12-15-2014

Diversity, Genetics, and Health Benefits of Sorghum Grain

Davina Rhodes
University of South Carolina - Columbia

DIVERSITY, GENETICS, AND HEALTH BENEFITS OF SORGHUM GRAIN

by

Davina Rhodes

Bachelor of Arts
New College of Florida, 1999

Master of Science
University of Illinois-Chicago, 2007

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Biological Sciences

College of Arts & Sciences

University of South Carolina

2014

Accepted by:

Stephen Kresovich, Major Professor

Erin Connolly, Committee Member

Jill Anderson, Committee Member

Lydia Matesic, Committee Member

Mitzi Nagarkatti, Committee Member

Lacy Ford, Vice Provost and Dean of Graduate Studies

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Stephen Kresovich, for giving me the opportunity to conduct a unique and truly integrative biological research project, and for encouraging the independence and confidence needed to advance in my research career. It has been an honor to train with such a talented, thoughtful, and generous teacher and mentor. I would also like to thank my committee members, Dr. Jill Anderson, Dr. Erin Connolly, Dr. Lydia Matesic, and Dr. Mitzi Nagarkatti, who each provided invaluable support and guidance throughout my research. Many thanks to my labmates, Nadia Shakoor, Rick Boyles, Zack Brenton, and Matt Myers, who have given me help at many points throughout this process, including extensive assistance with fieldwork. I am grateful to Dr. Giamila Fantuzzi, who gave me the opportunity to train in her lab for several years, encouraged me to pursue my PhD. I would like to express my appreciation to the United Sorghum Checkoff Program for showing an interest in my research and providing financial support. Finally, I would like to thank Geoffrey Morris, whose intellectual and emotional support has been invaluable throughout this process.

ABSTRACT

Staple cereal crops provide the majority of nutrients to the world's population, and thus, can significantly impact human nutrition and health. Phenotypic and genetic diversity within a crop can be useful for biofortification and crop improvement, but quantitative phenotyping is needed to identify varieties with high or low concentrations of a nutrient of interest, and to identify alleles responsible for quantitative trait variation of the nutrient. Sorghum [*Sorghum bicolor* (L.) Moench] is a diverse and widely adapted cereal crop that provides food for more than 500 million people in sub-Saharan Africa and Asia, and is becoming increasingly popular in specialty grain products in the United States. Sorghum is a valuable resource for nutrient diversity, as adaptation to different environments has led to extensive phenotypic and genetic diversity in the crop.

Many sorghum varieties are rich in flavonoids, primarily 3-deoxyanthocyanidins and proanthocyanidins, which appear to protect against chronic inflammatory diseases. Most studies have only explored the health benefits of a small number of sorghum accessions, but over 45,000 sorghum accessions exist in crop gene banks. A large genetically diverse sorghum panel can be used to identify varieties with high concentrations of flavonoids and to explore the effects of natural variation of sorghum flavonoids on inflammation. This same resource can also be used to identify varieties with high concentrations of protein, fat, or starch, which can lead to improved nutritional value of sorghum grain.

The overall aim of my dissertation project was to quantify sorghum flavonoids and identify allelic variants controlling them; quantify grain composition more broadly (protein, fat, and starch) and identify allelic variants controlling them; and investigate anti-inflammatory properties of sorghum extracts with contrasting levels of flavonoids. Using a large germplasm resource (USDA National Plant Germplasm System), high-throughput methods of phenotyping (near-infrared spectroscopy) and genotyping (genotyping-by-sequencing), association mapping (genome-wide association studies), and *in vitro* inflammation models, the work presented here provides new insights into the diversity, genetics, and anti-inflammatory properties of sorghum nutrients that are important to human health. It provides a survey of grain nutrient diversity in a large global panel of sorghum, identifies quantitative trait loci and candidate genes for underlying controls of these nutrients, and demonstrates that a larger variety of sorghum accessions than previously thought have anti-inflammatory properties.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

Undernutrition is present in many regions of the world, and leads to increased risk of infectious disease, stunted growth, and severe wasting. At the same time, overnutrition has also become prevalent in the global population, and is strongly correlated with chronic diseases such as type 2 diabetes, cardiovascular disease, and cancer. Staple cereal crops provide the majority of nutrients to the world's population, and thus have significant impact on human nutrition and the negative health effects of undernutrition and overnutrition.

Many studies are now focusing on the health benefits of whole grains, especially in relation to the chronic inflammatory diseases seen in overnutrition [4–10]. Flavonoids, a large diverse group of polyphenols comprised of more than 8,000 compounds, appear to contribute to the beneficial health effects of whole grains[11–13]. Most plant-based foods contain flavonoids, making them some of the most ubiquitous polyphenols in the human diet. Fruits, tea, chocolate, red wine, and coffee are rich sources of flavonoids, but are only small contributions to our daily calorie intake compared to grain, which provides between 24% and 80% of our daily energy[14]. In humans, dietary flavonoids are thought to act as antioxidants and signaling molecules, and their consumption is correlated with lower incidence of cardiovascular disease, cancer, type II diabetes, neurodegenerative disease, and other chronic diseases.[15] Potential anti-inflammatory effects of flavonoids have been studied extensively in the last decade, with particular focus on validating observed health benefits in green tea, grapes, and cranberries[16–20]. The anti-inflammatory mechanisms are not fully understood, but are thought to involve scavenging of free radicals, prevention of lipid peroxidation, inhibition of pro-inflammatory cytokines, and

modulation of gene expression [21–23]. Certain varieties of grains also contain polyphenols, including varieties of wheat, rice, maize, and sorghum [12,24–27].

Sorghum [*Sorghum bicolor* (L.) Moench] is one of the world's major cereal crops and a dietary staple for more than 500 million people in Asia and sub-Saharan Africa[28]. In the United States, it is used primarily as livestock feed and, increasingly for ethanol production. However, it is beginning to be used in food products, due to a rise in demand for specialty grains, especially those that are gluten free[29–33]. Sorghum's grain composition is similar to maize and wheat, providing, on average, 70% carbohydrate, 12% protein, and 3% fat. As in other cereals, the sorghum grain is predominantly starchy. The endosperm contains the majority of the starch and protein, while the germ contains the majority of the fat. Protein deficiency is a major cause of undernutrition in regions where a single cereal crop is the primary source of protein. Sorghum, as with other serial crops, does not provide adequate protein to meet nutritional needs on its own, so understanding the genetic controls of high protein could lead to improved nutritional quality of sorghum.

Sorghum's two major flavonoids—proanthocyanidins and 3-deoxyanthocyanidins—appear to have health-protective effects that may be superior to many of the more popularly consumed grains[34], fruits and vegetables [35]. This is possibly because sorghum, which evolved in a tropical climate with an exposed grain, contains some of the highest concentrations of proanthocyanidins in any plant-based food [36], and is the only known dietary source of 3-deoxyanthocyanidins[37–39]. Sorghum has the potential to alleviate negative health effects of obesity [40,41], diabetes [42,43], cancer [44–47], cardiovascular disease [48,49], and other chronic diseases [34,50]. The bulk of research on

sorghum health effects has been on its powerful antioxidant activity, but recent studies suggest that sorghum flavonoids also possess anti-inflammatory activity [34,50–52]. Some varieties of sorghum do not contain measurable amounts of polyphenols, while others contain high levels of polyphenols [35,53] Most studies have only explored the health benefits of a small number of sorghum accessions (distinct varieties of plants), but over 45,000 sorghum accessions are available from the U.S. National Plant Germplasm System's Germplasm Resources Information Network (GRIN) [54]. Utilizing accessions that are readily available from a crop gene bank allows for authentication of the accessions and reproducibility of the experiments. Using a large genetically diverse sorghum panel to explore the effects of natural variation of sorghum polyphenols on inflammation will help in discovering particularly beneficial accessions. Additionally, although several studies comparing health effects between sorghums with or without proanthocyanidins and 3-deoxyanthocyanidins have been conducted, none of them controlled for genetic background of the sorghums or utilized accessions that were readily available from crop gene banks [34,41,43,50]. Without adequate control of other genetic factors it may not be possible to attribute health effects to polyphenols *per se*.

Investigations into the health-benefits of food crops need to be conducted in parallel to an exploration of the natural diversity and genetic controls of important nutrients in food crops. Sorghum is a good system for cereal genomics, with a small genome (at ~730 Mb) that is fully sequenced. Crop improvement efforts aim to move desirable traits (such as high protein or flavonoids) found in underutilized germplasm into existing elite varieties that already contain traits needed for agricultural production (e.g., high yield). High concentrations of flavonoids are not found in many commonly

consumed cereals, such as wheat, rice, and maize [55], however, sorghum provides a valuable resource for flavonoids, as adaptation to different environments has led to extensive phenotypic and genetic diversity in the crop.[56,57] This diversity can be useful for crop improvement, but quantitative phenotyping is needed to identify accessions with high concentrations of flavonoids, as well as protein, and to identify quantitative trait loci (QTL; loci that are linked to the allele responsible for the trait variation) associated with variation in grain nutrients (reviewed by Flint-Garcia[58]). These QTL can be used in marker-assisted selection to accurately and efficiently breed for the trait of interest.

1.2 GOALS AND SIGNIFICANCE

The long-term goal of my research is to identify natural variation in food-plant nutrients that is useful for human health, specifically by connecting crop genomic resources with human nutrition research. The overall aim of my dissertation project is to quantify sorghum grain composition traits (protein, fat, starch, and polyphenols) and identify allelic variants controlling them (chapters 2 and 3), and to investigate the anti-inflammatory properties of sorghum extracts with contrasting levels of polyphenols (chapter 4).

1.3 CHAPTER SUMMARIES

In Chapter 2, the genetics of flavonoids are reviewed. I quantify total phenols, proanthocyanidins, and 3-deoxyanthocyanidins in a global sorghum diversity panel using near-infrared spectroscopy (NIRS) and characterize the patterns of variation with respect to geographic origin and botanical race. I identify novel quantitative trait loci for sorghum polyphenols, some of which colocalize with homologs of flavonoid pathway

genes from other crops, including an ortholog of maize (*Zea mays*) *Pr1* and a homolog of Arabidopsis (*Arabidopsis thaliana*) *TT16*. This survey of grain polyphenol variation in sorghum germplasm and catalog of flavonoid pathway-associated loci contributes toward the goal of producing sorghum crops that will contribute to marker-assisted breeding of sorghum crops that will benefit human health.

In Chapter 3, I quantify protein, fat, and starch in a global sorghum diversity panel using NIRS, identify novel QTL for sorghum grain composition using GWAS with 404,628 SNP markers, and use a published sorghum transcriptome atlas to identify candidate genes within the GWAS QTL regions, including *NAM-B1*, *AMY3*, and *SSIIb*. This survey of grain composition in sorghum germplasm and identification of QTL significantly associated with protein, fat, and starch, contributes to our understanding of the genetic basis of natural variation in sorghum grain composition.

In Chapter 4, features of inflammation are reviewed. I evaluate the anti-inflammatory effects of ethanol extracts from the bran of twenty sorghum accessions with comparable genetic backgrounds, using lipopolysaccharide (LPS)-induced mouse macrophage cells. The results demonstrate that sorghum accessions differentially modulate inflammation, with many accessions reducing the production of pro-inflammatory cytokines tumor necrosis factor (TNF)-α and interleukin (IL)-6, possibly by decreasing phosphorylation of NF-κB. Additionally, the results demonstrate that the RAW 264.7 model of inflammation is a good method for high throughput screening of sorghum accessions. The chapter on sorghum grain protein, fat and starch (chapter 3) was conducted with undernutrition in mind, while the chapters on sorghum grain polyphenols (chapters 2 and 4) were conducted with overnutrition in mind.

1.4 REFERENCES

1. Gross, L. S., Li, L., Ford, E. S. & Liu, S. Increased consumption of refined carbohydrates and the epidemic of type 2 diabetes in the united states: an ecologic assessment. *Am J Clin Nutr* **79,** 774–779 (2004).

2. Cordain, L. *et al.* Origins and evolution of the western diet: health implications for the 21st century. *Am J Clin Nutr* **81,** 341–354 (2005).

3. Slavin, J. L., Jacobs, D., Marquart, L. & Wiemer, K. The role of whole grains in disease prevention. *Journal of the American Dietetic Association* **101,** 780–785 (2001).

4. Slavin, J., Jacobs, D. & Marquart, L. Whole□grain consumption and chronic disease: Protective mechanisms. *Nutrition and Cancer* **27,** 14–21 (1997).

5. McKeown, N. M., Meigs, J. B., Liu, S., Wilson, P. W. & Jacques, P. F. Whole-grain intake is favorably associated with metabolic risk factors for type 2 diabetes and cardiovascular disease in the Framingham Offspring Study. *Am J Clin Nutr* **76,** 390–398 (2002).

6. Fung, T. T. *et al.* Whole-grain intake and the risk of type 2 diabetes: a prospective study in men. *Am J Clin Nutr* **76,** 535–540 (2002).

7. Liu, S. *et al.* Relation between changes in intakes of dietary fiber and grain products and changes in weight and development of obesity among middle-aged women. *Am J Clin Nutr* **78,** 920–927 (2003).

8. Cho, S. S., Qi, L., Fahey, G. C. & Klurfeld, D. M. Consumption of cereal fiber, mixtures of whole grains and bran, and whole grains and risk reduction in type 2 diabetes, obesity, and cardiovascular disease. *Am J Clin Nutr* ajcn.067629 (2013). doi:10.3945/ajcn.113.067629

9. He, M., Dam, R. M. van, Rimm, E., Hu, F. B. & Qi, L. Whole-grain, cereal fiber, bran, and germ intake and the risks of all-cause and cardiovascular disease–specific mortality among women with type 2 diabetes mellitus. *Circulation* **121,** 2162–2168 (2010).

10. Aune, D. *et al.* Dietary fibre, whole grains, and risk of colorectal cancer: systematic review and dose-response meta-analysis of prospective studies. *BMJ* **343,** d6617–d6617 (2011).

11. Fardet, A. New hypotheses for the health-protective mechanisms of whole-grain cereals: what is beyond fibre? *Nutrition Research Reviews* **23,** 65–134 (2010).

12. Liu, R. H. Whole grain phytochemicals and health. *Journal of Cereal Science* **46,** 207–219 (2007).

13. Borneo, R. & León, A. E. Whole grain cereals: functional components and health

benefits. *Food & Function* **3,** 110 (2012).

14. US Department of Agriculture. *Nutrient Content of the U.S. Food Supply, 1909-2010. Center for Nutrition Policy and Promotion. http://www.cnpp.usda.gov/USFoodSupply-1909-2010.htm*. (2011).

15. Del Rio, D. *et al.* Dietary (poly)phenolics in human health: structures, bioavailability, and evidence of protective effects against chronic diseases. *Antioxid. Redox Signaling* **18,** 1818–1892 (2013).

16. Dixon, R. A., Xie, D.-Y. & Sharma, S. B. Proanthocyanidins – a final frontier in flavonoid research? *New Phytol.* **165,** 9–28 (2005).

17. Aron, P. M. & Kennedy, J. A. Flavan-3-ols: nature, occurrence and biological activity. *Mol Nutr Food Res* **52,** 79–104 (2008).

18. Cabrera, C., Artacho, R. & Giménez, R. Beneficial effects of green tea—a review. *Journal of the American College of Nutrition* **25,** 79–99 (2006).

19. Neto, C. C. Cranberry and its phytochemicals: a review of in vitro anticancer studies. *J. Nutr.* **137,** 186S–193S (2007).

20. Dohadwala, M. M. & Vita, J. A. Grapes and cardiovascular disease. *J. Nutr.* **139,** 1788S–1793S (2009).

21. Rathee, P. *et al.* Mechanism of action of flavonoids as anti-inflammatory agents: a review. *Inflammation & Allergy - Drug Targets* **8,** 229–235 (2009).

22. Gomes, A., Fernandes, E., Lima, J. L. F. C., Mira, L. & Corvo, M. L. Molecular mechanisms of anti-inflammatory activity mediated by flavonoids. *Current Medicinal Chemistry* **15,** 1586–1605 (2008).

23. Pan, M.-H., Lai, C.-S. & Ho, C.-T. Anti-inflammatory activity of natural dietary flavonoids. *Food & Function* **1,** 15 (2010).

24. Winkel-Shirley, B. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol.* **126,** 485–493 (2001).

25. Himi, E. & Noda, K. Red grain colour gene (R) of wheat is a Myb-type transcription factor. *Euphytica* **143,** 239–242 (2005).

26. Abdel-Aal, E.-S. M., Young, J. C. & Rabalski, I. Anthocyanin composition in black, blue, pink, purple, and red cereal grains. *J. Agric. Food Chem.* **54,** 4696–4704 (2006).

27. Adom, K. K., Sorrells, M. E. & Liu, R. H. Phytochemical profiles and antioxidant activity of wheat varieties. *J. Agric. Food Chem.* **51,** 7825–7834 (2003).

28. FAO. Sorghum and millets in human nutrition. http://www.fao.org/docrep/T0818E/T0818E04.htm. (1995).

29. Janzen, Edward L., W. W. W. Cooperative marketing in specialty grains and identity

preserved grain markets. *North Dakota State University, Department of Agribusiness and Applied Economics, Agribusiness & Applied Economics Report* (2002).

30. Taylor, J. R. N., Schober, T. J. & Bean, S. R. Novel food and non-food uses for sorghum and millets. *J. Cereal Sci.* **44,** 252–271 (2006).

31. US Department of Agriculture, Center for Nutrition Policy and Promotion. *Trends in Dietary Fiber in the U.S. Food Supply; Sales of Grain Products. CNPP Fact Sheet No.2 http://www.cnpp.usda.gov/Publications/FoodSupply/FiberFactSheet.pdf.* (2007).

32. Cureton, P. & Fasano, A. in *Gluten-free food science and technology* (ed. Gallagher, E.) 1–15 (Wiley-Blackwell, 2009). at <http://onlinelibrary.wiley.com/doi/10.1002/9781444316209.ch1/summary>

33. Lemlioglu-Austin, D., Turner, N. D., McDonough, C. M. & Rooney, L. W. Effects of sorghum [*Sorghum bicolor* (L.) Moench] crude extracts on starch digestibility, estimated glycemic index (EGI), and resistant starch (RS) contents of porridges. *Molecules* **17,** 11124–11138 (2012).

34. Burdette, A. *et al.* Anti-inflammatory activity of select sorghum (*Sorghum bicolor*) Brans. *Journal of Medicinal Food* **13,** 879–887 (2010).

35. Awika, J. M. & Rooney, L. W. Sorghum phytochemicals and their potential impact on human health. *Phytochemistry* **65,** 1199–1221 (2004).

36. Gu, L. *et al.* Concentrations of proanthocyanidins in common foods and estimations of normal consumption. *J. Nutr.* **134,** 613–617 (2004).

37. Winefield, C. S. *et al.* Investigation of the biosynthesis of 3-deoxyanthocyanins in Sinningia cardinalis. *Physiol. Plant.* **124,** 419–430 (2005).

38. Sharma, M. *et al.* Expression of flavonoid 3'-hydroxylase is controlled by P1, the regulator of 3-deoxyflavonoid biosynthesis in maize. *BMC Plant Biol.* **12,** 196 (2012).

39. Malathi, P. *et al.* Differential accumulation of 3-deoxy anthocyanidin phytoalexins in sugarcane varieties varying in red rot resistance in response to *Colletotrichum falcatum* infection. *Sugar Tech* **10,** 154–157 (2008).

40. Park, M.-Y. *et al.* Effects of *Panicum miliaceum* L. extract on adipogenic transcription factors and fatty acid accumulation in 3T3-L1 adipocytes. *Nutrition Research and Practice* **5,** 192 (2011).

41. Moraes, É. A. *et al.* Sorghum genotype may reduce low-grade inflammatory response and oxidative stress and maintains jejunum morphology of rats fed a hyperlipidic diet. *Food Research International* **49,** 553–559 (2012).

42. Farrar, J. L., Hartle, D. K., Hargrove, J. L. & Greenspan, P. A novel nutraceutical property of select sorghum (*Sorghum bicolor*) brans: inhibition of protein glycation.

Phytother Res **22,** 1052–1056 (2008).

43. Hargrove, J. L., Greenspan, P., Hartle, D. K. & Dowd, C. Inhibition of aromatase and α-amylase by flavonoids and proanthocyanidins from Sorghum bicolor bran extracts. *J Med Food* **14,** 799–807 (2011).

44. Isaacson, C. The change of the staple diet of black South Africans from sorghum to maize (corn) is the cause of the epidemic of squamous carcinoma of the oesophagus. *Medical Hypotheses* **64,** 658–660 (2005).

45. Yang, L., Browning, J. D. & Awika, J. M. Sorghum 3-deoxyanthocyanins possess strong phase II enzyme inducer activity and cancer cell growth inhibition properties. *J. Agric. Food Chem.* **57,** 1797–1804 (2009).

46. Wu, L. *et al.* Chemical characterization of a procyanidin-rich extract from sorghum bran and its effect on oxidative stress and tumor inhibition in vivo. *J. Agric. Food Chem.* **59,** 8609–8615 (2011).

47. Devi, P. S., Kumar, M. S. & Das, S. M. Evaluation of antiproliferative activity of red sorghum bran anthocyanin on a human breast cancer cell line (mcf-7). *Int J Breast Cancer* **2011,** 891481 (2011).

48. Carr, T. P. *et al.* Grain sorghum lipid extract reduces cholesterol absorption and plasma non-HDL cholesterol concentration in hamsters. *J. Nutr.* **135,** 2236–2240 (2005).

49. Lee, S. M. & Pan, B. S. Effect of dietary sorghum distillery residue on hematological characteristics of cultured grey mullet (*Mugil cephalus*)—an animal model for prescreening antioxidant and blood thinning activities. *Journal of Food Biochemistry* **27,** 1–18 (2003).

50. Bralley, E., Greenspan, P., Hargrove, J. L. & Hartle, D. K. Inhibition of hyaluronidase activity by select sorghum brans. *J Med Food* **11,** 307–312 (2008).

51. Benson, K. F. *et al.* West African Sorghum bicolor leaf sheaths have anti-inflammatory and immune-modulating properties in vitro. *J Med Food* **16,** 230–238 (2013).

52. Shim, T.-J., Kim, T. M., Jang, K. C., Ko, J.-Y. & Kim, D. J. Toxicological evaluation and anti-inflammatory activity of a golden gelatinous sorghum bran extract. *Biosci. Biotechnol. Biochem.* **77,** 697–705 (2013).

53. Rhodes, D. H. *et al.* Genome-wide association study of grain polyphenol concentrations in global sorghum [*Sorghum bicolor* (L.) Moench] germplasm. *J. Agric. Food Chem.* (2014). doi:10.1021/jf503651t

54. USDA. GRIN National Genetic Resources Program. URL (http://www.ars-grin.gov) (2014-07-25). (2014).

55. Olsen, K. M. & Wendel, J. F. Crop plants as models for understanding plant

adaptation and diversification. *Front. Plant Sci.* **4,** (2013).

56. Harlan, J. R., Wet, D. & J, J. M. A simplified classification of cultivated sorghum. *Crop Science* **12,** 172–176 (1972).

57. Morris, G. P. *et al.* Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 453–458 (2013).

58. Flint-Garcia, S. A. Genetics and consequences of crop domestication. *J. Agric. Food Chem.* **61,** 8267–8276 (2013).

CHAPTER 2

GENOME-WIDE ASSOCIATION STUDY OF GRAIN POLYPHENOL CONCENTRATIONS IN GLOBAL SORGHUM [*SORGHUM BICOLOR* (L.) MOENCH] GERMPLASM[1]

[1] Reproduced with permission from Rhodes, D. H. *et al.* Genome-wide association study of grain polyphenol concentrations in global sorghum [*Sorghum bicolor* (L.) Moench] germplasm. *J. Agric. Food Chem.* (2014), *62*, 10916–10927. Copyright 2014 American Chemical Society.

2.1 ABSTRACT

Identifying natural variation of health-promoting compounds in staple crops and characterizing its genetic basis can help improve human nutrition through crop biofortification. Some varieties of sorghum, a staple cereal crop grown worldwide, have high concentrations of proanthocyanidins and 3-deoxyanthocyanidins, polyphenols with antioxidant and anti-inflammatory properties. We quantified total phenols, proanthocyanidins, and 3-deoxyanthocyanidins in a global sorghum diversity panel (n = 381) using near-infrared spectroscopy (NIRS), and characterized the patterns of variation with respect to geographic origin and botanical race. A genome-wide association study (GWAS) with 404,628 SNP markers identified novel quantitative trait loci for sorghum polyphenols, some of which colocalized with homologs of flavonoid pathway genes from other plants, including an ortholog of maize (*Zea mays*) *Pr1* and a homolog of Arabidopsis (*Arabidopsis thaliana*) *TT16*. This survey of grain polyphenol variation in sorghum germplasm and catalog of flavonoid pathway loci may be useful to guide future enhancement of cereal polyphenols.

2.2 INTRODUCTION

Polyphenols are a large diverse group of phytochemicals that include phenolic acids, stilbenes, lignans, isoflavonoids, and flavonoids.[1] All flavonoids share a common C6-C3-C6 backbone structure but differ in their oxidation level, glycosylation, acylation, and hydroxyl and methyl substitutions, allowing for an enormous variety of structure and function.[2] In plants, flavonoid secondary metabolites are involved in growth, pigmentation, pollination, and defense against pathogens, predators, and physical

factors.[3] In humans, dietary flavonoids are thought to act as antioxidants and signaling molecules, and their consumption is correlated with lower incidence of cardiovascular disease, cancer, type II diabetes, neurodegenerative disease, and other chronic illnesses.[4] Most plant-based foods contain flavonoids, making them some of the most ubiquitous polyphenols in the human diet. Polymerization of flavonoids yields complex compounds including proanthocyanidins, flavonoid polymers predominantly composed of flavan-3-ols, which are abundant in food plants. Proanthocyanidins contribute to the astringency and bitterness found in foods such as wine, cocoa, beans, and fruits, but they are not present in most commonly consumed vegetables and cereals.[5] They are also often considered anti-nutrients due to their nutrient binding capacity, especially to proteins and iron.[6] In the last decade, however, potential health protective effects of proanthocyanidins have been studied extensively, with particular focus on their contributions to observed health benefits of grape and cranberry. [7]

Sorghum is one of the world's major cereal crops and a dietary staple for more than 500 million people in sub-Saharan Africa and Asia.[8] In the Unites States, it is primarily used as animal feed, but is becoming more popular in food products due to a rise in demand for specialty grains, especially those that are gluten-free.[9–12] Domesticated sorghum has been classified into five major races (bicolor, guinea, caudatum, kafir, and durra) and 10 intermediate races (all combinations of the major races), based on morphological differences.[13] Two of the major polyphenol compounds in sorghum grain are proanthocyanidin and 3-deoxyanthocyanidin. Consumption of these two polyphenols has been correlated with several health benefits including protection against oxidative damage, inflammation, obesity, and diabetes.[14] Proanthocyanidins are constitutively

expressed, while 3-deoxyanthocyanidins are phytoalexins, expressed only in response to fungal infection.[15,16] Sorghum grain is the only known dietary source of 3-deoxyanthocyanidins, which otherwise have only been found in the flowers of sinningia (Sinningia cardinalis), the silk tissues of maize (Zea mays), and the stalks of sugarcane (Saccharum sp.).[17–19]

In sorghum grains, polyphenol compounds can be found in the pericarp (outer seed coat) and the testa (inner layer of tissue between the pericarp and the endosperm). A number of classical loci identified by their effects on grain color and testa presence control the presence or absence of polyphenol compounds in sorghum. [20] Genotypes with dominant alleles at the B1 and B2 loci have proanthocyanidins in the testa. Genotypes with a dominant allele at the spreader (S) locus, as well as dominant alleles at the B1 and B2 loci, have proanthocyanidins in both the pericarp and the testa, often, but not always, resulting in a brown appearance to the grain. The base pericarp color is red, yellow or white, and these colors are controlled by the R and Y loci. The S locus, and additional loci such as intensifier (I) and mesocarp thickness (Z), modify the base pericarp color, resulting in a range of colors from brilliant white to black with various shades of red, yellow, pink, orange, and brown among sorghum genotypes (see Figure 2.1). Using mutants for seed color traits, the biochemical and regulatory pathways underlying flavonoids and flavonoid products have been almost completely elucidated in Arabidopsis and maize, and extensively studied in other species (Table 2.1).[21] Therefore, homology can be used as a guide to discover genes involved in the sorghum flavonoid pathway. The gene underlying the B2 locus was recently cloned and designated *Tannin1*, along with two nonfunctional alleles of *Tannin1*, *tan1-a* and *tan1-b.*[22] *Tannin1* encodes a

WD40 protein homologous to the Arabidopsis proanthocyanidin regulator *transparent testa glabra1* (*TTG1).* The gene underlying the Y locus has also been cloned and designated *Yellow seed1*. *Yellow seed1* encodes a MYB protein, orthologous to the maize 3-deoxyanthocyanidin regulator *P1*, that is needed for accumulation of 3-deoxyanthocyanidins in the sorghum pericarp.[23] The R locus has been mapped to chromosome 3 between 57-59 Mb and the Z locus has been mapped to chromosome 2 between 56-57 Mb [24], but the underlying genes have not been identified.

While the genetic controls of polyphenol presence/absence have been well-studied using mutant lines and nonfunctional polymorphisms, there has been little study of quantitative natural variation in polyphenols.[25] Polyphenol nonfunctional mutations were strongly selected during cereal domestication, when bitter tasting and/or dark compounds were partly or completely lost in most cereals, including wheat, rice, and maize.[26] However, sorghum provides a valuable resource for polyphenol diversity, as adaptation to different environments has led to extensive phenotypic and genetic diversity in the crop.[13,27] This diversity can be useful for biofortification and crop improvement (e.g. desirable traits can be bred into existing elite varieties), but quantitative phenotyping is needed to identify alleles responsible for quantitative trait variation in grain polyphenols (reviewed by Flint-Garcia [28]). The goals of this study were to quantify the natural variation of two of the major sorghum grain polyphenols (proanthocyanidins and 3-deoxyanthocyanidins) and to identify single-nucleotide polymorphisms (SNPs) that are associated with low or high polyphenol concentrations using genome-wide association studies (GWAS). GWAS are used to map the genomic regions underlying phenotypic variation (known as quantitative trait loci) by scanning the genome for statistical

associations between genetic variation and phenotypic variation.[29] In contrast to the biparental linkage mapping approach, GWAS takes advantage of historical recombinations in a diverse panel and linkage disequilibrium between causal variants and nearby SNP markers. Although it has been used extensively to identify putative genetic controls of human disease, [30] it is a relatively new but promising tool in plant genomics.[27,31,32] Here we present a survey of the quantitative natural variation of polyphenols in a diverse worldwide panel of sorghum and a catalog of flavonoid-associated loci across the sorghum genome.

2.3 MATERIALS AND METHODS

2.3.1 Plant Materials

We investigated a total of 381 sorghum accessions, comprising 308 accessions from the Sorghum Association Panel (SAP) [33] and an additional 73 accessions selected based on presence of a pigmented testa using the U.S. National Plant Germplasm System's Germplasm Resources Information Network (GRIN).[34] The SAP includes accessions from all major cultivated races and geographic centers of diversity in sub-Saharan Africa and Asia, as well as important breeding lines from the United States. The 73 additional accessions were included to increase the proportion of accessions with high proanthocyanidins.

Seeds were obtained through GRIN and planted in late April 2012 at Clemson University Pee Dee Research and Education Center in Florence, SC. A twofold replicated complete randomized block design was used. Panicles from each plot were collected at physiological maturity (signified by a black layer at the base of the seed that normally

forms about 35 days after anthesis). Due to differences in maturity among these accessions, harvest occurred between September and October. Once harvested, panicles were air dried in a greenhouse and then mechanically threshed and any remaining glumes were removed with a Wheat Head Thresher (Precision Machine Company, Lincoln, NE).

2.3.2 Phenotyping

Twenty grams of cleaned whole grain from one replicate were scanned with a FOSS XDS spectrometer (FOSS North America, Eden Prairie, MN, USA) at a wavelength range of 400-2500 nm. To determine reproducibility, duplicates on a subset of 218 accessions available from replicate plots were also scanned. The NIR reflectance spectra were recorded using the ISIscan software (Version 3.10.05933) and converted to estimates of total phenol, proanthocyanidin, and 3-deoxyanthocyanidin concentrations. The spectrometer, software, and calibration curves used in this study were recently described.[35] Samples with unusual reflectance were visually inspected and near-infrared spectroscopy (NIRS), was repeated. Seventeen samples were removed from further analysis either because they contained mixed grain (mixed size, shape, or color) or because their readings were outside the range of the available NIRS calibration curve. Total phenol, proanthocyanidin, and 3-deoxyanthocyanidin data are expressed as mg gallic acid equivalent (GAE)/g, mg catechin equivalents (CE)/g, and absorbance (abs)/mL/g, respectively. These were the units used in creating the calibration curves, which measured total phenols with the Folin-Ciocalteu method, 3-deoxyanthocyanidins with the colorimetric method of Fuleki and Francis, and proanthocyanidins with the modified vanillin/HCl assay.[35] For the purposes of this study, we use a cutoff of greater

than 10.00 mg CE/g to define proanthocyanidin-containing varieties and greater than 50.00 abs/mL/g to define 3-deoxyanthocyanidin-containing varieties.

Visual appearance of grain was classified independently by two people by visually scoring three seeds per accession as white, yellow, red or brown. Testa presence was identified with three seeds per accession by cutting a thin layer off the pericarp and examining under a dissecting microscope. The total grain weight of 100 seeds per accession was recorded.

2.3.3 Genomic Analysis

Genotypes were available for the 324 accessions that were part of the SAP.[27] Genotyping-by-sequencing (GBS) was performed for the 73 additional accessions by the Institute for Genomic Diversity using the methods by Elshire et al.[36] Briefly, we provided seeds of the 73 additional accessions (the same seeds obtained from GRIN that we used to grow our panel) to the Institute for Genomic Diversity, where the following work was performed: Seedlings were grown to obtain tissue, DNA was isolated using the Qiagen DNeasy Plant kit, genomic DNA was digested individually using ApeKI, 96X multiplexed GBS libraries were constructed, and DNA sequencing was performed on the Illumina Genome Analyzer *IIx*. To extract SNP genotypes from sequence data, the GBS pipeline 3.0 in the TASSEL software package (Glaubitz, 2014) was used, with mapping to the BTx623 sorghum reference genome.[37] Missing genotype calls were imputed using the FastImputationBitFixedWindow plugin in TASSEL 4.0.[38]

GWAS was carried out on 404,628 SNP markers, using the statistical genetics package Genome Association and Prediction Integrated Tool (GAPIT).[39] with both a

general linear model (GLM) and a mixed linear model (MLM) with kinship. In a previous study we found that an MLM[40] with kinship (K), which controls for relatedness among the accessions in the panel, performs well to identify causative loci for sorghum polyphenols.[41] Bonferroni correction (Family-wise *P*-value of 0.01, $P < 10^{-6}$) was used to identify significant associations. Pseudo-heritability (proportion of phenotypic variation explained by genotype) was estimated from the kinship (K) model in GAPIT.[42] as the R-squared of a model with no SNP affects. A previously developed a priori candidate gene list was used and 35 additional candidate genes were added.[41]

2.4 RESULTS

2.4.1 Quantitative Variation in Grain Polyphenols

We first sought to determine the reliability of the NIRS estimates across the diverse material in the panel. Phenotypic variation for grain polyphenol concentrations was determined using a diverse association panel with 381 accessions (Figure 2.2). The standard deviation between the duplicates was similar across all concentrations of polyphenols ($r^2 = 0.06$, $P = 0.0001$) and proanthocyanidins ($r^2 = 0.01$, $P = 0.12$), with an average difference of 47% and 4%, respectively. However, the standard deviation between the 3-deoxyanthocyanidin duplicates becomes much larger for samples with higher 3-deoxyanthocyanidin concentrations ($r^2 = 0.32$, $P = 10^{-17}$), with an average difference of 72% (Figure 2.2C). To determine if the NIRS measurements of proanthocyanidin concentration were concordant with the known distribution of testa and *tan1-a* nonfunctional allele[22], we plotted proanthocyanidin concentration of accessions with or without a pigmented testa (Figure 2.3A), and accessions with the wild-type

Tannin1 allele or the *tan1-a* allele (Figure 2.3B). As expected, the absence of a testa and presence of *tan1-a* were primarily found in accessions containing less than 10 mg CE/g of proanthocyanidins. The mean proanthocyanidin concentrations in accessions with a pigmented testa were significantly higher than in accessions without a pigmented testa (18.17 versus 1.45 mg CE/g; $P = 10^{-17}$), and the mean proanthocyanidin concentrations in accessions with the wild-type *Tannin1* were significantly higher than in accessions with *tan1-a* (12.28 versus 0.86 mg CE/g; $P = 10^{-11}$).

Next we investigated the range of total phenol, proanthocyanidin, and 3-deoxyanthocyanidin concentrations and their covariation with each other and grain weight (Figure 2.4). Overall, proanthocyanidins were detected in 55% of the samples, while only 13% contained 3-deoxyanthocyanidins, and only 6% contained both polyphenols. The mean total polyphenol concentration was 7.00 mg (GAE)/g, the mean proanthocyanidin concentration was 7.73 mg CE/g, and the mean 3-deoxyanthocyanidin concentration was 27.40 abs/mL/g (Table 2.2 and Figure 2.4). Pearson's correlations were calculated between total phenols, proanthocyanidins, and 3-deoxyanthocyanidins. There was no significant correlation between proanthocyanidins and 3-deoxyanthocyanidins (0.02, $P = 0.7$), consistent with independent genetic control. In contrast, there was a strong positive correlation between total phenols and proanthocyanidins (0.95, $P < 10^{-15}$), and a weak positive correlation between total phenols and 3-deoxyanthocyanidins (0.12, $P = 0.02$). Variance in proanthocyanidins accounted for 90% of all the variance in total phenols (Figure 2.4). Since the seed coat (pericarp and testa) contains most of the polyphenols in the grain, and the ratio of seed coat (surface area) to endosperm is generally greater in smaller grains, we wondered if differences in grain size might be

underlying variation in polyphenol concentrations. In other words, are high grain polyphenol concentrations limited to small-grain varieties, which have a high proportion of seed coat to endosperm? No significant correlation was found between grain weight and either proanthocyanidins (-0.02, $P = 0.7$) or 3-deoxyanthocyanidins (-0.02, $P = 0.7$), and a small negative correlation was found between grain weight and total polyphenols (-0.10, $P = 0.04$). Pseudo-heritability was 81.7% for proanthocyanidins and 66.5% for 3-deoxyanthocyanidins.

2.4.2 Population Structuring of Polyphenol Concentrations

To determine the distribution of polyphenol traits with respect to global genetic diversity, we conducted a principal component analysis and highlighted the variation in polyphenol concentration (Figure 2.5A and Figure 2.5B), as well as morphological races (Figure 2.5C). At least some high proanthocyanidin accessions were found in most subpopulations, whereas high 3-deoxyanthocyanidin accessions were more restricted (Table 2.3). Bicolor (21.18 mg CE/g) and guinea-caudatum (17.89 mg CE/g) had the highest mean concentration of proanthocyanidins. Caudatum had moderate concentrations (13.20 mg CE/g) and the other botanical races and intermediate groups showed an average less than 10.00 mg CE/g. The highest mean concentrations of 3-deoxyanthocyanidins were found in bicolor-durra (36.95 abs/mL/g) and guinea (35.63 abs/mL/g) accessions (Table 2.3). We also determined the mean concentrations by country to better understand the geographic patterns for sorghum polyphenols (Table 2.4). Accessions from Uganda (19.03 mg CE/g) had the highest mean proanthocyanidin concentrations, accessions from South Africa (12.23 mg CE/g) and Sudan (10.33 mg CE/g) had moderate concentrations, while accessions from the other countries showed an

average less than 10.00 mg CE/g. The highest mean concentrations of 3-deoxyanthocyanidins were found in accessions from Nigeria (36.39 abs/mL/g) and Ethiopia 32.87 abs/mL/g).

