GENETIC ANALYSIS OF THE 'DRY' TRAIT, STEM ARCHITECTURE AND PLANT WATER CONTENT COMPONENTS IN SORGHUM

by

HUIZHE JIN

(Under the Direction of Andrew H. Paterson)

ABSTRACT

With increasing energy demand and diminishing petroleum reserves, energy sources derived from renewable plant-based biomass are an attractive substitute for fossil fuels. Being the fifth most widely grown cereal crop in the world, sorghum is among one of the most attractive bioenergy feedstocks currently under active study. In this thesis, we provide a valuable addition to the genetic resources of sorghum by describing a recombinant inbred line (RIL) population of 161 F5 genotypes for the widest euploid cross that can be made in eusorghums, of annual *S. bicolor* x perennial *S. propinquum*. It offers advantages over early-generation populations of this cross that will shed new light on genetic, environmental, and physiological / biochemical factors that regulate plant growth and development. A combination of linkage and association genetic approaches using a *S. bicolor* x *S. propinquum* F2 population, a *S. bicolor* BTx623 x IS3620c RIL population, and a *S. bicolor* diversity panel of 354 accessions characterized at 265,487 SNPs allows us to: tentatively identity and characterize the D gene for the 'dry' trait; and determine genomic regions corresponding to 3 stem diameter traits and 2 water content traits. We also revealed the evolutionary path of the D gene, as having evolved under

balancing selection during a recent tandem expansion of the BGLU gene family in sorghum after its divergence from maize. Co-localizations of stem diameter and water content traits with a number of other bioenergy traits, including plant height, flowering time, branching, stem volume and the ‘dry’ trait, were also observed, suggesting that their inheritance may be functionally (pleiotropy) or physically (linkage disequilibrium) linked. Increasing knowledge of the genetic control of these bioenergy traits in sorghum, and in identification of corresponding genes and their functions, may lead to tools and strategies for either enhancing or suppressing these traits, supporting advances toward greater use of plant-based biomass in biofuel production.

INDEX WORDS: QTL mapping, GWAS, comparative genomics, genetic correspondence, cereals

GENETIC ANALYSIS OF THE ‘DRY’ TRAIT, STEM ARCHITECTURE AND PLANT WATER CONTENT COMPONENTS IN SORGHUM

by

HUIZHE JIN

B.S., Huazhong University of Science and Technology, P.R. China, 2008

M.S., The University of Georgia, 2015

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2016

© 2016

Huizhe Jin

All Rights Reserved

GENETIC ANALYSIS OF THE 'DRY' TRAIT, STEM ARCHITECTURE AND PLANT WATER CONTENT COMPONENTS IN SORGHUM

by

HUIZHE JIN

Major Professor:	Andrew H. Paterson
Committee:	John M. Burke
	Russell L. Malmberg
	Paul Schliekelman

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2016

DEDICATION

To my parents Chenghong Jin and Chaoying Hu, my grandpa Zhimin Jin, and my husband Xu Tan.

ACKNOWLEDGEMENTS

I would like to express my utmost gratitude to my advisor Andrew Paterson, for his supervision and guidance throughout my research, and for his consistent support while I was facing obstacles in completing my Ph.D. study.

I would like to thank my other thesis committee members, John Burke, Russell Malmberg, and Paul Schliekelman, for their considerations and insightful advice.

I would like to thank all Plant Genome Mapping Laboratory members, Rosana Compton, Hugo Cuevas, Hui Guo, Changsoo Kim, Wenqian Kong, Uzay Sezen, Chengbo Zhou, Dong Zhang, for their generous help and constructive comments. I thank my former advisor Jeff Dean for supporting me in first 1.5 years of my research. I thank my former lab members John Michael, Ujwal Bagal, Walt Lorenz, for their innumerable encouragement when I first entered the U.S. And I thank my dearest friends Pei Xu, Xuan Wang for their company all these years.

Last but not least, I would like to thank my parents Chenghong Jin, Chaoying Hu, and my husband Xu Tan, for their endless love, support, patience and encouragement. All give me strength to support me in accomplishing this challenging journey.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

