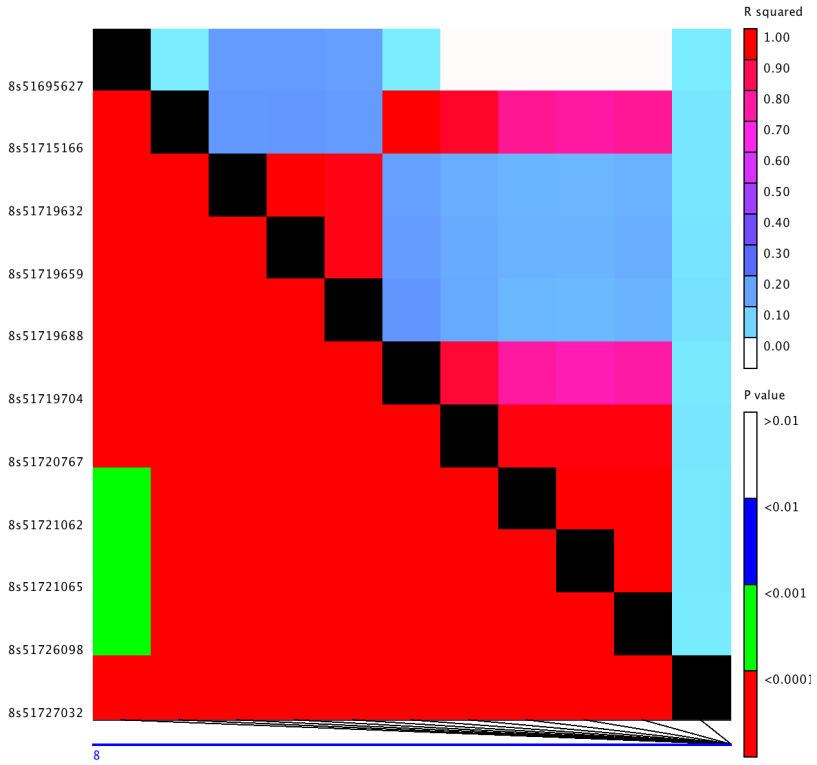


**Figure 3.** Linkage disequilibrium between significantly associated SNPs from chromosome 8. R-squared values to the top of the diagonal and associated p-values are at the bottom of the diagonal.

## Candidate genes

We identified potential candidate genes from significantly associated SNPs based on extent of LD between SNPs in the associated regions. Since all chromosome 8 SNPs were in strong LD with each other we identified genes within the range of 51715166 bp to 51726098 bp in chromosome 8 (*Supplementary Table S2*). Since chromosome 4 SNPs showed weak LD with neighboring SNPs, we identify genes within 2 Kb of the significant SNPs as potential candidate genes.

A total of five candidate genes had significantly associated SNPs that were within or in proximity of those genes (Table 1). All SNPs except one (S4_63400335) were localized within the mRNA of the associated genes. One SNP, S8_51715166, was located in coding sequences (CDS) of a gene encoding CASP like protein (*Sobic.008G111500*), whereas, the SNPs S8_51719704 and S8_51726098 were situated within the CDS of a heat shock protein (HSP90-6; *Sobic.008G111600*). Three signficant SNPs from ~64 Mb region of chromosome 4 were located in the 3'UTR region of a Ring-H2 finger protein, *Sobic.004G301300*. One SNP, S4_60623675, was localized in the 5'UTR region of fatty acid desaturase (FAD2) gene, *Sobic.004G260800*.

**Table 1.** Potential candidate genes from the significantly associated regions.

| Gene | Name* | Chr | Start | End | Associated_SNPs |
|---|---|---|---|---|---|
| Sobic.004G260700 | Uncharacterized protein | 4 | 60621941 | 60622538 | S4_60623675 |
| Sobic.004G260800 | FAD2 | 4 | 60623621 | 60625764 | S4_60623675 |
| Sobic.004G301300 | RING-H2 finger protein | 4 | 64018712 | 64019678 | 3 SNPs |
| Sobic.008G111500 | CASP-like protein 8 | 8 | 51714673 | 51715254 | S8_51715166 |
| Sobic.008G111600 | HSP-90-6 | 8 | 51719209 | 51726960 | 5 SNPs |

* Gene names based on annotated homologous maize genes..