2.4.3 Genome-Wide Association Studies

To investigate the genetic basis of natural variation in sorghum grain polyphenols, we conducted GWAS using 404,628 SNP markers. We were able to obtain genotype data for 373 out of the 381 phenotyped accessions. As a data quality check, we first collapsed the quantitative proanthocyanidin data to qualitative (presence or absence) data, and were able to repeat findings from previous GWAS and linkage studies (Figure 2.6 and Figure 2.7; Appendix B.1-B.2). Next, to identify novel alleles associated with quantitative variation of proanthocyanidins, we conducted a GWAS on the 373 accessions (Figure 2.8; Appendix B.3). A GLM identified 3,272 significant SNPs (Figure 2.8A), while the MLM identified 24 significant SNPs after accounting for population structure (Figure 2.8B). The genomic locations of the association peaks were generally similar between methods. A peak on chromosome 4 at ~61 Mb co-localized with *Tannin1* (Sb04g031730), as well as three a priori candidate genes in the region: a putative *Zm1* homolog (Sb04g031110), a putative *TTG1* homolog (Sb04g030840), and a putative *TT16* homolog (Sb04g031750) (Figure 2.8C). The GLM identified a peak at 58.6 Mb on chromosome 7 (S7_58603858; $P < 10^{-15}$), which was not present in the MLM.

In order to reduce the effects of known *Tannin1* nonfunctional alleles and identify additional quantitative loci, samples with the *tan1-a* and *tan1-b* alleles were removed and a GWAS was conducted on the remaining samples (Figure 2.9 and Appendix B.4). The

GLM identified 2,641 significant SNPs (Figure 2.9A). The association peak on chromosome 7 was again identified in the GLM and not in the MLM (Figure 5B). Additionally, there was a peak on chromosome 2 around 8 Mb (S2_8258226; $P < 10^{-11}$) identified in the GLM, near a putative *TT8* homolog (Sb02g006390). Both the GLM and the MLM identified a peak on chromosome 4, again around 61 Mb, and another peak on chromosome 4 between 53 Mb and 55 Mb, close to an F3'H *Pr1* coortholog.

To further map loci controlling quantitative proanthocyanidin variation, we ran a GWAS only on samples that contained proanthocyanidins (greater than 10.00 mg CE/g) and/or had a visible pigmented testa (Figure 2.10 and Appendix B.5). With this subset, there were 676 significant SNPs identified in the GLM, but association peaks were more diffuse (Figure 2.10A). The most significant SNP was on chromosome 6 (S6_56992521, $P < 3 \times 10^{-10}$) near a *TT16* a priori candidate (Sb06g028420). The MLM identified two significant SNPs, with a peak on chromosome 4, again around 61 Mb, and another peak on chromosome 4 between 53 Mb and 55 Mb (Figure 2.10B). Both the GLM and the MLM identified significant SNPs around 61.1 Mb on chromosome 1, which is near *yellow seed1*.

Next, a GWAS was conducted to identify genetic controls of 3-deoxyanthocyanidin variation among the 373 accessions (Figure 2.11 and Appendix B.6). The GLM identified 233 significant SNPs, with distinct association peaks on chromosomes 3 and 4 (Figure 2.11A). The peak on chromosome 3 was between 71-72 Mb and co-localized with a gene (Sb03g045170) homologous to both *TT18* (ANS) and *TT6* (F3H). The peak on chromosome 4 was between 53 Mb and 55 Mb, close to *TT1* and *TT2* homologs, and an F3'H *Pr1* coortholog. While there was not a distinct peak on

chromosome 1, the strongest association signal in the GWAS was found in a diffuse peak on chromosome 1 around 55 Mb ($P < 10^{-9}$). The closest a priori candidates were putative *TTG2* (Sb01g032120) and *TT2* (Sb01g032770) homologs. There were no distinct peaks or significant associations identified in the MLM (Figure 2.11B).

2.4.4 Grain Color

Since grain color is commonly used as a visual marker for sorghum polyphenol content, we used our data set to better understand both the correlation between visually scored grain color and polyphenol concentration, and the potential shared genetic basis for these traits. Based on visual assessment of grain appearance, we designated 142 white, 35 yellow, 48 red, and 152 brown grain accessions. An analysis of variance (ANOVA) showed significant variation among the grain color groups, so we conducted a post hoc Tukey test. Grain classified as red contained significantly more 3-deoxyanthocyanidins than brown ($P < 10^{-5}$) or white grain ($P < 10^{-5}$) accessions, but no significant difference was found between red and yellow accessions (Figure 2.12A and Table 2.5). Brown grain accessions contained significantly more proanthocyanidins than accessions with red ($P = 0.0001$), white ($P = 0.001$), or yellow ($P = 0.001$) grain (Figure 2.12B and Table 2.5). This was expected as most of the sorghums with testa layers were classified as brown (57%). We also compared proanthocyanidin concentrations between grain color in proanthocyanidin-containing (greater than 10.00 mg CE/g or presence of pigmented testa) accessions. Brown grain color classes contained significantly more proanthocyanidins than non-brown (brown $n = 120$, non-brown $n = 85$, $P < 10^{-13}$). However, when brown grain color classes were compared to each color class individually, they only contained significantly more proanthocyanidins than white color

classes ($P < 10^{-4}$). Red and yellow grain color classes also contained significantly more proanthocyanidins than white in the proanthocyanidin-containing accessions ($P = 0.001$ and $P = 0.02$).

To identify genes associated with brown grain, we conducted a presence/absence (brown versus non-brown) GWAS on all 373 of the accessions (Figure 2.13A-B and Appendix B.7) and another presence/absence (brown versus non-brown) GWAS on the 203 proanthocyanidin-containing accessions (Figures 2.13C-D; Appendix B.8). A distinct association peak on chromosome 8 at 52.9 Mb was observed in both GWAS. The nearest a priori candidate was a putative *TT12* homolog within 400 Kb (Sb08g021640). The GWAS conducted on all 373 accessions identified a peak on chromosome 3 at 63.6 Mb, within 100 kb of another putative *TT12* homolog (Sb03g035610), and also a peak on chromosome 6 (S6_56992521, $P < 3 \times 10^{-10}$) near a *TT16* a priori candidate (Sb06g028420) (Figures S2.4A and S2.4B). The GWAS conducted on the proanthocyanidin-containing accessions identified a peak on chromosome 2 around 69.6 Mb, very near another *TT12* homolog (Sb02g034720) (Figure 2.13C-D). This peak was also identified in the GWAS conducted on all 373 accessions, but was more diffuse. There were no peaks on chromosome 4 around *Tannin1* or on chromosome 2 around the *Z* locus.

To identify genes associated with red grain, we conducted a presence/absence (red versus non-red) GWAS on all of the samples (Figure 2.14 and Appendix B.9). Two association peaks on chromosome 4 were identified by both the GLM and MLM, in the same region as the peak in the 3-deoxyanthocyanidin GWAS. The first peak, at 54.5 Mb, colocalized with a priori candidate Sb04g024710, the F3'H *Pr1* coortholog that was also

in one of the 3-deoxyanthocyanidin GWAS peaks. The second peak, at 55.9 Mb, was very close to a priori candidate Sb04g026480, a putative MYB homolog. There was also a peak around 72 Mb on chromosome 3, in the same region as the peak in the 3-deoxyanthocyanidin GWAS, near a priori candidate Sb03g044980, a putative *TT19* homolog. A peak was identified on chromosome 6 between 7-8 Mb, which was not near any a priori genes, but was near a putative vacuolar sorting protein gene (Sb06g003780). There were no peaks on chromosome 3 around the R locus.

2.5 DISCUSSION

2.5.1 Genetic Controls of Sorghum Polyphenols

The genetic controls of the flavonoid pathway (Figure 2.15) have been well studied in many economically important food plants, including grape (*Vitis vinifera*), barley (*Hordeum vulgare*), maize (*Zea mays*), rice (*Oryza sativa*) and wheat (*Triticum* spp.).[43] Much of our understanding of flavonoid genetics, including biosynthetic enzymes, transporters, and regulatory proteins, come from analysis of Transparent Testa (TT) mutants in Arabidopsis.[44] Transcriptional regulation occurs through a ternary complex made up of *TT2*, *TT8*, and *TTG1*, which encode for MYB, bHLH and WD40 proteins (MBW complex), respectively.[44] This ternary complex is highly conserved among plant species.[45] In the sorghum proanthocyanidin pathway, the WD40 (*Tannin1*) component of the MBW complex has been identified, as well as a likely candidate for the bHLH; several studies have found a significant linkage and association on sorghum chromosome 2 around 8 Mb, near a putative bHLH transcription factor orthologous to Arabidopsis *TT8*.[22,24,41,46,47] The MYB transcription factor that would complete the

ternary complex has not been found in sorghum. The *Zm1* homolog on chromosome 4 at 61.1 Mb (Sb04g031110, 66.8% similarity), which was mapped in all of our proanthocyanidin GWAS, is a possible candidate for the missing MYB. The maize *Zm1* gene is a MYB transcription factor, homologous to classical maize grain pigmentation gene *C1* that can induce transcription of DFR, an essential structural enzyme in the flavonoid pathway.[48] Another possible explanation for the significant SNPs at this location is an indirect association with an undescribed allele at *Tannin-1*.

About two-thirds of the SAP accessions we studied were "converted" tropical accessions, meaning that alleles for reduced height and early flowering have been introgressed so they can be grown in temperate regions.[49] Surprisingly, the proanthocyanidin GWAS association peak on chromosome 7 (~58.6 Kb) precisely colocalizes with *dw3* (Sb07g023730), a dwarfing loci used in the conversion, in conjunction with *dw1*, *dw2*, and *dw4*.[27] Smaller peaks on chromosomes 6 (~39 Kb) and 9 (~57 Kb) were near the *dw2* and *dw1* loci. The association peaks on chromosomes 6, 7, and 9 may be artifacts arising from a lower mean proanthocyanidin concentration in the converted lines (4.4 mg CE/g) which all shared the same *dw* alleles, compared to the unconverted lines (11.0 mg CE/g). Accordingly, when we conducted a proanthocyanidin GWAS using only converted accessions to control for this spurious phenotypic covariation between proanthocyanidin and height, the peaks near *dw1*, *dw2*, and *dw3* disappeared, while the *Tannin1* peak remained (Figure 2.16).

As a phytoalexin.[15,16], the effect of the environment may make it more difficult to map the genetic basis of 3-deoxyanthocyanidins than the genetic basis of proanthocyanidins. Although the GLM was able to identify significant SNP associations

for 3-deoxyanthocyanidins, there were few peaks, and the MLM did not identify any significant associations. Detection of alleles contributing to variance of 3-deoxyanthocyanidins may require a larger sample size, additional replication, a biparental mapping population, or controlled fungal inoculations to induce biosynthesis of polyphenol compounds[23]. However, our results did provide a promising candidate for follow-up. A *Pr1* ortholog (Sb04g024750) lies within a distinct peak on chromosome 4, about 400 kb from the top SNP identified in the 3-deoxyanthocyanidin GWAS (S4_54975391; $P < 10^{-8}$), and 100 kb from the top SNP in the red grain GWAS (S4_54555458, $P < 10^{-13}$). *Pr1* is a maize F3'H enzyme, homologous to *TT7* in Arabidopsis. The F3'H enzyme is essential for production of 3-deoxyanthocyanidins, as well as the red phlobaphene pigments visible in maize[18], and has been implicated in production of these compounds in sorghum.[50] Overall, we observe a 1.6-fold difference in 3-deoxyanthocyanidin concentrations between accessions carrying the high concentration alleles and low concentration alleles for the top red grain-associated SNP ($P = 0.001$). F3'H is necessary for proanthocyanidin production as well, and, indeed, significant associations with SNPs in the ~54 Mb region on chromosome 4 were also identified in the GWAS with *tan1-a* and *tan1-b* samples removed, as well as the GWAS with only proanthocyanidin-containing samples.

Our study identified many peaks and SNPs significantly associated with proanthocyanidins and 3-deoxyanthocyanidins, hence there appear to be many small effect genes controlling natural variation of these traits. Consequently, a larger association panel, or a targeted biparental mapping population may be more effective in precisely identifying causal alleles. Moving forward, sequence analysis and expression

analysis of the candidate genes are needed to identify causal polymorphisms, and lay the groundwork for the use of polyphenol genetic variation in crop improvement.

2.5.2 Crop Improvement for Sorghum Polyphenols

Efforts to characterize polyphenols, with the goal of producing high polyphenol specialty varieties, have been undertaken in several grain crops, including purple wheat,[51] black rice,[52] multi-colored maize,[53] multi-colored barley,[54] and black sorghum.[55] Our diverse association panel contained a wide range of proanthocyanidin and 3-deoxyanthocyanidin concentrations, and this genetic variation may be useful in breeding programs to produce high polyphenol specialty varieties. Bicolor sorghums had the highest mean proanthocyanidin concentrations, but their grain weight is significantly less (20% less) than non-bicolor sorghums ($P < 10^{-9}$). Combined with low yield potential, the small grain size makes it difficult to use bicolor race sorghums in a grain sorghum breeding program, but may still be of interest to breeders wanting to produce specialty varieties. In addition to bicolor sorghums, caudatum and guinea-caudatum sorghums also had high mean proanthocyanidin concentrations, and are promising sources for increasing proanthocyanidin concentrations in sorghum. In particular, among the caudatum and guinea-caudatum sorghums, caudatum sorghums from tropical climates such as Uganda had the highest mean proanthocyanidin concentrations, so may be good material for breeding high polyphenol sorghums. While bicolor-durra and guinea sorghums had the highest mean 3-deoxyanthocyanidin concentrations, the difference among all the races was not significant, so it may be more important to simply identify unique genotypes across the sorghum collection. Chemical analysis is underway on the samples that were

outside of the NIRS calibration curves, and true biological outliers may open up new avenues for future work on sorghum varieties with extreme polyphenol concentrations.

Increasing 3-deoxyanthocyanidin production may be challenging, since, as phytoalexins, they are not constitutively expressed, but rather synthesized by plants under pathogen attack.[15,16] We note in our comparison of 3-deoxyanthocyanidin concentrations from duplicate samples that the difference between duplicates becomes larger for accessions with higher 3-deoxyanthocyanidin concentrations. One possibility is that there is greater technical variation in the 3-deoxyanthocyanidin NIRS estimates, but Dykes et al.[35] demonstrated the same correlation coefficient between the NIRS-predicted values and the values in the validation set for proanthocyanidins (r = 0.81) and 3-deoxyanthocyanidins (r = 0.82). Therefore, we would not expect to see differences in accuracy of the NIRS predictions for proanthocyanidins and 3-deoxyanthocyanidins in our study. As this was a field study, another possibility is that uncontrolled environmental variation may have contributed to the difference between the duplicate samples. Accessions with the genetic capability to produce grain 3-deoxyanthocyanidins may be producing low or high 3-deoxyanthocyanidin concentrations depending on the exposure to inducing agents on a given panicle. Controlled inoculation studies are needed to further explore this possibility[23].

The spreader gene is a promising target for increasing grain proanthocyanidin concentrations, and a previous report using a small number of varieties has shown higher proanthocyanidin concentrations in varieties with a functional spreader.[56] Given that three peak SNP associations in the brown grain GWAS were near putative MATE transporter *TT12* homologs, we propose that the spreader gene may be a *MATE*

transporter. A biparental mapping population segregating the spreader gene would be needed to confirm this hypothesis. To get a sense of the effect these loci may have on proanthocyanidin concentrations, we compared concentrations of each allele in proanthocyanidin-containing accessions. There was a 1.8-fold (S3_63633634, $P = 0.04$), a 1.5-fold (S2_69656067, $P = 0.0003$), and a 1.7-fold (S8_52906014, $P = 0.0002$) difference between accessions carrying the high concentration alleles and low concentration alleles. When the three polymorphisms are considered together, accessions with all three high-alleles (S2_69656067 = "A", S3_63633634 = "A", S8_52906014 = "G") have 1.7 to 2.7-fold higher proanthocyanidin concentrations ($P = 10^{-8}$), consistent with an additive effect more than doubling the concentration of proanthocyanidins in sorghum grain.

Appearance of grain color is predominantly due to polyphenols, but can also be influenced by endosperm color and grain weathering. Taken in total, the color classes used for our analysis represent general groups and are not definitive descriptors of any specific trait. For example, it is possible to have a sorghum classified as brown that does not have a testa layer, as well to have a sorghum classified as white that has a testa layer (see Figure 2.1). However, our results support the use of visual categorization of grain color as a simple assessment of polyphenol concentrations in crop improvement programs; brown grain has significantly higher proanthocyanidin concentrations than non-brown, red grain has significantly higher 3-deoxyanthocyanidin concentrations than non-red, and white grain has significantly lower concentrations of these polyphenols than non-white. Additionally, the genetic architecture of grain color reflects, to an extent, that of the polyphenols with which they are associated. For instance, the red grain GWAS and

the 3-deoxyanthocyanidin GWAS produced similar association peaks on chromosomes 4 (~54 Mb), which may map to the sorghum *Pr1* ortholog, and chromosome 3 (~72 Mb), which colocalizes with putative homologs of *ANS*, *F3H*, and *TT19*. The brown grain GWAS and the proanthocyanidin-containing GWAS produced similar association peaks on chromosome 6 (~57 Mb) near a priori candidate *TT16*, a key regulatory protein in the proanthocyanidin branch of the flavonoid pathway. Overall, to increase sorghum proanthocyanidin and 3-deoxyanthocyanidin concentrations quantitatively, there are many associated alleles available, but none of them have large effect. This survey of grain polyphenol variation in sorghum germplasm and catalog of flavonoid pathway-associated loci contributes toward the goal of producing sorghum crops that will contribute to marker-assisted breeding of sorghum crops that will benefit human health.

2.6 TABLES

Table 2.1 Summary of flavonoid pathway genes[a]

Function of reference gene	Reference gene name *A. thaliana*	Reference gene name *Z. mays*	Reference gene name other species	Functional category
Chalcone synthase (CHS)	*TT4*	*C2*		Biosynthesis
Chalcone isomerase (CHI)	*TT5*	*CHI1*		Biosynthesis
Flavone 3-hydroxylase (F3H)	*TT6*	*F3H*		Biosynthesis
Flavone 3'-hydroxylase (F3'H)	*TT7*	*Pr1*		Biosynthesis
Dihydroflavonol reductase (DFR)	*TT3*	*A1*		Biosynthesis
Anthocyanidin synthase (ANS/LDOX)	*TT18*	*A2*		Biosynthesis
UDP-flavonoid glucosyl transferase (UFGT)	*TT15*	*Bz1*		Biosynthesis
Anthocyanidin reductase (ANR)	*Banyuls (BAN)*			Biosynthesis
Flavonoid oxidase	*TT10*			Biosynthesis
Leucoanthocyanin reductase (LAR)			*VvLAR*[a]	Biosynthesis
MYB transcription factor	*TT2, MYB11/12/111*	*P1, C1, Zm1*	*Yellow seed1 (Y locus)*[b]	Regulation
bHLH transcription factor	*TT8*	*B1*		Regulation
WD40 repeat protein	*TTG1*		*Tannin1 (B2 locus)*[b]	Regulation
WRKY transcription factor	*TTG2*			Regulation
MADS-box transcription factor	*TT16*			Regulation
Zn-finger transcription factor	*TT1*			Regulation
MATE vacuolar transport	*TT12*			Transport
Glutathione-S-transferase	*TT19*	*BZ2*		Transport
H+-ATPase proton pump	*aha10*			Transport
MRP anthocyanin transporter		*ZmMRP3*		Transport

[a]*Vitis vinifera;* [b]*Sorghum bicolor*

Table 2.2 Polyphenol concentrations in 373 sorghum varieties

constituent	mean	range	SD
total phenols (mg GAE/g)	7.00	ND – 37.46	± 5.92
proanthocyanidins (mg CE/g)	7.73	ND – 78.51	± 15.45
3-deoxyanthocyanidins (abs/mL/g)	27.40	ND - 149.21	± 24.05

Table 2.3 Polyphenol concentrations by race

race[a]	*n*	total phenols mean (mg GAE/g)	total phenols range (mg GAE/g)	PA mean (mg CE/g)	PA range (mg CE/g)	3-DA mean (abs/mL/g)	3-DA range (abs/mL/g)
bicolor	15	13.68 ± 6.69	0.74 - 24.49	21.18 ± 17.68	ND - 50.16	26.91 ± 33.65	ND - 102.96
bicolor-durra	19	6.59 ± 4.28	ND - 13.38	3.89 ± 12.06	ND - 23.35	36.95 ± 28.24	1.30 - 113.42
caudatum	86	9.08 ± 5.86	ND - 27.32	13.20 ± 14.15	ND - 52.83	28.22 ± 21.06	ND - 110.73
caudatum-kafir	20	6.27 ± 5.41	ND - 15.68	7.00 ± 15.13	ND - 31.98	26.65 ± 16.87	6.70 - 58.25
durra	15	2.17 ± 3.61	ND - 11.68	ND	ND - 17.64	22.17 ± 21.33	ND - 71.10
guinea	11	1.95 ± 5.25	ND - 15.44	ND	ND - 33.45	35.63 ± 36.88	0.93 - 135.34
guinea-caudatum	15	10.01 ± 3.13	2.54 - 15.87	17.89 ± 9.76	ND - 34.92	19.72 ± 15.69	0.40 - 60.10
kafir	29	6.02 ± 4.05	1.32 - 14.71	6.50 ± 10.20	ND - 28.72	17.59 ± 20.65	ND - 94.49

[a]If a race contained a small sample size (less than 10 accessions), it was not included in this analysis. PA, proanthocyanidins; 3-DA, 3-deoxyanthocyanidins; ND, not detected (absorbance was less than 0.001)

Table 2.4 Polyphenol concentrations by geographic origin

country[a]	*n*	total phenols mean (mg GAE/g)	total phenols range (mg GAE/g)	PA mean (mg CE/g)	PA range (mg CE/g)	3-DA mean (abs/mL/g)	3-DA range (abs/mL/g)
Uganda	44	10.99 ± 5.17	1.17 - 27.32	19.03 ± 12.02	ND - 52.83	27.37 ± 20.8	1.30 - 110.73
South Africa	31	9.11 ± 5.21	1.11 - 20.63	12.23 ± 12.37	ND - 43.75	13.52 ± 14.1	ND - 38.82
Sudan	31	7.50 ± 3.34	ND - 14.67	10.33 ± 8.93	ND -25.26	27.15 ± 15.1	4.13 - 60.10
Nigeria	21	5.0 ± 6.46	ND - 24.49	1.21 ± 21.36	ND - 50.16	36.39 ± 35.8	ND - 135.34
Ethiopia	29	5.71 ± 5.43	ND - 15.94	1.53 ± 13.13	ND - 23.53	32.87 ± 21.1	ND - 77.59
India	21	3.90 ± 5.09	ND - 16.98	ND	ND - 32.13	28.74 ± 28.7	ND - 113.42
USA	71	5.09 ± 5.25	ND - 29.93	3.6 ± 12.55	ND - 63.80	27.50 ± 24.2	ND - 95.20

[a]If a country contained a small sample size (less than 10 accessions), it was not included in this analysis. PA, proanthocyanidins; 3-DA, 3-deoxyanthocyanidins; ND, not detected (absorbance was less than 0.001).

Table 2.5 Polyphenol concentrations by color

color	*n*	total phenols mean (mg GAE/g)	total phenols range (mg GAE/g)	PA mean (mg CE/g)	PA range (mg CE/g)	3-DA mean (abs/mL/g)	3-DA range (abs/mL/g)
white	142	4.0 ± 3.10	ND – 14.67	2.00 ± 8.84	ND – 25.26	22.74 ± 14.03	ND – 58.41
yellow	35	6.0 3± 6.18	ND – 23.69	4.60 ± 15.98	ND – 42.30	29.30 ± 27.89	ND – 98.90
red	48	6.97 ± 7.30	ND – 27.32	4.48 ± 21.10	ND – 52.83	42.21 ± 30.43	ND - 135.34
brown	152	10.01 ± 6.01	ND – 37.46	14.74 ± 15.63	ND – 78.51	26.46 ± 26.64	ND - 149.21

proanthocyanidins; 3-DA, 3-deoxyanthocyanidins; ND, not detected (absorbance was less than 0.001).

2.7 FIGURES

Figure 2.1 Natural variation in sorghum grain color. Three accessions (with three seeds of each accession) of grain with the appearance of (A) brown (PI597965, PI533927, PI35038), (B) white (PI533755, PI533845, PI534028), (C) yellow (PI659691, PI656011, PI533776), and (D) red (PI576418, PI534047, PI564165) pericarps. The outer coat has been scraped off of some samples, revealing the presence or absence of a pigmented testa.

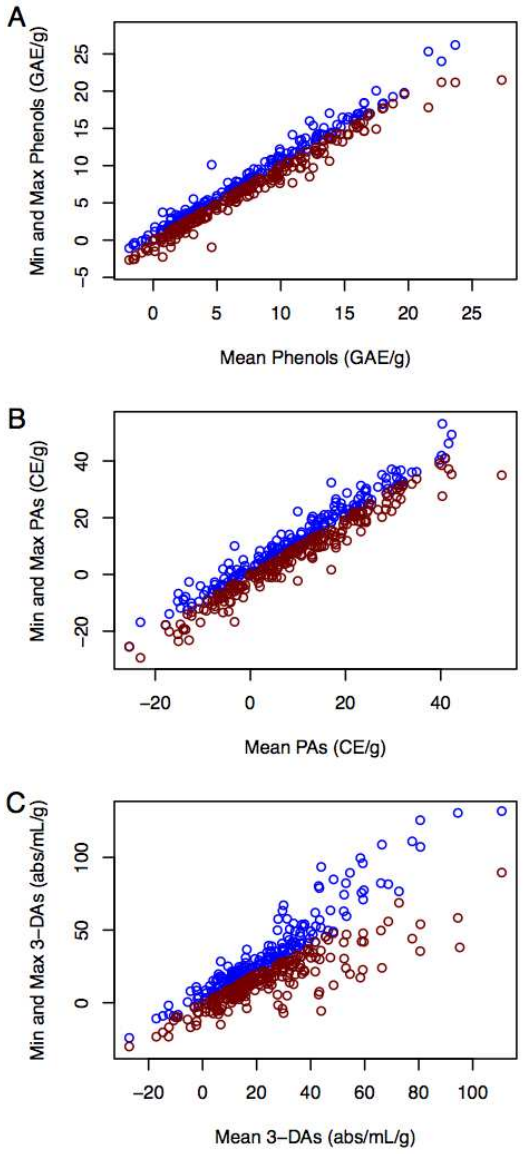


Figure 2.2 Phenotypic variation of grain polyphenol concentrations in 381 sorghum varieties. Samples are ordered on the x-axis according to their mean value for the accession. The observed value for each replicate is given on the y-axis, with the higher value of the duplicates in red and the lower value of the duplicates in blue. (A) total polyphenols, (B) proanthocyanidins (PAs), and (C) 3-deoxyanthocyanidins (3-DAs).

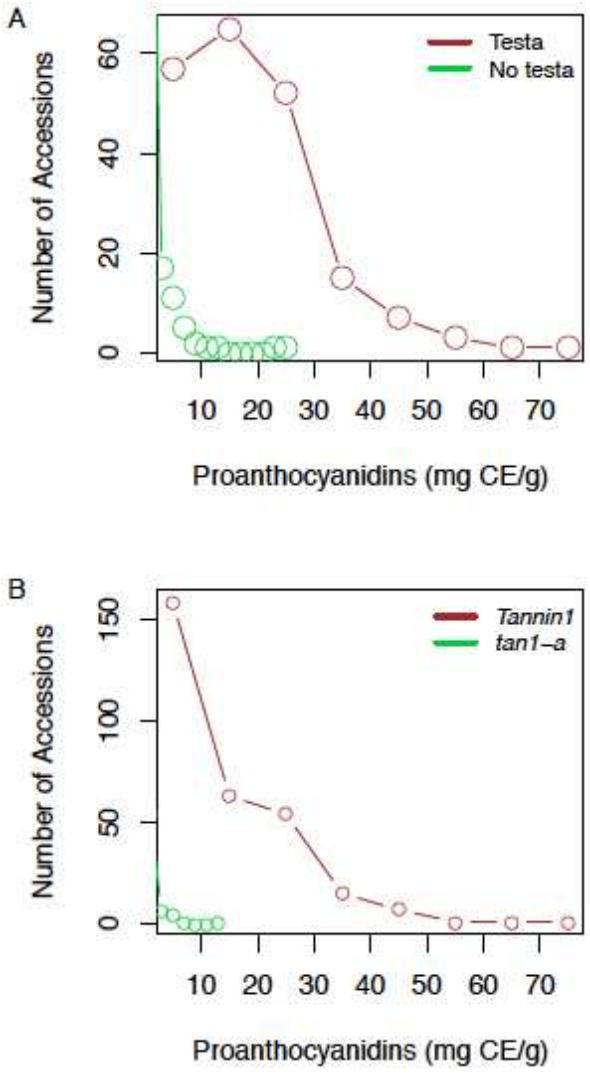


Figure 2.3 Variation of proanthocyanidin concentrations in testa phenotype and *Tannin1* genotype. Comparison of estimates of proanthocyanidin concentration (A) between accessions with and without a pigmented testa, and (B) between accessions containing the wild type *Tannin1* allele or the *tan1-a* null allele.

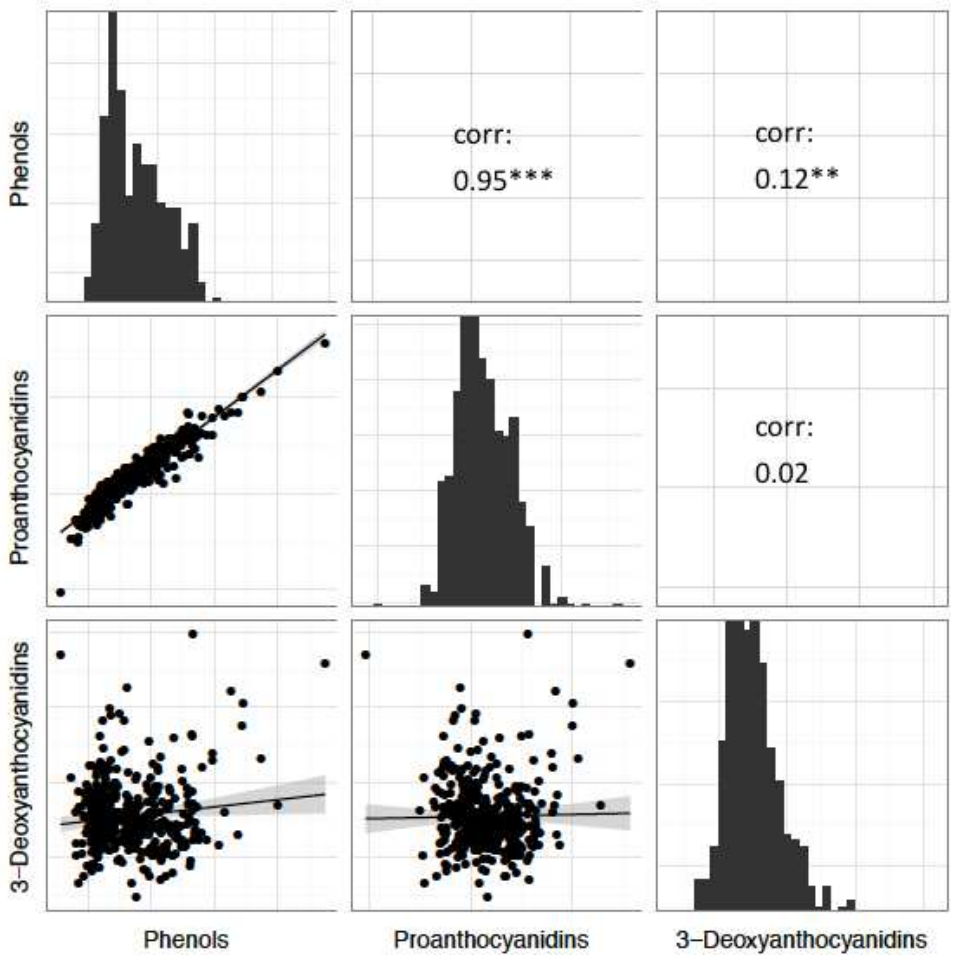


Figure 2.4 Relationship within and between grain polyphenol traits in a global sorghum germplasm collection. The center diagonal presents histograms of the mean concentrations of each trait. The lower corner contains scatter plots with regression lines showing the relationships between the traits. The upper corner shows Pearson's correlations between the traits. Units are mg GAE/g for total phenols, mg CE/g for proanthocyanidins, and abs/mL/g for 3-deoxyanthocyanidins. (*n* =381)

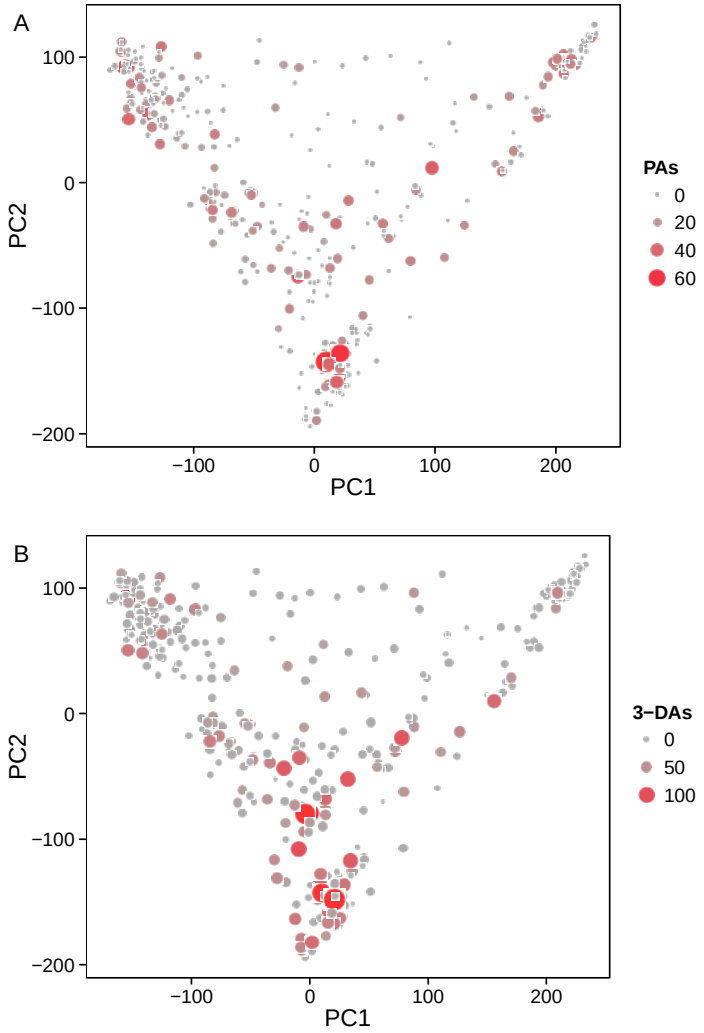


Figure 2.5 Population structure of grain polyphenol traits in a global sorghum germplasm collection. Accessions plotted according to the first two principal components of sorghum population structure based on the SNP data, showing (A) proanthocyanidin concentration (mg CE/g), (B) 3-deoxyanthocyanidin concentrations (abs/mL/g), and (C) morphological race, where the SAP are squares, the 74 additional accessions are circles, and accessions of unknown races are in gray

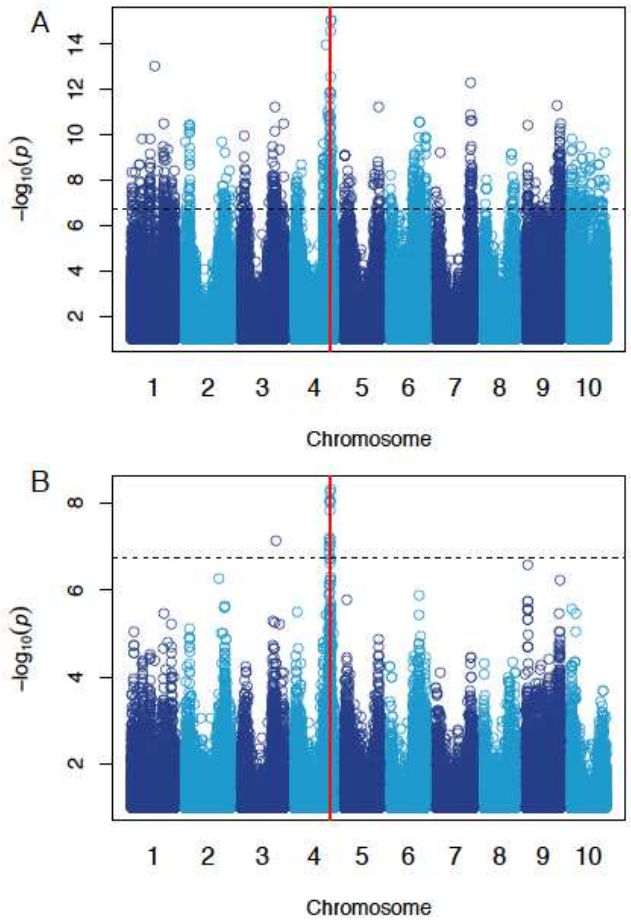


Figure 2.6 GWAS for proanthocyanidin presence/absence in sorghum grain. Manhattan plot of association results from (A) a GLM analysis and (B) an MLM analysis using ~404,628 SNP markers and 373 accessions (146 proanthocyanidin accessions, 227 non-proanthocyanidin accessions). Presence is defined as proanthocyanidins greater than 10.00 mg CE/g and absence is defined as proanthocyanidins less than 10.00 mg CE/g.

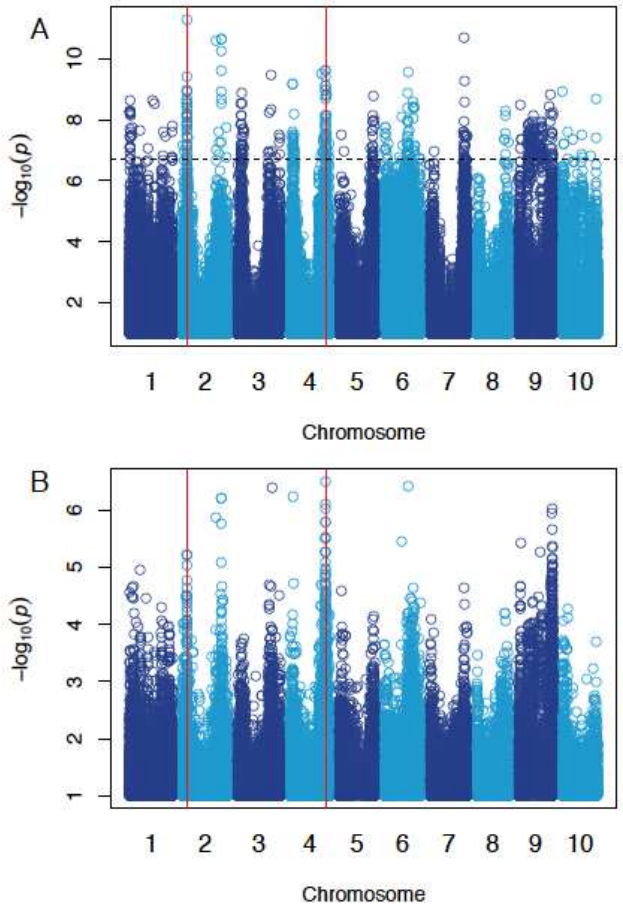


Figure 2.7 GWAS for proanthocyanidin presence/absence in sorghum grain with accessions containing *tan1-a* and *tan1-b* removed. Manhattan plot of association results from (A) a GLM analysis and (B) an MLM analysis using ~404,628 SNP markers and 312 accessions (150 proanthocyanidin accessions, 162 non-proanthocyanidin accessions). Presence is defined as proanthocyanidins greater than 10.00 mg CE/g and absence is defined as proanthocyanidins less than 10.00 mg CE/g.