LIST OF TABLES viii

LIST OF FIGURES xi

CHAPTER

1 INTRODUCTION AND LITERATURE REVIEW 1

1.1 Sorghum for Biofuel Production 1

1.2 Bioenergy Traits in Sorghum 2

1.3 Gene Mapping in Sorghum 4

1.4 Overview of Dissertation Chapters 5

1.5 References 7

2 GENETIC ANALYSIS OF RECOMBINANT INBRED LINES FOR *SORGHUM BICOLOR* X *SORGHUM PROPINQUUM* 10

2.1 Abstract 11

2.2 Introduction 12

2.3 Materials and Methods 14

2.4 Results 16

2.5 Discussion 22

2.6 Acknowledgements 24

2.7 References 32

3 IDENTIFICATION AND CHARACTERIZATION OF THE SORGHUM 'DRY' GENE 37

3.1 Abstract 38

3.2 Introduction 38

3.3 Materials and Methods 41

3.4 Results and Discussion 47

3.5 Conclusion 65

3.6 References 107

4 GENETIC ANALYSIS OF STEM DIAMETERS AND WATER CONTENTS TO IMPROVE BIOENERGY TRAITS IN SORGHUM 116

4.1 Abstract 117

4.2 Introduction 117

4.3 Materials and Methods 121

4.4 Results and Discussion 125

4.5 Conclusion 134

4.6 References 158

5 CONCLUSIONS 164

5.1 References 168

LIST OF TABLES

Page

Table 2.1: Comparison of regions of segregation distortion between *S. bicolor* (SB) x *S. propinquum* (SP) F5 RIL and F2 populations 25

Table 2.2: Comparison of regions showing over-/under-representations of residual heterozygosity between *S. bicolor* x *S. propinquum* F5 RIL and F2 populations .26

Table 2.3: Biometric parameters of QTL for days to first flower in the *S. bicolor* x *S. propinquum* RILs 28

Table 3.1: Sorghum accessions and countries of origin in the 'D' gene sub-panel.....66

Table 3.2: Locations of markers flanking the "Dry" trait 67

Table 3.3: SNPs with significant association to the 'dry' trait 68

Table 3.4: Candidate genes for re-sequencing.....69

Table 3.5: Non-synonymous substitution at the 'dry' locus 70

Table 3.6: Ontology annotations of Sobic.006G145300 71

Table 3.7: General information for the Dry trait across maturity dates.....72

Table 3.8: Regression models of the 'dry' trait versus moisture content, and the estimated mean moisture content of 'dry' or 'juicy' phenotypes considering the impact of maturity date, for the overall plant, stalks, leaves and inflorescence 73

Table 3.9: Stalk and overall moisture content for each subgroup.....74

Table 3.10: Distribution of the 'dry' trait across races and geographic regions.....75

Table 4.1: Descriptive statistics for stem and water content traits 136

Table 4.2: Correlation coefficients of stem and water content traits 137
Table 4.3: Heritability and variance components for stem diameter and water content traits, based on genotype, year, and genotype by year interaction percentages...138
Table 4.4: QTLs affecting stem diameter traits using overall BLUP values in BTx623 x IS3620c RILs 139
Table 4.5: Overlapping QTL regions among different measures of stem diameter 140
Table 4.6: Stem diameter QTLs with corresponding significant SNPs 141
Table 4.7: QTLs affecting water content traits using overall BLUP and single year values in BTx623 x IS3620c RILs 142
Table 4.8: Overlapping QTL regions between water content and stem diameter traits ..143
Table 4.9: Water content QTLs with corresponding significant SNPs.....144
Table 4.10: QTLs affecting stem volume trait in BTx623 x IS3620c RILs 145
Table S3.1: Primer pairs of selected exons for targeted re-sequencing.....91
Table S3.2: Significant association variants characterized in targeted re-sequencing 92
Table S3.3: Synonym for gene model v1.4 in gene model v3.1 at the selected candidate gene region.....93
Table S3.4: Possible causative loci of the 'dry' trait 94
Table S3.5: Distribution of individuals with W and G phenotypes in each country based on the Sorghum diversity panel 95
Table S3.6: Phylogeny of beta-glucosidase gene family.....96
Table S4.1: QTLs affecting stem diameter traits using single year values in BTx623 x IS3620c RILs 151
Table S4.2: QTLs in Shiringani et al.'s study.....152

Table S4.3: SNPs with significant association to stem diameter traits.....153

Table S4.4: Synteny block contained in qSWC12_1.1 and qSWC6.1 154

Table S4.5: SNPs with significant association to water content traits.....155

LIST OF FIGURES

Page

Figure 2.1: A genetic map of the *S. bicolor* x *S. propinquum* RILs 29

Figure 2.2: QTLs for days to first flower in the *S. bicolor* x *S. propinquum* RILs.....31

Figure 3.1: The classification of midrib color 78

Figure 3.2: Initial mapping of the 'Dry' gene on sorghum chromosome 6 79

Figure 3.3: Mapping of the 'dry' gene in the BTx623 × IS3620c RIL population.....80

Figure 3.4: GWAS of the 'dry' trait with Manhattan plots and Quantile-Quantile plots for 2009 and 2010.....81