## Gene network and expression

We used the string-db (or whichever) to identify high confidence (0.6) first interactors of candidate genes. The two genes in chromosome 4 had a total of 10 first interactors (Fig 4). HSP90-6 (*Sobic.008G111600*), the only chromosome 8 gene with first neighbors, had a total of 142 interactors. The gene interaction network for both sets of genes had higher number of protein-protein interaction (PPI) than expected (PPI enrichment p-value $<1e^{-16}$). Gene interaction networks from chromosome 4 and chromosome 8 were sign ificantly enriched (FDR<0.001) for three and six biochemical pathways, respectively (Supplementary Table S3). The chromosome 4 genes were enriched for biosynthesis of unsaturated fatty acids, fatty acid metabolism and nitrogen metabolism pathways, whereas, chromosome 8 genes were enriched for protein processing, protein export, spliceosome, endocytosis, RNA degradation, and plant-pathogen interaction. Figure 5 shows expression atlas of chromosome 4 and chromosome 8 group of genes and interactors across various tissue types. While clear clustering of genes with differential expression across reproductive and vegetative tissues was not seen, different clusters of genes showed varying degree of transcript abundance across tissue types.

The FAD2 gene (*Sobic.004G260800*) in chromosome 4 interacted with another fatty acid desaturase gene (DES2, *Sobic.004G260600*) which is located ∼570 Kb upstream from the FAD2 gene. Both FAD2 and DES2 strongly interact with three nitrate reductase (NADH) genes, two of which located in chromosome 4 at ∼55 Mb (*Sobic.004G196101*) and ∼65 Mb (*Sobic.004G312500*) while the other is located around 58-59 Mb in chromosome 7 (*Sobic.007G153900*) (Figure 4). The FAD2 gene was highly expressed in flower, embryo and shoot, whereas, its interactor DES2 had higher expression in root tissues (Fig 6). The heat shock proteins are known to be molecular chaperones primarily involved in drought and stress response but can also be involved in other molecular processes during plant development (Yu et al., 1998; Khan et al., 2019). The gene expression results showed HSP-90 (*Sobic.008G111600*) to be highly expressed in floral meristem, plant embryo and vegetative meristem compared to root, shoot, and flower tissues (Fig 6). The interactors of HSP90-6 included several

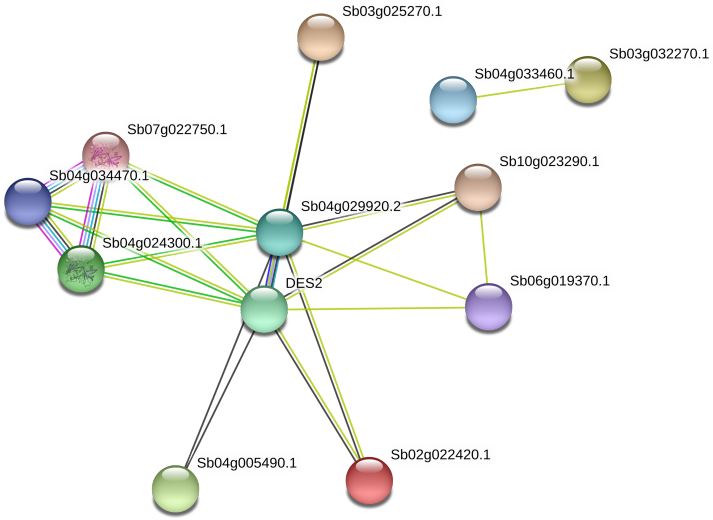


**Figure 4.** Network of candidates genes (*Sobic.004G260800.1* and *Sobic.004G033460.1*) and interactors from associated chromosome 4 SNPs.

HSP70 genes which were also highly expressed in reproductive tissues compared to root and shoot tissues (Fig 6).