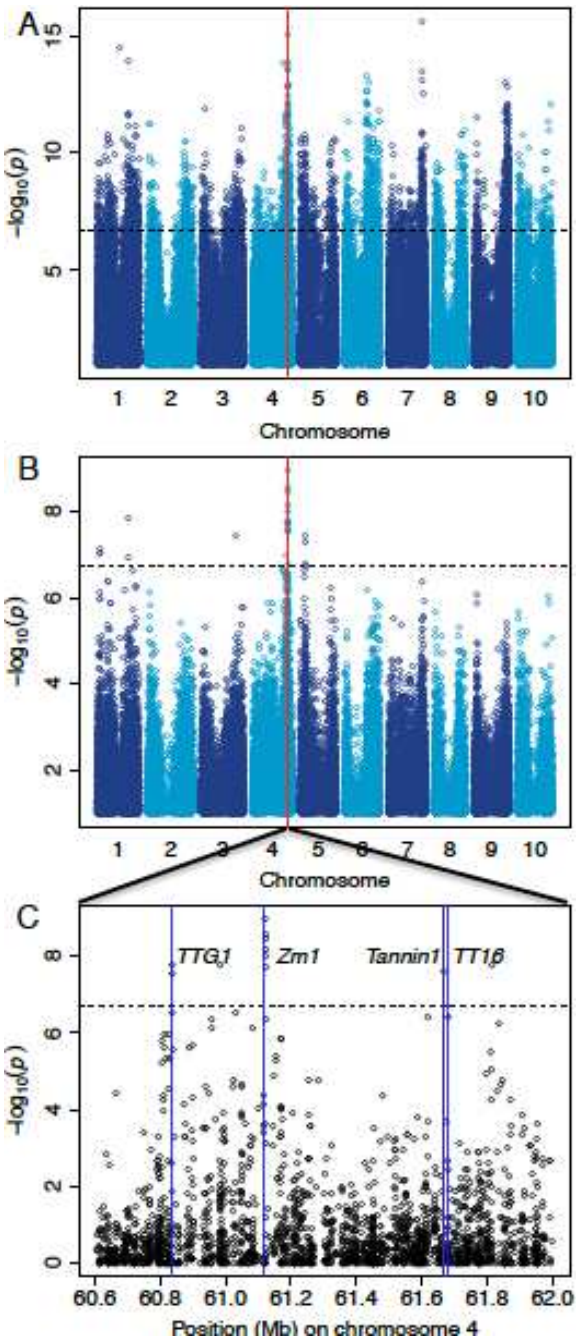


Figure 2.8 GWAS for proanthocyanidin concentration in sorghum grain. Manhattan plot of association results from (A) a GLM analysis, (B) an MLM analysis, and (C) a closeup of the peak on chromosome 4 showing *Tannin1* and other candidate genes in the region, using 404,628 SNP markers and 373 accessions. Axes: the $-\log_{10}$ p-values (y axis) plotted against the position on each chromosome (x axis). Each circle represents a SNP. The dashed horizontal line represents the genome-wide significance threshold as determined by Bonferroni correction. Regions with $-\log_{10}$ *p*-values above the threshold are candidates. The vertical lines indicate the location of *Tannin-1* and *a priori* candidate genes in the *Tannin-1* region (~61 Mb).

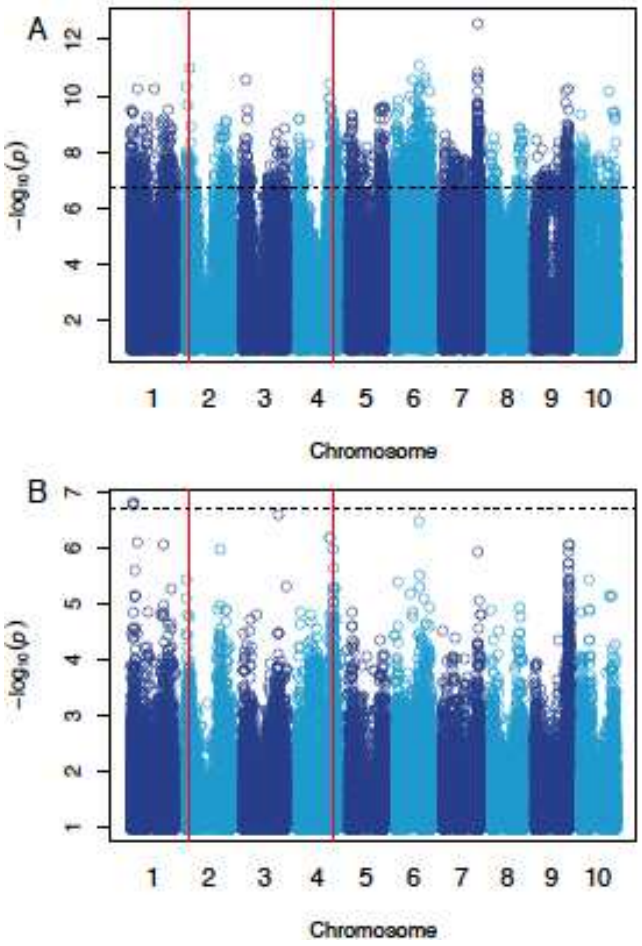


Figure 2.9 GWAS for proanthocyanidin concentration in sorghum grain with accessions containing *tan1-a and tan1-b* nonfunctional alleles removed. Manhattan plot of association results from (A) a GLM analysis, and (B) an MLM analysis, using 404,628 SNP markers and 312 accessions. Axes: the $-\log_{10}$ *p*-values (y axis) plotted against the position on each chromosome (x axis). Each circle represents a SNP. The dashed horizontal line represents the genome-wide significance threshold as determined by Bonferroni correction. Regions with $-\log_{10}$ *p*-values above the threshold are candidates. The red vertical lines highlight the location of candidate genes (TT8 on chrm. 2 and TTG1, Zm1, and TT16 on chrm. 4).

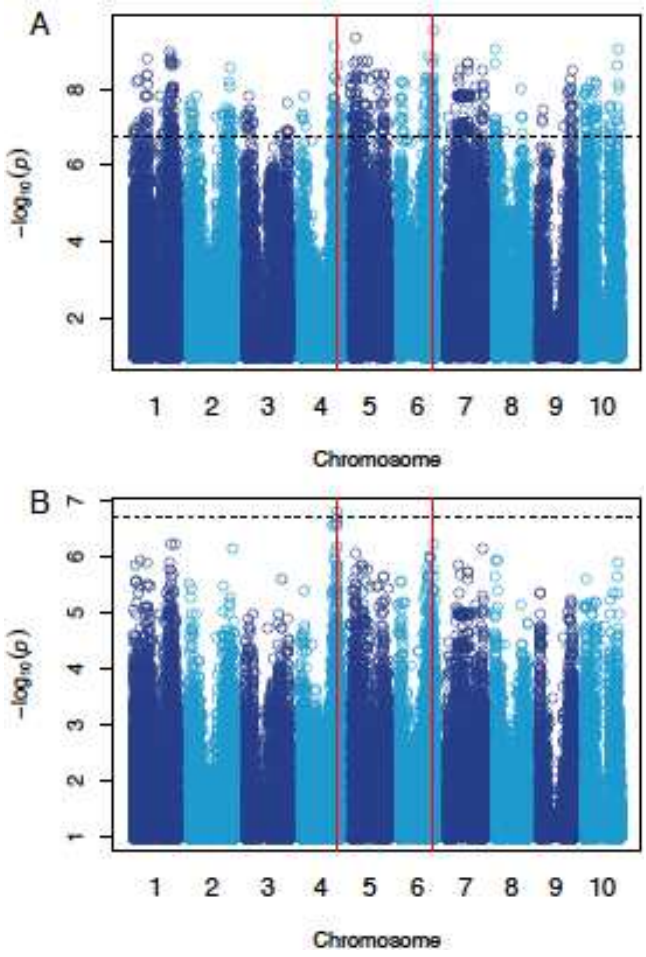


Figure 2.10 GWAS for proanthocyanidin concentration in proanthocyanidin-containing sorghum grain (greater than 10.00 mg CE/g or pigmented testa). Manhattan plot of association results from (A) a GLM analysis, and (B) an MLM analysis, using 404,628 SNP markers and 208 accessions. Axes: the $-\log_{10}$ *p*-values (y axis) plotted against the position on each chromosome (x axis). Each circle represents a SNP. The dashed horizontal line represents the genome-wide significance threshold as determined by Bonferroni correction. Regions with $-\log_{10}$ *p*-values above the threshold are candidates. The red vertical lines highlight the location of candidate genes (TT16, Tannin1 region, Pr1/TT7).

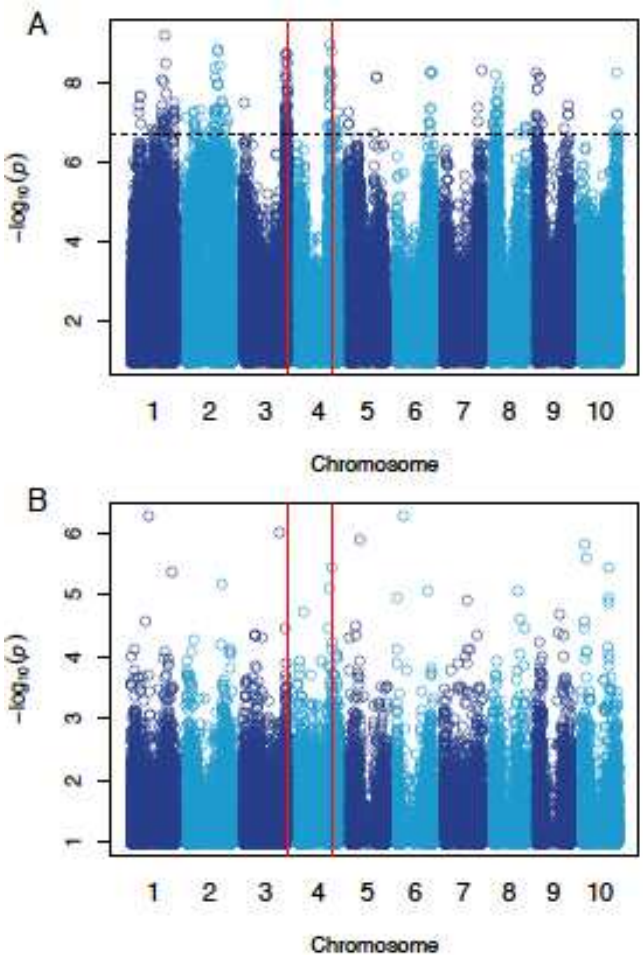


Figure 2.11 GWAS for 3-deoxyanthocyanidin concentration in sorghum grain. Manhattan plot of association results from (A) a GLM analysis, and (B) an MLM analysis, using 404,628 SNP markers and 373 accessions. Axes: the $-\log_{10}$ *p*-values (y axis) plotted against the position on each chromosome (x axis). Each circle represents a SNP. The dashed horizontal line represents the genome-wide significance threshold as determined by Bonferroni correction. Regions with $-\log_{10}$ *p*-values above the threshold are candidates. The red vertical lines highlight the location of candidate genes (TT18/ANS, TT6/F3H, Pr1/TT7).

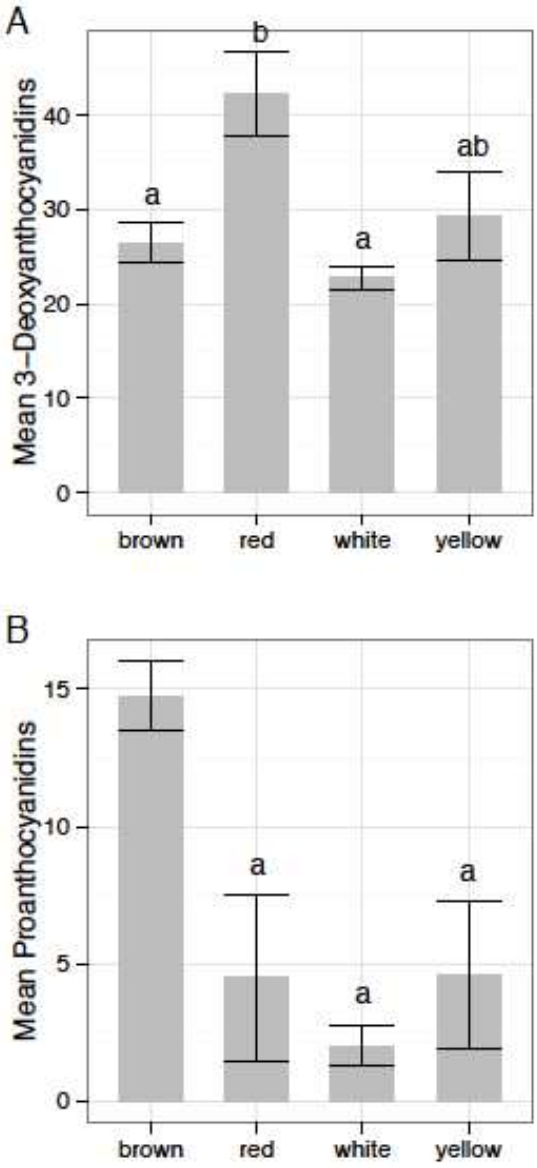


Figure 2.12 Polyphenol differences between grain colors. Mean concentrations of (A) proanthocyanidins and (B) 3-deoxyanthocyanidins in accessions of each grain color. Color categories share the same letter if they are not significantly different from each other, based on a post hoc Tukey HSD test (brown, n = 152; red, n = 48; white, n = 142; yellow,n=35)

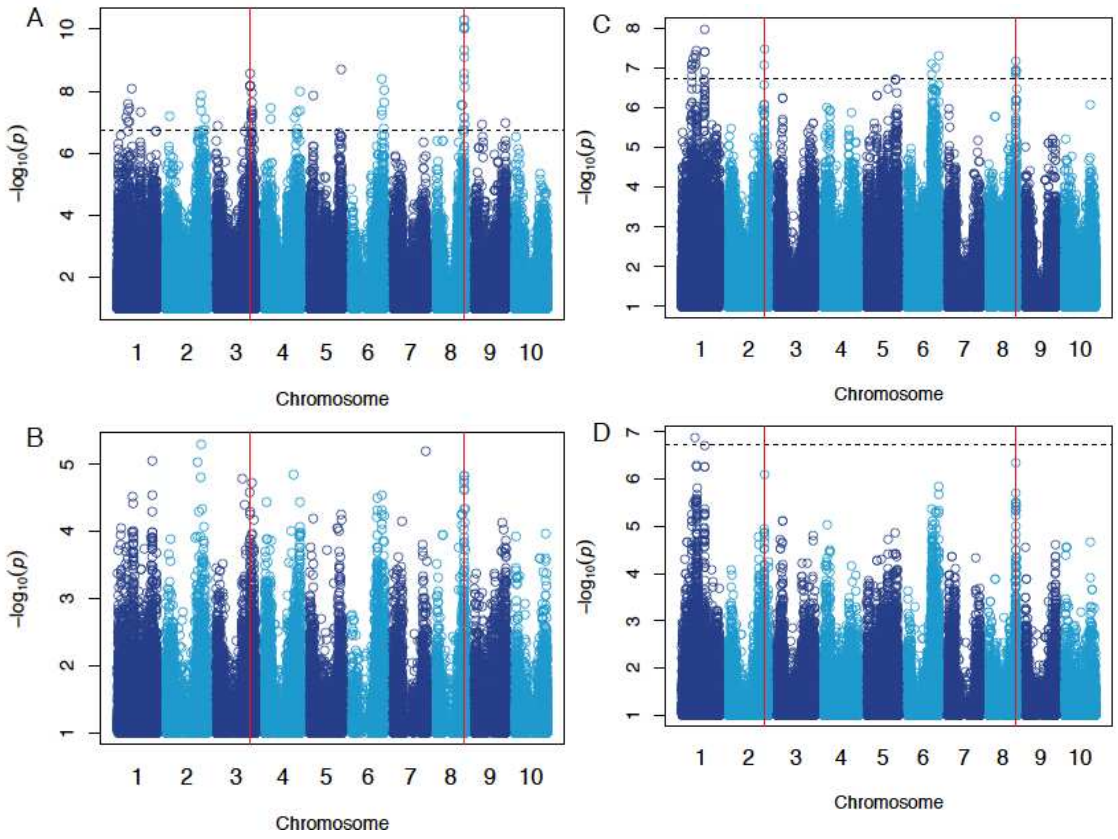


Figure 2.13 GWAS for brown grain sorghum. Manhattan plot of association results from (A) a GLM analysis in all accessions, (B) an MLM analysis in all accessions, (C) a GLM analysis in proanthocyanidin-containing accessions, (D) and an MLM analysis in proanthocyanidin-containing accessions, using ~404,628 SNP markers and 373 (148 brown, 225 not brown) accessions for A and B, and 203 (116 brown, 87 not brown) accessions for C and D. Proanthocyanidin-containing sorghum grain is defined as proanthocyanidins greater than 10.00 mg CE/g or having a pigmented testa.

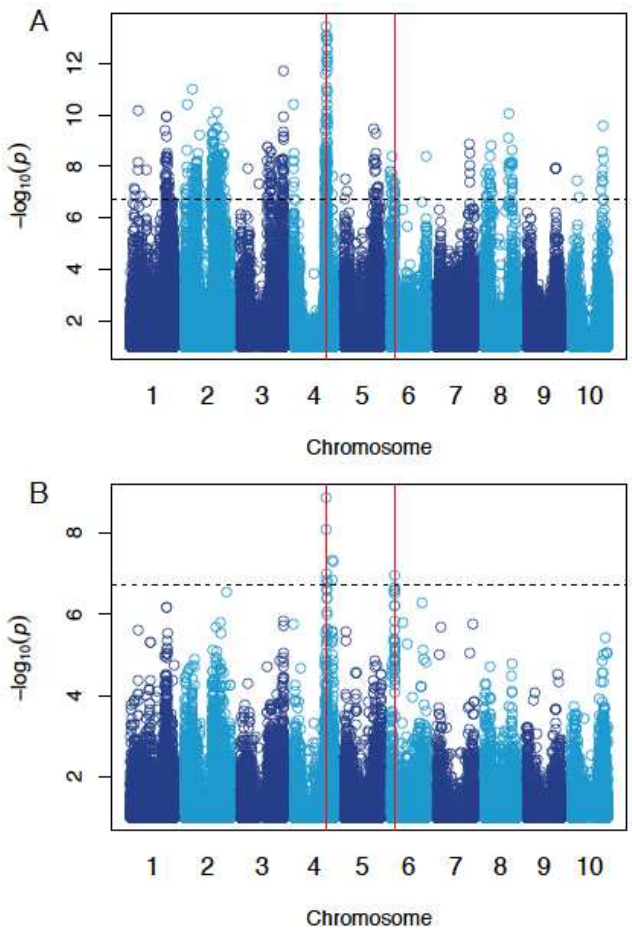


Figure 2.14 GWAS for red grain sorghum. Manhattan plot of association results from (A) a GLM analysis, and (B) an MLM analysis using ~404,628 SNP markers and 373 (48 red, 325 non-red) accessions.

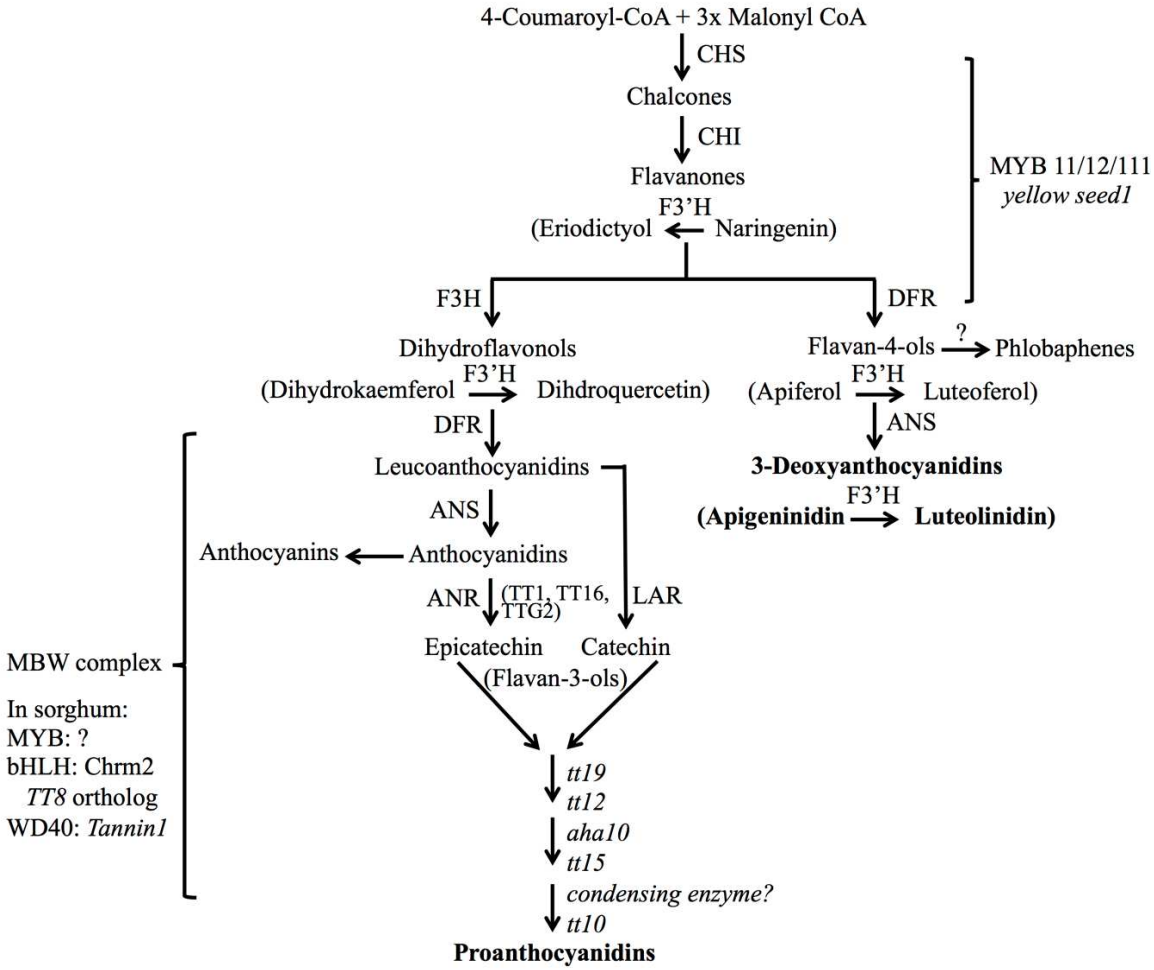


Figure 2.15 Simplified scheme of flavonoid biosynthetic pathway. Enzyme abbreviations are in uppercase letters, while gene abbreviations are in italics. Question marks depict unknown steps. Chalcone synthase (CHS), chalcone-flavanone isomerase (CHI), flavanone 3-hydroxylase (F3H), flavanone 3'-hydroxylase (F3'H), dihydroflavonol-4-reductase (DFR), anthocyanidin synthase (ANS), anthocyanidin reductase (ANR), leucoanthocyanidin reductase (LAR); MYB-bHLH-WD40 (MBW).

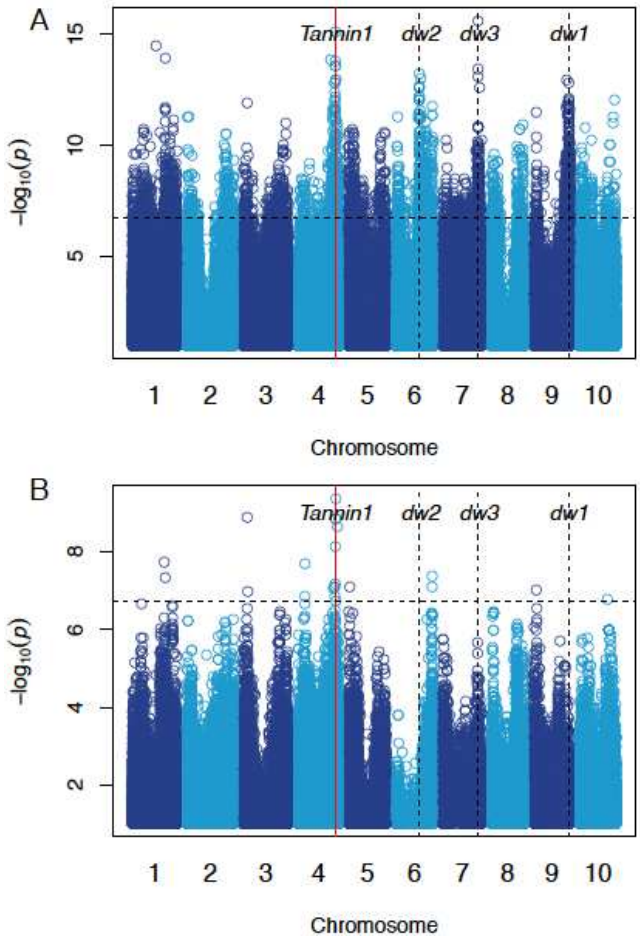


Figure 2.16 GWAS for proanthocyanidins in entire panel versus converted lines. Manhattan plot of association results from (A) a GLM analysis using all 373 accessions, and (B) a GLM analysis using only the 190 converted accessions.

2.8 REFERENCE

1. Tsao, R. Chemistry and biochemistry of dietary polyphenols. *Nutrients* **2,** 1231–1246 (2010).

2. Hichri, I. *et al.* Recent advances in the transcriptional regulation of the flavonoid biosynthetic pathway. *J. Exp. Bot.* **62,** 2465–2483 (2011).

3. Buer, C. S., Imin, N. & Djordjevic, M. A. Flavonoids: new roles for old molecules. *J. Integr. Plant Biol.* **52,** 98–111 (2010).

4. Del Rio, D. *et al.* Dietary (poly)phenolics in human health: structures, bioavailability, and evidence of protective effects against chronic diseases. *Antioxid. Redox Signaling* **18,** 1818–1892 (2013).

5. Hellstrom, J. K., Törrönen, A. R. & Mattila, P. H. Proanthocyanidins in common food products of plant origin. *J. Agric. Food Chem.* **57,** 7899–7906 (2009).

6. Santos-Buelga, C. & Scalbert, A. Proanthocyanidins and tannin-like compounds – nature, occurrence, dietary intake and effects on nutrition and health. *J. Sci. Food Agric.* **80,** 1094–1117 (2000).

7. Dixon, R. A., Xie, D.-Y. & Sharma, S. B. Proanthocyanidins – a final frontier in flavonoid research? *New Phytol.* **165,** 9–28 (2005).

8. FAO. Sorghum and millets in human nutrition. http://www.fao.org/docrep/T0818E/T0818E04.htm. (1995).

9. Janzen, Edward L., W. W. W. Cooperative marketing in specialty grains and identity preserved grain markets. *North Dakota State University, Department of Agribusiness and Applied Economics, Agribusiness & Applied Economics Report* (2002).

10. Taylor, J. R. N., Schober, T. J. & Bean, S. R. Novel food and non-food uses for sorghum and millets. *J. Cereal Sci.* **44,** 252–271 (2006).

11. Elbehri, A. *The changing face of the u.s. grain system: differentiation and identity preservation trends*. (United States Department of Agriculture, Economic Research Service, 2007). at http://ideas.repec.org/p/ags/uersrr/7185.html

12. Cureton, P. & Fasano, A. in *Gluten-free food science and technology* (ed. Gallagher, E.) 1–15 (Wiley-Blackwell, 2009). at http://onlinelibrarywiley.com/doi/10.1002/9781444316209.ch1/summary

13. Harlan, J. R., Wet, D. & J, J. M. A simplified classification of cultivated sorghum. *Crop Science* **12,** 172–176 (1972).

14. Awika, J. M. & Rooney, L. W. Sorghum phytochemicals and their potential impact on human health. *Phytochemistry* **65,** 1199–1221 (2004).

15. Nicholson, R. L., Kollipara, S. S., Vincent, J. R., Lyons, P. C. & Cadena-Gomez, G.

Phytoalexin synthesis by the sorghum mesocotyl in response to infection by pathogenic and nonpathogenic fungi. *Proc. Natl. Acad. Sci. U.S.A.* **84,** 5520–5524 (1987).

16. Dixon, R. A. Natural products and plant disease resistance. *Nature* **411,** 843–847 (2001).

17. Winefield, C. S. *et al.* Investigation of the biosynthesis of 3-deoxyanthocyanins in *Sinningia cardinalis*. *Physiol. Plant.* **124,** 419–430 (2005).

18. Sharma, M. *et al.* Expression of flavonoid 3'-hydroxylase is controlled by P1, the regulator of 3-deoxyflavonoid biosynthesis in maize. *BMC Plant Biol.* **12,** 196 (2012).

19. Malathi, P. *et al.* Differential accumulation of 3-deoxy anthocyanidin phytoalexins in sugarcane varieties varying in red rot resistance in response to Colletotrichum falcatum infection. *Sugar Tech* **10,** 154–157 (2008).

20. Rooney, W. L. in *Sorghum: origin, history, technology, and production* (eds. Smith, C. W. & Frederiksen, R. A.) (John Wiley & Sons, 2000).

21. Morohashi, K. *et al.* A genome-wide regulatory framework identifies maize pericarp color1 controlled genes. *Plant Cell* (2012). doi:10.1105/tpc.112.098004

22. Wu, Y. *et al.* Presence of tannins in sorghum grains is conditioned by different natural alleles of *Tannin1*. *Proc. Natl. Acad. Sci. U.S.A.* (2012).

23. Ibraheem, F., Gaffoor, I. & Chopra, S. Flavonoid phytoalexin-dependent resistance to anthracnose leaf blight requires a functional *yellow seed1* in Sorghum bicolor. *Genetics* **184,** 915–926 (2010).

24. Mace, E. S. & Jordan, D. R. Location of major effect genes in sorghum (*Sorghum bicolor* (L.) Moench). *Theor Appl Genet* **121,** 1339–1356 (2010).

25. Routaboul, J.-M. *et al.* Metabolite profiling and quantitative genetics of natural variation for flavonoids in arabidopsis. *J. Exp. Bot.* (2012). doi:10.1093/jxb/ers067

26. Olsen, K. M. & Wendel, J. F. Crop plants as models for understanding plant adaptation and diversification. *Front. Plant Sci.* **4,** (2013).

27. Morris, G. P. *et al.* Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 453–458 (2013).

28. Flint-Garcia, S. A. Genetics and consequences of crop domestication. *J. Agric. Food Chem.* **61,** 8267–8276 (2013).

29. Myles, S. *et al.* Association Mapping: Critical considerations dhift from genotyping to experimental design. *Plant Cell* **21,** 2194–2202 (2009).

30. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6,** 95–108 (2005).

31. Huang, X. *et al.* Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* **42,** 961–967 (2010).

32. Shu, X., Backes, G. & Rasmussen, S. K. Genome-wide association study of resistant starch (RS) phenotypes in a barley variety collection. *J. Agric. Food Chem.* **60,** 10302–10311 (2012).

33. Casa, A. M. *et al.* Community resources and strategies for association mapping in sorghum. *Crop Sci.* **48,** 30 (2008).

34. USDA. GRIN National Genetic Resources Program. URL (http://www.ars-grin.gov) (2014-07-25). (2014).

35. Dykes, L., Hoffmann Jr., L., Portillo-Rodriguez, O., Rooney, W. L. & Rooney, L. W. Prediction of total phenols, condensed tannins, and 3-deoxyanthocyanidins in sorghum grain using near-infrared (NIR) spectroscopy. *J. Cereal Sci.* (2014). doi:10.1016/j.jcs.2014.02.002

36. Kaluza, W. Z., McGrath, R. M., Roberts, T. C. & Schroeder, H. H. Separation of phenolics of sorghum bicolor (L.) Moench grain. *J. Agric. Food Chem.* **28,** 1191–1196 (1980).

37. Fuleki, T. & Francis, F. J. Quantitative methods for anthocyanins. *Journal of Food Science* **33,** 72–77 (1968).

38. Price, M. L., Van Scoyoc, S. & Butler, L. G. A critical evaluation of the vanillin reaction as an assay for tannin in sorghum grain. *J. Agric. Food Chem.* **26,** 1214–1218 (1978).

39. Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6,** e19379 (2011).

40. Paterson, A. H. *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* **457,** 551–556 (2009).

41. Buckler Lab for Maize Genetics and Diversity. TASSEL. URL(http://sourceforge.net/projects/tassel) (7/25/14). (2014).

42. Lipka, A. E. *et al.* GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28,** 2397–2399 (2012).

43. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38,** 203–208 (2006).

44. Morris, G. P. *et al.* Dissecting genome-wide association signals for loss-of-function phenotypes in sorghum flavonoid pigmentation traits. *G3: Genes, Genomes, Genet.* **3,** 2085–2094 (2013).

45. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42,** 355–360 (2010).

46. Liu, Z. *et al.* Regulation, evolution, and functionality of flavonoids in cereal crops. *Biotechnol. Lett.* **35,** 1765–1780 (2013).

47. Lepiniec, L. *et al.* Genetics and biochemistry of seed flavonoids. *Annu. Rev. Plant Biol.* **57,** 405–430 (2006).

48. Petroni, K. & Tonelli, C. Recent advances on the regulation of anthocyanin synthesis in reproductive organs. *Plant Science* **181,** 219–229 (2011).

49. Nesi, N. *et al.* The TT8 gene encodes a basic helix-loop-helix domain protein required for expression of DFR and BAN genes in arabidopsis siliques. *Plant Cell* **12,** 1863–1878 (2000).

50. Furukawa, T. *et al.* The Rc and Rd genes are involved in proanthocyanidin synthesis in rice pericarp. *The Plant Journal* **49,** 91–102 (2007).

51. Franken, P., Schrell, S., Peterson, P. A., Saedler, H. & Wienand, U. Molecular analysis of protein domain function encoded by the myb-homologous maize genes C1, Zm 1 and Zm 38. *The Plant Journal* **6,** 21–30 (1994).

52. *Sorghum: Origin, History, Technology, and Production*. (John Wiley & Sons, 2000).

53. Boddu, J. *et al.* Expression of a putative flavonoid 3′-hydroxylase in sorghum mesocotyls synthesizing 3-deoxyanthocyanidin phytoalexins. *Physiol. Mol. Plant Pathol.* **65,** 101–113 (2004).

54. Chen, W., Müller, D., Richling, E. & Wink, M. Anthocyanin-rich purple wheat prolongs the life span of *caenorhabditis elegans* probably by activating the DAF-16/FOXO transcription factor. *J. Agric. Food Chem.* **61,** 3047–3053 (2013).

55. Sriseadka, T., Wongpornchai, S. & Rayanakorn, M. Quantification of flavonoids in black rice by liquid chromatography-negative electrospray ionization tandem mass spectrometry. *J. Agric. Food Chem.* **60,** 11723–11732 (2012).

56. Zilic, S., Serpen, A., Akıllıoğlu, G., Gökmen, V. & Vančetović, J. Phenolic compounds, carotenoids, anthocyanins, and antioxidant capacity of colored maize (*Zea mays* L.) kernels. *J. Agric. Food Chem.* **60,** 1224–1231 (2012).

57. Kim, M.-J. *et al.* Relationship between phenolic compounds, anthocyanins content and antioxidant activity in colored barley germplasm. *J. Agric. Food Chem.* **55,** 4802–4809 (2007).

58. Dykes, L., Rooney, W. L. & Rooney, L. W. Evaluation of phenolics and antioxidant activity of black sorghum hybrids. *J. Cereal Sci.* **58,** 278–283 (2013).

59. Dykes, L., Rooney, L. W., Waniska, R. D. & Rooney, W. L. Phenolic compounds and antioxidant activity of sorghum grains of varying genotypes. *J. Agric. Food Chem.* **53,** 6813–6818 (2005).

CHAPTER 3

NATURAL VARIATION AND GENOME-WIDE ASSOCIATION STUDY OF GRAIN COMPOSITION IN GLOBAL SORGHUM GERMPLASM[2]

[2] Davina H. Rhodes*, Leo Hoffmann Jr, William L. Rooney, Matt Myers, Richard Boyles, Zach Brenton, Geoffrey Morris, Stephen Kresovich. To be submitted to *G3: Genes, Genomes, Genetics.*

3.1 ABSTRACT

Sorghum [*Sorghum bicolor* (L.) Moench] is an important cereal crop for dryland areas in the United States and for small-holder farmers in Africa. Natural variation of sorghum grain composition (protein, fat, and starch) between accessions can be used for crop improvement, but the genetic controls are still unresolved. The goals of this study were to quantify natural variation of sorghum grain composition and to identify single-nucleotide polymorphisms (SNPs) associated with variation in their concentrations. In this study, we quantified protein, fat, and starch in a global sorghum diversity panel (n = 381) using near-infrared spectroscopy (NIRS). Protein content ranged from 7.5% to 20.9%, fat content ranged from 1.1% to 4.9%, and starch content ranged from 60.8% to 73.2%. Among the sorghum races, bicolor accessions had the highest mean protein (14.7%) and fat (3.7%), and the lowest mean starch (65%). Kafir accessions had the lowest mean protein (10.5%) and fat (2.6%), and the highest mean starch (68.3%). A genome-wide association study (GWAS) with 404,628 SNP markers identified 81, 81, and 11 significant single nucleotide polymorphism (SNP) markers for sorghum protein, fat, and starch, respectively. Published RNAseq data, generated as a community resource for transcriptomic analyses, was used to identify candidate genes within a GWAS quantitative trait loci (QTL) region. Candidate genes identified include *NAM-B1*, *AMY3*, and *SSIIb*, genes previously shown to be associated with grain composition traits. This survey of grain composition in sorghum germplasm and identification of QTL significantly associated with protein, fat, and starch, contributes to our understanding of the genetic basis of natural variation in sorghum grain composition.

3.2 INTRODUCTION

The 1996 World Food summit announced a goal of halving the number of undernourished people in the world by the year 2015. Although much progress has been made towards this goal, one in eight people still suffer from chronic hunger [1]. This can be alleviated by improving the nutrition of staple cereal crops, which provide the majority of nutrients to the world's population, especially in developing countries. Sorghum, one of the world's most important cereal crops, feeds millions of people in sub-Saharan Africa [2], where the highest prevalence of undernourishment in the world is found [1]. Understanding the natural variation of protein, fat, and starch, and identifying quantitative trait loci (QTL) associated with their natural variation in sorghum grain can help improve its nutritional quality through crop improvement programs and marker assisted selection.