Figure 3.5: LD blocks with significant SNPs aligned to sorghum chromosome 6.....83

Figure 3.6: The distribution of overall plant, stalk, leaf, and inflorescence moisture content of the sorghum genotypes over the growing season 84

Figure 3.7: Stem moisture QTL mapping.....85

Figure 3.8: K-means clustering using the first three PCA components.....86

Figure 3.9: Geographic distribution of the 'dry' trait based on sorghum diversity panel .87

Figure 3.10: Dendogram of 9 immediately adjacent sorghum beta-glucosidase genes on chromosome 6.....88

Figure 3.11: Phylogenetic tree of the beta-glucosidase gene family 89

Figure 4.1: QTLs and significant SNPs for stem diameter and water content traits along the published *S. bicolor* reference genome.....146

Figure 4.2: Manhattan plots for stem diameter traits.....148

Figure 4.3: Manhattan plots for water content traits.....150
Figure S3.1: Comparison of the D protein sequence in *Sorghum bicolor* genome annotation version 1.4 and version 3.1 99
Figure S3.2: Alignment of 9 beta-glucosidase proteins at the D candidate gene regions100
Figure S3.3: Six branches of the tree corresponding to group I to VI.....102
Figure S4.1: QQ-plots for stem diameter traits of the sorghum diversity panel in year 2009 and 2010.....156
Figure S4.2: QQ-plots for water content traits of the sorghum diversity panel in year 2009 and 2010.....157

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Cultivated sorghum (*Sorghum bicolor*) is an important and widely distributed cereal that can be used for food, feed, fodder, and ethanol fermentation. In the past 50 years, the area planted with sorghum worldwide has increased 66% while yield has increased by 244 percent (STROADE *et al.* 2016). Grain sorghum has been used for making wine and gluten-free food. Sweet sorghum has been used for making syrup. Forage sorghum has been used as silage for feeding livestock. More recently, to meet increasing energy demand, biomass sorghum has been actively under study for the production of bioenergy.

1.1 Sorghum for Biofuel Production

The US Energy Information Administration projects a 48% increase in energy consumption between 2012 and 2040 (EIA 2016). With increasing energy needs and diminishing petroleum reserves, energy sources derived from renewable plant-based biomass are suggested to be the only direct substitute for fossil fuels available on a significant scale and with less pollution (TOMES *et al.* 2011). Plants having potential as biofuel feedstocks should require limited inputs (irrigation, fertilizer, pesticides), produce high yields of biomass, and be convertible to bio-based products in a relatively efficient and economical manner (VERMERRIS 2011). Being the fifth most widely grown cereal

crop in the world, sorghum was proposed as a potential biofuel feedstock over 30 years ago (BURTON 1986). As a C4 crop that converts energy more efficiently into biomass than C3 plants, sorghum also has many advantages that make it an attractive biofuel feedstock, with lower need for fertilizers and pesticides than many crops, high water-use efficiency, drought tolerance, and wide adaptability to a variety of climates and soil conditions (SABALLOS 2008). In addition to its resilient nature, sorghum is appealing for biofuel production also for its short life cycle (4 months on average), abundant genetic diversity, and a history of improvement of lignocellulose, sugar and starch yields (ROONEY *et al.* 2007).

1.2 Bioenergy Traits in Sorghum

To maximize biomass yield per geographic unit and production system, genetic improvement needs to target bioenergy traits such as the 'dry' trait, stem architecture (diameters), and plant water content components in sorghum.

The 'dry' trait, with reduced moisture content being preferred for biomass destined for biofuel production, is largely controlled by a single gene difference. In sorghum, the 'dry' phenotype, manifested by white leaf midribs and dry, pithy stalks, is genetically dominant to the alternative 'juicy' phenotype, manifested by green midribs and dull, juicy stalks. To our knowledge, no 'dry/juicy' trait variation in other cereals has been noted. Previous studies have mapped the 'dry' gene (*D*) to co-segregate with an SSR marker Xtxp97 (HART *et al.* 2001), or reside between RFLP marker Xtxs1030 and Xumc34 on sorghum chromosome 6 (XU *et al.* 2000), but no precise location had been reported.

Stem diameter, with thicker being preferred for biomass yield, is itself a measurement to monitor plant growth. Previous studies have mapped quantitative trait loci (QTLs) for stem diameter on chromosome 1, 2, 3, 4, 6, 7, 8, 9 (SHIRINGANI *et al.* 2010). However, their measurements were only performed at the base of the plant, e.g., 5 cm above the ground in Miscanthus (ATIENZA *et al.* 2003), 20 cm above the ground in sorghum (SHIRINGANI *et al.* 2010), or 40 cm above the ground in rice (KASHIWAGI and ISHIMARU 2004), lacking a comprehensive picture of stem architecture.