## Discussion

Despite being two of the most studied grain quality phenotypes, large proportion of genetic variance in starch and protein remains unexplained. In this study, we aimed to identify genetic loci associated with starch and protein content in sorghum grain. We observed strong genotypic effect and some genotype $\times$ environment effect for starch and protein in our population (Supplementary Table S1). Previous studies have also reported high heritability for starch and protein in different populations (Rami et al., 1998; Murray et al., 2008; Rhodes et al., 2017).

In our genome-wide association study, we used only the random genetic effects (BLUPs) to identify marker trait association for strictly genetic effects. The lack of genetic association for protein content could be due to smaller genetic effects and larger environmental effects on this trait. Starch, with no environmental effect and larger genotypic effects, had genetic variants significantly associated with the phenotype. Five SNPs (in strong LD with each other) from a single locus that encodes for a heat shock protein (HSP) 90 (*Sobic.008G111600*), with two SNPs located on the coding sequence, showed significant association. This loci was not identified during association mapping of starch in previous studies using this population (Rhodes et al., 2017; Boyles et al., 2017). HSPs are common group of

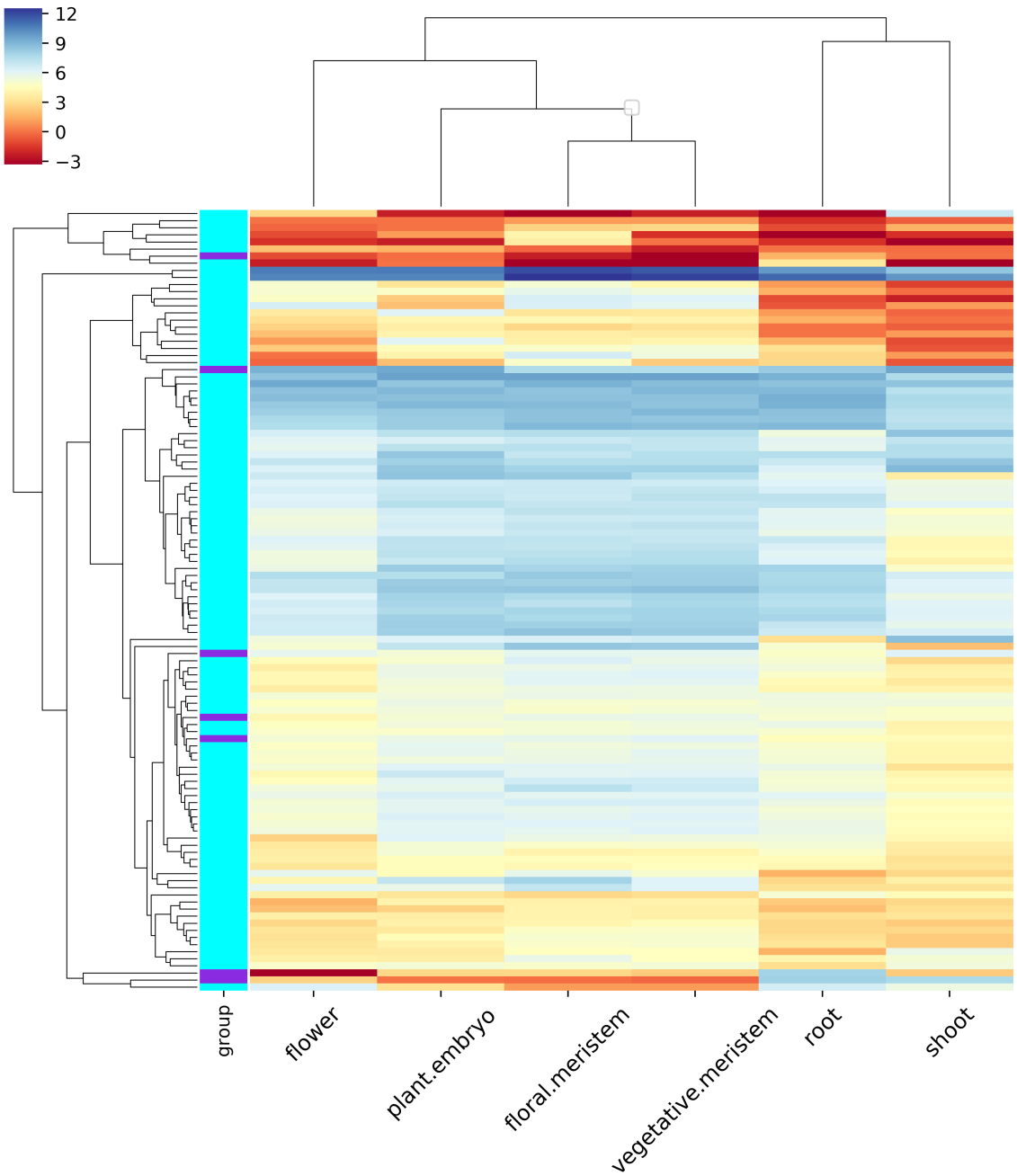


**Figure 5.** Heatmap showing gene expression analysis of interactors of candidate genes. The row colors represents chromosome 'group': purple and cyan represent chromosome 4 and 8 related genes, respectively.