Seeds contain protein, fat, and starch stores in order to support the developing seedling until it can sustain itself. Since these nutrient stores are also critical components of the human diet, many researchers have focused on improving the nutrient composition of seeds from food plants [3]. For instance, the Illinois long-term selection experiment, which began in 1896, has increased the oil and protein content of maize inbred lines to 20% and 27%, respectively, compared to ~6% and ~12%, in an average maize line [4–7]. The composition of grain is controlled by complex regulation that takes place during the seed filling stage of seed maturation, when protein, fat, and starch storage compounds accumulate [8]. Due to the importance that grain holds in the world food system, this process has been extensively studied in cereal crops [9,10]. Key insights have been discovered through several rice and maize mutations with altered grain composition,

including *opaque-2* and *floury-2*, which affect protein content [11–14]; *linoleic1* and *fad2*, which affect fat content [15–17]; and *shrunken1* and *amylose extender1*, which affect starch content [18–20]. Mutations that modify sorghum grain composition include *waxy*, which lacks amylose and has increased protein and starch digestibility [21,22]; *sugary*, which has increased sucrose content [23,24]; and *high-lysine*, which has increased lysine content and protein digestibility [25].

QTL and association studies have detected several loci controlling sorghum grain composition [26–30], and the *waxy* mutation has been mapped to 1.8 Mb on chromosome 10 [31], but more work needs to be done to precisely identify genes responsible for natural variation of grain composition. Recently, GWAS studies have been successful in identifying allelic polymorphisms for important agronomic traits in cereal crops [32–35], including alleles responsible for variation in grain composition [33,35–39], but a GWAS study on sorghum grain composition has not been conducted.

Surveying the natural variation of grain composition in the sorghum germplasm and finding the molecular basis underlying the variation are necessary for understanding how to improve the nutritional value of sorghum. New sources of genetic variation can be used for crop improvement, especially in developing countries where technologies that exist for improving the nutritional value of grain, such as commercial fortification, are not accessible and/or affordable [40–42]. The goals of this study were to quantify natural variation of sorghum grain composition and to identify SNPs that are associated with variation in grain composition. Here, we characterize the natural variation of sorghum grain composition in a global diversity panel of ~400 sorghum varieties, and use GWAS

with 404,628 SNP markers to identify allelic variation associated with variation in grain composition.

3.3 MATERIALS AND METHODS

3.3.1 Plant Materials

We investigated a total of 381 sorghum accessions, comprising 308 accessions from the Sorghum Association Panel (SAP) [43] and an additional 73 accessions selected to supplement the panel. The panel includes domesticated sorghum from all five major races (bicolor, guinea, caudatum, kafir, and durra) and 10 intermediate races (all combinations of the major races), which are based on morphological differences [44], as well as important breeding lines from the United States. Seeds were obtained from the U.S. National Plant Germplasm System's Germplasm Resources Information Network (GRIN) [45] and planted in late April 2012 at Clemson University Pee Dee Research and Education Center in Florence, SC. A two-fold replicated complete randomized block design was used. Panicles from each plot were collected at physiological maturity, which occurs once grain filling is complete. Due to differences in maturity among these accessions, harvest occurred between September and October. Once harvested, panicles were air dried in a greenhouse and then mechanically threshed and any remaining glumes were removed with a Wheat Head Thresher (Precision Machine Company, Lincoln, NE). This panel is referred to as SC2012.

3.3.2 Phenotyping

Protein, fat, and starch content were predicted using NIRS. Twenty grams of cleaned whole grain from one replicate were scanned with a FOSS XDS spectrometer

(FOSS North America, Eden Prairie, MN, USA). To determine reproducibility, duplicates on a subset of 218 accessions available from replicate plots were also scanned. The NIR reflectance spectra were recorded using the ISIscan software (Version 3.10.05933) and converted to estimates of protein, fat, and starch concentrations. Samples with unusual reflectance were visually inspected and NIRS was repeated. Seventeen samples were removed from further analysis either because they contained mixed grain (mixed size, shape, or color) or because their readings were outside the range of the available NIRS calibration curve. Flowering-date was determined by the number of days from planting until the start of anthesis. The total grain weight of 100 grains per accession was recorded. Chemical analysis for protein, fat, and starch concentrations in a subset of 34 samples (17 accessions with duplicates) was performed by Ward Laboratories, Inc. (Kearney, NE).

3.3.3 Genomic Analysis

Genotypes were available for all of the accessions [32,46]. GWAS was carried out on 404,628 SNP markers, using the statistical genetics package Genome Association and Prediction Integrated Tool (GAPIT) [47]. A standard mixed linear model (MLM) [48] with kinship (K), which controls for relatedness among the accessions in the panel, was performed [49]. GAPIT corrected for multiple testing error by controlling the false discovery rate (FDR) at 5% using the Benjamini and Hochberg procedure [50]. Pseudo-heritability (proportion of phenotypic variation explained by genotype) was estimated from the kinship (K) model in GAPIT [49] as the R-squared of a model with no SNP affects. An *a priori* candidate gene list with 521 candidates was developed.

3.3.4 Expression data

To identify candidate genes within the GWAS QTL regions, we used a published sorghum transcriptome atlas that included tissues from young leaves, primordial inflorescences, inflorescences, anthers, pistils, whole seeds 5 days after pollination, whole seeds 10 days after pollination, developing embryo, and developing endosperm [51] (Appendix C). We used the definitions of Davidson et al, as follows: FPKM < or = 1 = "not expressed"; FPKM < or = 4 = "low-expressed"; FPKM between 4 and 24 = "intermediate-expressed"; and FPKM > or = 24 = "high-expressed".

3.4 RESULTS

3.4.1 Phenotypic variation of sorghum grain composition

We first investigated the range of protein, fat, and starch content and their covariation with each other. We found that the germplasm showed a wide range of diversity in grain composition. Protein content ranged from 7.5% to 20.9%, fat content ranged from 1.1% to 4.9%, and starch content ranged from 60.8% to 73.2% (Figure 3.1). Pearson's correlations were calculated between protein, fat, and starch (Figure 3.1). There was a strong negative correlation between starch and both protein (r = -0.88, $p < 10^{-17}$) and fat (r = -0.73, $p < 10^{-17}$), and a strong positive correlation between protein and fat (r = 0.75, $p < 10^{-17}$). Grain composition concentrations are expressed as percentage by total seed weight, therefore an increase in one component necessitates a decrease in another component. The negative correlations with starch may, in part, be driven by this method. In order to account for differences in seed weight, we multiplied the percent concentration by the seed weight of each accession to get absolute estimates of the mass

of each constituent per grain, and recalculated Pearson's correlations. Using these estimates, there was a moderate positive correlation between starch and both protein (r = 0.58, $p < 10^{-17}$) and fat (r = 0.50, $p < 10^{-17}$), and a strong positive correlation between protein and fat (r = 0.84, $p < 10^{-17}$). These positive correlations between the traits reflect that total amounts of protein, fat, and starch increase with increases in total seed weight.

Next we investigated grain protein, fat, and starch covariation with factors that could reduce their biological availability for human consumption. Since the digestibility of protein and starch can be decreased by proanthocyanidins, and possibly other polyphenols[52], it is useful to know if there is a pattern of covariation between grain composition traits and polyphenol content. To this end, we used polyphenol data previously generated by our group[46] to calculate Pearson's correlations with protein, fat, and starch concentrations (using the weight adjusted concentrations). Total phenolics had a small positive correlation with protein (0.13, P = 0.01) and a small negative correlation with starch (-0.13, P = 0.01). The 3-deoxyanthocyanidins had a small positive correlation with protein (0.14, P = 0.01) and fat (0.12, P = 0.02). Proanthocyanidins were not significantly correlated with protein, fat or starch.

Since NIRS estimates rely on predictive equations developed through chemical analysis of a calibration population, concentrations that are outside of the range of the calibration population, or at the high or low extremes of the calibration population, may not be accurately predicted. Therefore, in order to verify the accuracy of the NIRS estimates, chemical analyses were conducted on a subset of 34 samples (17 accessions in duplicate) with very high or very low estimates of protein, fat, and starch. Pearson’s correlations between the NIRS and chemical analyses results found that there were

significant correlations with protein (0.43, P = 0.01) and with starch (0.56, P = 0.001), but not with fat (-0.02, P = 0.91; Figure 3.2). These results suggest that NIRS predictions may not be as accurate when measuring high or low extremes of protein and starch concentrations, and may not be at all accurate when measuring fat concentrations. Absolute levels of fat are much lower than protein and starch (on average, fat made up only 2.9% of the grain constituents, compared to 11.8% protein and 67.1% starch), which may be the cause of the measurement error in fat.

3.4.2 Population structure of grain composition traits

Knowledge of variation in grain composition across the sorghum races can be applied to germplasm utilization. Among the sorghum races (Figure 3.3A), bicolor accessions had the highest mean protein (14.7%) and fat (3.7%), and the lowest mean starch (65%). Bicolor-durra (12.9%) and durra (12.7%) accessions also had high mean protein. Guinea (3.2%) and durra (3.2%) accessions also had high mean fat. Kafir accessions had the lowest mean protein (10.5%) and fat (2.6%), and the highest mean starch (68.3%). Guinea (67.2%) and caudatum-kafir (67.2%) accessions also had high mean starch.

We also determined the mean concentrations by country to better understand the geographic patterns for protein, fat, and starch in sorghum grain (Figure 3.3B). Accessions from Ethiopia had the highest mean protein (13.8%) and the lowest mean starch (65.8%). There were no significant differences in fat content in accessions between countries.

3.4.3 Genome wide association study

We had GBS data for 373 out of the 381 phenotyped accessions. Pseudo-

heritability, the proportion of variance explained by genotype in the mixed model, was 95.7% for protein, 73% for fat, and 91.2% for starch. The lower heritability of fat may be due to the NIRS measurement error discussed in the previous section.

Prior to running GWAS, we conducted an extensive literature search to identify potential candidate genes, and compiled a list of previously identified candidate genes associated with grain composition [37,38,53], as well as genes known to be involved in grain maturation and grain filling [8,10,54,55,55] in Arabidopsis, rice, and maize, resulting in a list of 520 *a priori* candidate genes. To investigate the genetic basis of natural variation of protein, fat, and starch in sorghum grain, we conducted a GWAS using the diverse association panel with 404,628 SNP markers. Again, we first multiplied the percent concentration by the seed weight of each accession in order to control for differences in seed weight. The MLM identified 81, 81, and 11 significant SNPs for protein, fat, and starch, respectively, at a genome-wide FDR of 5% (Appendix D). To identify candidate genes within a GWAS QTL region, we used RNAseq data that was generated as a community resource for transcriptomic analyses [51]. Genes in a QTL region that were expressed during grain maturation were considered good candidates.

The MLM for both protein and fat identified 81 significant SNPs at a genome-wide FDR of 5% (Figure 3.4A-B, Appendix D.1-D.2), with two highly significant association peaks. There was a large association peak on chromosome 2 at ~57.7 Mb. Close to this peak is an *a priori* candidate gene that is a putative homolog of alpha-amylase 3 (*AMY3*, Sb02g023790; 57,701,214-57,703,517 bp). The expression data for this gene shows no induction in the leaves or in the day 5 seeds, but a low expression (2.1 FPKM) in the day 10 seeds and in the endosperm (5.8 FPKM; Appendix C.1). Also near

this peak is an *a priori* candidate gene that is a putative homolog of *NAC2* (Sb02g023960; 57,931,636-57,703,517 bp). The expression data shows this gene is only expressed in the seed, with no induction on day 5, but highly expressed by day 10 at 62.4 FPKM. It is also highly expressed in the endosperm (65.3 FPKM), but not in the embryo (1.6 FPKM; Appendix C.1).

The second highly significant association peak in the protein and fat GWAS was on chromosome 4 at ~57.7 Mb (Figure 3.4A-B, Appendix D.1-D.2). It was much more significant in the protein GWAS. The closest *a priori* candidate is a putative *wrinkled1* homolog (Sb04g027940; 57,859,449-57,863,521 bp). This gene has moderate expression in the leaves (13.5 FPKM), and high expression in the day 5 seeds (27.4 FPKM) with a decrease by day 10 (14.1 FPKM; Appendix C.2). The embryo has moderate levels (18.7 FPKM), while the endosperm has high levels (31.2 FPKM). The peak is also near a gene that has homology to *starch synthase IIb* (*SSIIb*, Sb04g028060; 57,999,747-58,003,544 bp). Expression is particularly high in leaves (80.7 FPKM) and still elevated in day 5 seeds (21.1 FPKM), but lower by day 10 (3.5 FPKM). The embryo and endosperm have the same levels at ~ 5 FPKM (Appendix C.2).

The starch GWAS identified 11 significant SNP associations (Figure 3.4C, Appendix D.3). The top SNP was on chromosome 6 at 48.9 Mb. The *a priori* candidate is another putative *NAC homolog* at 48.6 Mb (Sb06g019010; 48,600,551-48,601,945 bp), which has high expression in the day 5 (78.7 FPKM) and day 10 (62.1 FPKM) seeds, as well as in the endosperm (39.3 FPKM) (Appendix C.3). The most defined peak in the GWAS was on chromosome 2, with SNP associations from 66.2 Mb to 68.2 Mb. The closest a priori candidate was a chromatin remodeling factor gene (PICKLE;

Sb02g033850) at 68.4 Mb, with moderate expression in all tissues.

Since starch makes up the majority of the grain composition, it is possible that some variation in protein and fat are driven by variation in starch. To determine if starch could be influencing the values, we ran two linear models in which we fit either protein or fat as the dependent variable and starch as the independent variable (using the weight adjusted values). We hypothesized that natural variation in starch pathways might be affecting protein and fat content in the grain due to a limited pool of carbon. If we assume that patterns in protein and fat are driven by starch, then starch could account for a significant proportion of the variance—34% of all the variance in protein ($p < 10^{-17}$) and 21% of the variance in fat ($p < 10^{-17}$)—but there is a large portion of variance in protein and fat is still unexplained. Therefore, we conducted GWAS on the residuals (the amount of variation in fat and protein that could not be explained by starch) from the linear models to determine if there was anything left to map after accounting for covariation in starch (Figure 3.5). The GWAS on protein residuals identified 82 significant SNPs at the FDR adjusted significance threshold, with a peak on chromosome 2 at ~57.6 Mb and chromosome 4 at ~57.8 Mb (Figure 3.5A). The fat residuals GWAS identified 73 significant SNPs at the FDR adjusted significance threshold, also with a large peak on chromosome 2 at ~57.6 Mb and a smaller peak on chromosome 4 at ~57.8 Mb (Figure 3.5B).

3.4.4 Control Analysis on GWAS QTL

To test if the GWAS QTL are stable across environments, we conducted a GWAS using phenotype data from a sorghum panel grown in Kansas in 2007 (hereafter referred

to as KS2007) that primarily consisted of the SAP [26]. No SNPs reached the FDR adjusted significance threshold and there were no obvious association peaks (Figure 3.6). The replicate samples from our dataset were grown in a two-fold block design, so as a control analysis, we conducted a GWAS separately on data from each block. We had genotype data for 213 of the 218 duplicate accessions. The GWAS identified the same association peaks when run separately on each block (Figure 3.7). Phenotypic covariates are another potential source of misleading associations [56]. Maturity differences across the panel can potentially lead to grain composition differences. If maturity was a confounding factor in the panel, then we could expect that one or more of the QTL identified in the SC2012 GWAS was actually maturity loci instead of grain composition loci. With this in mind, we conducted a GWAS using flowering time data for the SC2012 panel. We had genotype data for 230 of the 234 phenotyped accessions. The major peak in the GWAS mapped to the previously identified maturity locus, *ma1* [32,57], and, importantly, did not map to significant associations identified in the SC2012 GWAS (Figure 3.8).

3.5 DISCUSSION

3.5.1 Covariation of starch fat and protein in sorghum grain

GWAS revealed that protein, fat, and starch variation in the sorghum global diversity panel appear to be controlled by many small effect genes, some of which are significantly associated with more than one grain composition trait. GWAS for protein and fat identified two major peaks in common, one on chromosome 2 at 57.7 Mb and the other on chromosome 4 at 57.7 Mb. The starch GWAS only identified 11 significant associations with small peak, none of which were in common with protein and fat.

We believe that the large peak on chromosome 2 at 57.7 Mb is a true association. The peak remained when GWAS was performed on the individual biological replicates, suggesting that, given that environment, we have the correct phenotypes and associations. Additionally, the peak does not appear to be related to flowering time differences among the accessions in the panel. The peak is near a QTL that was significantly associated with fat in a sorghum linkage study that used a biparental population derived from the cultivar Rio and BTx623, which was grown in Texas (hereafter referred to as TX2008) [28]. TX2008 identified a QTL on chromosome 2 near the genetic marker txp298, which is located at ~57.1 Mb [58]. Promising *a priori* candidates near this peak are the *AMY3* and *NAM-B1* homologs. AMY3 is an alpha-amylase debranching enzyme that hydrolyzes the glucosidic bonds that make up starch. *AMY1* was previously identified as a candidate gene in a maize grain composition GWAS study [38]. A recent study using AMY3 overexpression lines found that the increased levels of AMY3 did not significantly affect starch content, but fat content was increased in the mature endosperm where starch had been partially degraded [59]. The authors suggested that starch degradation during grain maturation led to the release of sucrose that was then shunted into the Kennedy pathway for fat synthesis [59]. The other candidate genes near the peak on chromosome 2 is a putative *NAC* gene with homology to *NAM-B1*. *NAM-B1* is a wheat gene that was found to be involved in nutrient remobilization from senescing leaves to the developing grain, leading to alterations in grain protein, iron, and zinc content [60]. In this same study, two stay-green plants showed significant reduction of RNA levels in different *NAM* homologs, compared to control lines, and these stay-green plants exhibited delayed chlorophyll degradation in flag leaves [60]. Allelic variation in several other *NAC* genes has

been implicated in senescence regulation [61]. Interestingly, a functional sorghum stay-green gene (*SG3a*) has been mapped to a region near the txp298 genetic marker, (which is located on chromosome 2 at ~57.1 Mb) [62–64]. *Stg3* is related to delayed onset of leaf senescence during post-anthesis water deficit, as well as lower rates of leaf senescence [65].

The significant association peak on chromosome 4, at 57.7 Mb also colocalized with a QTL identified in the TX2008 study, which was significantly associated with protein and corneous endosperm [28]. The TX2008 QTL was near the genetic markers txp41 located on chromosome 4 at ~58.6 Mb [58], which is near an *SSIIb* gene (Sb04g028060; 57,999,747-58,003,544 bp). Studies in both maize and rice have found that *SSIIb*, a starch branching enzyme, is primarily expressed in the leaves, with weaker expression in the seeds, while *SSIIa* is primarily expressed in the endosperm [66,67]. The sorghum expression data for this gene is consistent with these patterns, with very high expression in the leaves and moderate expression in the seeds, embryo and endosperm (Appendix C.2). The KS2007 study used the QTL identified in the TX2008 study to conduct a candidate gene assay, in which they looked for SNP associations with grain composition traits. The KS2007 study, primarily composed of SAP lines, found a significant association between starch and the SNP (58,000,108 bp) within the *SSIIb* gene [26], suggesting that this may be the gene responsible for the peak in the SC2012 GWAS. Another candidate gene possibility is a *wrinkled1* (Sb04g027940; 57,859,449-57,863,521) *a priori* candidate that is 140 kb closer to the significant SNP identified in the SC2012 GWAS. *Wrinkled1* is a key regulator controlling seed oil biosynthesis, and has been found to alter fatty acid and amino acid content in maize when overexpressed [68].

We have identified many candidate genes for the peaks shared between grain

composition traits, but further studies are required to validate their involvement in grain composition variation between sorghum varieties. Since sorghum grain composition traits appear to be controlled by many small effect genes, biparental mapping or nested association mapping may be helpful in further refining candidate genes [38]. Additionally, sequence analysis of the candidate genes is needed to identify causal polymorphisms.

3.5.2 Improvement of sorghum grain composition for human nutrition

The range of protein, fat, and starch content found in our diverse association panel may be useful for sorghum improvement. Bicolor sorghums had significantly higher mean protein levels (14.7%) than any other sorghum race, and are promising sources of genetic material for high protein sorghums. Cereals are predominantly used as sources of starch. Bicolor is the least derived race (i.e., retains most similarity to wild ancestors among the races), and high protein varieties may have been inadvertently counter-selected during cereal domestication when high starch varieties were selected. It may be that human selection for different food uses influenced the patterns of grain composition distribution among the races (e.g., thick porridge in one region requires a certain grain composition, while flat bread in another region requires a different grain composition).

This study provides genetic trait association loci that can be explored further for their potential use in molecular breeding to modify the composition of grain sorghum. The high heritability of each trait suggests the genetic contribution to variation is strong, however, the GWAS with the KS2007 SAP accessions did not identify the same large association peaks identified in the GWAS with the SC2012 SAP accessions, suggesting that a year-to-year or site-to-site environmental effect may be responsible for the

difference. This is not surprising since many studies have found grain composition variation between environments, indicating that at least some genes may only be significantly influential in a particular environment [28,69]. For example, in one study, fifty-one sorghum cultivars grown in five locations over two years exhibited protein and fat concentrations that were inconsistent across environments and years [70]. In another study, nine sorghum cultivars grown in three locations (two in Kansas and one in Texas) in one year were found to have significantly higher starch and lower protein and fat concentrations in Kansas compared to Texas, but composition was not affected by irrigation differences [71]. However, in another study that investigated grain composition differences between differing irrigation levels in ten sorghum cultivars, significant differences were found, with starch increasing as irrigation levels increased, and protein increasing as irrigation levels decreased [72]. In a study that evaluated waxy sorghum hybrids in two locations in Nebraska over two years, a significant difference was found in starch concentrations between locations and years [73]. NIRS and GWAS on SAP accessions grown in two subsequent years is currently underway and may help to confirm the results presented here, as well as provide a greater understanding of the heritability of protein, fat, and starch in sorghum grain.

Overall, we have identified promising sources of genetic material for manipulation of grain composition traits, and several loci and candidate genes that may control sorghum grain composition. Identification of SNPs that were previously found to have significant associations with protein, fat, and starch in sorghum grain suggests that GWAS is capable of detecting functional polymorphisms associated with sorghum grain composition traits. This survey of grain composition in sorghum germplasm and

identification of QTL significantly associated with protein, fat, and starch, contributes to our understanding of the genetic basis of natural variation in sorghum grain composition.

3.6 FIGURES

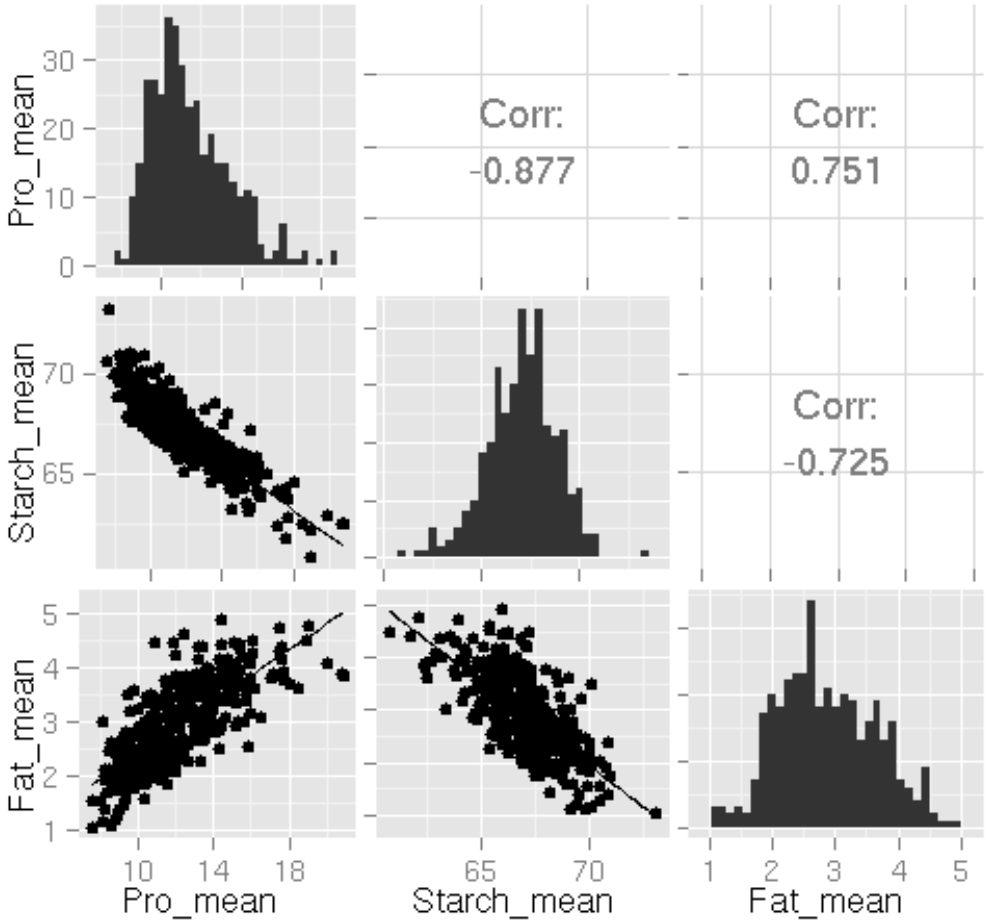


Figure 3.1 Relationship within and between grain composition traits in a global sorghum germplasm collection. The center diagonal presents histograms of each trait. The scatter plots with regression lines show the relationships between the traits. (*n* = 373)

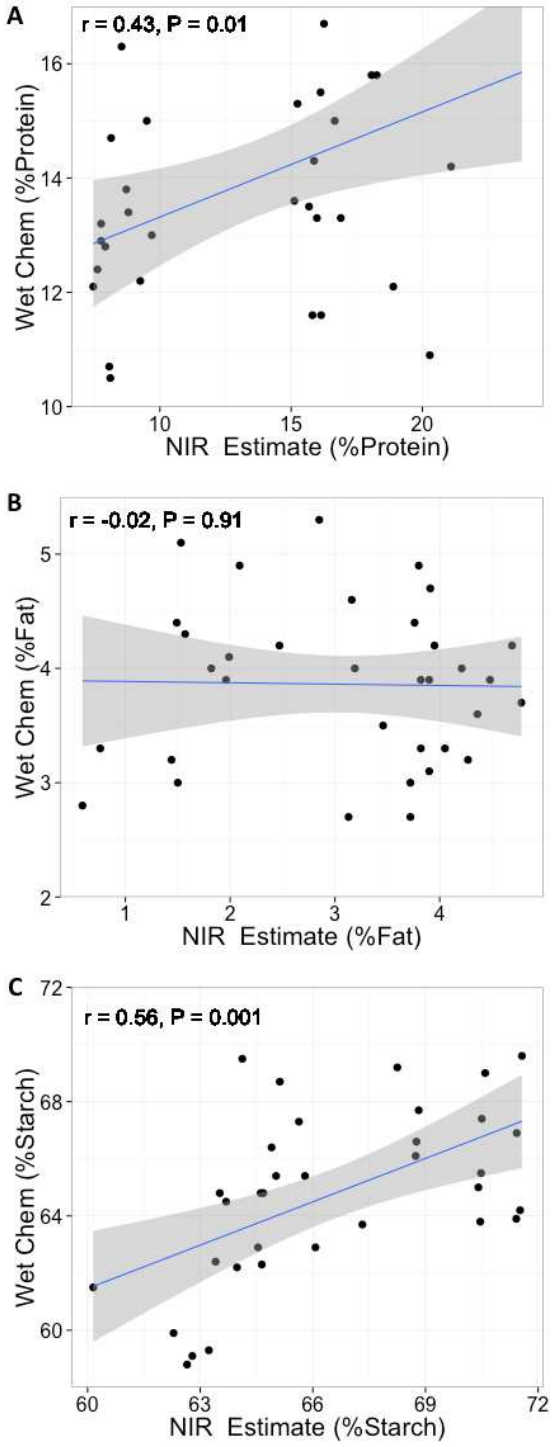


Figure 3.2 Correlations between NIRS estimates and chemical analysis. 34 sorghum grain samples (17 accessions in duplicate) were analyzed by chemical analysis and results were compared to NIRS results for concentrations of (A) protein, (B) fat, and (C) starch.

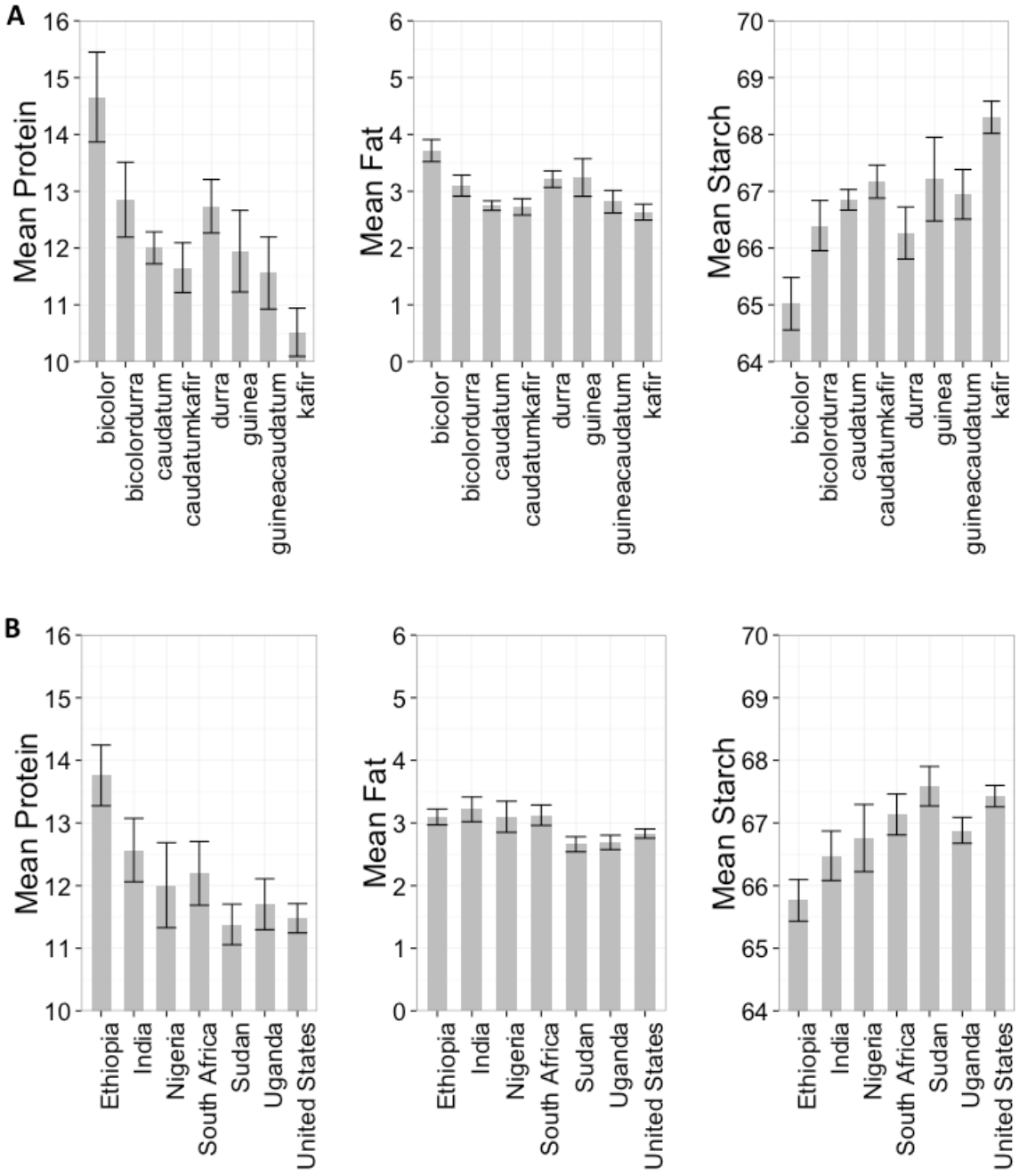


Figure 3.3 Population structure of grain composition traits in a global sorghum germplasm collection. Mean grain composition concentrations among (A) races and (B) geographic origin.

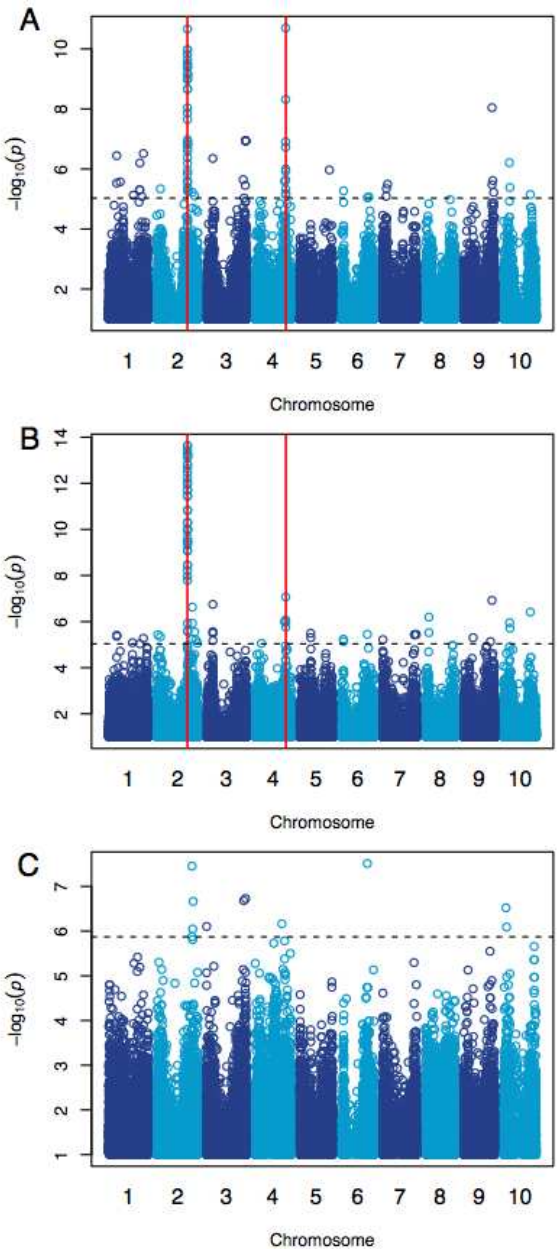


Figure 3.4 GWAS for protein, fat, and starch content in sorghum grain. Manhattan plots of association results from a MLM analysis using 404,627 SNP markers and 373 accessions. Each point represents a SNP, with the -log10 *p*-values plotted against the position on each chromosome. The red vertical lines indicate the positions of candidate genes. The horizontal dashed line represents the genome-wide significance threshold at 5% FDR. (A) protein; (B) fat; (C) starch.

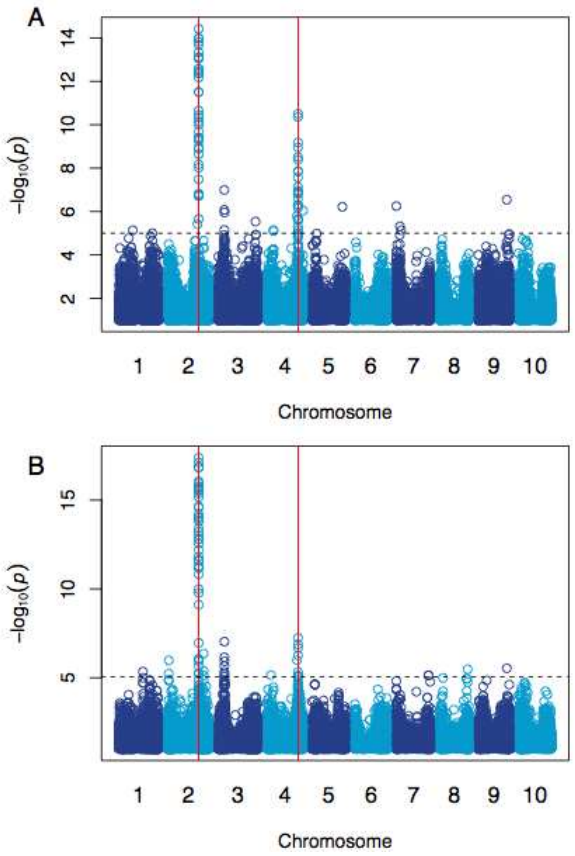


Figure 3.5 Residuals GWAS for protein and fat content in sorghum grain. Manhattan plots of association results from a MLM analysis using 404,627 SNP markers and 373 accessions. Each point represents a SNP, with the -log10 *p*-values plotted against the position on each chromosome. The red vertical lines indicate the positions of candidate genes. The horizontal dashed line represents the genome-wide significance threshold at 5% FDR. (A) protein and (B) fat.

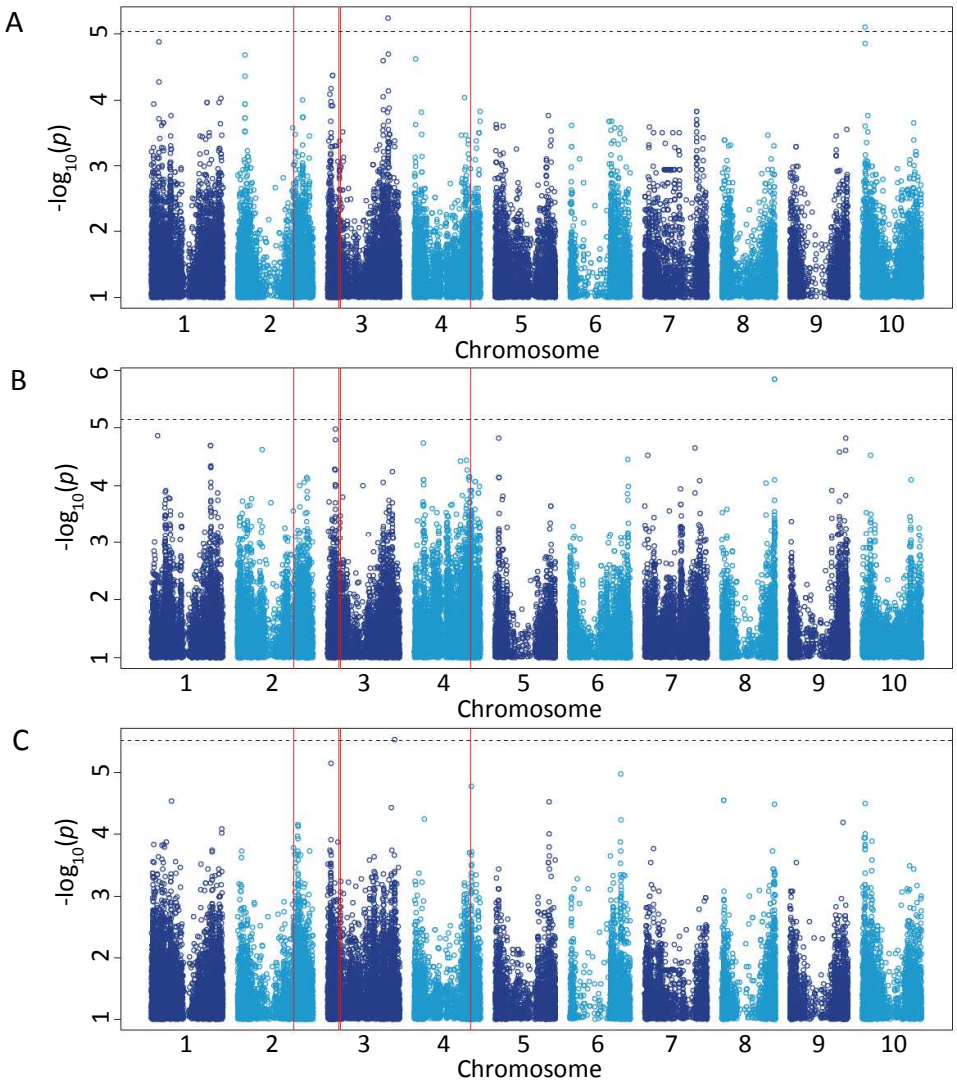


Figure 3.6 GWAS for protein, fat, and starch content in sorghum grain grown in Kansas in 2007[26]. Manhattan plots of association results from a MLM analysis using 404,627 SNP markers and 239 accessions. Each point represents a SNP, with the -log10 *p*-values plotted against the position on each chromosome. The red vertical lines indicate the positions of the major peaks that were identified with the data from the South Carolina panel. The horizontal dashed line represents the genome-wide significance threshold at 5% FDR. (A) protein; (B) fat; (C) starch.