Plant water content, with less being preferred for biomass yield and storage (GORDON 1967), is a useful indicator of plant moisture level. Previous studies have reported QTLs in a number of sorghum mapping populations and with different measurement statistics (SHIRINGANI *et al.* 2010; GUAN *et al.* 2011; LV *et al.* 2013; HAN *et al.* 2015). One common drawback of these studies is that juiciness was not generally segregating well in mapping populations. For example, in one study (SHIRINGANI *et al.* 2010), both parents of the studied RILs were juicy -- although transgressive segregation was observed, the power and accuracy of QTL detection could be reduced (LI *et al.* 2005). Another drawback is that some measurement statistics adopted, e.g., juice weight: wet weight of stems and leaves (GUAN *et al.* 2011), or relative juice weight: the weight difference between wet and dry stems (SHIRINGANI *et al.* 2010), may cause some bias in measuring the plant moisture level. Water content, the percent difference between wet and dry biomass weight, may be a better estimate than juice weight or relative juice weight, as it normalizes the moisture level data to a scale of 0 to 1. Consistent with this definition, Han et al. (HAN *et al.* 2015) reported a major QTL (qSW6) for stem water content on chromosome 6. However, there is still a lack of knowledge about the genetic

basis of leaf water content and the interaction between water contents of stems and leaves.

Increasing knowledge of the genetic control of these biomass-related traits in sorghum, and identification of corresponding genes and their functions, may lead to tools and strategies for either enhancing or suppressing these traits, supporting advances toward greater use of plant-based biomass in biofuel production.

1.3 Gene Mapping in Sorghum

Linkage mapping and association mapping are two important complementary approaches for investigating the genetic basis of bioenergy traits. Conventional biparental linkage mapping using molecular markers has been widely used in plant breeding with successes in many organisms (TANKSLEY *et al.* 1989; PATERSON 1996). However, biparental linkage mapping is constrained to assess only the alleles that differ between the two parents of a designed cross, and to offer relatively limited resolution. In contrast, association mapping can improve resolution by taking advantage of long-term accumulation of historical recombination events, but finds many spurious associations due to genetic relatedness and population structure. Also, the precision available in association mapping can be limited for particular local regions with strong LD in inbreeding organisms such as sorghum (MORRELL *et al.* 2012). Thus, using a combination of linkage and association mapping approaches can be beneficial to gene identification by taking advantage of the relatively high detection power of linkage mapping and the relatively high resolution of association mapping, mitigating constraints of each approach (TANG *et al.* 2013).

1.4 Overview of Dissertation Chapters

The aim of this dissertation is to explore in sorghum several biomass-related traits, including the 'dry' trait, stem architecture and water content components using a combination of linkage and association genetic approaches. Our analysis provided a comprehensive description of genes and genomic regions responsible for these biomass-related traits as well as their interactions.

In Chapter 2, we described a *S. bicolor* x *S. propinquum* recombinant inbred line (RIL) population of 161 F5 genotypes that segregate for a wide range of traits. This genetic map provides a valuable addition to the genetic resources currently available for this botanical and genomic model.

In Chapter 3, we determined the precise location of the D gene for the 'dry' trait in sorghum by a combination of linkage mapping from a *S. bicolor* x *S. propinquum* F2 population and a *S. bicolor* BTx623 x IS3620c RIL population, and genome-wide association studies (GWAS) from a *S. bicolor* diversity panel (CASA *et al.* 2008) of 354 accessions; characterized the candidates for the D gene by targeted re-sequencing and functional analysis; and explored the evolutionary path of the preferred candidate by population genomics and phylogenetic approaches. The relationship between the 'dry' trait and plant moisture content as well as the race-specific pattern and geographic distribution of the 'dry' trait were also discussed.

In Chapter 4, we identified quantitative trait loci (QTLs) for three stem diameters and stem/leaf water contents in a *S. bicolor* BTx623 x IS3620c RIL population, and validated the QTLs in a *S. bicolor* diversity panel (CASA *et al.* 2008) using GWAS. The relationships of stem diameter and water content to other important biomass-related traits,

including plant height, flowering time, branching, and the ‘dry’ midrib trait were also discussed.

Chapter 5 concludes the dissertation by summarizing the results, and discussing the impacts of this work on future genomic studies. Taken together, these studies provide in sorghum a new genetic resource (RIL map) and a comprehensive analysis of the genetic basis of several important bioenergy traits as well as their interactions.