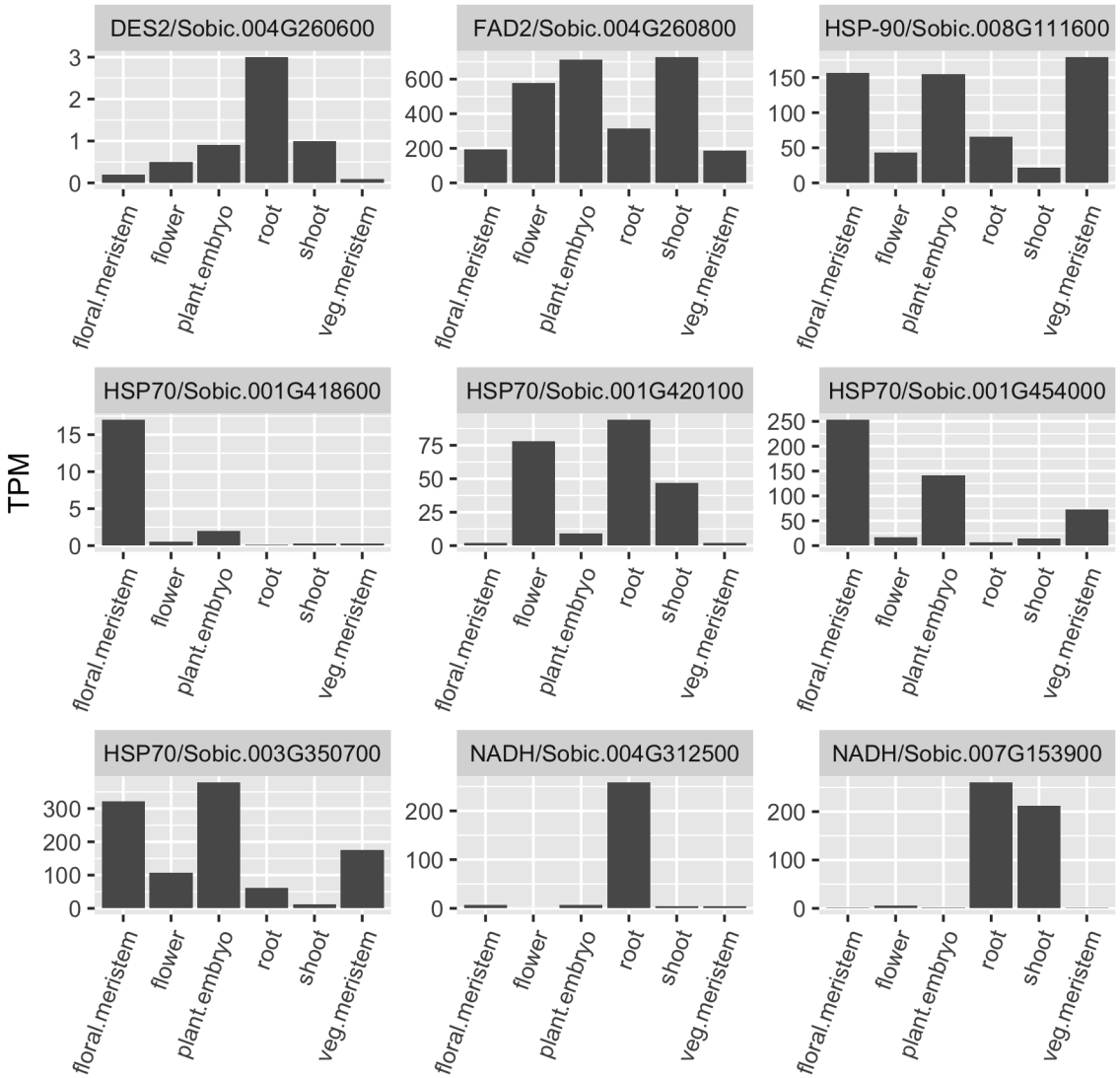


**Figure 6.** Gene expression of candidate genes and some of their interactors. X-axis represents various tissue types, and y-axis shows relative expression (TPM: transcript per million). HSP: heat shock protein, FAD/DES: fatty acid desaturase, NADH: nitrate reductase.

protein found in eukaryotes and function as molecular chaperones that help in refolding proteins denatured by heat and keep them from aggregating (Vierling, 1991; Boston et al., 1996). Two distinct members of the Hsp70 family of stress-related protein were localized in the maize amyloplast and form transient complexes with starch synthase 1 (SSI) and other stromal enzymes (Yu et al., 1998). In Japanese sake-brewing rice rich in starch content, the HSP70 protein was highly abundant in amyloplast compared to cytosol and its concentration was elevated during the later stages of grain development (Kamara et al., 2009). The HSP90 gene in this study was seen to be strongly interacting with numerous molecular chaperones including HSP70. Since the HSP90 and interacting HSP70 proteins showed higher expression in the embryo and floral meristem compared to root and shoot tissues, they are likely important candidates responsible for protein processing and export during grain filling in sorghum.

For complex traits, understanding if a genetic variant affects multiple phenotypes simultaneously (pleiotropy) or affects one phenotype through affecting another phenotype is one of the major challenge (Yang et al., 2012). Starch and protein constitute most of the grain composition and display a strong negative correlation. We identified additional marker trait associations when: 1) starch and protein were fit as dependent variables in a multivariate mixed model, or 2) when protein was fit as independent variable for a model with starch as dependent variable. This approach helped us identify an important variant, S4_60623675, which showed significant association in both of the above mentioned models and was located in the 5' UTR of a fatty acid