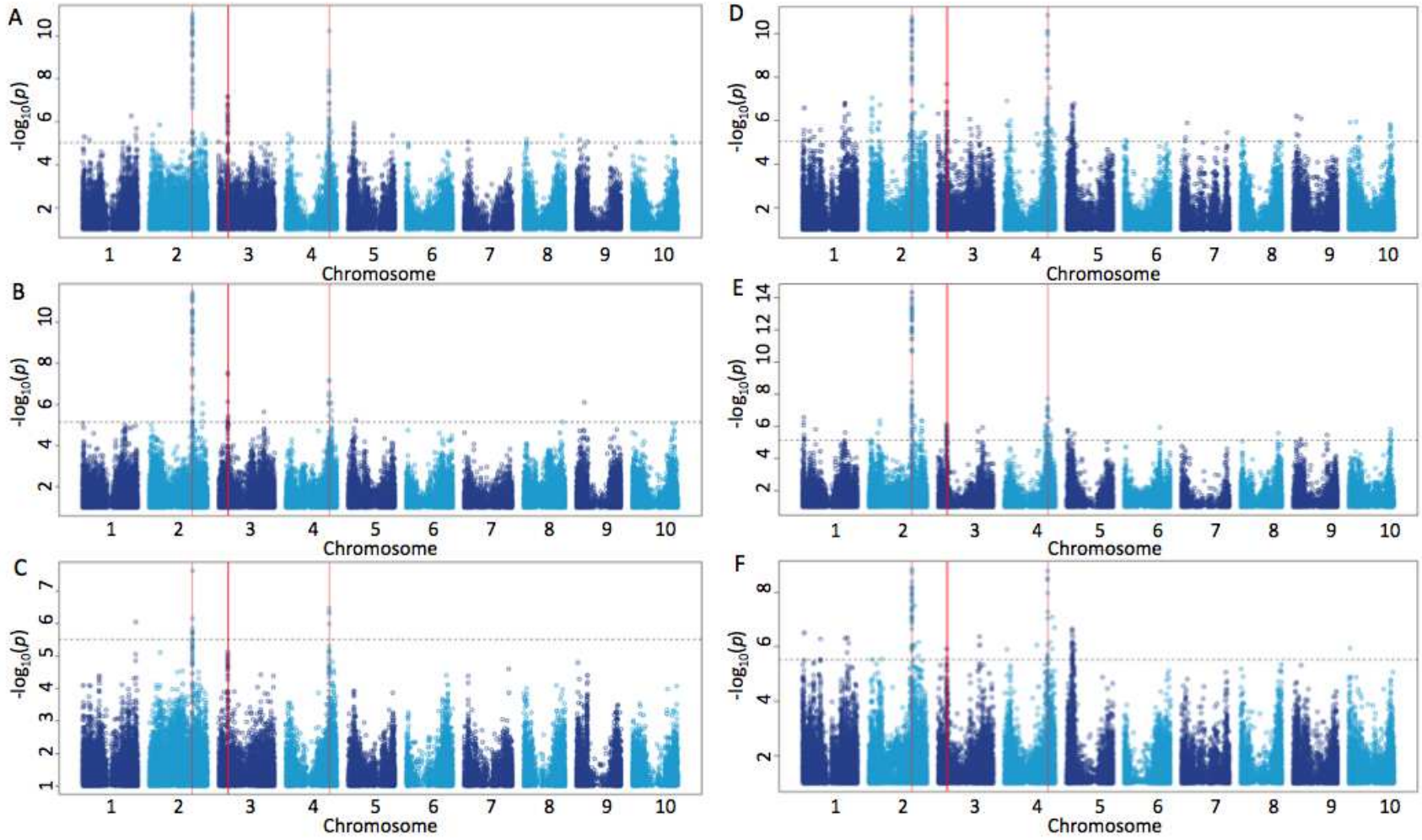

A
B
C
D
E
F
-log10(p)
Chromosome
1
2
3
4
5
6
7
8
9
10

Figure 3.7 GWAS for protein, fat, and starch content in replicate sets 1 and 2. Manhattan plots of association results from a CMLM analysis using 404,627 SNP markers and 213 accessions. Each point represents a SNP, with the -log10 *p*-values plotted against the position on each chromosome. The red vertical lines indicate the positions of peaks that were common between protein, fat, and starch. The horizontal dashed line represents the genome-wide significance threshold at 5% FDR. (A) protein replicate 1; (B) fat replicate 1; (C) starch replicate 1; (D) protein replicate 2; (E) fat replicate 2; (F) starch replicate 2.

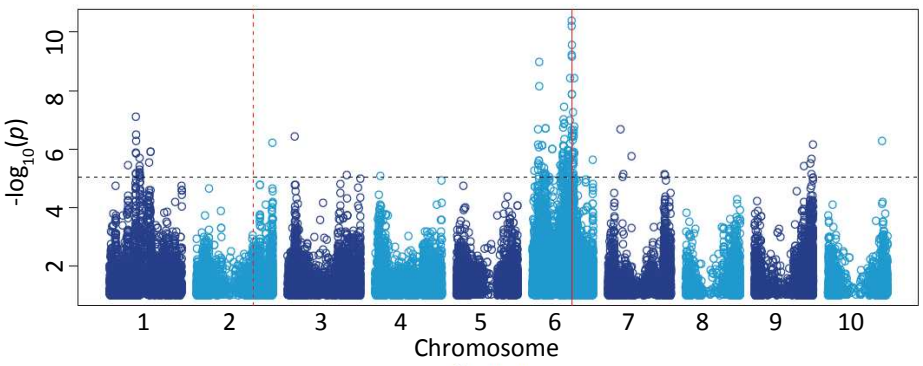


Figure 3.8 GWAS for flowering time in sorghum grain. Manhattan plot of association results from a MLM analysis using 404,627 SNP markers and 230 accessions. Each point represents a SNP, with the -log10 *p*-values plotted against the position on each chromosome. The red vertical non-dashed line indicates the positions of *ma1*. The red vertical dashed line indicates the position of the highly significant peak identified in the grain composition GWAS. The horizontal dashed line represents the genome-wide significance threshold at 5% FDR.

3.7 REFERENCES

1. FAO. *State of food insecurity in the world, 2013 the multiple dimensions of food security.* (Food and Agriculture Organization, 2013).

2. FAO. Sorghum and millets in human nutrition. http://www.fao.org/docrep/T0818E/T0818E04.htm. (1995).

3. Grusak, M. A. & DellaPenna, D. Improving the nutrient composition of plants to enhance human nutrition and health. *Annual Review of Plant Physiology and Plant Molecular Biology* **50,** 133–161 (1999).

4. Hopkins, C. G. Improvement in the chemical composition of the corn kernel. *J. Am. Chem. Soc.* **21,** 1039–1057 (1899).

5. Goldman, I. L., Rocheford, T. R. & Dudley, J. W. Quantitative trait loci influencing protein and starch concentration in the Illinois long term selection maize strains. *Theor. Appl. Genet.* **87,** 217–224 (1993).

6. Dudley, J. W. & Lambert, R. J. in *Plant Breeding Reviews* (ed. Janick, J.) 79–110 (John Wiley & Sons, Inc., 2003). at <http://onlinelibrary.wiley.com/doi/10.1002/9780470650240.ch5/summary>

7. Moose, S. P., Dudley, J. W. & Rocheford, T. R. Maize selection passes the century mark: a unique resource for 21st century genomics. *Trends in Plant Science* **9,** 358–364 (2004).

8. Gutierrez, L., Van Wuytswinkel, O., Castelain, M. & Bellini, C. Combined networks regulating seed maturation. *Trends in Plant Science* **12,** 294–300 (2007).

9. Baud, S., Dubreucq, B., Miquel, M., Rochat, C. & Lepiniec, L. Storage reserve accumulation in arabidopsis: metabolic and developmental control of seed filling. *The Arabidopsis Book* e0113 (2008). doi:10.1199/tab.0113

10. Vicente-Carbajosa, J. & Carbonero, P. Seed maturation: developing an intrusive phase to accomplish a quiescent state. *The International Journal of Developmental Biology* **49,** 645–651 (2005).

11. Mertz, E. T., Bates, L. S. & Nelson, O. E. Mutant gene that changes protein composition and increases lysine content of maize endosperm. *Science* **145,** 279–280 (1964).

12. Nelson, O. E., Mertz, E. T. & Bates, L. S. Second mutant gene affecting the amino acid pattern of maize endosperm proteins. *Science* **150,** 1469–1470 (1965).

13. Schmidt, R. J., Burr, F. A., Aukerman, M. J. & Burr, B. Maize regulatory gene opaque-2 encodes a protein with a ‘leucine-zipper’ motif that binds to zein DNA. *Proc Natl Acad Sci U S A* **87,** 46–50 (1990).

14. Coleman, C. E., Lopes, M. A., Gillikin, J. W., Boston, R. S. & Larkins, B. A. A defective signal peptide in the maize high-lysine mutant floury 2. *Proceedings of the National Academy of Sciences of the United States of America* **92,** 6828 (1995).

15. Poneleit, C. G. & Alexander, D. E. Inheritance of linoleic and oleic acids in maize. *Science* **147,** 1585–1586 (1965).

16. Mikkilineni, V. & Rocheford, T. R. Sequence variation and genomic organization of fatty acid desaturase-2 (fad2) and fatty acid desaturase-6 (fad6) cDNAs in maize. *Theor Appl Genet* **106,** 1326–1332 (2003).

17. Wassom, J. J., Mikkelineni, V., Bohn, M. O. & Rocheford, T. R. QTL for fatty acid composition of maize kernel oil in Illinois High Oil × B73 backcross-derived lines. *Crop Sci.* **48,** 69 (2008).

18. Chourey, P. S. & Nelson, O. E. The enzymatic deficiency conditioned by the shrunken-1 mutations in maize. *Biochem Genet* **14,** 1041–1055 (1976).

19. Shure, M., Wessler, S. & Fedoroff, N. Molecular identification and isolation of the Waxy locus in maize. *Cell* **35,** 225–233 (1983).

20. Wilson, L. M. *et al.* Dissection of maize kernel composition and starch production by candidate gene association. *Plant Cell* **16,** 2719–2733 (2004).

21. Karper, R. E. Inheritance of waxy endosperm in sorghum. **Journal of Heredity,** 257–262 (1933).

22. Lichtenwalner, R. E., Ellis, E. B. & Rooney, L. W. Effect of incremental dosages of the waxy gene of sorghum on digestibility. *J ANIM SCI* **46,** 1113–1119 (1978).

23. Martin, J. H. in *YearBook of Agriculture* 523 (USDA, 1936).

24. Boyer, C. D. & Liu, K.-C. Starch and water-soluble polysaccharides from sugary endosperm of sorghum. *Phytochemistry* **22,** 2513–2515 (1983).

25. Singh, R. & Axtell, J. D. High lysine mutant gene (hi) that improves protein quality and biological value of grain sorghum. *Crop Sci.* **13,** 535 (1973).

26. Sukumaran, S. *et al.* Association Mapping for grain quality in a diverse sorghum collection. *The Plant Genome Journal* **5,** 126 (2012).

27. Figueiredo, L. F. de A. *et al.* Variability of grain quality in sorghum: association with polymorphism in Sh2, Bt2, SssI, Ae1, Wx and O2. *Theor Appl Genet* **121,** 1171–1185 (2010).

28. Murray, S. C. *et al.* Genetic improvement of sorghum as a biofuel feedstock: I. QTL for stem sugar and grain nonstructural carbohydrates. *Crop Sci.* **48,** 2165 (2008).

29. Hamblin, M. T., Salas Fernandez, M. G., Tuinstra, M. R., Rooney, W. L. & Kresovich, S. Sequence variation at candidate loci in the starch metabolism pathway in sorghum: prospects for linkage disequilibrium mapping. *Crop Sci.* **47,** S–125

(2007).

30. Rami, J.-F. *et al.* Quantitative trait loci for grain quality, productivity, morphological and agronomical traits in sorghum (*Sorghum bicolor* L. Moench). *Theor Appl Genet* **97,** 605–616 (1998).

31. McIntyre, C. L., Drenth, J., Gonzalez, N., Henzell, R. G. & Jordan, D. R. Molecular characterization of the waxy locus in sorghum. *Genome* **51,** 524–533 (2008).

32. Morris, G. P. *et al.* Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 453–458 (2013).

33. Huang, X. *et al.* A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490,** 497–501 (2012).

34. Jiao, Y. *et al.* Genome-wide genetic changes during modern breeding of maize. *Nat Genet* **44,** 812–815 (2012).

35. Zhao, K. *et al.* Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. *Nat Commun* **2,** 467 (2011).

36. Rasmussen, S. K. & Shu, X. Quantification of amylose, amylopectin, and β-glucan in search for genes controlling the three major quality traits in barley by genome-wide association studies. *Front. Plant Sci.* **5,** 197 (2014).

37. Li, H. *et al.* Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat Genet* **45,** 43–50 (2013).

38. Cook, J. P. *et al.* Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. *Plant Physiol.* **158,** 824–834 (2012).

39. Shu, X., Backes, G. & Rasmussen, S. K. Genome-wide association study of resistant starch (RS) phenotypes in a barley variety collection. *J. Agric. Food Chem.* **60,** 10302–10311 (2012).

40. Seleka, T. B., Jackson, J. C., Batsetswe, L. & Kebakile, P. G. Small-scale milling and the feasibility of mandatory fortification of sorghum and maize flour in Botswana. *Development Southern Africa* **28,** 461–476 (2011).

41. Nestel, P., Bouis, H. E., Meenakshi, J. V. & Pfeiffer, W. Biofortification of staple food crops. *J. Nutr.* **136,** 1064–1067 (2006).

42. Horton, S. The Economics of Food Fortification. *J. Nutr.* **136,** 1068–1071 (2006).

43. Casa, A. M. *et al.* Community resources and strategies for association mapping in sorghum. *Crop Sci.* **48,** 30 (2008).

44. Harlan, J. R., Wet, D. & J, J. M. A simplified classification of cultivated sorghum. *Crop Science* **12,** 172–176 (1972).

45. USDA. GRIN National Genetic Resources Program. URL (http://www.ars-grin.gov) (2014-07-25). (2014).

46. Rhodes, D. H. *et al.* Genome-wide association study of grain polyphenol concentrations in global sorghum [*Sorghum bicolor* (L.) Moench] germplasm. *J. Agric. Food Chem.* (2014). doi:10.1021/jf503651t

47. Lipka, A. E. *et al.* GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28,** 2397–2399 (2012).

48. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38,** 203–208 (2006).

49. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42,** 355–360 (2010).

50. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57,** 289–300 (1995).

51. Davidson, R. M. *et al.* Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *The Plant Journal* **71,** 492–502 (2012).

52. Duodu, K. ., Taylor, J. R. ., Belton, P. . & Hamaker, B. . Factors affecting sorghum protein digestibility. *J. Cereal Sci.* **38,** 117–131 (2003).

53. Wang, E. *et al.* Control of rice grain-filling and yield by a gene with a potential signature of domestication. *Nat. Genet.* **40,** 1370–1374 (2008).

54. Holdsworth, M. J., Bentsink, L. & Soppe, W. J. J. Molecular networks regulating Arabidopsis seed maturation, after-ripening, dormancy and germination. *New Phytol.* **179,** 33–54 (2008).

55. Santos-Mendoza, M. *et al.* Deciphering gene regulatory networks that control seed development and maturation in Arabidopsis. *The Plant Journal* **54,** 608–620 (2008).

56. Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6,** e19379 (2011).

57. Murphy, R. L. *et al.* Coincident light and clock regulation of pseudoresponse regulator protein 37 (PRR37) controls photoperiodic flowering in sorghum. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 16469–16474 (2011).

58. Mace, E. S. & Jordan, D. R. Integrating sorghum whole genome sequence information with a compendium of sorghum QTL studies reveals uneven distribution of QTL and of gene-rich regions with significant implications for crop improvement. *Theor Appl Genet* **123,** 169–191 (2011).

59. Whan, A. *et al.* Engineering α-amylase levels in wheat grain suggests a highly sophisticated level of carbohydrate regulation during development. *J. Exp. Bot.*

eru299 (2014). doi:10.1093/jxb/eru299

60. Uauy, C., Distelfeld, A., Fahima, T., Blechl, A. & Dubcovsky, J. A NAC gene regulating senescence improves grain protein, zinc, and iron content in wheat. *Science* **314,** 1298–1301 (2006).

61. Thomas, H. & Ougham, H. The stay-green trait. *J. Exp. Bot.* **65,** 3889–3900 (2014).

62. Tuinstra, M. R., Grote, E. M., Goldsbrough, P. B. & Ejeta, G. Identification of quantitative trait loci associated with pre-flowering drought tolerance in sorghum. *Crop science (USA)* (1996).

63. Sanchez, A. C., Subudhi, P. K., Rosenow, D. T. & Nguyen, H. T. Mapping QTLs associated with drought resistance in sorghum (Sorghum bicolor L. Moench). *Plant Mol Biol* **48,** 713–726 (2002).

64. BORRELL, A. K., Jordan, D. R., Mullet, J. & Klein, P. Drought tolerant plants produced by modification of the stay-green stgx locus. (2013). at <http://www.google.com/patents/WO2013071366A1>

65. Harris, K. *et al.* Sorghum stay-green QTL individually reduce post-flowering drought-induced leaf senescence. *J. Exp. Bot.* **58,** 327–338 (2007).

66. Harn, C. *et al.* Isolation and characterization of the zSSIIa and zSSIIb starch synthase cDNA clones from maize endosperm. *Plant Mol. Biol.* **37,** 639–649 (1998).

67. Zhang, G. *et al.* Double repression of soluble starch synthase genes SSIIa and SSIIIa in rice (Oryza sativa L.) uncovers interactive effects on the physicochemical properties of starch. *Genome* **54,** 448–459 (2011).

68. Pouvreau, B. *et al.* Duplicate maize wrinkled1 transcription factors activate target genes involved in seed oil biosynthesis. *Plant Physiol.* **156,** 674–686 (2011).

69. Beta, T. & Corke, H. Genetic and environmental variation in sorghum starch properties. *Journal of Cereal Science* **34,** 261–268 (2001).

70. Saeed, M., Francis, C. A., Rajewski, J. F. & Maranville, J. W. Genotype environment interaction and stability analysis of protein and oil in grain sorghum. *Crop Science* **27,** 169 (1987).

71. Wu, X. *et al.* Effects of growing location and irrigation on attributes and ethanol yields of selected grain sorghums. *Cereal Chemistry* **85,** 495–501 (2008).

72. Liu, L. *et al.* Impact of deficit irrigation on sorghum physical and chemical properties and ethanol yield. *Transactions of the ASABE* **56,** 1541–1549 (2014).

73. Wu, X. *et al.* Evaluation of Nebraska waxy sorghum hybrids for ethanol production. *Cereal Chemistry Journal* **90,** 198–203 (2013).

CHAPTER 4

SORGHUM [*SORGHUM BICOLOR* (L.) MOENCH] GENOTYPE DETERMINES DEGREE OF ANTI-INFLAMMATORY PROPERTIES OF SORGHUM BRAN[3]

[3] Davina H. Rhodes, Stephen Kresovich. To be submitted to *Journal of Nutrition and Food Science.*

4.1 ABSTRACT

Inflammation is the underlying cause of many chronic diseases, including obesity, type 2 diabetes, cardiovascular disease, and cancer. Identifying foods with anti-inflammatory properties may help to prevent or attenuate damage caused by inflammation. Grain makes up the majority of the human diet, so identifying grain varieties with significant anti-inflammatory effects can aid in the selection of grains for a health-promoting diet. Sorghum, a major cereal crop grown worldwide, has been reported to have anti-inflammatory properties related to its polyphenol content. There are over 45,000 sorghum accessions (distinct varieties of plants) available through the USDA's National Plant Germplasm System, providing an enormous resource for screening the anti-inflammatory properties of the natural variation of sorghum polyphenols.

This study evaluated the anti-inflammatory effects of ethanol extracts from the bran of twenty sorghum accessions with comparable genetic backgrounds. Correlations were calculated between anti-inflammatory effects and total polyphenol, proanthocyanidin, and 3-deoxyanthocyanidin concentrations. Cell viability, tumor necrosis factor (TNF)-α production and interleukin (IL)-6 production were measured using lipopolysaccharide (LPS)-stimulated RAW 264.7 mouse macrophage cells. Using a subset of five sorghum extracts, nuclear transcription factor kappaB (NF-κB) phosphorylation was measured in LPS-stimulated RAW 264.7 cells.

The addition of varying concentrations of sorghum extracts, both with and without LPS stimulation, did not reduce viability of RAW 264.7 cells. Thirteen of the sorghum extracts significantly reduced TNF-α and/or IL-6 at varying extract

concentrations. One of the extracts significantly increased TNF-α and IL-6 at concentrations of 60 ug/mL. Two accessions had no effect on cytokine levels. NF-κB phosphorylation was significantly reduced by extracts at concentrations of 30 ug/mL and 15 ug/mL. Averaging results from all of the sorghum accessions, there was a negative correlation between IL-6 and 3-deoxyanthocyanidins at extract concentrations of 60 ug/mL. In contrast, there was a positive correlation between TNF-α and both total polyphenols and proanthocyanidins at concentrations of 60 ug/mL. Our results demonstrate that sorghum accessions differentially modulate inflammation, with many accessions reducing pro-inflammatory cytokines, possibly by decreasing phosphorylation of NF-κB. Additionally, we demonstrate that the RAW 264.7 model of inflammation is a good method for high throughput screening of anti-inflammatory effects of sorghum extracts.

4.2 INTRODUCTION

Grain makes up the majority of the American diet, contributing 24% of our daily energy.[1] Consumption of whole grain has been correlated with protective health effects related to several chronic inflammatory diseases, including obesity, type 2 diabetes, cancer, and cardiovascular disease.[2–7] However, the protective mechanisms involved in these beneficial effects are still unresolved. Inflammation is known to be the underlying cause of many chronic diseases [8,9] and identifying foods with anti-inflammatory properties may help to prevent or attenuate the damage caused by inflammation. There is a large body of research demonstrating the anti-inflammatory effects of a variety of fruits [10–12], but these foods are a small contribution to daily food intake compared to grain products. Therefore, understanding the anti-inflammatory effects of cereal grains can help

in the selection of foods for a health-promoting diet. Some studies suggest that many of the beneficial health effects of whole grain may be due to polyphenols in the bran.[13–15] Polyphenols are a large diverse group of phytochemicals found in abundance in fruits, vegetables, tea, chocolate, red wine, and coffee. Certain varieties of grains also contain polyphenols, including varieties of wheat, rice, maize, and sorghum.[14,16–18]

Based on worldwide production, sorghum is one of the world's major cereal crops.[19] Half of the sorghum produced is used for human food consumption, feeding millions of people in Asian and sub-Saharan Africa.[19] In the United States, it is used primarily as livestock feed, but it is also used in many specialty grain products and gluten-free food products.[20–23] Flavonoids, primarily 3-deoxyanthocyanidins and proanthocyanidins, are the major polyphenols found in sorghum.[24] The 3-deoxyanthocyanidins are not widely found in nature, and sorghum is their only known dietary source.[25–27] Proanthocyanidins are not commonly found in high concentrations in cereal crops, but many sorghum varieties are rich sources of this flavonoid.[28] Polyphenols are predominantly located in the outer seed coat (the bran) of the sorghum seed. The majority of research on sorghum polyphenol health benefits has been on its high antioxidant activity compared to commonly consumed fruits[29,30], but some studies have also suggested that sorghum grain may have anti-inflammatory activity.[31, 32]

Inflammation is a complex physiological response to harmful stimuli such as pathogens, damaged cells, or irritants. The mediators of inflammation are involved in defense and repair mechanisms, but in some instances dysregulation of their production can lead to chronic inflammation, which is implicated in the pathophysiology of most chronic diseases, including cardiovascular disease, cancer, obesity and type 2 diabetes.[33,

[34] A key feature of inflammation is the activation of inflammatory cells, especially monocytes and macrophages, which produce pro-inflammatory cytokines, including TNF-α and IL-6. The RAW 264.7 mouse macrophage cell line is commonly used to screen natural products for potential anti-inflammatory properties.[35,36–40] Lipopolysaccharide (LPS), the outer coat of Gram-negative bacteria, is applied to the RAW 264.7 cells to induce an array of inflammatory responses. Upon macrophage activation by LPS, cytoplasmic NF-κB is phosphorylated and translocates to the nucleus, where it binds to promoter and enhancer regions of target genes, inducing transcription of key mediators of inflammation, including IL-6 and TNF-α. The NF-κB signal transduction pathway plays a crucial role in inflammation, and excessive activation of the pathway can lead to chronic inflammation.[41,42] Although the RAW 264.7 inflammation model is a reductive one, it provides useful information about the potential health benefits of a test compound and is a good high-throughput screening method for anti-inflammatory effects of natural variation within a food plant species. This can act as a guide in the selection of a subset of varieties to use in more complex disease models.

In vitro, sorghum bran extracts, especially polyphenol-rich varieties, inhibit hyaluronidase, an enzyme that is increased in certain inflammatory diseases[32]; decrease TNF-α and IL-1β in LPS-challenged human peripheral blood mononuclear cells (PBMC) [31]; and reduce production of nitric oxide in RAW 264.7 cells.[43] *In vivo,* red sorghum grain reduces production of TNF-α when consumed by male Wistar rats on a high fat diet[44]; and sorghum extracts significantly reduce inflammatory molecules, including inducible nitric oxide (iNOS) and cyclooxygenase (COX)-2, in 12-O-tetradecanoylphorbol-13-acetate (TPA)-induced ear models of inflammation, and the anti-inflammatory activity

correlates with phenolic level and antioxidant level.[31,43] While there is evidence of benefits of sorghum polyphenols on human health, more studies are needed to characterize the physiological effects and mechanisms of action.

Some varieties of sorghum do not contain measurable amounts of polyphenols, while others contain high levels of polyphenols.[24,45] Most studies have only explored the health benefits of a small number of sorghum accessions (distinct varieties of plants), but over 45,000 sorghum accessions are available from the U.S. National Plant Germplasm System's Germplasm Resources Information Network (GRIN).[46] Utilizing accessions that are readily available from a crop gene bank allows for authentication of the accessions and reproducibility of the experiments. Using a large genetically diverse sorghum panel to explore the effects of natural variation of sorghum polyphenols on inflammation will help in discovering particularly beneficial varieties. Additionally, although several studies comparing health effects between sorghums with or without proanthocyanidins and 3-deoxyanthocyanidins have been conducted, none of them controlled for genetic background of the sorghums or utilized accessions that were readily available from crop gene banks.[31,32,44,47] Without adequate control of other genetic factors it may not be possible to attribute health effects to polyphenols *per se*. The goals of this study were to identify and compare the anti-inflammatory effects of twenty genetically similar sorghum varieties with contrasting grain flavonoid concentrations, and to gain a broader understanding of the diversity of anti-inflammatory effects available among sorghum accessions.

4.3 MATERIALS AND METHODS

4.3.1 Plant Materials

We selected 20 sorghum accessions from a panel of 381 sorghum accessions that we previously evaluated for flavonoid concentrations [45]. The panel primarily consisted of the Sorghum Association Panel (SAP)[48], which includes accessions from all major cultivated races and geographic centers of diversity in sub-Saharan Africa and Asia, as well as important breeding lines from the United States. Also included were 73 accessions selected based on the presence of proanthocyanidins using GRIN. Seeds for all the sorghum accessions came from GRIN and are readily available through GRIN. To select the subset of 20 accessions from the 381 accessions that had been grown, we identified accessions with high concentrations of proanthocyanidins and/or 3-deoxyanthocyanidins and used a kinship matrix to identify accessions with similar genetic background (high kinship value) but contrasting flavonoid content.

The grain samples used for this experiment have previously been described.[45] Briefly, the panel was planted in late April 2012 at Clemson University Pee Dee Research and Education Center in Florence, SC, in a twofold replicated complete randomized block design. Panicles were collected at physiological maturity between September and October, and mechanically threshed. Samples were phenotyped by near infrared spectroscopy (NIRS) as previously described.[45] Total phenol, proanthocyanidin, and 3-deoxyanthocyanidin data are expressed as mg gallic acid equivalent (GAE)/g, mg catechin equivalents (CE)/g, and absorbance (abs)/mL/g, respectively. Data are presented as the mean of the replicates.

4.3.2 Genomic Analysis

To select accessions with comparable genetic backgrounds, we used the genotypes of each accession to assess relatedness. Genotypes were available for the 381 accessions.[45,49] Based on 404,628 SNP markers, cryptic relatedness (kinship among the sorghum accessions unknown to the investigator)[50] between accessions was calculated in a kinship matrix in a unified mixed linear model[51] using the statistical genetics package Genome Association and Prediction Integrated Tool (GAPIT).[52]

4.3.3 Preparation of sorghum bran extracts

A tangential abrasive dehulling device (TADD; Venables Machine Works, Saskatoon, Canada) equipped with an 80-grit abrasive disk was used to remove the bran from the grain.[53] Bran was mixed with 50% ethanol (1g/mL) and placed on a shaker at room temperature for three hours. Samples were then centrifuged at 5000 rpm for 15 minutes and supernatant was poured through a 0.2 micromolar filter into a sterile container. Samples were refrigerated and protected from light until ready to use.

4.3.4 Cell Cultures

The mouse macrophage cell line RAW 264.7 (TIB-71 from American Type Culture Collection (ATCC)) was cultured on 100 mm culture dishes and maintained in Dulbecco's modified Eagle's medium (DMEM, ATCC), supplemented with 10% fetal bovine serum (ATCC) and 100 I.U./mL penicillin and 100 ug/mL streptomycin (ATCC) at 37 □ in a humidified incubator with 5% CO_2.

4.3.5 Cell Viability Assay

Cell viability was measured using the MTT Cell Proliferation Assay (R&D Systems), an indirect method of measuring metabolically active cells. RAW 264.7 cells were seeded in a 96-well plate (1 x 10^5 cells/well) and incubated for two hours to allow cells to recover and adhere to the cell culture plate. Cells were pretreated for one hour with sorghum extracts at concentrations of 125 ug/mL, 60 ug/mL, 30 ug/mL, and 15 ug/mL and then activated with LPS at 1 μg/mL or vehicle for an additional 18 hours. The MTT reagent was added to each well and cells were incubated for an additional 2 hours until purple dye was visible under the microscope. Detergent Reagent was added and the plates were left in the dark at room temperature for 4 hours. Absorbance was measured at 570 nm in a Synergy H1 Hybrid Multi-Mode Microplate Reader (BioTek). Results are expressed as the ratio of absorbance in extract treated cells versus untreated cells.

4.3.6 Cytokine assays

Cells were seeded in 12-well plates at 1 x 10^6 and incubated for 2 hours to allow time to recover and adhere to the substrate. Cells were pretreated for 1 hour with sorghum bran extracts at concentrations of 60 ug/mL, 30 ug/mL, and 15 ug/mL, or with a negative control (50% EtOH or sorghum extracts without LPS) and then stimulated with LPS at 1 μg/mL or vehicle for an additional 18 hours. Cell culture medium was collected and tested using TNF-α and IL-6 ELISA Ready-Set-Go! kits purchased from eBioscience. Assays were carried out according to kit instructions. Absorbance was measured at 450 nm with wavelength subtraction at 570 nm in a Synergy H1 Hybrid Multi-Mode Microplate Reader (BioTek). Results are expressed as the percent of cytokine level in

extract-treated cells versus LPS-only treated cells.

4.3.7 NF-κB assay

Phospho-RelA/NF-κB p65 (S536) Cell-Based fluorogenic ELISA kit was purchased from R&D Systems. Cells were seeded in 96-well plates at 1 x 10^6 and incubated for 2 hours to allow time to recover and adhere to the substrate. Cells were pretreated for 1 hour with sorghum bran extracts, at concentrations of 60 ug/mL, 30 ug/mL, and 15 ug/mL, and then stimulated with LPS at 1 μg/mL for 1 hour. Cells were fixed and permeabilized in the 96-well plate and the assay was carried out according to kit instructions. Using a Synergy H1 Hybrid Multi-Mode Microplate Reader (BioTek), fluorescence for phosphorylated NF-κB was measured with excitation at 540 nm and emission at 600 nm, and fluorescence for total NF-κB was measured with excitation at 360 nm and emission at 450 nm. Results were normalized by dividing the phosphorylated NF-κB fluorescence by the total NF-κB fluorescence. Results are expressed as the percent of phosphorylated NF-κB in extract-treated cells versus LPS-only treated cells.

4.3.8 Statistical Analysis

Differences were assessed using analysis of variance (ANOVA) followed by post hoc Tukey HSD test. Pearson's correlation coefficient was also used. Results are expressed as mean values ± standard deviation (SD). All calculations were performed using R.[54]

4.4 RESULTS

4.4.1 Selection of target sorghum accessions

Twenty sorghum accessions with varying polyphenol concentrations were chosen to investigate the anti-inflammatory properties of sorghum bran extract (Figure 4.1). The panel contained eleven proanthocyanidin-containing accessions (based on NIRS values greater than 10 mg CE/g or presence of a pigmented testa), seven 3-deoxyanthocyanidin-containing accessions (based on NIRS values greater than 50 abs/mL/g), and five low polyphenol accessions that did not contain either of the flavonoids (Table 4.1). Three of the accessions contained both flavonoids. Total polyphenol concentrations in the panel ranged from 0 to 24 GAE/g, proanthocyanidin concentrations ranged from 0 to 42 mg CE/g, and 3-deoxyanthocyanidin concentrations ranged from 0 to 110 abs/mL/g, and (Figure 4.2).

4.4.2 Sorghum extracts improve viability in LPS-stimulated RAW 264.7 cells

To investigate the anti-inflammatory effects of sorghum bran extracts, we used LPS to induce an inflammatory state in RAW 264.7 mouse macrophage cells. We first conducted an MTT assay to assess the effects of varying concentrations of sorghum bran extracts on cell toxicity. RAW 264.7 macrophages were pretreated with sorghum extracts or vehicle for one hour, followed by LPS or vehicle for 18 hours. Averaged over all extracts, cell viability was not significantly different for cells treated with extracts at all concentrations compared to cells treated with cell media vehicle (Figure 4.3A). Additionally, when cells were treated with both LPS and the extracts, cell viability was not significantly different for cells treated with extracts at all concentrations compared to

cells treated with LPS alone (Figure 4.3B).

4.4.3 Sorghum extracts differentially modulate IL-6 and TNF-α

We examined the effects of varying concentrations of sorghum extracts on the secretion of pro-inflammatory cytokines TNF-α and IL-6 in LPS-stimulated RAW 264.7 macrophage cells. We first tested the effects of sorghum extracts without the addition of LPS, and found that on average IL-6 and TNF-α were induced at extract concentrations of 125 ug/mL and above. Therefore, in subsequent experiments, we used extract concentrations of 15 ug/mL, 30 ug/mL, and 60 ug/mL. Next, we pretreated RAW 264.7 macrophages with the twenty sorghum bran extracts for one hour, followed by stimulation with LPS for 18 hours. There was a range of responses among the accessions (Figure 4.4A-B). Thirteen of the sorghum accessions significantly inhibited TNF-α and/or IL-6 at varying extract concentrations. One of the accessions (PI656038) significantly increased both IL-6 (P = 0.001) and TNF-α (P = 0.001) at extract concentrations of 60 ug/mL. Two accessions (PI221619, PI533991, and PI533957) did not significantly affect cytokine levels at any extract concentration. Averaged over all of the sorghum accessions, cells treated with 30 ug/mL of sorghum extract produced significantly less IL-6 ($P = 6 \times 10^{-4}$) and TNF-α (P = 0.002) than those treated with LPS alone (Figure 4.4C-D). In contrast, TNF-α was significantly increased in cells with extract concentrations of 60 ug/mL (P = 0.02) compared to cells treated with LPS alone (Figure 4.4D).

We hypothesized that the flavonoid composition of the sorghum extracts was influencing cytokine levels in LPS-stimulated RAW 264.7 cells, so Pearson's correlations

were calculated between concentrations of cytokines and flavonoids. There was a significant negative correlation between IL-6 levels and 3-deoxyanthocyanidins (-0.44, *P* = 0.05) when extract concentrations of 60 ug/mL were used. In contrast, there was a significant positive correlation between TNF-α levels and both total polyphenols (0.52, *P* = 0.02) and proanthocyanidins (0.60, *P* = 0.006) when extract concentrations of 60 ug/mL were used, and a significant positive correlation between TNF-α levels and proanthocyanidins when extract concentrations of 30 ug/mL were used (0.51, *P* = 0.02). Figure 4.4E-F, shows the effects of the sorghum extracts, grouped by their flavonoid content, on TNF-α and IL-6 secretions, but there were no significant differences in cytokine levels found between the flavonoid groups.

4.4.4 Sorghum extracts suppress NF-κB activation

To determine if sorghum extracts might be reducing IL-6 and TNF-α by suppressing NF-κB activation, we examined the effects of extracts from five sorghum accessions of varying flavonoid concentrations (Figure 4.5A-B) on NF-κB phosphorylation in LPS-stimulated RAW 264.7 macrophages. Averaged over all extracts, NF-κB phosphorylation was significantly decreased, compared to LPS alone, at extract concentrations of 30 ug/mL (*P* = 0.04) and 15 ug/mL (*P* = 0.04; Figure 4.6A). Among individual extracts, NF-κB phosphorylation was significantly decreased by PI221619, a high proanthocyanidin accession, at a concentration of 30 ug/mL (*P* = 0.02), and PI656079, a high 3-deoxyanthocyanidin accession, at a concentration of 15 ug/mL (*P* = 0.05; Figure 4.6B).

4.5 DISCUSSION

In this study, sorghum bran extracts from several different sorghum genotypes attenuated cytokine production in RAW 264.7 macrophage cells. IL-6 and TNF-α were, on average, significantly reduced at a sorghum extract concentration of 30 ug/mL. Among the individual accessions tested, there was large variation in cytokine inhibition, but the majority of extracts exhibited anti-inflammatory properties. NF-κB phophorylation was significantly decreased when LPS-activated RAW 264.7 cells were pretreated with sorghum bran extracts at concentrations of 30 ug/mL and 15 ug/mL. Sorghum 3-deoxyanthocyanidin concentrations had a significant negative correlation with IL-6 levels when extract concentrations of 60 ug/mL were used. In contrast, proanthocyanidin and total polyphenol concentrations were positively correlated with TNF-α at extract concentrations of 60 ug/mL. This suggests that sorghum proanthocyanidins can be pro-inflammatory at higher concentrations. This is in agreement with a study that found that high concentrations of a proanthocyanidin-containing sorghum slightly induced COX-2 production in PBMC cells [31]. Taken all together, this data suggests that sorghum extract concentrations of 30 ug/mL generally have the most inhibitory effect on inflammation, but the large variation between accessions indicate that accessions need to be tested individually in order to determine the most effective concentration.