1.5 References

ATIENZA, S. G., Z. SATOVIC, K. K. PETERSEN, O. DOLSTRA and A. MARTIN, 2003 Identification of QTLs influencing agronomic traits in Miscanthus sinensis Anderss. I. Total height, flag-leaf height and stem diameter. Theoretical and Applied Genetics **107:** 123-129.

BURTON, G. W., 1986 Biomass production from herbaceous plant, pp. xiv, 667 p. in *Biomass energy development*, edited by W. H. SMITH. Plenum Press, New York.

CASA, A. M., G. PRESSOIR, P. J. BROWN, S. E. MITCHELL, W. L. ROONEY *et al.*, 2008 Community resources and strategies for association mapping in sorghum. Crop Science **48:** 30-40.

EIA, 2016 International Energy Outlook 2016, pp.

GORDON, C. H., 1967 Storage Losses in Silage as Affected by Moisture Content and Structure. Journal of Dairy Science **50:** 397-&.

GUAN, Y. A., H. L. WANG, L. QIN, H. W. ZHANG, Y. B. YANG *et al.*, 2011 QTL mapping of bio-energy related traits in Sorghum. Euphytica **182:** 431-440.

HAN, Y. C., P. LV, S. L. HOU, S. Y. LI, G. S. JI *et al.*, 2015 Combining Next Generation Sequencing with Bulked Segregant Analysis to Fine Map a Stem Moisture Locus in Sorghum (Sorghum bicolor L. Moench). Plos One **10**.

HART, G. E., K. F. SCHERTZ, Y. PENG and N. H. SYED, 2001 Genetic mapping of Sorghum bicolor (L.) Moench QTLs that control variation in tillering and other morphological characters. Theoretical and Applied Genetics **103:** 1232-1242.

KASHIWAGI, T., and K. ISHIMARU, 2004 Identification and functional analysis of a locus for improvement of lodging resistance in rice. Plant Physiology **134:** 676-683.

LI, R. H., M. A. LYONS, H. WITTENBURG, B. PAIGEN and G. A. CHURCHILL, 2005 Combining data from multiple inbred line crosses improves the power and resolution of quantitative trait loci mapping. Genetics **169:** 1699-1709.

LV, P., G. S. JI, Y. C. HAN, S. L. HOU, S. Y. LI *et al.*, 2013 Association analysis of sugar yield-related traits in sorghum [Sorghum bicolor (L.)]. Euphytica **193:** 419-431.

MORRELL, P. L., E. S. BUCKLER and J. ROSS-IBARRA, 2012 Crop genomics: advances and applications. Nature Reviews Genetics **13:** 85-96.

PATERSON, A. H., 1996 *Genome mapping in plants*. Academic Press R.G. Landes Co.

ROONEY, W. L., J. BLUMENTHAL, B. BEAN and J. E. MULLET, 2007 Designing sorghum as a dedicated bioenergy feedstock. Biofuels Bioproducts & Biorefining-Biofpr **1:** 147-157.

SABALLOS, A., 2008 Development and utilization of sorghum as a bioenergy crop., pp. xxi, 449 p. in *Genetic improvement of bioenergy crops*, edited by W. VERMERRIS. Springer, New York.

SHIRINGANI, A. L., M. FRISCH and W. FRIEDT, 2010 Genetic mapping of QTLs for sugar-related traits in a RIL population of Sorghum bicolor L. Moench. Theoretical and Applied Genetics **121:** 323-336.

STROADE, J., M. BOLAND and M. TAYLOR, 2016 AGMRC Sorghum profile, pp.

TANG, H. B., H. E. CUEVAS, S. DAS, U. U. SEZEN, C. B. ZHOU *et al.*, 2013 Seed shattering in a wild sorghum is conferred by a locus unrelated to domestication. Proceedings of the National Academy of Sciences of the United States of America **110:** 15824-15829.

TANKSLEY, S. D., N. D. YOUNG, A. H. PATERSON and M. W. BONIERBALE, 1989 Rflp Mapping in Plant-Breeding - New Tools for an Old Science. Bio-Technology **7:** 257-264.

TOMES, D. T., P. LAKSHMANAN and D. SONGSTAD, 2011 *Biofuels : global impact on renewable energy, production agriculture and technological advancements*. Springer, New York.

VERMERRIS, W., 2011 Survey of Genomics Approaches to Improve Bioenergy Traits in Maize, Sorghum and SugarcaneFree Access. Journal of Integrative Plant Biology **53:** 105-119.

XU, W., P. K. SUBUDHI, O. R. CRASTA, D. T. ROSENOW, J. E. MULLET *et al.*, 2000 Molecular mapping of QTLs conferring stay-green in grain sorghum (Sorghum bicolor L. Moench). Genome **43:** 461-469.