desaturase gene (*Sobic.004G260800*). The fatty acid desaturase gene interacted with two more desaturase genes and three nitrate reductase genes, forming a network that is highly enriched for biochemical pathways for fatty acid and nitrogen. One of the interacting desaturase gene (DES2, *Sobic.004G260600*) is involved in the biosynthesis of aliphatic side chain of sorgoleone by converting palmitoleic acid to hexadecadienoic acid (Pan et al., 2007). Sorgoleone is a phytotoxic secondary metabolite that plays a direct role in allelopathic interactions. Sorgoleone is known to inhibit photosynthesis, but the relationship between Sorgoleone biosynthesis pathway and seed development is unclear (Einhellig et al., 1993). Rhodes et al. (2017) has previously reported significantly associated variants for protein and fat around 57-58 Mb of chromosome 4 which is ∼2-3 Mb from our associated SNPs. The high expression of HSP90 gene in plant embryo combined with a candidate SNPs in the fatty acid desaturase gene identified using MV-LMM hints at possible connection between the biochemical pathways for starch, protein and fat content during grain development. The enrichment of biochemical pathways involving these genes and their high expression in reproductive tissues warrants further characterization and functional

analysis of these candidates.

In conclusion, we were able to identify a previously uncharacterized genomic region associated with starch content using univariate LMM. The genomic region harbored a heat shock protein which shows strong protein-protein interaction and its network is enriched for several biochemical pathways. Additionally, we also showed that use of MV-LMM for correlated traits can help identify additional genomic regions that go undetected with the univariate GWAS of single traits. The candidates of this study might be involved in intricate metabolic pathway and represent possible pleiotropic targets for source-sink activities during grain filling.

## Supporting Information

Supplementary information is included in Appendix C.