Given that most of the sorghum accessions possessed some degree of anti-inflammatory properties, including the low polyphenol sorghums, it is likely that constituents in the bran other than the flavonoids that were measured are also contributing to the anti-inflammatory effects of sorghum extracts. In a recent study,

phenolic acid derivatives isolated from sorghum grains decreased LPS-stimulated NO, iNOS, and COX-2 in RAW 264.7 cells[55]. Other phenolic compounds have been identified in sorghum bran, including flavones, flavanones, phlobaphenes, and anthocyanins[24,56], which may be contributing to the anti-inflammatory effects demonstrated in this study. It is interesting that the proanthocyanidin-containing accession and the 3-deoxyanthocyanidin-containing accession were the only two accessions out of the five that were tested that significantly reduced NF-κB phosphorylation, despite the fact that other accessions decreased IL-6 and TNF- α to a greater degree. It may be that sorghum proanthocyanidins and 3-deoxyanthocyanidins are able to attenuate inflammation through this pathway, while samples containing other types of polyphenols attenuate inflammation through different signaling pathways. Other pathways found to be inhibited by flavonoids include the signal transducer and activator of transcription (STAT)-1, activated protein (AP)-1, and mitogen-activated protein kinases (MAPK). High-performance liquid chromatography and mass spectrometry (HPLC-MS) is currently underway to identify the precise polyphenol content of each of the twenty sorghum extracts, which may provide more information as to what compounds are responsible for the anti-inflammatory effects. If there is a particular polyphenol identified that appears to be responsible for the greatest effect, it would be interesting to phenotype the entire SAP (~400 sorghum accessions) for this polyphenol to investigate its natural variation.

Though there is some debate as to the biological relevance of *in vitro* anti-inflammatory studies, such as the common RAW 264.7 model, many studies have found similar effects in animal models.[57–60] The negative correlation between 3-deoxyanthocyanidin concentrations in sorghum extracts and IL-6 levels in LPS-

stimulated cells makes 3-deoxyanthocyanidin-containing sorghum accessions attractive candidates for *in vivo* follow up studies. Questions to be addressed for these and proanthocyanidin-containing accessions, in addition to their anti-inflammatory effects, are degree of intestinal absorption and pre-and post-absorption modifications. Little is known about absorption of 3-deoxyanthocyanidins, but in proanthocyanidins, degree of polymerization highly influences absorption. Small proanthocyanidin compounds are absorbed in the small intestine, while large ones pass through the small intestine into the large intestine where they are catabolized by intestinal bacteria before they are absorbed.[61] For this reason, it has been suggested that health benefits derived from proanthocyanidins may be largely due to their effects on intestinal bacteria.[62] Therefore, the proanthocyanidin-containing sorghum accessions that had anti-inflammatory effects in this study are good candidates for in *vivo* follow up studies in disease models such as ulcerative colitis. In fact, several studies have found that proanthocyanidins from grape seeds attenuate inflammation in colitis animal models by modulating NF-κB pathways.[63,63]

The sorghum panel was planted in two independent field blocks. It is interesting to note that there was a block effect between the sorghum replicates. IL-6, TNF-α, and NF-κB were significantly different between the replicates. One possible explanation is that there may have been some differences in growing environment between the blocks. These were field-grown samples, so there was weathering (i.e. fungus on the surface of the grains) that could contribute to variation of anti-inflammatory effects. Another possibility is that the duplicates may have inadvertently been treated differently during preparation of extracts, leading to differences in composition of the final extracts. This

difference between the duplicates can be investigated further through repeat experiments, and through greenhouse experiments to test environmental effects.

This study provides evidence of sorghum grain anti-inflammatory activity through modulation of IL-6, TNF-α, and NF-κB, which was partly related to flavonoid content. Additionally, it shows that sorghum bran extracts possess anti-inflammatory properties that vary by genotype, demonstrating the importance of exploring genetic diversity within a crop to discover its full anti-inflammatory potential.

4.6 TABLES

Table 4.1 Polyphenol concentrations and categories for 20 sorghum accessions[a]

Taxa	Total Phenols (GAE/g)	PAs (CE/g)	3-DAs (abs/mL/g	Flavonoid Category[b,c,d,e]
PI221610	23.69 ± 3.56	42.30 ± 9.91	14.57 ± 10.8	PA
PI221619	11.49 ± 0.08	28.59 ± 0.05	0	PA
PI221723	18.06 ± 0.45	40.19 ± 2.42	59.42 ± 16.4	PA + 3DA
PI229830	16.54 ± 2.65	29.69 ± 10.42	11.18 ± 1.64	PA
PI229838	9.31 ± 0.96	7.19 ± 4.94	0	low
PI229875	7.31 ± 0.61	11.12 ± 0.60	19.86 ± 0.49	PA
PI297139	22.60 ± 1.99	41.66 ± 6.38	110.73 ± 29.9	PA + 3DA
PI329440	0	0	19.36 ± 15.3	Low
PI35038	18.79 ± 0.64	30.78 ± 4.84	11.35 ± 9.37	PA
PI533792	0	0	41.71 ± 28.66	3DA
PI533902	9.69 ± 0.64	5.89 ± 1.68	77.59 ± 47.3	PA + 3DA
PI533957	16.48 ± 0.99	24.91 ± 0.04	80.63 ± 37.6	PA + 3DA
PI533991	2.54 ± 1.49	0	9.51 ± 4.32	low
PI542718	15.93 ± 0.54	40.94 ± 0.15	34.39 ± 9.91	PA
PI561072	2.53 ± 2.23	0.50 ± 0.62	19.35 ± 3.03	low
PI576426	2.48 ± 0.78	0	72.69 ± 5.58	3DA
PI655978	4.99 ± 1.36	2.42 ± 3.16	95.20 ± 80.80	3DA
PI656007	1.95 ± 1.69	2.80 ± 1.58	3.98 ± 5.59	low
PI656038	13.36 ± 1.12	28.84 ± 1.94	0.72 ± 2.29	PA
PI656079	4.25 ± 1.63	1.67 ± 2.69	42.97 ± 52.76	3DA

[a] Concentrations are the mean of NIR values on accession grown in duplicate plots ± SD

[b] If one of the replicates had a 3DA NIRS value >50 abs/mL/g, then it was designated as a "3DA" flavonoid category, even if the average of the replicates was <50 abs/mL/g.

[c] If the accession contained a pigmented testa, then it was designated as a "PA" flavonoid category, even if the average of the replicates was <10 CE/g.

[d] "low" indicates that proanthocyanidins and 3-deoxyanthocyanidins were not detected by NIR

[e] Proanthocyanidin (PA); 3-deoxyanthocyanidin (3DA)

4.7 FIGURES

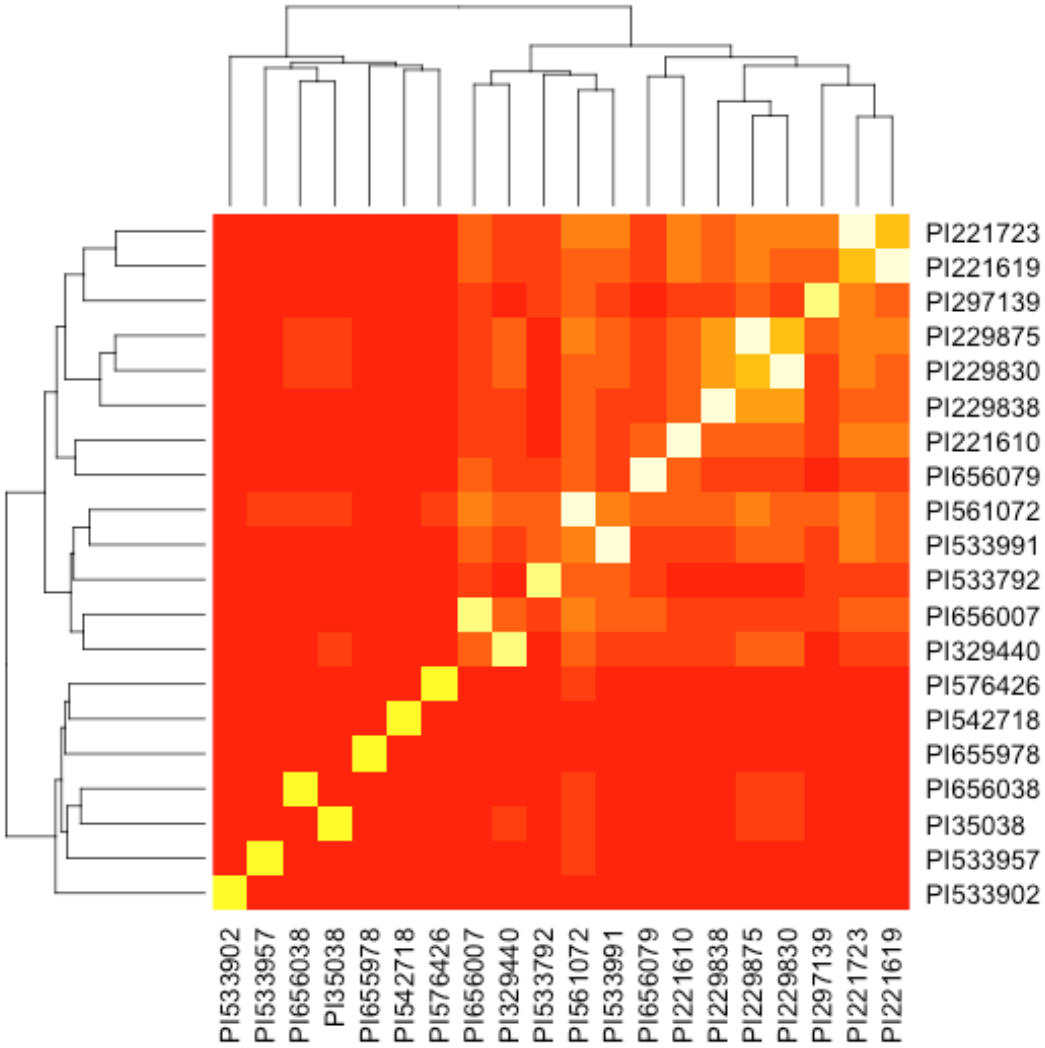


Figure 4.1 Heatmap and dendrogram of hierarchical clustering showing the estimated kinship among 20 sorghum accessions. Based on 404,628 SNP markers, cryptic relatedness between accessions was calculated in a kinship matrix using a unified mixed linear model in GAPIT.

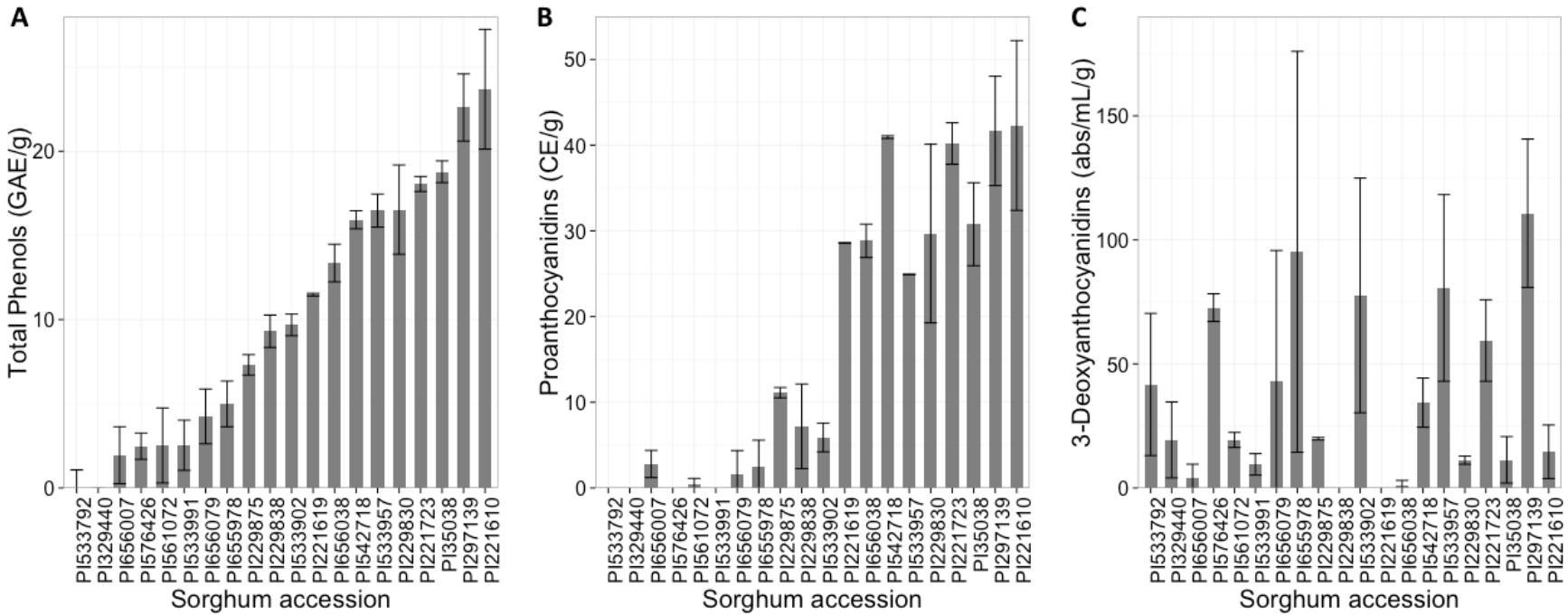


Figure 4.2 Polyphenol concentrations in the grain of 20 sorghum accessions. NIRS estimates of (A) total phenol concentrations (GAE/g), (B) proanthocyanidin concentrations (CE/g), and (C) 3-deoxyanthocyanidin concentrations (abs/mL/g), ordered from lowest total phenols to highest. Error bars represent the mean of accession grown in duplicate plots ± SD.

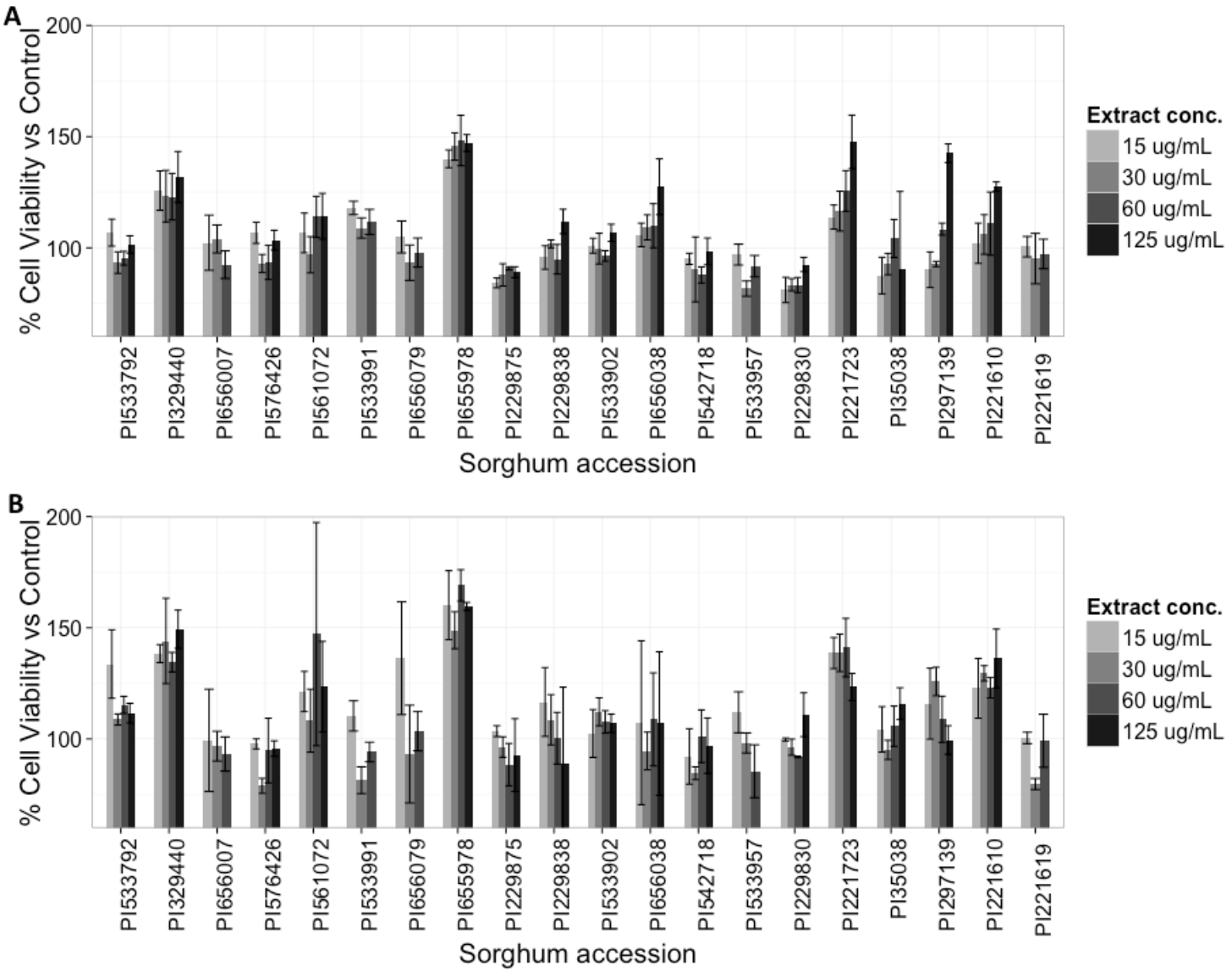
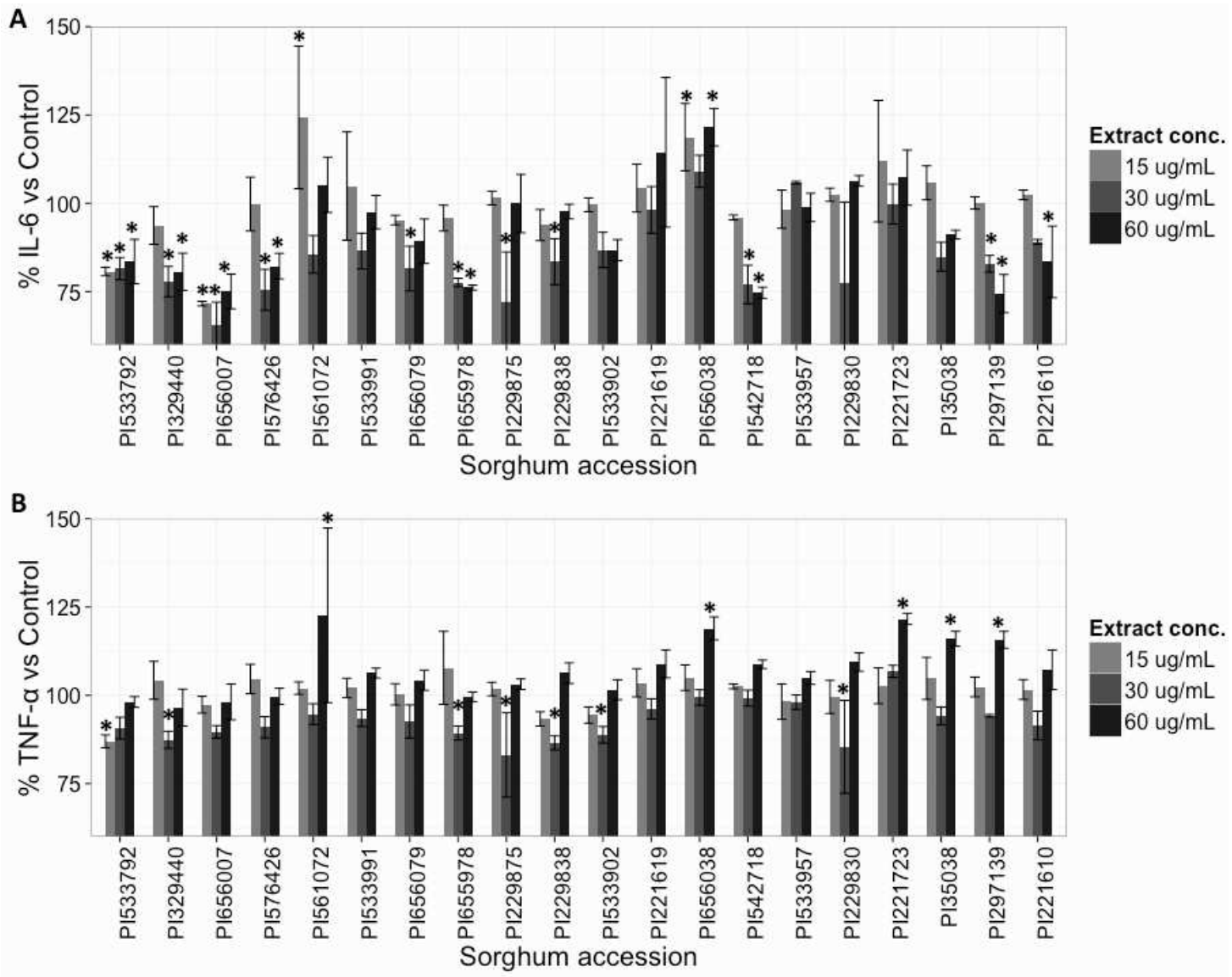

A
% Cell Viability vs Control
200
150
100
PI533792
PI329440
PI656007
PI576426
PI561072
PI533991
PI656079
PI655978
PI229875
PI229838
PI533902
PI656038
PI542718
PI533957
PI229830
PI221723
PI35038
PI297139
PI221610
PI221619
Sorghum accession
Extract conc.
15 ug/mL
30 ug/mL
60 ug/mL
125 ug/mL
B
% Cell Viability vs Control
200
150
100
PI533792
PI329440
PI656007
PI576426
PI561072
PI533991
PI656079
PI655978
PI229875
PI229838
PI533902
PI656038
PI542718
PI533957
PI229830
PI221723
PI35038
PI297139
PI221610
PI221619
Sorghum accession
Extract conc.
15 ug/mL
30 ug/mL
60 ug/mL
125 ug/mL

Figure 4.3 MTT cell viability assays of RAW 264.7 cells treated with sorghum bran extracts. Cell were seeded in 96-well plates, incubated for two hours, and then treated with sorghum extracts at concentrations of 125 ug/mL, 60 ug/mL, 30 ug/mL, and 15 ug/mL. Error bars represent the means of triplicate experiments ± SD. (A) Cells incubated with sorghum extracts for 18 hours, with results expressed as percent of absorbance in extract-treated cells versus untreated cells. (B) Cells pretreated with sorghum extracts for 1 hour and then stimulated with LPS for an additional 18 hours, with results expressed as percent of absorbance in extract-treated cells versus LPS-only treated cells. Accessions are ordered from lowest total phenols to highest.


A
% IL-6 vs Control
150
125
100
75
PI533792
PI329440
PI656007
PI576426
PI561072
PI533991
PI656079
PI655978
PI229875
PI229838
PI533902
PI221619
PI656038
PI542718
PI533957
PI229830
PI221723
PI35038
PI297139
PI221610
Sorghum accession
Extract conc.
15 ug/mL
30 ug/mL
60 ug/mL
B
% TNF-α vs Control
150
125
100
75
PI533792
PI329440
PI656007
PI576426
PI561072
PI533991
PI656079
PI655978
PI229875
PI229838
PI533902
PI221619
PI656038
PI542718
PI533957
PI229830
PI221723
PI35038
PI297139
PI221610
Sorghum accession
Extract conc.
15 ug/mL
30 ug/mL
60 ug/mL

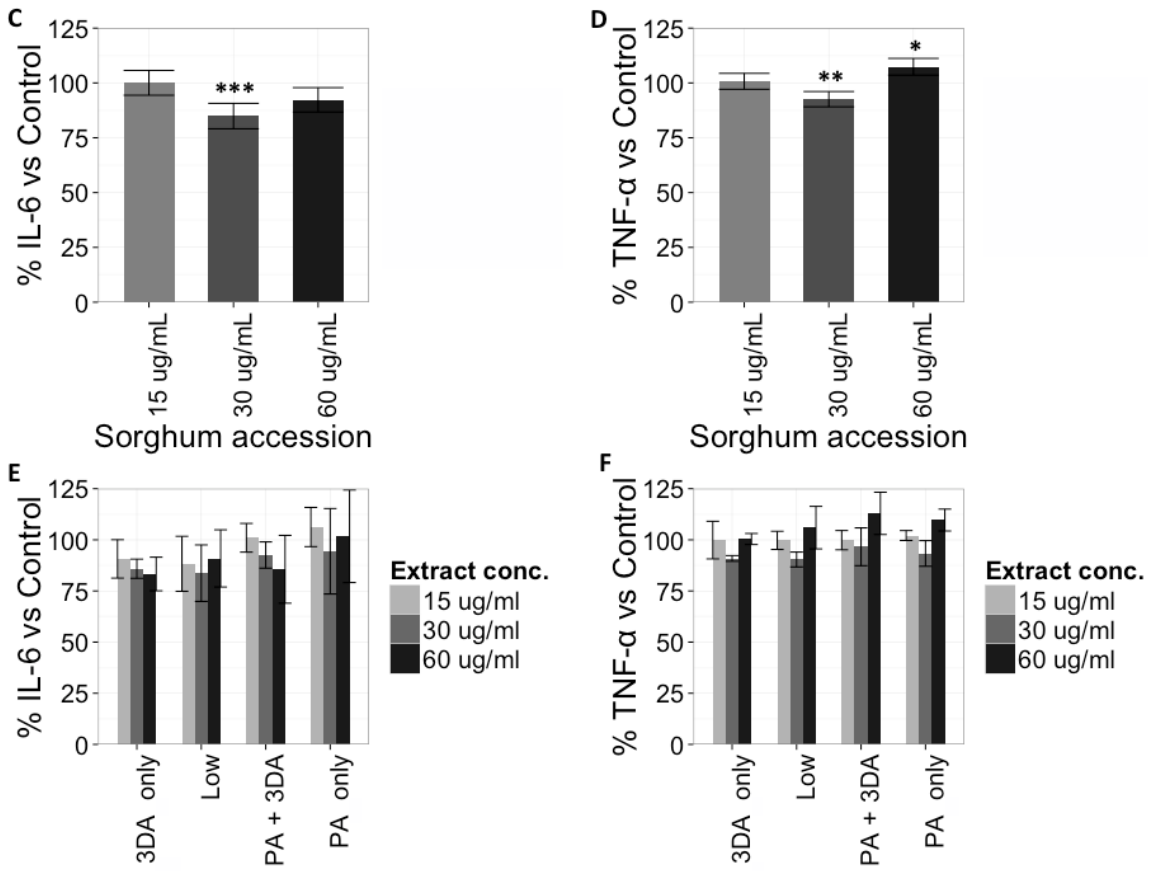


Figure 4.4 Sorghum bran extracts differentially modulate TNF-α and IL-6 production in RAW 264.7 cells. Cells were seeded in 12-well plates, incubated for 2 hours, and pretreated with sorghum bran extracts for 1 hour before LPS stimulation (1 μg/mL) for 18 hours. (A) IL-6 and (B) TNF-α response for each of the twenty sorghum accessions (* = P-values < 0.05). (C) Average IL-6 and (D) TNF-α response with varying concentrations of extracts, (E) Average IL-6 and (F) TNF-α response for polyphenol group. Results expressed as ratio of extract treated cells versus LPS-only treated cells. Error bars represent the means of triplicate experiments ± SD. Accessions are ordered from lowest total phenols to highest.

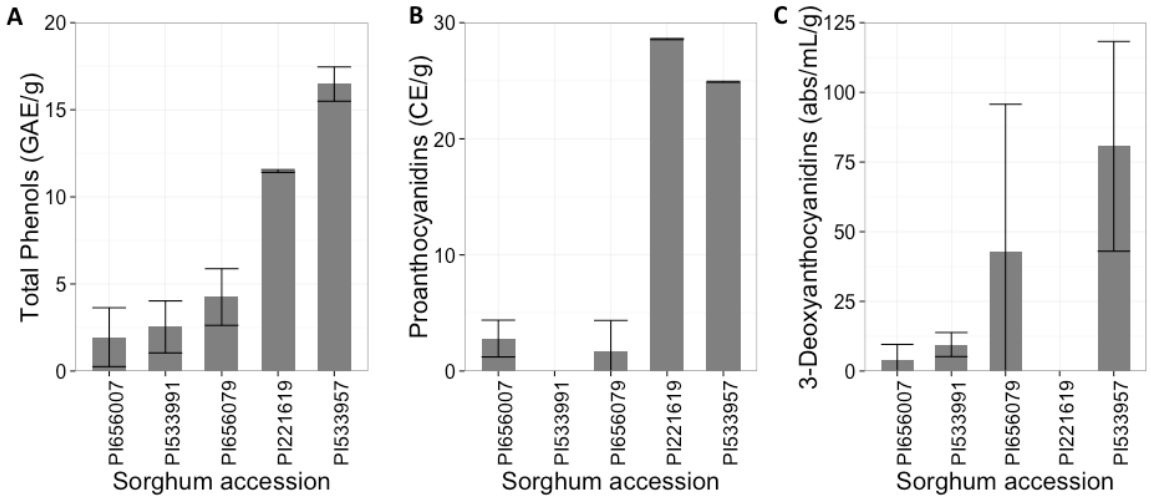


Figure 4.5 Polyphenol concentrations in the grain of five sorghum accessions. NIRS estimates of (A) total polyphenol concentrations (GAE/g), (B) proanthocyanidin concentrations (CE/g), and (C) 3-deoxyanthocyanidin concentrations (abs/mL/g).

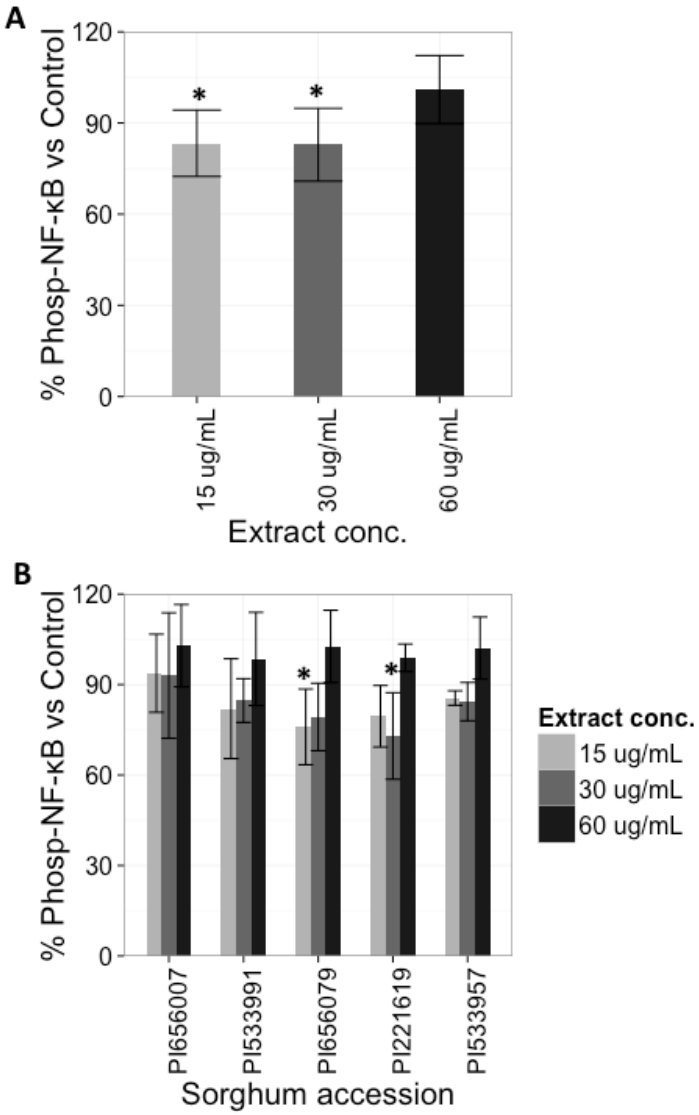


Figure 4.6 Sorghum bran extracts reduce NF-κB activation in RAW 264.7 cells. Cells were seeded in 96-well plates, incubated for 2 hours, and pretreated with sorghum bran extracts for 1 hour before LPS stimulation (1 μg/mL) for 1 hour. (A) Average NF-κB response with varying concentrations of extracts, (B) NF-κB response for each of the five sorghum accessions (* = P-values < 0.05). Results expressed as percentage of extract treated cells versus LPS-only treated cells. Error bars represent the means of triplicate experiments ± SD.

4.8 REFERENCES

1. US Department of Agriculture. *Nutrient Content of the U.S. Food Supply, 1909-2010. Center for Nutrition Policy and Promotion. http://www.cnpp.usda.gov/USFoodSupply-1909-2010.htm*. (2011).

2. McKeown, N. M., Meigs, J. B., Liu, S., Wilson, P. W. & Jacques, P. F. Whole-grain intake is favorably associated with metabolic risk factors for type 2 diabetes and cardiovascular disease in the Framingham Offspring Study. *Am J Clin Nutr* 76, 390–398 (2002).

3. Fung, T. T. *et al.* Whole-grain intake and the risk of type 2 diabetes: a prospective study in men. *Am J Clin Nutr* **76,** 535–540 (2002).

4. Liu, S. *et al.* Relation between changes in intakes of dietary fiber and grain products and changes in weight and development of obesity among middle-aged women. *Am J Clin Nutr* **78,** 920–927 (2003).

5. He, M., Dam, R. M. van, Rimm, E., Hu, F. B. & Qi, L. Whole-grain, cereal fiber, bran, and germ intake and the risks of all-cause and cardiovascular disease–specific mortality among women with type 2 diabetes mellitus. *Circulation* **121,** 2162–2168 (2010).

6. Aune, D. *et al.* Dietary fibre, whole grains, and risk of colorectal cancer: systematic review and dose-response meta-analysis of prospective studies. *BMJ* **343,** d6617–d6617 (2011).

7. Cho, S. S., Qi, L., Fahey, G. C. & Klurfeld, D. M. Consumption of cereal fiber, mixtures of whole grains and bran, and whole grains and risk reduction in type 2 diabetes, obesity, and cardiovascular disease. *Am J Clin Nutr* ajcn.067629 (2013).

8. Couzin-Frankel, J. Inflammation bares a dark side. *Science* **330,** 1621–1621 (2010).

9. Manabe, I. Chronic inflammation links cardiovascular, metabolic and renal diseases. *Circulation Journal* **75,** 2739–2748 (2011).

10. Jensen, G. S. *et al.* In vitro and in vivo antioxidant and anti-inflammatory capacities of an antioxidant-rich fruit and berry juice blend. Results of a pilot and randomized, double-blinded, placebo-controlled, crossover study. *J. Agric. Food Chem.* **56,** 8326–8333 (2008).

11. Basu, A., Rhone, M. & Lyons, T. J. Berries: emerging impact on cardiovascular health. *Nutrition Reviews* **68,** 168–177 (2010).

12. Szajdek, A. & Borowska, E. J. Bioactive compounds and health-promoting properties of berry fruits: A Review. *Plant Foods Hum Nutr* **63,** 147–156 (2008).

13. Fardet, A. New hypotheses for the health-protective mechanisms of whole-grain cereals: what is beyond fibre? *Nutrition Research Reviews* **23,** 65–134 (2010).

14. Liu, R. H. Whole grain phytochemicals and health. *Journal of Cereal Science* **46,** 207–219 (2007).

15. Borneo, R. & León, A. E. Whole grain cereals: functional components and health benefits. *Food & Function* **3,** 110 (2012).

16. Winkel-Shirley, B. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol.* **126,** 485–493 (2001).

17. Himi, E. & Noda, K. Red grain colour gene (R) of wheat is a Myb-type transcription factor. *Euphytica* **143,** 239–242 (2005).

18. Abdel-Aal, E.-S. M., Young, J. C. & Rabalski, I. Anthocyanin composition in black, blue, pink, purple, and red cereal grains. *J. Agric. Food Chem.* **54,** 4696–4704 (2006).

19. FAO. Sorghum and millets in human nutrition. http://www.fao.org/docrep/T0818E/T0818E04.htm. (1995)

20. Janzen, Edward L., W. W. W. Cooperative marketing in specialty grains and identity preserved grain markets. *North Dakota State University, Department of Agribusiness and Applied Economics, Agribusiness & Applied Economics Report* (2002).

21. Taylor, J. R. N., Schober, T. J. & Bean, S. R. Novel food and non-food uses for sorghum and millets. *J. Cereal Sci.* **44,** 252–271 (2006).

22. US Department of Agriculture, Center for Nutrition Policy and Promotion. *Trends in Dietary Fiber in the U.S. Food Supply; Sales of Grain Products. CNPP Fact Sheet No.2* at http://www.cnpp.usda.gov/Publications/FoodSupply/FiberFactSheet.pdf. (2007).

23. Cureton, P. & Fasano, A. in *Gluten-Free Food Science and Technology* (ed. Gallagher, E.) 1–15 (Wiley-Blackwell, 2009). at http://onlinelibrary.wiley.com/doi/10.1002/9781444316209.ch1/summary

24. Awika, J. M. & Rooney, L. W. Sorghum phytochemicals and their potential impact on human health. *Phytochemistry* **65,** 1199–1221 (2004).

25. Winefield, C. S. *et al.* Investigation of the biosynthesis of 3-deoxyanthocyanins in *Sinningia cardinalis*. *Physiol. Plant.* **124,** 419–430 (2005).

26. Sharma, M. *et al.* Expression of flavonoid 3'-hydroxylase is controlled by P1, the regulator of 3-deoxyflavonoid biosynthesis in maize. *BMC Plant Biol.* **12,** 196 (2012).

27. Malathi, P. *et al.* Differential accumulation of 3-deoxy anthocyanidin phytoalexins in sugarcane varieties varying in red rot resistance in response to *Colletotrichum falcatum* infection. *Sugar Tech* **10,** 154–157 (2008).

28. Gu, L. *et al.* Concentrations of proanthocyanidins in common foods and estimations of normal consumption. *J. Nutr.* **134,** 613–617 (2004).

29. Awika, J. M., Rooney, L. W., Wu, X., Prior, R. L. & Cisneros-Zevallos, L. Screening methods to measure antioxidant activity of sorghum (*Sorghum bicolor*) and sorghum

products. *J. Agric. Food Chem.* **51,** 6657–6662 (2003).

30. Carbonneau, M.-A. *et al.* Antioxidant properties of 3-deoxyanthocyanidins and polyphenolic extracts from côte d'ivoire's red and white sorghums assessed by OrAC and in vitro LDL oxidisability tests. *Food Chemistry* **145,** 701–709 (2014).

31. Burdette, A. *et al.* Anti-inflammatory activity of select sorghum (*Sorghum bicolor*) Brans. *Journal of Medicinal Food* **13,** 879–887 (2010).

32. Bralley, E., Greenspan, P., Hargrove, J. L. & Hartle, D. K. Inhibition of hyaluronidase activity by select sorghum brans. *J Med Food* **11,** 307–312 (2008).

33. Park, E. J. *et al.* Dietary and genetic obesity promote liver inflammation and tumorigenesis by enhancing IL-6 and tnf expression. *Cell* **140,** 197–208 (2010).

34. Festa, A. *et al.* Chronic subclinical inflammation as part of the Insulin Resistance Syndrome The Insulin Resistance Atherosclerosis Study (IRAS). *Circulation* **102,** 42–47 (2000).

35. Liang, Y.-C. *et al.* Suppression of inducible cyclooxygenase and inducible nitric oxide synthase by apigenin and related flavonoids in mouse macrophages. *Carcinogenesis* **20,** 1945–1952 (1999).