CHAPTER 2

GENETIC ANALYSIS OF RECOMBINANT INBRED LINES FOR *SORGHUM BICOLOR* X *SORGHUM PROPINQUUM* [1]

[1] Kong W*, Jin H*, Franks CD et al. 2013. *G3: Genes | Genomes | Genetics* 3: 101-108. (*co-first authors)

Reprinted here with permission of the publisher.

2.1 Abstract

We describe a recombinant inbred line (RIL) population of 161 F5 genotypes for the widest euploid cross that can be made to cultivated sorghum (*S. bicolor*) using conventional techniques, *S. bicolor* x *S. propinquum*, that segregates for many traits related to plant architecture, growth and development, reproduction and life history. The genetic map of the *S. bicolor* x *S. propinquum* RILs contains 141 loci on 10 linkage groups collectively spanning 773.1 cM. Although the genetic map has DNA marker density well-suited to quantitative trait loci mapping and samples most of the genome, our previous observations that sorghum pericentromeric heterochromatin is recalcitrant to recombination is highlighted by the finding that the vast majority of recombination in sorghum is concentrated in small regions of euchromatin that are distal to most chromosomes. The advancement of the RIL population in an environment to which the *S. bicolor* parent was well adapted (indeed bred for), but the *S. propinquum* parent was not, largely eliminated an allele for short-day flowering that confounded many other traits, for example permitting us to map new quantitative trait loci for flowering that previously eluded detection. Additional recombination that has accrued in the development of this RIL population may have also improved resolution of apices of heterozygote excess, accounting for their greater abundance in the F5 than the F2 generation. The *S. bicolor* x *S. propinquum* RIL population offers advantages over early-generation populations that will shed new light on the genetic, environmental, and physiological/biochemical factors that regulate plant growth and development.

2.2 Introduction

As a botanical and genomic model for grasses, *Sorghum bicolor* L. Moench. (sorghum), a native of tropical Africa that is the most drought-resistant of the world's top five cereal crops, is a logical complement to the largely-sequenced genome of rice (*Oryza*). Sorghum has biochemical and morphological specializations to improve carbon assimilation at high temperatures (C4 photosynthesis), whereas rice uses C3 photosynthesis more typical of temperate grasses. Like rice, the most recent genome duplication in sorghum appears to be ~70 million years ago (PATERSON *et al.* 2004), simplifying its comparative and functional genomics. With a high-quality DNA sequence (PATERSON *et al.* 2009), the ~740 megabase pair sorghum genome is of high value for better understanding the genome of maize (SCHNABLE *et al.* 2009), and in particular the impact of an ancient tetraploidy in maize shortly after its lineage diverged from that of sorghum (SWIGONOVA *et al.* 2004). Sorghum is of particular importance as a diploid model for the Saccharinae clade of grasses that includes recently formed complex polyploids such as *Saccharum* (sugarcane, currently the worlds no. 1 biofuel crop), and *Miscanthus*, among the greatest-yielding of biomass crops in the U.S. Midwest (HEATON *et al.* 2008). Each of these polyploids shares substantial genetic colinearity and synteny with sorghum (MING *et al.* 1998; KIM *et al.* 2012), and *Saccharum* quantitative trait loci (QTL) often show positional correspondence to those of sorghum (MING *et al.* 2001; MING *et al.* 2002). One of the few crops suited to all proposed approaches for renewable fuel production. i.e. from starch, sugar, and/or cellulose, sorghum itself is presently the no. 2 U.S. source of fuel ethanol from grain (after maize), and is a promising cellulosic biofuel crop (ROONEY *et al.* 2007).

Sorghum bicolor x *Sorghum propinquum* is thought to be the widest euploid cross that can be made with the cultigen (*S. bicolor*) by conventional means, and interspecific populations from these species offer opportunities to genetically dissect a wide range of traits related to plant domestication and crop productivity, some of which have begun to receive attention (CHITTENDEN *et al.* 1994; LIN *et al.* 1995; PATERSON *et al.* 1995a; PATERSON *et al.* 1995b; LIN *et al.* 1999; HU *et al.* 2003; FELTUS *et al.* 2006). The opportunities offered by comparison of *S. bicolor* and *S. propinquum* have led to much effort to develop genomics resources including a detailed genetic map (CHITTENDEN *et al.* 1994; BOWERS *et al.* 2003), bacterial artificial chromosome-based physical maps for both species (LIN *et al.* 1999; DRAYE *et al.* 2001; BOWERS *et al.* 2005), expressed sequence tag (EST) resources (PRATT *et al.* 2005), and a genome sequence (PATERSON *et al.* 2009).