## Data availability

The codes and data used in the study are available at github.com/sirjansapkota/StarchProtein.

## References


Aschard, Hugues, Bjarni J Vilhjálmsson, Nicolas Greliche, Pierre-Emmanuel Morange, David-Alexandre Trégouët, and Peter Kraft (2014). "Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies". In: *The American Journal of Human Genetics* 94.5, pp. 662–676.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

Boston, Rebecca S, Paul V Viitanen, and Elizabeth Vierling (1996). "Molecular chaperones and protein folding in plants". In: *Post-transcriptional control of gene expression in plants*. Springer, pp. 191–222.

Boyles, Richard E, Elizabeth A Cooper, Matthew T Myers, Zachary Brenton, Bradley L Rauh, Geoffrey P Morris, and Stephen Kresovich (2016). "Genome-wide association studies of grain yield components in diverse sorghum germplasm". In: *The plant genome* 9.2.

Boyles, Richard E, Brian K Pfeiffer, Elizabeth A Cooper, Bradley L Rauh, Kelsey J Zielinski, Matthew T Myers, Zachary Brenton, William L Rooney, and Stephen Kresovich (2017). "Genetic dissection of sorghum grain quality traits using diverse and segregating populations". In: *Theoretical and applied genetics* 130.4, pp. 697–716.

Carlson, Maryn O, Gracia Montilla-Bascon, Owen A Hoekenga, Nicholas A Tinker, Jesse Poland, Matheus Baseggio, Mark E Sorrells, Jean-Luc Jannink, Michael A Gore, and Trevor H Yeats (2019). "Multivariate genome-wide association analyses reveal the genetic basis of seed fatty acid composition in oat (Avena sativa L.)" In: *G3: Genes, Genomes, Genetics* 9.9, pp. 2963–2975.

Casa, Alexandra M, Gael Pressoir, Patrick J Brown, Sharon E Mitchell, William L Rooney, Mitchell R Tuinstra, Cleve D Franks, and Stephen Kresovich (2008). "Community resources and strategies for association mapping in sorghum". In: *Crop science* 48.1, pp. 30–40.

Cook, Jason P, Michael D McMullen, James B Holland, Feng Tian, Peter Bradbury, Jeffrey Ross-Ibarra, Edward S Buckler, and Sherry A Flint-Garcia (2012). "Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels". In: *Plant physiology* 158.2, pp. 824–834.

Einhellig, Frank A, James A Rasmussen, Angela M Hejl, and Itamar F Souza (1993). "Effects of root exudate sorgoleone on photosynthesis". In: *Journal of chemical ecology* 19.2, pp. 369–375.

Gaire, Rupesh, Mao Huang, Clay Sneller, Carl Griffey, Gina Brown-Guedira, and Mohsen Mohammadi (2019). "Association Analysis of Baking and Milling Quality Traits in an Elite Soft Red Winter Wheat Population". In: *Crop Science* 59.3, pp. 1085–1094.

Gallagher, Michael D and Alice S Chen-Plotkin (2018). "The post-GWAS era: from association to function". In: *The American Journal of Human Genetics* 102.5, pp. 717–730.

Gieger, Christian, Ludwig Geistlinger, Elisabeth Altmaier, Martin Hrabé De Angelis, Florian Kronenberg, Thomas Meitinger, Hans-Werner Mewes, H-Erich Wichmann, Klaus M Weinberger, Jerzy Adamski, et al. (2008). "Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum". In: *PLoS genetics* 4.11.

Glaubitz, Jeffrey C, Terry M Casstevens, Fei Lu, James Harriman, Robert J Elshire, Qi Sun, and Edward S Buckler (2014). "TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline". In: *PloS one* 9.2, e90346.

Kamara, Joseph S, Miki Hoshino, Yuki Satoh, Nasrin Nayar, Motoko Takaoka, Tsuneo Sasanuma, and Toshinori Abe (2009). "Japanese sake-brewing rice cultivars show high levels of globulin-like protein and a chloroplast stromal HSP70". In: *Crop science* 49.6, pp. 2198–2206.