36. Wadsworth, T. L. & Koop, D. R. Effects of the wine polyphenolics quercetin and resveratrol on pro-inflammatory cytokine expression in RAW 264.7 macrophages. *Biochemical Pharmacology* **57,** 941–949 (1999).

37. Wang, J. & Mazza, G. Inhibitory effects of anthocyanins and other phenolic compounds on nitric oxide production in LPS/IFN-γ-activated RAW 264.7 macrophages. *J. Agric. Food Chem.* **50,** 850–857 (2002).

38. Terra, X. *et al.* Grape-seed procyanidins act as antiinflammatory agents in endotoxin-stimulated RAW 264.7 macrophages by inhibiting NFkB signaling pathway. *J. Agric. Food Chem.* **55,** 4357–4365 (2007).

39. Youn, C. K. *et al.* Silibinin inhibits LPS-stimulated macrophage activation by blocking p38 MAPK in RAW 264.7 cells. *Biomol Ther (Seoul)* **21,** 258–263 (2013).

40. Esposito, D., Chen, A., Grace, M. H., Komarnytsky, S. & Lila, M. A. Inhibitory effects of wild blueberry anthocyanins and other flavonoids on biomarkers of acute and chronic inflammation in vitro. *J. Agric. Food Chem.* (2014).

41. Barnes, P. J. & Karin, M. Nuclear Factor-κB — A pivotal transcription factor in chronic inflammatory diseases. *New England Journal of Medicine* **336,** 1066–1071 (1997).

42. Tak, P. P. & Firestein, G. S. NF-κB: a key role in inflammatory diseases. *Journal of Clinical Investigation* **107,** 7–11 (2001).

43. Shim, T.-J., Kim, T. M., Jang, K. C., Ko, J.-Y. & Kim, D. J. Toxicological evaluation

and anti-inflammatory activity of a golden gelatinous sorghum bran extract. *Biosci. Biotechnol. Biochem.* **77,** 697–705 (2013).

44. Moraes, É. A. *et al.* Sorghum genotype may reduce low-grade inflammatory response and oxidative stress and maintains jejunum morphology of rats fed a hyperlipidic diet. *Food Research International* **49,** 553–559 (2012).

45. Rhodes, D. H. *et al.* Genome-wide association study of grain polyphenol concentrations in global sorghum [*Sorghum bicolor* (L.) Moench] germplasm. *J. Agric. Food Chem.* (2014).

46. USDA. GRIN National Genetic Resources Program. URL (http://www.ars-grin.gov) (2014-07-25). (2014).

47. Hargrove, J. L., Greenspan, P., Hartle, D. K. & Dowd, C. Inhibition of aromatase and α-amylase by flavonoids and proanthocyanidins from Sorghum bicolor bran extracts. *J Med Food* **14,** 799–807 (2011).

48. Casa, A. M. *et al.* Community resources and strategies for association mapping in sorghum. *Crop Sci.* **48,** 30 (2008).

49. Morris, G. P. *et al.* Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 453–458 (2013).

50. Voight, B. F. & Pritchard, J. K. Confounding from cryptic relatedness in case-control association studies. *PLoS Genet.* **1,** e32 (2005).

51. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38,** 203–208 (2006).

52. Lipka, A. E. *et al.* GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28,** 2397–2399 (2012).

53. Awika, J. M., McDonough, C. M. & Rooney, L. W. Decorticating sorghum to concentrate healthy phytochemicals. *J. Agric. Food Chem.* **53,** 6230–6234 (2005).

54. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/., 2013).

55. Nguyen, P.-H., Lee, J. H., Min, B. S. & Woo, M. H. Isolation of benzoic and cinnamic acid derivatives from the grains of *Sorghum bicolor* and their inhibition of lipopolysaccharide-induced nitric oxide production in RAW 264.7 cells. *Food Chemistry* (2014).

56. Boddu, J., Svabek, C., Ibraheem, F., Jones, A. D. & Chopra, S. Characterization of a deletion allele of a sorghum Myb gene *yellow seed1* showing loss of 3-deoxyflavonoids. *Plant Science* **169,** 542–552 (2005).

57. Hasko, G. *et al.* Adenosine receptor agonists differentially regulate IL-10, TNF-alpha, and nitric oxide production in RAW 264.7 macrophages and in endotoxemic mice. *J Immunol* **157,** 4634–4640 (1996).

58. Wang, H. *et al.* HMG-1 as a late mediator of endotoxin lethality in mice. *Science* **285,** 248–251 (1999).

59. Mezei, O. *et al.* Soy isoflavones exert antidiabetic and hypolipidemic effects through the PPAR pathways in obese Zucker rats and murine RAW 264.7 cells. *J. Nutr.* **133,** 1238–1243 (2003).

60. Xu, G.-L. *et al.* Preventive effect of crocin in inflamed animals and in LPS-challenged RAW 264.7 cells. *J. Agric. Food Chem.* **57,** 8325–8330 (2009).

61. Deprez, S., Mila, I., Huneau, J. F., Tome, D. & Scalbert, A. Transport of proanthocyanidin dimer, trimer, and polymer across monolayers of human intestinal epithelial Caco-2 cells. *Antioxid. Redox Signaling* **3,** 957–967 (2001).

62. Ou, K. & Gu, L. Absorption and metabolism of proanthocyanidins. *Journal of Functional Foods* **7,** 43–53 (2014).

63. Li, X. *et al.* Proanthocyanidins from grape seeds modulate the NF-κB signal transduction pathways in rats with TNBS-induced ulcerative colitis. *Molecules* **16,** 6721–6731 (2011).

CHAPTER 5

CONCLUSION

The work presented here provides new insights into the diversity, genetics, and anti-inflammatory properties of sorghum nutrients that are important to human health. It provides a survey of flavonoid (chapter 2) and protein, fat, and starch (chapter 3) diversity in a large global panel of sorghum; identifies QTL and candidate genes for underlying controls of these nutrients (chapters 2 and 3); and demonstrates that a larger variety of sorghum accessions than previously thought have anti-inflammatory properties (chapter 4).

This project also sought to address a broader question on how to navigate a large germplasm collection in order to investigate a trait of interest. To answer questions about the anti-inflammatory effects of sorghum grain in specific disease states, it is only feasible to test two or three accessions in a single *in vivo* experiment. How then do we decide which two to three accessions out of over 45,000 to use? Taking advantage of tools designed for high throughput phenotyping and trait mapping (e.g., GBS, NIRS, RAW 264.7 inflammation model), this daunting task becomes potentially feasible. In the work presented here, hundreds of sorghum accessions were screened relatively quickly and cost-effectively, providing one of the largest surveys of natural variation of grain composition traits and flavonoids in sorghum, the first GWAS of quantitative natural variation of grain composition traits and flavonoids in sorghum, and permitted targeted studies on anti-inflammatory properties of sorghum grain.

These techniques, however, are not without their limitations. Although GBS is a powerful and cost-effective sequencing tool, it can produce datasets with missing SNPs due to low coverage of sequencing, thus, reducing power in association studies[1]. The

GBS for the work presented here dealt with this issue through imputation, a common method by which the missing nucleotides are replaced with predictions through statistical inference [2]. This method has been shown to have relatively high levels of accuracy in specific plant systems, but not all[3]. With the ever decreasing cost of high throughput sequencing, follow-up studies with more complete SNP coverage will be possible. GWAS also has its limitations. The two major problems in GWAS are 1) false positives dues to population structure, and 2) allelic heterogeneity (multiple independent alleles at the same gene) or genetic heterogeneity (multiple genes controlling the trait) which may interfere with the detection of SNPs linked to phenotypic variation [4,5]. Following up a GWAS with biparental linkage mapping can help resolve the issue by breaking up the genotype and phenotype covariance[6]. Nested association mapping (NAM) also addresses these issues by combining the advantages of linkage mapping and association mapping. This approach increases statistical power and mapping resolution, while generating fewer false positives and false negatives [7–9]. NIRS is cost-effective, rapid, and non-destructive, making it ideal for high-throughput phenotyping[10]. However, the tradeoff is reduced accuracy and a reliance on only the variation found within the calibration population used to produce the predictive equations[11]. Samples that are outside of the range of the calibration population, or at the high or low extremes of the calibration population, may not have accurate predictions. These accessions may be the most valuable accessions for crop improvement and human nutrition, and need to be validated through chemical analysis. Finally, a concern with the RAW 264.7 model of inflammation is its ability to accurately predict effects of biological relevance to human health and disease. Results can provide information about the potential of the test agent, but *in vivo* follow up studies

must be performed to verify the effects seen *in vitro*.

The results and limitations of this dissertation provide a guide for future research in several focus areas, including plant genetics, plant breeding, chemistry, and human nutrition. GWAS identifies SNPs that are in linkage disequilibrium with functional variants, but is not expected to identify the functional alleles themselves. To identify putative functional alleles for the high or low levels of grain composition traits in sorghum grain, sequencing of candidate genes should be performed. Previously identified functional variants in sorghum genes include "obvious" loss-of-function mutations such as premature stop codons or missense mutations. Additional methods for confirmation of candidate genes could include the use of Arabidopsis (*Arabidopsis thaliana*) knockouts, or, as transformation in sorghum becomes more routine, knockdown studies with RNA interference (RNAi) can be conducted.

Plant breeding can help to reduce covarying or confounding factors in grain composition. A surprising result from the *in vitro* studies in chapter 4 was that the majority of sorghum accessions, regardless of flavonoid content, demonstrated anti-inflammatory properties. Several previous studies, conducted with a small number of sorghum varieties, did not find significant anti-inflammatory effects in their low-polyphenol controls[12,13]. A powerful way to reduce confounding factors is to control for differences in genetic background using near isogenic lines. These are pairs of lines, developed through backcrossing, that only differ in the genomic region of interest.

Chemical analysis of the sorghum bran extracts to identify additional polyphenol compounds will provide more clues regarding what might be driving the anti-

inflammatory effects seen in the RAW 264.7 macrophage cells. We might find that a particular set of sorghum bran polyphenols modulate one pro-inflammatory signal transduction pathway, while another set modulates a different pathway. This could present the opportunity to breed sorghum varieties for the purpose of targeting specific pro-inflammatory signaling pathways in human disease.

Follow up studies *in vivo* will be the key for discovering the anti-inflammatory effects of sorghum grain in specific disease states. Proanthocyanidins have been shown to be beneficial in colon disease[14], so testing proanthocyanidin-containing sorghum accessions in a colitis model would be a good starting point. Additionally, sorghum 3-deoxyanthocyanidins have shown potential anti-cancer properties[15–17], so testing 3-deoxyanthocyanidin-containing accessions in a cancer model is another route to take.

This research brings together new tools in order to gain insight into the health benefits of bioactive compounds in plants through the use of crop genomic diversity, and helps lay the groundwork for the use of natural variation of sorghum nutrients in crop and human nutrition improvement.

REFERENCES

1. Poland, J. A. & Rife, T. W. Genotyping-by-sequencing for plant breeding and genetics. *The Plant Genome Journal* **5,** 92 (2012).

2. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11,** 499–511 (2010).

3. Hickey, J. M., Crossa, J., Babu, R. & de los Campos, G. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. *Crop Science* **52,** 654 (2012).

4. Platt, A., Vilhjálmsson, B. J. & Nordborg, M. Conditions under which genome-wide association studies will be positively misleading. *Genetics* **186,** 1045–1052 (2010).

5. Segura, V. *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* **44,** 825–830 (2012).

6. Myles, S. *et al.* Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* **21,** 2194–2202 (2009).

7. Buckler, E. S. *et al.* The genetic architecture of maize flowering time. *Science* **325,** 714–718 (2009).

8. Yu, J., Holland, J. B., McMullen, M. D. & Buckler, E. S. Genetic design and statistical power of nested association mapping in maize. *Genetics* **178,** 539–551 (2008).

9. Stich, B. Comparison of mating designs for establishing nested association mapping populations in maize and *Arabidopsis thaliana*. *Genetics* **183,** 1525–1534 (2009).

10. Cabrera-Bosquet, L., Crossa, J., von Zitzewitz, J., Serret, M. D. & Luis Araus, J. High-throughput phenotyping and genomic selection: the frontiers of crop breeding converge. *Journal of Integrative Plant Biology* **54,** 312–320 (2012).

11. Montes, J. M., Melchinger, A. E. & Reif, J. C. Novel throughput phenotyping platforms in plant genetic studies. *Trends in Plant Science* **12,** 433–436 (2007).

12. Burdette, A. *et al.* Anti-inflammatory activity of select sorghum (*Sorghum bicolor*) Brans. *Journal of Medicinal Food* **13,** 879–887 (2010).

13. Bralley, E., Greenspan, P., Hargrove, J. L. & Hartle, D. K. Inhibition of hyaluronidase activity by select sorghum brans. *J Med Food* **11,** 307–312 (2008).

14. Li, X. *et al.* Proanthocyanidins from grape seeds modulate the NF-κB signal transduction pathways in rats with TNBS-induced ulcerative colitis. *Molecules* **16,** 6721–6731 (2011).

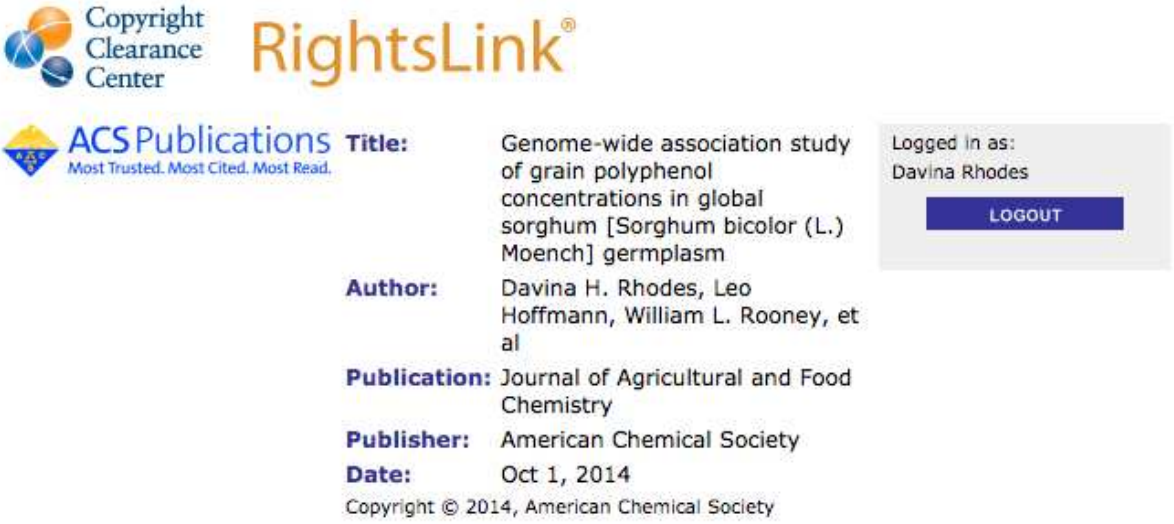
15. Yang, L., Browning, J. D. & Awika, J. M. Sorghum 3-deoxyanthocyanins possess strong phase II enzyme inducer activity and cancer cell growth inhibition properties.

J. Agric. Food Chem. **57,** 1797–1804 (2009).

16. Devi, P. S., Kumar, M. S. & Das, S. M. Evaluation of antiproliferative activity of red sorghum bran anthocyanin on a human breast cancer cell line (mcf-7). *Int J Breast Cancer* **2011,** 891481 (2011).

17. Suganyadevi, P., Saravanakumar, K. M. & Mohandas, S. The antiproliferative activity of 3-deoxyanthocyanins extracted from red sorghum (Sorghum bicolor) bran through P53-dependent and Bcl-2 gene expression in breast cancer cell line. *Life Sciences* (2013).

APPENDIX A: PERMISSION TO REPRINT



Copyright Clearance Center RightsLink®

ACS Publications Most Trusted. Most Cited. Most Read.

Title:	Genome-wide association study of grain polyphenol concentrations in global sorghum [Sorghum bicolor (L.) Moench] germplasm
Author:	Davina H. Rhodes, Leo Hoffmann, William L. Rooney, et al
Publication:	Journal of Agricultural and Food Chemistry
Publisher:	American Chemical Society
Date:	Oct 1, 2014

Copyright © 2014, American Chemical Society

Logged in as: Davina Rhodes

LOGOUT

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

APPENDIX B: FLAVONOID SNP ASSOCIATIONS

Table B.1. The 20 most statistically significant SNPs associated with proanthocyanidins using qualitative (presence/absence) phenotype[a,b]

GLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog
S4_62353772	8.46E-16	Sb04g032140 (62,146,623-62,155,642)	*TT1*	22.8
S4_62353785	8.46E-16	Sb04g032140 (62,146,623-62,155,642)	*TT1*	22.8
S4_61815549	2.55E-15	Sb04g031750 (61,676,174-61,682,528)	*TT16*	42.9
S4_54197180	1.02E-14	Sb04g024710 (54,540,924-54,542,759)	*Pr1*	69
S1_39378307	8.29E-14	Nothing close		
S4_61619739	**2.68E-13**	**Sb04g031730 (61,667,040-61,668,067)**	***Tan1***	**66.6**
S7_58603858	4.99E-13	Sb07g023840 (58,814,239-58,816,555)	*TT12*	51.3
S4_61121403	**1.16E-12**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_61122392	**1.32E-12**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_61835734	1.36E-12	Sb04g031750 (61,676,174-61,682,528)	*TT16*	42.9
S4_61122503	**2.36E-12**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S9_52833941	**5.16E-12**	**Sb09g023270 (52,888,054-52,890,562)**	***TTG2***	**31.5**
S4_61122769	**5.26E-12**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S3_58240976	5.52E-12	Sb03g029820 (58,069,805-58,070,945)	*VvLAR1*	43.9
S5_59459035	5.57E-12	Sb05g026490 (60,376,700-60,377,964)	*BZ2*	55.6
S4_57401071	8.46E-12	Sb04g027540 (57461668-57463298)	*TT2*	37.6
S7_58598713	1.33E-11	Sb07g023840 (58,814,239-58,816,555)	*TT12*	51.3
S4_57440178	1.37E-11	Sb04g027540 (57461668-57463298)	*TT2*	37.6
S4_61122741	**1.39E-11**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_61170405	**1.43E-11**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
MLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog

S4_61815549	4.82E-09	Sb04g031750 (61,676,174-61,682,528)	*TT16*	42.9
S4_61121403	**5.50E-09**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_61122392	**8.01E-09**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_61122503	**9.11E-09**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_61122769	**9.62E-09**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_61122463	**1.50E-08**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_61170405	**6.32E-08**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_59629619	6.50E-08	Sb04g030570 (60,539,878-60,541,963)	*VvLAR1*	49.7
S3_59791550	7.44E-08	Sb03g031780 (60,152,738-60,153,745)	*BZ2*	49.4
S4_60811893	**7.97E-08**	**Sb04g030840 (60,836,442-60,839,338)**	***TTG1***	**32.6**
S4_61170190	**8.07E-08**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_60825369	**9.69E-08**	**Sb04g030840 (60,836,442-60,839,338)**	***TTG1***	**32.6**
S4_59071280	9.81E-08	Sb04g030570 (60,539,878-60,541,963)	*VvLAR1*	49.7
S4_59070937	1.33E-07	Sb04g030570 (60,539,878-60,541,963)	*VvLAR1*	49.7
S4_61122741	**1.92E-07**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_62353772	1.95E-07	Sb04g032140 (62,146,623-62,155,642)	*TT1*	22.8
S4_62353785	1.95E-07	Sb04g032140 (62,146,623-62,155,642)	*TT1*	22.8
S4_60812946	**2.47E-07**	**Sb04g030840 (60,836,442-60,839,338)**	***TTG1***	**32.6**
S9_4414223	2.58E-07	Sb09g003750 (4,251,107-4,252,160)	*TT19*	42.1
S4_61667908	**4.84E-07**	**Sb04g031730 (61,667,040-61,668,067)**	***Tan1***	**66.6**

[a]*n* = 373; 204 proanthocyanidin accessions, 169 non-proanthocyanidin accessions

[b]SNPs within 100kb of the candidate gene are in bold text.

Table B.2. The 20 most statistically significant SNPs associated with proanthocyanidins, with accessions containing *tan1-a* and *tan1-b* null alleles removed, using qualitative (presence/absence) phenotype [a,b]

GLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog
S2_8258226	5.02E-12	Sb02g006390 (8,003,227-8,008,714)	*TT8*	19.3
S7_58390034	2.00E-11	Sb07g023840 (58,814,239-58,816,555)	*TT12*	51.3
S2_66685741	**2.17E-11**	**Sb02g031190 (66,677935-66682300)**	***TT18/ANS***	**48.1**
S2_66685788	**2.17E-11**	**Sb02g031190 (66,677935-66682300)**	***TT18/ANS***	**48.1**
S2_57862726	2.34E-11	Sb02g024250 (58453163-58454123)	*MYB5*	41.4
S2_66684971	**5.53E-11**	**Sb02g031190 (66,677935-66682300)**	***TT18/ANS***	**48.1**
S2_66680971	**2.31E-10**	**Sb02g031190 (66,677935-66682300)**	***TT18 (ANS)***	**48.1**
S4_61115792	**2.35E-10**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_61115781	**2.44E-10**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S6_41070539	2.66E-10	Sb06g014550 (40,216,040-40,217,587)	*TT6*	42.7
S4_54197180	2.84E-10	Sb04g024710 (54,540,924-54,542,759)	*Pr1*	69
S3_58240976	3.27E-10	Sb03g029820 (58,069,805-58,070,945)	*VvLAR1*	43.9
S2_8182066	3.51E-10	Sb02g006390 (8,003,227-8,008,714)	*TT8*	19.3
S4_61074352	3.60E-10	Sb04g031110 (61115048-61116356)	*Zm1/TT2*	66.8
S7_59242217	**5.46E-10**	**Sb07g024260 (59,291,132-59,293,418)**	***TT4***	**62.3**
S4_5016773	6.43E-10	Sb04g004736 (4,527,943-4,530,433)	*TT16*	53.6
S4_5016793	6.43E-10	Sb04g004736 (4,527,943-4,530,433)	*TT16*	53.6
S4_61077385	7.45E-10	Sb04g031110 (61115048-61116356)	*Zm1/TT2*	66.8
S4_61077412	7.45E-10	Sb04g031110 (61115048-61116356)	*Zm1/TT2*	66.8
S4_61100543	1.04E-09	Sb04g031110 (61115048-61116356)	*Zm1/TT2*	66.8
MLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog

S4_61115792	3.14E-07	Sb04g031110 (61115048-61116356)	*Zm1/TT2*	66.8
S6_41070539	3.69E-07	Sb06g014550 (40,216,040-40,217,587)	*TT6*	42.7
S3_59791550	4.06E-07	Sb03g031780 (60,152,738-60,153,745)	*BZ2*	49.4
S4_6213041	5.55E-07	Nothing close		
S2_66685741	**5.94E-07**	**Sb02g031190 (66,677935-66682300)**	***TT18/ANS***	**48.1**
S2_66685788	5.94E-07	**Sb02g031190 (66,677935-66682300)**	***TT18/ANS***	**48.1**
S4_61115781	7.88E-07	Sb04g031110 (61115048-61116356)	*Zm1/TT2*	66.8
S4_61060973	9.33E-07	Sb04g031110 (61115048-61116356)	*Zm1/TT2*	66.8
S9_58158124	9.33E-07	Sb09g028860 (57,658,680-57,661,148)	*TT4*	39.7
S9_57272292	1.10E-06	Sb09g028860 (57,658,680-57,661,148)	*TT4*	39.7
S2_57862726	1.34E-06	Sb02g024250 (58453163-58454123)	*MYB5*	41.4
S4_61100543	1.56E-06	Sb04g031110 (61115048-61116356)	*Zm1/TT2*	66.8
S4_61074352	1.57E-06	Sb04g031110 (61115048-61116356)	*Zm1/TT2*	66.8
S2_66684971	1.67E-06	Sb02g031190 (66,677935-66682300)	*TT18/ANS*	48.1
S9_58474437	2.09E-06	Sb09g028860 (57,658,680-57,661,148)	*TT4*	39.7
S4_60811893	2.92E-06	Sb04g030840 (60,836,442-60,839,338)	*TTG1*	32.6
S4_61023448	2.99E-06	Sb04g031110 (61115048-61116356)	*Zm1/TT2*	66.8
S4_60825369	3.10E-06	Sb04g031110 (61115048-61116356)	*Zm1/TT2*	66.8
S6_29795989	3.53E-06	Nothing close		
S9_5199160	3.73E-06	Sb09g003750 (4,251,107-4,252,160)	*TT19*	42.1

[a]*n* = 304; 146 proanthocyanidin accessions, 158 non-proanthocyanidin accessions

[b]SNPs within 100kb of the candidate gene are in bold text, and SNPs within the candidate gene are in blue text.

Table B.3. The 20 most statistically significant SNPs associated with quantitative proanthocyanidins (n = 373), identified by GLM and MLM analysis. SNPs within 100kb of the candidate gene are in bold text, and SNPs within the candidate gene are in blue text.

GLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog
S7_58603858	2.49E-16	Sb07g023840 (58,814,239-58,816,555)	*TT12*	51.3
S4_61815549	8.32E-16	Sb04g031750 (61,676,174-61,682,528)	*TT16*	42.9
S1_39378307	3.23E-15	Nothing close		
S1_53978849	1.10E-14	Sb01g031050 (53,615,438-53,616,246)	*BZ2*	47.5
S4_54197180	1.37E-14	Sb04g024710 (54,540,924-54,542,759)	*Pr1*	69
S4_61619739	**1.63E-14**	**Sb04g031730 (61,667,040-61,668,067)**	***Tan1***	**66.6**
S4_61667908	**2.56E-14**	**Sb04g031730 (61,667,040-61,668,067)**	***Tan1***	**66.6**
S4_61835734	2.60E-14	Sb04g031750 (61,676,174-61,682,528)	*TT16*	42.9
S7_58598713	3.59E-14	Sb07g023840 (58,814,239-58,816,555)	*TT12*	51.3
S6_39193058	5.82E-14	Sb06g014250 (39,313,831-39,320,550)	*MRP3*	57.2
S7_58598684	8.01E-14	Sb07g023840 (58,814,239-58,816,555)	*TT12*	51.3
S6_41988201	9.92E-14	Sb06g014550 (40,216,040-40,217,587)	*TT6*	42.7
S9_54293541	1.11E-13	Sb09g024300 (53,852,476-53,858,301)	*aha10*	31.8
S4_62353772	1.19E-13	Sb04g032140 (62,146,623-62,155,642)	*TT1*	22.8
S4_62353785	1.19E-13	Sb04g032140 (62,146,623-62,155,642)	*TT1*	22.8
S9_57185867	1.56E-13	Sb09g028860 (57,658,680-57,661,148)	*TT4*	39.7
S4_61122392	**1.57E-13**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_61121403	**1.58E-13**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S6_39309082	**2.22E-13**	**Sb06g014250 (39,313,831-39,320,550)**	***MRP3***	**57.2**
S6_39193160	2.36E-13	Sb06g014250 (39,313,831-39,320,550)	*MRP3*	**57.2**
MLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog

S4_61121403	**1.14E-09**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_61122392	**2.76E-09**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_61122503	**3.54E-09**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_61122769	**7.05E-09**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_61122463	**1.04E-08**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S1_53978857	1.50E-08	Sb01g031050 (53,615,438-53,616,246)	*BZ2*	**47.5**
S4_60837496	**1.69E-08**	**Sb04g030840 (60,836,442-60,839,338)**	***TTG1***	**32.6**
S4_60985684	1.73E-08	Sb04g030840 (60,836,442-60,839,338)	*TTG1*	**32.6**
S4_61815549	1.76E-08	Sb04g031750 (61,676,174-61,682,528)	*TT16*	**42.9**
S4_61122741	**1.95E-08**	**Sb04g031110 (61115048-61116356)**	***Zm1/TT2***	**66.8**
S4_61667908	**2.60E-08**	**Sb04g031730 (61,667,040-61,668,067)**	***Tan1***	**66.6**
S4_60837673	**2.83E-08**	**Sb04g030840 (60,836,442-60,839,338)**	***TTG1***	**32.6**
S4_62933630	2.94E-08	Sb04g032140 (62,146,623-62,155,642)	*TT1*	**22.8**
S3_59791550	3.47E-08	Sb03g031780 (60,152,738-60,153,745)	*BZ2*	**49.4**
S5_8500513	3.67E-08	Nothing close		
S5_8500525	5.45E-08	Nothing close		
S1_4155554	7.23E-08	Nothing close		
S1_4155549	9.62E-08	Nothing close		
S1_4155558	9.62E-08	Nothing close		
S4_59070803	1.10E-07	Sb04g030570 (60,539,878-60,541,963)	*VvLAR1*	49.7

Table B.4. The 20 most statistically significant SNPs associated with quantitative proanthocyanidins, with accessions containing *tan1-a* and *tan1-b* null alleles removed[a,b]

GLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog
S7_58390034	2.44E-13	Sb07g023840 (58,814,239-58,816,555)	*TT12*	51.3
S6_40279893	**8.28E-12**	**Sb06g014550 (40,216,040-40,217,587)**	***TT6***	**42.7**
S2_8258226	9.23E-12	Sb02g006390 (8,003,227-8,008,714)	*TT8*	19.3
S7_59242217	**1.55E-11**	**Sb07g024260 (59,291,132-59,293,418)**	***TT4***	**62.3**
S7_59533160	2.10E-11	Sb07g024550 (59,614,356-59,621,625)	*TT4*	62.3
S6_50528364	2.21E-11	Sb06g019650 (49209408-49210630)	*MYBL2*	34.9
S3_7160483	2.62E-11	Sb03g008740 (9403274-9404860)	*VvLAR*	46.2
S6_29795989	2.68E-11	Nothing close		
S7_58603858	2.78E-11	Sb07g023840 (58,814,239-58,816,555)	*TT12*	51.3
S6_49582290	3.22E-11	Sb06g019710 (49,279,763-49,283,771)	*TTG2*	31.5
S4_54197180	3.67E-11	Sb04g024710 (54,540,924-54,542,759)	*Pr1*	69
S2_2647622	4.17E-11	Sb02g001870 (1,832,991-1,837,334)	*TT4*	56.5
S6_40783545	4.57E-11	Sb06g014550 (40,216,040-40,217,587)	*TT6*	42.7
S9_57185867	5.49E-11	Sb09g028860 (57,658,680-57,661,148)	*TT4*	39.7
S6_39331101	**5.62E-11**	**Sb06g014250 (39,313,831-39,320,550)**	***ZmMRP3***	**57.2**
S6_39331116	**5.62E-11**	**Sb06g014250 (39,313,831-39,320,550)**	***ZmMRP3***	**57.2**
S1_39378307	5.63E-11	Nothing close		
S1_11176525	5.73E-11	Sb01g011610 (10,430,260-10,439,085)	*aha10*	58.2
S6_39193058	6.36E-11	Sb06g014250 (39,313,831-39,320,550)	*ZmMRP3*	57.2
S6_56992521	6.38E-11	Sb06g028420 (57,199,780-57,202,049)	*TT16*	47.6
MLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog

S1_4155549	1.54E-07	Nothing close		
S1_4155558	1.54E-07	Nothing close		
S1_4155554	1.73E-07	Nothing close		
S3_59791550	2.52E-07	Sb03g031780 (60,152,738-60,153,745)	*BZ2*	49.4
S6_40279893	**3.41E-07**	**Sb06g014550 (40,216,040-40,217,587)**	***TT6***	**42.7**
S4_54183440	6.29E-07	Sb04g024710 (54,540,924-54,542,759)	*Pr1*	69
S4_54183441	6.29E-07	Sb04g024710 (54,540,924-54,542,759)	*Pr1*	69
S1_11176525	7.59E-07	Sb01g011610 (10,430,260-10,439,085)	*aha10*	58.2
S1_53978857	8.37E-07	Sb01g031050 (53,615,438-53,616,246)	*BZ2*	47.5
S9_58927171	8.61E-07	Sb09g028860 (57,658,680-57,661,148)	*TT4*	39.7
S9_57272292	9.01E-07	Sb09g028860 (57,658,680-57,661,148)	*TT4*	39.7
S2_57780837	1.06E-06	Sb02g024250 (58453163-58454123)	*MYB5*	41.4
S4_61048857	1.09E-06	Sb04g031110 (61115048-61116356)	*Zm1/TT2*	66.8
S9_57231987	1.19E-06	Sb09g028860 (57,658,680-57,661,148)	*TT4*	39.7
S7_58390034	1.20E-06	Sb07g023840 (58,814,239-58,816,555)	*TT12*	51.3
S9_57236791	1.93E-06	Sb09g028860 (57,658,680-57,661,148)	*TT4*	39.7
S9_57236778	1.96E-06	Sb09g028860 (57,658,680-57,661,148)	*TT4*	39.7
S4_60956965	2.30E-06	Sb04g030840 (60,836,442-60,839,338)	*TTG1*	32.6
S9_57240634	2.47E-06	Sb09g028860 (57,658,680-57,661,148)	*TT4*	39.7
S1_6923595	2.61E-06	Sb01g007780 (6,694,707-6,701,002)	*TT16*	40.1

[a]*n* = 312

[b]SNPs within 100kb of the candidate gene are in bold text

Table B.5. The 20 most statistically significant SNPs associated with proanthocyanidins in proanthocyanidin-containing samples[a,b,c]

GLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog
S6_56992521	2.79E-10	Sb06g028420 (57199780-57202049)	*TT16*	47.6
S5_7528053	4.45E-10	Nothing close		
S4_55064203	7.50E-10	Sb04g024750 (54577319-54579415)	*Pr1*	69.8
S10_55871365	9.22E-10	Sb10g025470 (54782097-54786374)	*Aha10*	79.2
S8_1556705	9.67E-10	Sb08g001710 (1717743-1722003)	*TT12*	53.1
S1_58929618	1.09E-09	Sb01g034730 (58188566-58192353)	*TT2*	52.4
S1_61195857	1.46E-09	Sb01g037670 (61237360-61241520)	*TT2 (Y1)*	33.3
S6_49644262	1.49E-09	Sb06g019710 (49,279,763-49,283,771)	*TTG2*	31.5
S6_56992652	1.51E-09	Sb06g028420 (57199780-57202049)	*TT16*	47.6
S1_22691799	1.55E-09	Nothing close		
S1_59704922	1.74E-09	Sb01g034730 (58188566-58192353)	*TT2*	52.4
S5_8884346	1.74E-09	Nothing close		
S1_61135449	1.74E-09	Sb01g037670 (61237360-61241520)	*TT2 (Y1)*	33.3
S5_18851072	1.74E-09	Nothing close		
S5_26558086	1.74E-09	Nothing close		
S6_56342445	2.04E-09	Sb06g027180 (56113254-56114668)	*TT2 (Zm1)*	
S8_1556679	2.06E-09	Sb08g001710 (1717743-1722003)	*TT12*	53.1
S1_61786623	**2.06E-09**	**Sb01g038250 (61790324- 61792661)**	***MYBL2***	**31.8**
S7_35431682	2.10E-09	Nothing close		
S7_35751604	2.10E-09	Nothing close		

MLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog
S4_60949342	1.46E-07	Sb04g030840 (60836442-60839338)	*TTG1*	32.6
S4_60957038	1.93E-07	Sb04g030840 (60836442-60839338)	*TTG1*	32.6
S4_61078555	2.35E-07	Sb04g031110 (61115048-61116356)	*Zm1 (TT2)*	66.8
S4_54183440	2.54E-07	Sb04g024710 (54540924-54542759)	*Pr1*	69
S4_54183441	2.54E-07	Sb04g024710 (54540924-54542759)	*Pr1*	69
S4_61048857	2.75E-07	Sb04g031110 (61115048-61116356)	*Zm1 (TT2)*	66.8
S1_61195857	5.57E-07	Sb01g037670 (61237360-61241520)	*TT2 (Y1)*	33.3
S1_68346744	5.58E-07	Sb01g045000 (68161437- 68166405)	*Aha10*	65.4
S6_56992521	5.95E-07	Sb06g028420 (57199780-57202049)	*TT16*	47.6
S4_61115792	5.99E-07	Sb04g031110 (61115048-61116356)	*Zm1*	66.8
S2_72791558	**6.56E-07**	**Sb02g038530 (72749789- 72752961)**	***BAN***	**47.6**
S7_59864685	7.04E-07	Sb07g024890 (59864926-59866468)	*TT2 (Zm38)*	34.1
S4_57593469	7.75E-07	Sb04g027540 (57461668- 57463298)	*TT2*	37.6
S5_7528053	8.22E-07	Nothing close		
S6_50558023	9.18E-07	Sb06g019650 (49209408-49210630)	*MybL2*	34.9
S4_55064203	9.64E-07	Sb04g024710 (54540924-54542759)	*Pr1*	69
S6_50623578	1.02E-06	Sb06g019650 (49209408-49210630)	*MybL2*	34.9
S6_50623586	1.02E-06	Sb06g019650 (49209408-49210630)	*MybL2*	34.9
S6_50623588	1.02E-06	Sb06g019650 (49209408-49210630)	*MybL2*	34.9
S6_50623581	1.02E-06	Sb06g019650 (49209408-49210630)	*MybL2*	34.9

[a]n = 204

[b]Proanthocyanidin-containing is defined as > 10 mg CE/g or pigmented testa

[c]SNPs within 100kb of the candidate gene are in bold text.