Among many other aspects of growth and development, *S. bicolor* and *S. propinquum* differ in characteristics related to perenniality, a life history strategy for which the *Sorghum* genus has become a model (PATERSON *et al.* 1995b; HU *et al.* 2003; JANG *et al.* 2009). Both consideration of how to expand agriculture to provide plant biomass for production of fuels or chemical feedstocks (TILMAN *et al.* 2009), and strategies to rebalance food production with preservation of ecological capital (GLOVER *et al.* 2010), focus heavily on perenniality. Perenniality may also be a curse -- *Sorghum halepense*, a wild perennial polyploid resulting from natural hybridization between *S. bicolor* and *S. propinquum*, finds occasional use as forage and even food (seed/flour), but is most noted as one of the world's most noxious weeds, having spread from its west Asian center of diversity across much of Asia, Africa, Europe, North and South America,

and Australia. Demonstration that most genes responsible for variations in size and number in *Sorghum* and *Oryza* of an important perennation organ, the rhizome, map to corresponding chromosomal locations (HU *et al.* 2003), suggests that information about rhizomatousness from a few models (that are also major crops) may extrapolate broadly to a wide range of taxa.

By single-seed descent from the same *S. bicolor* x *S. propinquum* F2 population used in early-generation genetic analysis (LIN *et al.* 1995), we have produced and describe here a recombinant inbred line (RIL) population of 161 F5 genotypes that segregate for a wide range of traits, providing a valuable addition to the genetic resources available for this botanical and genomic model. The genetic control of flowering provides an example of how the RIL population contributes to improved knowledge of trait inheritance.

2.3 Materials and Methods

2.3.1 Genotyping and data analysis

The mapping population comprised 161 F5 RILs derived by selfing of single F2 plants described previously (LIN *et al.* 1995), from a controlled cross between single plants of *S. bicolor* BTx623, and *S. propinquum* (unnamed accession). Leaf samples were frozen at -80°C and lyophilized for 48 hours. Genomic DNA was extracted from the lyophilized leaf sample based on Aljanabi et al (1999). Polymerase chain reactions for simple sequence repeat (SSR) analysis were carried out under standard conditions for all primer pairs using 1 U of Taq polymerase with 10X polymerase chain reaction buffer (100 mM Tris-HCl at pH 9, 500 mM KCl, and 15 mM $MgCl_2$), 2 mM dNTP, 3 mM

$MgCl_2$, 0.2 mM of each primer, and 20 ng of DNA template with a final reaction volume of 10 mL. The thermo-cycling was performed with the following program: (1) Preheat at 95°C for 3 minutes, (2) denaturation at 95°C for 30 seconds, (3) annealing at 65°C for 1 minute (-1°C/ cycle), (4) extension at 72°C for 1 minute, (5) 10 cycles of steps (2) ~ (4), (6) denaturation at 95°C for 30 seconds, (7) annealing at 55°C for 1 minute, (8) extension at 72°C for 1 minute, (9) 32 cycles of steps (6) ~ (7), and (10) final extension at 72°C for 5 minutes. The amplified products were visualized in 10% polyacrylamide gels with silver staining.

2.3.2 Linkage and QTL analysis

A total of 161 F5 individuals were genotyped. MAPMAKER (LANDER *et al.* 1987) was used for map construction with the data type 'ri self', which is suitable for the RIL configuration. Heterozygosity in codominant markers was treated as missing data by MAPMAKER because the 'ri self' configuration does not recognize it. Map distances, cM, were calculated using the Kosambi function (KOSAMBI 1944). Marker loci were grouped by two-point linkage analysis with a logarithm of odds ratio (LOD) threshold of 4.0 and a maximum distance of 30 cM. Local maximum likelihood orders of marker loci were confirmed using the 'ripple' command. The map was drawn using Adobe Illustrator. In 2009, 2010, and 2011, single 1.5 m plots of each RIL were transplanted (2009, 2011) or direct seeded at the University of Georgia Plant Science Farm, Watkinsville, GA, in a completely randomized design. Flowering dates were recorded for the first five flowers per plot. The average of the first five flowering days was calculated in Microsoft Excel. The means of the flowering dates over years were estimated using best linear unbiased

prediction (BLUP) with SAS PROC MIXED. Lines, environmental effect, and their interaction were treated as random. The broad sense heritability (H) was calculated using the variance component method ($H = V_G/(V_G + \frac{V_{GE}}{e} + \frac{V_{residual}}{re})$). Heritability = 60.822/ (60.822 + 102.57/3 + 1.5848/3) = 63.66. QTL analysis used composite interval mapping method in Windows QTL Cartographer V2.5_010 (WANG *et al.* 2011).