Khan, Abid, Muhammad Ali, Abdul Mateen Khattak, Wen-Xian Gai, Huai-Xia Zhang, Ai-Min Wei, Zhen-Hui Gong, et al. (2019). "Heat Shock Proteins: Dynamic Biomolecules to Counter Plant Biotic and Abiotic Stresses". In: *International journal of molecular sciences* 20.21, p. 5321.

Korte, Arthur and Ashley Farlow (2013). "The advantages and limitations of trait analysis with GWAS: a review". In: *Plant methods* 9.1, p. 29.

Korte, Arthur, Bjarni J Vilhjálmsson, Vincent Segura, Alexander Platt, Quan Long, and Magnus Nordborg (2012). "A mixed-model approach for genome-wide association studies of correlated traits in structured populations". In: *Nature genetics* 44.9, p. 1066.

Li, Heng and Richard Durbin (2010). "Fast and accurate long-read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 26.5, pp. 589–595.

Mace, Emma S, Shuaishuai Tai, Edward K Gilding, Yanhong Li, Peter J Prentis, Lianle Bian, Bradley C Campbell, Wushu Hu, David J Innes, Xuelian Han, et al. (2013). "Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum". In: *Nature communications* 4, p. 2320.

Morris, Geoffrey P, Punna Ramu, Santosh P Deshpande, C Thomas Hash, Trushar Shah, Hari D Upadhyaya, Oscar Riera-Lizarazu, Patrick J Brown, Charlotte B Acharya, Sharon E Mitchell, et al. (2013). "Population genomic and genome-wide association studies of agroclimatic traits in sorghum". In: *Proceedings of the National Academy of Sciences* 110.2, pp. 453–458.

Murray, Seth C, Arun Sharma, William L Rooney, Patricia E Klein, John E Mullet, Sharon E Mitchell, and Stephen Kresovich (2008). "Genetic improvement of sorghum as a biofuel feedstock: I. QTL for stem sugar and grain nonstructural carbohydrates". In: *Crop Science* 48.6, pp. 2165–2179.

Myles, Sean, Jason Peiffer, Patrick J Brown, Elhan S Ersoz, Zhiwu Zhang, Denise E Costich, and Edward S Buckler (2009). "Association mapping: critical considerations shift from genotyping to experimental design". In: *The Plant Cell* 21.8, pp. 2194–2202.

Olson, Andrew, Robert R Klein, Diana V Dugas, Zhenyuan Lu, Michael Regulski, Patricia E Klein, and Doreen Ware (2014). "Expanding and vetting Sorghum bicolor gene annotations through transcriptome and methylome sequencing". In: *The Plant Genome* 7.2.

Pan, Zhiqiang, Agnes M Rimando, Scott R Baerson, Mark Fishbein, and Stephen O Duke (2007). "Functional characterization of desaturases involved in the formation of the terminal double bond of an unusual 16: 3Δ9, 12, 15 fatty acid isolated from Sorghum bicolor root hairs". In: *Journal of Biological Chemistry* 282.7, pp. 4326–4335.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American journal of human genetics* 81.3, pp. 559–575.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Rami, J-F, Philippe Dufour, Gilles Trouche, Geneviève Fliedel, Christian Mestres, Fabrice Davrieux, P Blanchard, and Perla Hamon (1998). "Quantitative trait loci for grain quality, productivity, morphological and agronomical traits in sorghum (Sorghum bicolor L. Moench)". In: *Theoretical and applied genetics* 97.4, pp. 605–616.

Reif, Jochen C, Manje Gowda, Hans P Maurer, CFH Longin, Viktor Korzun, Erhard Ebmeyer, Reiner Bothe, Christof Pietsch, and Tobias Würschum (2011). "Association mapping for quality traits in soft winter wheat". In: *Theoretical and Applied Genetics* 122.5, pp. 961–970.

Rhodes, Davina H, Leo Hoffmann, William L Rooney, Thomas J Herald, Scott Bean, Richard Boyles, Zachary W Brenton, and Stephen Kresovich (2017). "Genetic architecture of kernel composition in global sorghum germplasm". In: *BMC genomics* 18.1, p. 15.