Table B.6. The 20 most statistically significant SNPs associated with 3-deoxyanthocyanidins

GLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog
S1_55090837	5.54E-10	Sb01g032120 (54,964,329-54,971,032)	*TTG2*	31.5
S4_53602929	**1.05E-09**	**Sb04g024000 (53677619-53679332)**	***TT2***	**39.1**
S2_52398162	1.36E-09	Sb02g020840 (51,245,183-51,246,425)	*TT1*	44.9
S2_52428322	1.48E-09	Sb02g020840 (51,245,183-51,246,425)	*TT1*	44.9
S4_54975391	1.49E-09	Sb04g024750 (54,577,319-54,579,415)	*Pr1*	69.8
S3_71992658	1.77E-09	Sb03g044310 (71,642,586-71,647,590)	*TTG1*	22.3
S3_71451623	1.77E-09	Sb03g044310 (71,642,586-71,647,590)	*TTG1*	22.3
S3_71451622	1.77E-09	Sb03g044310 (71,642,586-71,647,590)	*TTG1*	22.3
S3_71451639	2.76E-09	Sb03g044310 (71,642,586-71,647,590)	*TTG1*	22.3
S1_56106667	2.81E-09	Sb01g032770 (55,688,190-55,689,966)	*TT2*	32.2
S3_72439585	**3.19E-09**	**Sb03g045170 (72,438,235-72,439,721)**	***TT18 (ANS) or TT6 (F3H)***	**53.7 and 43.6**
S2_55517998	3.42E-09	Sb02g020840 (51,245,183-51,246,425)	*TT1*	44.9
S2_47975089	3.84E-09	Sb02g019110 (46,364,670-46,367,910)	*TT1*	44.2
S7_62456631	**4.32E-09**	**Sb07g027340 (62,417,921-62,419,759)**	***TT3 (DFR) or ban (ANR)***	**53.1 and 53.2**
S4_53207632	4.70E-09	Sb04g023670 (53,325,993-53,331,790)	*TT1*	26.1
S4_53238118	4.93E-09	Sb04g023670 (53,325,993-53,331,790)	*TT1*	26.1
S3_71344598	4.95E-09	Sb03g044310 (71,642,586-71,647,590)	*TTG1*	22.3
S9_2058373	4.95E-09	Sb09g002350 (2,560,223-2,564,930	*TT18*	53.4
S10_57946114	5.23E-09	Sb10g029090 (58,901,962-58,902,681)	*BZ2*	44.4
S6_57486103	5.32E-09	Sb06g028630 (57,389,363-57,390,325)	*TT19*	51.9
MLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog
S1_28347083	5.36E-07	Sb01g022080 (27,064,644-27,065,491)	*BZ2*	77.2

S6_11701409	5.55E-07	Nothing close		
S3_59641176	9.77E-07	Sb03g031780 (60,152,738-60,153,745)	*BZ2*	49.4
S5_18600463	1.27E-06	Sb05g008850 (17,836,989-17,839,307)	*TT4*	60.3
S10_5605417	1.59E-06	Sb10g005940 (5379135-5380662)	*TT6*	46.1
S10_7508355	2.48E-06	Sb10g006700 (6330484- 6331579)	*ZmMRP3*	56.1
S10_44404082	3.72E-06	Nothing close		
S4_55097941	3.77E-06	Sb04g024750 (54,577,319-54,579,415)	*Pr1*	69.8
S1_66585370	4.49E-06	Sb01g043620 (66764553- 66770308)	*aha10*	34.6
S2_59315745	7.00E-06	Sb02g024260 (58500434-58501503)	*MYB5*	43.8
S4_51221699	8.08E-06	Sb04g022250 (51691231-51692495)	*TT19*	53.3
S6_50558023	8.37E-06	Sb06g019650 (49209408-49210630)	*MybL2*	34.9
S8_41806375	8.84E-06	Sb08g016160 (42830918- 42833876)	*VvLAR1*	46.2
S6_1689167	1.09E-05	Sb06g001270 (1828896-1841084)	*TT15*	61.5
S10_44365942	1.17E-05	Nothing close		
S7_38576434	1.22E-05	Nothing close		
S10_44535863	1.38E-05	Nothing close		
S4_9998328	1.93E-05	Sb04g008710 (10243964- 10245183)	*TTG1*	34.3
S9_39939122	2.00E-05	Sb09g015900 (39917155- 39919621)	*TTG2*	29.1
S8_45086484	2.54E-05	Sb08g016160 (42830918- 42833876)	*VvLAR1*	46.2

[a]*n* = 373

[b]SNPs within 100kb of the candidate gene are in bold text, and SNPs within the candidate gene are in blue text.

Table B.7. The 20 most statistically significant SNPs associated with brown grain in all samples[a,b]

GLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog
S8_52906292	4.88E-11	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S8_52906290	4.88E-11	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S8_52905638	8.65E-11	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S8_52905789	9.40E-11	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S8_52905746	9.40E-11	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S8_52905782	9.40E-11	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S8_52906014	4.47E-10	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S8_52905903	7.61E-10	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S5_58872647	1.90E-09	Nothing close		
S8_52905887	2.56E-09	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S3_63,632,556	**2.56E-09**	**Sb03g035610 (63,636,898-63,639,676)**	***TT12***	**52.3**
S6_57354543	**4.09E-09**	**Sb06g028630 (57389363-57390325)**	***TT19***	**51.9**
S8_52905918	4.18E-09	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S3_63633611	**6.46E-09**	**Sb03g035610 (63636898-63639676)**	***TT12***	**52.3**
S3_63633634	**6.46E-09**	**Sb03g035610 (63636898-63639676)**	***TT12***	**52.3**
S3_63633534	**6.46E-09**	**Sb03g035610 (63636898-63639676)**	***TT12***	**52.3**
S3_63632616	**6.70E-09**	**Sb03g035610 (63636898-63639676)**	***TT12***	**52.3**
S8_53641265	7.07E-09	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S1_25948962	7.75E-09	Sb01g021480 (25195290-25198691)	*TT1*	23.4
S6_61020370	9.05E-09	Sb06g031790 (60106108-60107732)	*TT6 SbF3H*	76.8
MLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog
S2_65858236	**5.06E-06**	**Sb02g030900 (65890970-65895810)**	*MYB5*	41
S7_61285102	6.20E-06	Sb07g026200 (61402853-61410316)	*TT16*	47.6

S1_64705875	8.92E-06	Sb01g039690 (63184945-63187783)	*TT10*	56.6
S2_59208527	9.06E-06	Sb02g024260 (58500434-58501503)	*MYB5*	43.8
S4_54284947	1.40E-05	Sb04g024750 (54,577,319-54,579,415)	*Pr1*	69.8
S8_52906292	1.51E-05	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S8_52906290	1.51E-05	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S2_65506207	1.58E-05	Sb02g030900 (65890970-65895810)	*MYB5*	41
S3_49074666	1.63E-05	Sb03g024610 (49271787-49272639)	*BZ2*	51.4
S8_52496098	1.71E-05	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S3_66976497	1.84E-05	Nothing close		
S8_52905638	1.88E-05	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S8_52905789	2.39E-05	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S8_52905746	2.39E-05	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S8_52905782	2.39E-05	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S3_62988188	2.64E-05	Sb03g035420 (63530165- 63532952)	*TT19*	52.8
S6_57354543	**2.84E-05**	**Sb06g028630 (57,389,363-57,390,325)**	***TT19***	**51.9**
S1_64705774	2.89E-05	Sb01g039690 (63184945-63187783)	*TT10*	56.6
S1_28300277	2.94E-05	Sb01g022080 (270064644-27065491)	*BZ2*	77.2
S6_48708854	3.15E-05	Sb06g019710 (49,279,763-49,283,771)	*TTG2*	31.5

[a]*n* = 373

[b]SNPs within 100kb of the candidate gene are in bold text.

Table B.8. The 20 most statistically significant SNPs associated with brown grain in proanthocyanidin-containing samples[a,b]

GLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog
S1_43750629	1.05E-08	Not close		
S2_70534763	3.23E-08	Sb02g036250 (70669178-70671026)	*TT2*	36.9
S1_28300277	3.52E-08	Sb01g022080 (270064644-27065491)	*BZ2*	77.2
S1_43795821	3.76E-08	Nothing close		
S1_43795823	3.76E-08	Nothing close		
S1_25645929	4.47E-08	Sb01g021480 (25195290-25198691)	*TT1*	23.4
S6_60932972	4.79E-08	Sb06g031790 (60106108-60107732)	*TT6 SbF3H*	76.8
S1_27137629	5.39E-08	Sb01g022080 (270064644-27065491)	*BZ2*	77.2
S1_21753836	5.81E-08	Sb01g019270 (20498918-20501026)	*TT2*	32.1
S8_52905638	6.84E-08	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S1_19240247	7.64E-08	Sb01g018950 (19886314-19990009)	*TT4*	53.7
S6_47977310	7.84E-08	Sb04g021220 (49989704-49993123)	*TTG2*	31.5
S1_27140285	7.96E-08	Sb01g022080 (270064644-27065491)	*BZ2*	77.2
S1_19240248	8.26E-08	Sb01g018950 (19886314-19990009)	*TT4*	53.7
S2_69656067	8.42E-08	Sb02g034720 (69218784-69221644)	*TT12*	51.1
S1_25790756	9.26E-08	Sb01g021480 (25195290-25198691)	*TT1*	23.4
S6_55368461	**9.51E-08**	**Sb06g026350 (55388316-55390216)**	***TT6***	**49.4**
S1_19194280	1.06E-07	Sb01g018950 (19886314-19990009)	*TT4*	53.7
S8_52905746	1.12E-07	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S8_52905782	1.12E-07	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
MLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog

S1_25645929	1.34E-07	Sb01g021480 (25195290-25198691)	*TT1*	23.4
S1_43750629	1.94E-07	Nothing close		
S8_52713649	4.57E-07	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S1_28300277	5.17E-07	Sb01g022080 (270064644-27065491)	*BZ2*	77.2
S1_43795821	5.40E-07	Nothing close		
S1_43795823	5.40E-07	Nothing close		
S1_29200037	5.57E-07	Sb01g022080 (270064644-27065491)	*BZ2*	77.2
S2_70534763	7.86E-07	Sb02g036250 (70669178-70671026)	*TT2*	36.9
S6_60932972	1.46E-06	Sb06g031790 (60106108-60107732)	*TT6*	76.8
S1_29531896	1.54E-06	Sb01g022080 (270064644-27065491)	*BZ2*	77.2
S8_52905638	2.00E-06	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S6_61020370	2.14E-06	Sb06g031790 (60106108-60107732)	*TT6*	76.8
S1_29462834	2.15E-06	Sb01g022080 (270064644-27065491)	*BZ2*	77.2
S1_28844833	2.62E-06	Sb01g022080 (270064644-27065491)	*BZ2*	77.2
S1_25790756	2.63E-06	Sb01g022080 (270064644-27065491)	*BZ2*	77.2
S1_28582280	2.71E-06	Sb01g022080 (270064644-27065491)	*BZ2*	77.2
S1_29420103	2.87E-06	Sb01g022080 (270064644-27065491)	*BZ2*	77.2
S1_28554765	3.11E-06	Sb01g022080 (270064644-27065491)	*BZ2*	77.2
S8_52906290	3.13E-06	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6
S8_52906292	3.13E-06	Sb08g021640 (53,297,815-53299964)	*TT12*	60.6

[a]*n* = 204

[b]Proanthocyanidin-containing is defined as > 10 mg CE/g or pigmented testa

[c]SNPs within 100kb of the candidate gene are in bold text.

Table B.9. The 20 most statistically significant SNPs associated with red grain[a,b]

GLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog
S4_54493561	3.49E-14	Sb04g024750 (54,577,319-54,579,415)	*Pr1*	69.8
S4_55902677	8.00E-14	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S4_54555458	**8.10E-14**	**Sb04g024750 (54,577,319-54,579,415)**	***Pr1***	**69.8**
S4_55747640	1.11E-13	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S4_55882791	1.26E-13	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S4_55867630	2.70E-13	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S4_55867633	2.70E-13	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S4_55713265	5.04E-13	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S4_55900173	6.32E-13	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S4_55747610	7.91E-13	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S4_55747632	7.91E-13	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S4_55747565	1.23E-12	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S3_72346264	**1.82E-12**	**Sb03g044980 (72307409-72308922)**	***TT19***	**54.7**
S4_55900636	1.88E-12	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S4_53815136	2.56E-12	Sb04g024000 (53677619-53679332)	*Pr1*	39.1
S4_55156807	8.35E-12	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S2_14401715	1.01E-11	Sb02g010030 (14563011-14570104)	*TT15*	59.7
S4_55760426	1.09E-11	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S4_55710493	1.33E-11	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S4_55156795	2.24E-11	Sb04g024750 (54,577,319-54,579,415)	*MYB111*	38.8
MLM				
SNP	p-value	closest *a priori* gene (location)	homolog	% similarity to homolog
S4_54555458	**1.37E-09**	**Sb04g024750 (54,577,319-54,579,415)**	***Pr1***	**69.8**

S4_54493561	8.30E-09	Sb04g024750 (54,577,319-54,579,415)	*Pr1*	69.8
S4_64587640	4.63E-08	Sb04g034620 (64455176- 64457720)	*TT10*	57
S4_65817192	**4.99E-08**	**Sb04g036040 (65831134-65834278)**	***aha10***	**75.7**
S4_55156807	9.65E-08	Sb04g024750 (54,577,319-54,579,415)	*Pr1*	69.8
S6_8129134	1.07E-07	Nothing close		
S4_64635899	1.45E-07	Sb04g034620 (64455176- 64457720)	*TT10*	57
S4_55747640	1.46E-07	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S4_55882791	1.64E-07	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S6_8336655	2.21E-07	Nothing close		
S4_55156795	2.32E-07	Sb04g024750 (54,577,319-54,579,415)	*Pr1*	69.8
S6_7640589	2.63E-07	Nothing close		
S6_7539209	2.66E-07	Nothing close		
S4_55902677	2.67E-07	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S2_70842527	2.84E-07	Sb02g036250 (70669178-70671026)	*TT2*	36.9
S6_7726594	3.11E-07	Nothing close		
S4_55867630	3.85E-07	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S4_55867633	3.85E-07	Sb04g026480 (56291313-56292251)	*MYB111*	38.8
S6_53856417	5.15E-07	Sb06g025020 (54009597- 54014513)	*TT8*	50.2
S6_7640690	5.90E-07	Nothing close		

[a]*n* = 373

[b]SNPs within 100kb of the candidate gene are in bold text.

APPENDIX C: EXPRESSION DATA

Table C.1. Expression data for candidate genes near the significant SNP on Chrm2 , 7.7Mb

gene_ID	leaves	inflor1	inflor2	anther	pistil	seed5	seed10	embr	endosperm
Sb02g023680	0	2.33069	0	0	0	0.670766	1.55299	4.296	0
Sb02g023690	5.9953	20.3048	34.2138	7.9720	19.4546	14.0674	17.3198	37.83	16.3391
Sb02g023700	0	0	19.9321	252.24	2.06251	6.61183	16.9653	0	12.1629
Sb02g023710	8.00116	10.6773	17.5759	41.157	15.6181	15.9528	12.4378	22.28	12.723
Sb02g023720	14.727	37.443	62.3528	3.6441	28.7085	31.3496	13.167	108.6	10.4572
Sb02g023730	0	0.784739	3.47564	3.3886	0	0	0	0	0
Sb02g023740	10.0312	4.35748	7.36153	40.462	3.0888	2.36443	7.05754	15.35	8.27954
Sb02g023750	0	1.86859	1.94506	0.9864	2.8889	3.61537	0.360628	2.786	0.317303
Sb02g023755	0	0	0	0	0	0	0	0	0
Sb02g023760	2.8041	9.74789	9.39109	0	5.51813	5.48551	3.42333	2.168	0
Sb02g023765	0	0	0	0	0	0	0	0	0
Sb02g023770	22.8519	0	0.151196	0.3753	0.17611	0.319654	0.228533	2.554	0.920474
Sb02g023780	0	2.11993	1.58853	2.9521	1.42408	1.5162	0.420499	1.455	0
Sb02g023790	**0**	**2.44529**	**1.02772**	**0**	**0**	**0**	**2.07284**	**0**	**5.75006**
Sb02g023800	0	12.6653	2.28424	19.878	5.77448	1.40064	0.726751	2.025	0
Sb02g023810	8.55662	16.1504	11.3857	3.8117	15.1159	18.6172	5.40596	6.086	10.4291
Sb02g023820	22.1742	44.8399	56.0609	84.116	43.6109	82.664	26.3481	27.17	19.3189
Sb02g023830	0	6.59185	2.79825	3.7731	8.61071	7.47094	1.43201	4.797	0.788552
Sb02g023840	1.903	9.7713	13.3864	11.431	8.79515	8.22447	4.55728	7.288	3.92795
Sb02g023850	0	0	0	0	0	0	0	0	0
Sb02g023860	0	0	0	0	0	0	0	0	0
Sb02g023870	0	0	0	0	0	0	0	0	0.700349
Sb02g023880	0	0	0	0	0	0	0	0	0

gene_ID	leaves	inflor1	inflor2	anther	pistil	seed5	seed10	embr	endosperm
Sb02g023890	0	0	0	0	0	0	0	0	0
Sb02g023892	0	0	0	0	0	0	0	0	0
Sb02g023895	0	0	0	0	0	0	0	0	0
Sb02g023897	0	0	0	0	0	0	0	0	0
Sb02g023900	0	0	0	0	0	0	0	0	0
Sb02g023910	1.25733	1.96357	2.64888	4.7739	1.09286	1.73258	0	0	7.18001
Sb02g023920	1.12847	28.0138	4.99207	0.7111	20.2752	5.3753	0.702854	1.193	5.05168
Sb02g023930	5.70017	12.8674	18.062	7.5488	17.3888	22.7002	12.4516	16.91	8.4293
Sb02g023940	8.43646	38.287	28.7814	9.2484	41.3633	34.9809	9.87075	22.87	8.06112
Sb02g023950	0	0	0	0	0	0	0	0	0
Sb02g023955	0	0	0	0	0	0	0	0	0
Sb02g023960	**0**	**0**	**0**	**0**	**0**	**0.676488**	**62.4068**	**1.642**	**65.289**

Table C.2. Expression data for candidate genes near the significant SNP on Chrm4, 57.7Mb

gene_ID	leaves	inflor1	inflor2	anther	pistil	seed5	seed10	embry	endosperm
Sb04g027650	0.73133	4.26973	1.63284	0	7.13977	5.4527	0	0	0
Sb04g027660	188.546	57.1422	112.676	89.421	87.8698	86.4094	86.3187	7.5989	57.8603
Sb04g027670	2.53335	2.31188	1.16451	0	1.46615	0.631974	0.760041	6.0354	0
Sb04g027680	0	1.68389	0	0	1.68558	0	0	6.7550	0
Sb04g027690	2.0943	18.5169	12.7806	13.117	23.9244	18.5843	4.22904	13.091	4.77923
Sb04g027700	3.16185	156.809	33.2842	1.5776	94.5522	40.7492	27.7802	100.06	5.12776
Sb04g027705	0	0	0	0	0	0	0	0	0
Sb04g027710	7.74616	16.2274	23.3784	4.9658	21.3814	14.0704	3.19389	32.023	4.93014
Sb04g027720	4.4102	7.71401	9.46648	24.659	11.7321	17.6894	5.24778	16.355	3.87536
Sb04g027730	108.669	3.72498	32.7003	2.7696	3.41543	7.23709	2.03196	6.6726	0.798253
Sb04g027740	0	53.9924	54.9759	19.719	32.5054	35.3426	7.72419	96.532	21.8408
Sb04g027750	0	6.81907	0	0	1.59609	0	0	0	0
Sb04g027760	0	37.4204	13.3888	1.8410	29.5497	27.2916	7.71662	28.394	2.8771
Sb04g027763	0	5.27185	10.2028	2.1452	24.3643	33.2616	6.86241	4.5695	4.60036
Sb04g027766	0	7.24242	10.0652	1.2731	33.9948	23.3235	7.64669	5.5024	3.31526
Sb04g027770	0	0	1.2577	8.1781	0	0	0	0	0
Sb04g027771	0	0	0	0	0	0	0	0	0
Sb04g027773	0	0	0	0	0	0	0	0	0
Sb04g027775	0	0	0	0	0	0	0	0	0
Sb04g027776	5.83346	0	0.140652	0	0	0	0	0.2815	0
Sb04g027778	0	0	0	0	0	0	0	0	0
Sb04g027780	61.7812	27.0869	34.1836	21.025	22.8289	27.6869	7.75636	14.163	9.44638
Sb04g027790	8.01227	10.5371	11.5947	7.1947	11.5555	10.5041	5.69906	12.518	5.98454
Sb04g027800	8.5806	10.8173	11.7671	0	12.7604	7.8703	2.42028	10.861	2.70414
Sb04g027810	1278.6	27.2237	88.867	8.7663	46.2582	69.68	39.0455	3.8307	18.0047

gene_ID	leaves	inflor1	inflor2	anther	pistil	seed5	seed10	embry	endosperm
Sb04g027815	1.94015	6.12821	4.43369	2.3419	7.01663	4.46855	2.85394	5.4223	1.32529
Sb04g027820	2.30924	6.63729	5.10568	6.1364	4.61184	5.43002	6.83531	9.9250	5.45184
Sb04g027830	8.96736	3.30603	3.70324	1.5768	4.62597	3.81609	2.83086	1.7211	1.63032
Sb04g027840	2.50108	0.836148	1.49807	1.2026	0.92270	0.744119	0	0	0
Sb04g027843	0	0	0	0	0	0	0	0	0
Sb04g027846	0	0	0	0	0	0	0	0	0
Sb04g027850	0	0	0	0	0	0	0	0	0
Sb04g027860	0.56449	0	0	0	0	1.96159	8.73977	0	2.80423
Sb04g027870	7.09031	30.8035	36.8057	52.770	24.1789	31.6686	12.0743	27.696	11.2593
Sb04g027880	198.428	58.9054	71.1705	73.714	74.7104	118.005	26.2183	49.648	26.7136
Sb04g027890	16.1457	20.4031	23.4735	30.302	19.5986	25.1393	14.7792	25.428	12.215
Sb04g027900	0.88662	15.2006	4.39256	1.0410	34.1873	31.1038	6.60782	4.4459	12.4031
Sb04g027910	2.28452	24.6923	58.8269	51.617	30.6533	44.4812	8.43038	5.6706	10.6943
Sb04g027920	3.53631	7.41364	11.0236	10.088	6.89128	8.51964	6.6762	7.4032	5.55642
Sb04g027925	0	0	0	0	0	0	0	0	0
Sb04g027930	0	89.0026	19.0844	25.066	195.795	67.5163	11.7598	1.5445	2.75919
Sb04g027940	**13.4855**	**25.3132**	**24.6645**	**14.683**	**27.0335**	**27.3917**	**14.149**	**18.742**	**31.1764**
Sb04g027950	13.2928	65.9596	20.2752	2.5892	68.4315	57.8658	32.4856	64.309	63.9801
Sb04g027960	0	15.3189	1.81792	0	18.4058	5.88915	0.4053	3.5933	0
Sb04g027970	16.2721	30.5997	36.3126	30.066	22.4896	31.2146	29.1253	20.113	37.942
Sb04g027980	0.93355	2.9024	2.64631	7.5892	1.92674	2.97485	4.01138	2.3287	1.91179
Sb04g027990	14.8272	8.59989	14.6025	26.760	8.99914	9.25894	7.73246	8.538	7.1487
Sb04g028000	15.8315	35.8274	37.2271	23.727	39.9063	24.241	15.8366	40.901	16.3728
Sb04g028010	0	0.559678	1.35391	0	1.21904	1.79474	0	0	0
Sb04g028015	0	0.676973	0.719732	0.3896	1.18907	0.653495	0.420828	2.1321	0
Sb04g028020	214.137	0.774813	1.55421	0.9933	0	0.790386	0	0	0
Sb04g028030	39.2943	7.42981	13.8686	9.9932	11.399	22.6709	12.2489	21.272	12.2932
Sb04g028040	0	0	0	0	0	0	0	0	0

gene_ID	leaves	inflor1	inflor2	anther	pistil	seed5	seed10	embry	endosperm
Sb04g028050	767.24	36.2032	40.2015	13.657	140.866	82.0592	18.432	86.977	19.9151
Sb04g028060	**80.6905**	**17.2847**	**4.47903**	**3.9450**	**40.9066**	**21.0613**	**3.50636**	**5.3521**	**5.08887**

Table C.3. Expression data for candidate genes near the significant SNP on Chrm6, 48.8Mb

gene_ID	leaves	inflor1	inflor2	anther	pistil	seed5	seed10	embryo	endosperm
Sb06g019010	**1.0787**	**19.8095**	**23.5318**	**1.4969**	**5.63451**	**78.6608**	**62.0977**	**5.11923**	**39.2647**
Sb06g019015	0	0	0	0	0	0	0	0	0
Sb06g019020	0	28.0475	2.34575	0.8735	29.5048	26.15	1.93895	1.73359	0.869728
Sb06g019030	167.60	35.9534	65.5943	1.5072	27.9022	52.4701	9.83391	13.8034	10.1455
Sb06g019040	0	1.89623	0	0	0	0	0	0	0
Sb06g019043	0	0	0	0	0	0	0	0	0
Sb06g019046	0	1.39149	1.28464	1.4948	0	0	0	2.01371	0
Sb06g019050	0	0	0	0	0	1.97889	10.3375	56.0965	17.3933
Sb06g019060	0	9.13901	2.62463	0	13.6169	7.1539	1.45194	4.93426	0.461682
Sb06g019070	0	88.1934	1.07513	0	0	8.76246	0	0	0
Sb06g019080	13.942	4.27086	3.32156	1.7409	3.88611	4.16478	1.09393	1.15886	1.57501
Sb06g019085	0	0	0	0	0	0	0	0	0
Sb06g019090	0.5816	1.11565	0	0	0	0	0	0	0
Sb06g019100	379.77	653.703	367.613	95.533	652.354	489.726	329.187	341.131	291.376
Sb06g019105	0	4.38968	1.46853	0	6.87291	2.51216	0	4.97549	0
Sb06g019110	11.474	111.522	46.628	32.543	110.383	112.309	81.1536	130.975	60.9299
Sb06g019120	1.1712	46.63	14.9212	1.2467	14.3896	5.41831	0.779662	1.48891	0.879481
Sb06g019130	37.560	4.18819	3.82466	1.8824	11.4045	6.20356	22.3028	3.42221	48.068
Sb06g019140	0.8430	2.46611	2.13152	1.4875	3.66841	1.40035	0.323714	2.13213	0.379765
Sb06g019150	15.647	38.1188	18.8009	17.872	16.2661	13.6899	6.71654	8.57036	6.61489
Sb06g019160	37.416	22.6099	21.2094	163.46	21.2621	35.9644	31.6271	59.8787	17.6973
Sb06g019170	12.443	11.4223	15.2383	10.039	15.1396	12.4181	6.62713	6.76693	13.9943
Sb06g019180	32.346	42.3676	38.1915	25.224	64.0169	53.6291	27.3556	29.2114	37.0469
Sb06g019190	0	1.51381	0.96479	0.8702	1.22313	1.01062	0.359188	0.85791	0
Sb06g019200	10.35	55.6172	73.4476	55.148	60.7769	55.7958	34.7427	31.8607	48.6426

gene_ID	leaves	inflor1	inflor2	anther	pistil	seed5	seed10	embryo	endosperm
Sb06g019210	2.4253	26.4266	21.2916	29.917	26.5723	7.07552	18.1732	16.7602	15.1294
Sb06g019215	0	0	0	0	0	0	0	0	0
Sb06g019220	0	0	0	0.2518	0	0	0	0	0
Sb06g019230	37.222	73.4043	81.8889	71.077	91.3867	82.3839	40.6072	48.5786	47.8507
Sb06g019240	0	5.25338	1.80915	0	8.99107	6.18722	0.506669	1.87243	0.804186
Sb06g019245	6.6781	20.2955	23.9005	5.2305	16.7144	32.7837	10.5675	18.3567	7.67451
Sb06g019250	1.7942	8.40965	8.60721	2.5451	8.99324	4.89734	2.4994	7.0212	2.93217
Sb06g019260	7.6546	18.102	20.9162	15.099	29.3711	23.9525	11.7477	21.0342	15.7889
Sb06g019270	6.4553	14.5577	24.126	21.037	31.2323	21.4287	7.57784	25.2134	8.88994
Sb06g019275	0	1.47963	1.95627	2.9293	1.63872	2.59347	2.38775	3.66054	0
Sb06g019280	2.8713	57.3062	22.2902	29.048	62.4483	70.1156	28.4198	22.4816	27.8322
Sb06g019290	0	166.793	3.95845	0	81.1638	16.3163	11.3494	93.5597	9.00988
Sb06g019300	2.3507	0.837201	1.68086	0	0.788441	1.2417	1.06488	3.63468	0
Sb06g019310	9.5576	7.54925	5.627	2.9306	8.0575	7.68804	2.88069	3.48233	3.37948
Sb06g019320	**3.0550**	**1223.47**	**1720.13**	**29.206**	**1873.29**	**4979.89**	**806.416**	**799.895**	**293.703**

APPENDIX D: GRAIN COMPOSITION SNP ASSOCIATIONS

Table D.1 Statistically significant SNPs associated with protein

SNP[a]	p-value[b]	MAF[c]	R^2 [d]	closest *a priori* gene (location)	Homolog	Homolog description	% [e]
S4_57657983	4.43E-06	0.17	0.29	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57656443	4.43E-06	0.41	0.29	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57656473	1.09E-05	0.41	0.28	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57656457	1.09E-05	0.41	0.28	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57663731	1.22E-05	0.41	0.28	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57645574	1.51E-05	0.41	0.28	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57663557	1.69E-05	0.41	0.28	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57645884	1.69E-05	0.44	0.28	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57609482	1.69E-05	0.40	0.27	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57662361	1.69E-05	0.41	0.27	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57662551	2.19E-05	0.38	0.27	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57662409	2.19E-05	0.40	0.27	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57662424	2.19E-05	0.40	0.27	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57645089	2.62E-05	0.39	0.27	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57644830	2.62E-05	0.40	0.27	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57663773	2.65E-05	0.40	0.27	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57663934	4.87E-05	0.40	0.27	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57663936	4.87E-05	0.40	0.27	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S4_57641319	1.02E-04	0.15	0.26	Sb04g027940 (57,859,449-57,863,521)	AT3G54320	*WRL1*	30.6
S9_53422385	1.79E-04	0.06	0.26				
S2_57679376	1.79E-04	0.22	0.26	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57610561	2.49E-04	0.39	0.26	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57645073	3.93E-04	0.38	0.26	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57678951	1.66E-03	0.44	0.25	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29

SNP[a]	p-value[b]	MAF[c]	R^{2} [d]	closest *a priori* gene (location)	Homolog	Homolog description	% [e]
S3_72664832	1.66E-03	0.40	0.25				
S3_72664834	1.66E-03	0.40	0.25				
S3_72664833	1.66E-03	0.40	0.25				
S3_72664835	1.66E-03	0.40	0.25				
S2_57729868	1.68E-03	0.34	0.25	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S4_57619802	1.68E-03	0.36	0.25	Sb04g027940 (57,859,449-57,863,521)	AT3G54320	*WRL1*	30.6
S2_57605117	1.83E-03	0.39	0.25	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57605132	1.83E-03	0.39	0.25	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57605139	1.83E-03	0.39	0.25	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S4_57619784	2.25E-03	0.36	0.25	Sb04g027940 (57,859,449-57,863,521)	AT3G54320	*WRL1*	30.6
S2_57679247	2.39E-03	0.34	0.25	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57721916	2.86E-03	0.39	0.24	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S1_64583413	3.34E-03	0.02	0.24				
S1_15018832	3.81E-03	0.06	0.24				
S2_57728934	4.33E-03	0.37	0.24	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S3_11828757	4.48E-03	0.37	0.24	None close			
S10_11107326	6.08E-03	0.02	0.24				
S1_58114305	6.09E-03	0.01	0.24				
S2_57726523	6.23E-03	0.41	0.24	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S4_57594856	9.12E-03	0.41	0.24	Sb04g027940 (57,859,449-57,863,521)	AT3G54320	*WRL1*	30.6
S5_55916389	9.69E-03	0.01	0.24				
S2_57679565	9.89E-03	0.30	0.24	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	
S4_57635555	1.02E-02	0.42	0.24	Sb04g027940 (57,859,449-57,863,521)	AT3G54320	*WRL1*	30.6
S2_57726615	1.26E-02	0.40	0.24	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	30.6
S3_68275529	1.86E-02	0.05	0.23				
S4_57619661	1.87E-02	0.35	0.23	Sb04g027940 (57,859,449-57,863,521)	AT3G54320	*WRL1*	30.6
S9_54885754	1.92E-02	0.22	0.23				

SNP[a]	p-value[b]	MAF[c]	R^{2} [d]	closest *a priori* gene (location)	Homolog	Homolog description	% [e]
S2_57727196	1.92E-02	0.40	0.23	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S4_55666106	1.97E-02	0.12	0.23				
S1_21499901	2.01E-02	0.03	0.23				
S2_57727421	2.12E-02	0.40	0.23	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S1_15018913	2.14E-02	0.05	0.23				
S7_9235696	2.28E-02	0.02	0.23				
S9_53407206	2.33E-02	0.01	0.23				
S3_72152710	2.41E-02	0.38	0.23				
S10_12029907	2.77E-02	0.01	0.23				
S2_57728715	2.84E-02	0.27	0.23	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S7_7073656	2.93E-02	0.11	0.23				
S2_7407872	2.93E-02	0.02	0.23				
S4_57774278	2.97E-02	0.21	0.23				
S1_58060924	3.04E-02	0.35	0.23				
S1_57287903	3.04E-02	0.07	0.23				
S6_4717038	3.20E-02	0.00	0.23				
S2_57695323	3.20E-02	0.49	0.23	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_58954943	3.27E-02	0.21	0.23				
S9_56662775	3.43E-02	0.19	0.23				
S2_66625727	3.43E-02	0.43	0.23				
S4_57794439	3.69E-02	0.22	0.23	Sb04g027940 (57,859,449-57,863,521)	AT3G54320	*WRL1*	30.6
S10_50000241	3.91E-02	0.05	0.23				
S4_57593430	3.93E-02	0.39	0.23	Sb04g027940 (57,859,449-57,863,521)	AT3G54320	*WRL1*	30.6
S1_45833251	4.02E-02	0.02	0.23				
S2_70911685	4.18E-02	0.01	0.23				
S7_6730549	4.23E-02	0.06	0.23				
S1_61652914	4.28E-02	0.06	0.23				

SNP[a]	p-value[b]	MAF[c]	R^2 [d]	closest *a priori* gene (location)	Homolog	Homolog description	% [e]
S6_51716811	4.36E-02	0.09	0.23				
S6_48889838	4.58E-02	0.10	0.23				
S3_69921718	4.66E-02	0.08	0.23				

[a] 81 significant SNPs found using MLM

[b] FDR adjusted P-value

[c]Minor allele frequency

[d] R^2 of model with SNP

[e]Percent similarity to homolog

Table D.2 Statistically significant SNPs associated with fat

SNP[a]	p-value[b]	MAF[c]	R^{2} [d]	closest *a priori* gene (location)	Homolog	Homolog description	% [e]
S2_57645574	3.73E-09	0.41	0.29	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57663731	3.73E-09	0.41	0.29	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57656443	3.73E-09	0.41	0.29	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57645089	3.73E-09	0.39	0.29	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57656473	3.73E-09	0.41	0.29	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57656457	3.73E-09	0.41	0.29	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57663557	3.73E-09	0.41	0.28	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57663773	3.73E-09	0.40	0.28	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57662409	6.26E-09	0.40	0.28	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57662424	6.26E-09	0.40	0.28	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57609482	8.68E-09	0.40	0.28	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57662551	1.14E-08	0.38	0.28	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57645884	1.86E-08	0.44	0.27	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57662361	2.57E-08	0.41	0.27	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57645073	3.06E-08	0.38	0.27	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57663934	4.77E-08	0.40	0.27	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57663936	4.77E-08	0.40	0.27	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57644830	7.66E-08	0.40	0.26	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57610561	7.97E-08	0.39	0.26	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57729868	2.93E-07	0.34	0.26	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57679247	2.99E-07	0.34	0.26	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57728934	8.84E-07	0.37	0.25	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57678951	8.84E-07	0.44	0.25	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57721916	1.58E-06	0.39	0.25	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29

SNP[a]	p-value[b]	MAF[c]	R^{2} [d]	closest *a priori* gene (location)	Homolog	Homolog description	% [e]
S2_57727196	1.58E-06	0.40	0.25	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57679376	1.82E-06	0.22	0.25	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57726523	3.61E-06	0.41	0.24	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57605117	4.68E-06	0.39	0.24	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57605132	4.68E-06	0.39	0.24	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57605139	4.68E-06	0.39	0.24	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57679565	6.28E-06	0.30	0.24	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57726615	1.02E-05	0.40	0.24	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57727421	1.06E-05	0.40	0.24	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57727066	4.09E-05	0.41	0.23	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57727065	4.09E-05	0.41	0.23	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57732524	6.81E-05	0.28	0.23	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57695323	1.16E-04	0.49	0.22	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57728715	1.74E-04	0.27	0.22	Sb02g023690 (57,547,760-57,552,542)	AT5G36880	Acyl coA synthetase	34.3
S4_57657983	8.90E-04	0.17	0.21	Sb04g027940 (57,859,449-57,863,521)	AT3G54320	*WRL1*	30.6
S9_53422385	1.22E-03	0.06	0.21				
S3_11828757	1.77E-03	0.37	0.21	None close			
S2_66671179	2.26E-03	0.45	0.21				
S10_50000241	3.63E-03	0.05	0.21				
S8_6611311	5.85E-03	0.13	0.20				
S4_57619802	7.55E-03	0.36	0.20	Sb04g027940 (57,859,449-57,863,521)	AT3G54320	*WRL1*	30.6
S4_55666106	9.00E-03	0.12	0.20				
S4_57641319	9.20E-03	0.15	0.20	Sb04g027940 (57,859,449-57,863,521)	AT3G54320	*WRL1*	30.6
S10_11463179	9.42E-03	0.18	0.20				
S2_66625727	9.64E-03	0.43	0.20				
S2_57703534	9.64E-03	0.23	0.20				
S4_57619784	1.38E-02	0.36	0.20	Sb04g027940 (57,859,449-57,863,521)	AT3G54320	*WRL1*	30.6

SNP[a]	p-value[b]	MAF[c]	R^{2} [d]	closest *a priori* gene (location)	Homolog	Homolog description	% [e]
S10_12029907	1.51E-02	0.01	0.20				
S2_57721957	1.85E-02	0.23	0.20	Sb02g023790 (57,701,214-57,703,517)	AT1G69830	AMY3	29
S2_57678376	2.01E-02	0.19	0.20				
S3_11909508	2.01E-02	0.36	0.20				
S2_70911685	2.03E-02	0.01	0.20				
S8_6611198	2.13E-02	0.13	0.20				
S3_11944753	2.20E-02	0.05	0.20	None close			
S5_21101672	2.21E-02	0.08	0.20				
S6_48889838	2.35E-02	0.10	0.20				
S7_62603146	2.35E-02	0.17	0.20				
S7_59483342	2.35E-02	0.02	0.20				
S7_59483390	2.35E-02	0.02	0.20				
S2_2696401	2.35E-02	0.40	0.19				
S1_15096821	2.42E-02	0.01	0.19				
S1_15018832	2.64E-02	0.06	0.19				
S2_7407872	2.72E-02	0.02	0.19				
S5_21322403	2.92E-02	0.07	0.19				
S9_18260324	2.92E-02	0.01	0.19				
S1_64583413	3.05E-02	0.02	0.19				
S6_4442676	3.31E-02	0.00	0.19				
S2_66670880	3.31E-02	0.39	0.19				
S7_549037	3.38E-02	0.40	0.19				
S2_73097110	3.41E-02	0.01	0.19				
S3_11960725	3.59E-02	0.34	0.19				
S6_4717038	3.83E-02	0.00	0.19				
S9_50532933	3.85E-02	0.01	0.19				
S1_44209436	4.27E-02	0.04	0.19				

SNP[a]	p-value[b]	MAF[c]	R^{2} [d]	closest *a priori* gene (location)	Homolog	Homolog description	% [e]
S2_76371136	4.37E-02	0.02	0.19				
S4_12581504	4.47E-02	0.13	0.19				
S1_58114305	4.58E-02	0.01	0.19				

[a] 81 significant SNPs found using MLM

[b] FDR adjusted P-value

[c]Minor allele frequency

[d] R^{2} of model with SNP

[e]Percent similarity to homolog

Table D.3 Statistically significant SNPs associated with starch

SNP[a]	p-value[b]	MAF[c]	R^2 [d]	closest *a priori* gene (location)	Homolog	Homolog description	% [e]
S6_48889838	0.007	0.10	0.33				
S2_66166782	0.007	0.04	0.33				
S3_72588175	0.018	0.31	0.33				
S3_69127635	0.018	0.05	0.32				
S2_68167599	0.018	0.02	0.32				
S10_4811966	0.020	0.01	0.32				
S4_50643504	0.036	0.26	0.32				
S3_309287	0.036	0.14	0.32				
S10_5870897	0.036	0.01	0.32				
S2_67321333	0.036	0.01	0.32				
S2_66166766	0.049	0.04	0.32				

[a] 11 significant SNPs found using MLM

[b] FDR adjusted P-value

[c] Minor allele frequency

[d] R^2 of model with SNP

[e] Percent similarity to homolog