Rice, Brian R, Samuel B Fernandes, and Alexander E Lipka (2020). "Multi-Trait Genome-wide Association Studies Reveal Loci Associated with Maize Inflorescence and Leaf Architecture". In: *Plant and Cell Physiology*.

Sapkota, Sirjan, Rick Boyles, Elizabeth Cooper, Zachary Brenton, Matthew Myers, and Stephen Kresovich (Jan. 2020). "Impact of sorghum racial structure and diversity on genomic prediction of grain yield components". In: *Crop Science*. DOI: `10.1002/csc2.20060`.

Sukumaran, Sivakumar, Wenwen Xiang, Scott R Bean, Jeffrey F Pedersen, Stephen Kresovich, Mitchell R Tuinstra, Tesfaye T Tesso, Martha T Hamblin, and Jianming Yu (2012). "Association mapping for grain quality in a diverse sorghum collection". In: *The Plant Genome* 5.3, pp. 126–135.

Swarts, Kelly, Huihui Li, J Alberto Romero Navarro, Dong An, Maria Cinta Romay, Sarah Hearne, Charlotte Acharya, Jeffrey C Glaubitz, Sharon Mitchell, Robert J Elshire, et al. (2014). "Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants". In: *The Plant Genome* 7.3.

Thoen, Manus PM, Nelson H Davila Olivas, Karen J Kloth, Silvia Coolen, Ping-Ping Huang, Mark GM Aarts, Johanna A Bac-Molenaar, Jaap Bakker, Harro J Bouwmeester, Colette Broekgaarden,

et al. (2017). “Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping”. In: *New Phytologist* 213.3, pp. 1346–1362.

VanRaden, Paul M (2008). “Efficient methods to compute genomic predictions”. In: *Journal of dairy science* 91.11, pp. 4414–4423.

Vierling, Elizabeth (1991). “The roles of heat shock proteins in plants”. In: *Annual review of plant biology* 42.1, pp. 579–620.

Wang, Xiaoqian, Yunlong Pang, Jian Zhang, Zhichao Wu, Kai Chen, Jauhar Ali, Guoyou Ye, Jianlong Xu, and Zhikang Li (2017). “Genome-wide and gene-based association mapping for rice eating and cooking characteristics and protein content”. In: *Scientific reports* 7.1, pp. 1–10.

Wilson, Larissa M, Sherry R Whitt, Ana M Ibáñez, Torbert R Rocheford, Major M Goodman, and Edward S Buckler (2004). “Dissection of maize kernel composition and starch production by candidate gene association”. In: *The Plant Cell* 16.10, pp. 2719–2733.

Yang, Qiong and Yuanjia Wang (2012). “Methods for analyzing multivariate phenotypes in genetic association studies”. In: *Journal of probability and statistics* 2012.

Yang, Qiong, Hongsheng Wu, Chao-Yu Guo, and Caroline S Fox (2010). “Analyze multivariate phenotypes in genetic association studies by combining univariate association tests”. In: *Genetic epidemiology* 34.5, pp. 444–454.

Yu, Ying, Helen He Mu, Chen Mu-Forster, and Bruce P Wasserman (1998). “Polypeptides of the maize amyloplast stroma: stromal localization of starch-biosynthetic enzymes and identification of an 81-kilodalton amyloplast stromal heat-shock cognate”. In: *Plant physiology* 116.4, pp. 1451–1460.

Zhao, Keyan, Chih-Wei Tung, Georgia C Eizenga, Mark H Wright, M Liakat Ali, Adam H Price, Gareth J Norton, M Rafiqul Islam, Andy Reynolds, Jason Mezey, et al. (2011). “Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa”. In: *Nature communications* 2.1, pp. 1–10.

Zhou, Xiang and Matthew Stephens (2012). “Genome-wide efficient mixed-model analysis for association studies”. In: *Nature genetics* 44.7, pp. 821–824.

Zhou, Xiang and Matthew Stephens (2014). “Efficient multivariate linear mixed model algorithms for genome-wide association studies”. In: *Nature methods* 11.4, p. 407.

Zhu, Fan (2014). “Structure, physicochemical properties, modifications, and uses of sorghum starch”. In: *Comprehensive Reviews in Food Science and Food Safety* 13.4, pp. 597–610.