Seed of the RIL population are distributed by the U.S. Department of Agriculture-Agricultural Research Service (USDA-ARS), Lubbock, TX (J. Burke).

2.4 Results

2.4.1 DNA markers and map construction

A total of 203 SSRs initially were selected and scored, derived from sugarcane ESTs (prefix "CA" or "TC"), previously mapped RFLP probe sequences ["Xcup" (SCHLOSS *et al.* 2002)], sorghum sequenced genomic clones ["Xtxp" (KONG *et al.* 2000)], sorghum EST sequences ["Xisep" (RAMU *et al.* 2009)], previously developed SSRs ["Xgap" (BROWN *et al.* 1996)], unpublished SSRs from Agropolis-Cirad-Genoplante ("mSbCIR"), and an unmapped scaffold in the genome sequence. Of those 203 markers, 135 segregating for 141 marker loci were mapped into 10 linkage groups corresponding to the 10 sorghum chromosomes. The remaining markers were excluded due to redundancy (i.e. cosegreration of multiple bands from the same primer), and weak and/or apparent artifactual amplifications. Among the 141 loci mapped in the F5 RILs, there is an average of 9 (5.6%) missing genotypes per locus, with 95% of the loci having less than 29 (18%) missing genotypes. Among 95 loci mapped in the F2 population, there is

an average of 25 (6.8%) missing genotypes per locus, with 95% of the loci having less than 103 (27.8%) missing genotypes.

The genetic map of the RILs derived from annual *S. bicolor* and perennial *S. propinquum* (Figure 2.1) contains 141 loci on 10 linkage groups collectively spanning 773.1 cM. A total of 35 (24.8%) loci have dominant inheritance, with null alleles from *S. propinquum* at 14 loci and from *S. bicolor* at 21 loci, which is not a significant difference (χ^2=1.4, 1 d.f, P=0.2367). The average interval between consecutive loci is 5.48 cM, ranging from 0.0 cM between cosegregating markers to 25.7 cM in the largest gap (on chromosome 5). Construction of the map used a two-step strategy. First, to minimize ambiguity caused by distorted loci, we constructed a framework map by selecting a subset of clearly scored markers that also did not deviate significantly from the expected Mendelian ratio (1:1) at $P<10^{-5}$ after Bonferroni correction. To assign linkage groups to chromosomes, we anchored framework markers to physical locations by blasting against the sorghum genome sequence. We then assigned and placed additional markers to the framework at LOD score of $\geq$3.0 and carefully checked for double recombination events in the original scoring data.

2.4.2 Comparison of genetic map to physical positions

Comparison of the genetic map to physical positions of the mapped loci reveals the relationship between genetic distances and physical distances, and the physical distribution of markers along the genome. Each marker on the genetic map was aligned to its corresponding physical position by virtue of the published *S. bicolor* genome sequence (PATERSON *et al.* 2009) (Figure 2.1). The sorghum genome sequence information was

given priority in ordering markers that were indistinguishable genetically. Overall, a total of 110 of 141 markers in 10 linkage groups are well aligned to their physical positions. The marker order in the genetic map occasionally deviates from that in the physical map. Distal markers on a linkage group tend to disagree more with the physical map than markers in the middle of a group, presumably due to lack of flanking markers at the ends of chromosomes. A small number of markers show best matches to sequences that are on different chromosomes from where they map genetically. Factors that may contribute to the discrepancies between the genetic and physical maps include multiple amplifications of paralogous loci; sequence assembly errors; or cryptic structural differences between *S. bicolor* and *S. propinquum*.

The mapped SSR marker loci provide substantial coverage of the genetic map, with the exception of chromosome 6 for which markers only cover the lower one third of the chromosome (Figure 2.1). The unmapped region of this chromosome includes a large heterochromatic block (about 34 Mb) that contains the *S. propinquum Ma1* allele conferring short-day flowering (LIN *et al.* 1995). There is ample polymorphism between the parental genotypes in this region and we mapped the region in the F2 population (LIN *et al.* 1995). However, the RIL population was advanced in a temperate latitude, and artificial selection has largely eliminated photoperiodic flowering. This selection, together with limited recombination in this heterochromatic region, account for it being under-populated with DNA markers in the RIL map.

Marker distribution is not even along the physical map: markers are concentrated in distal regions and sparse in central regions of the chromosomes. In an extreme case, chromosome 8, a recombinational distance of 4.7 cM spans a physical distance of

approximately 46.3 Mb, covering a remarkable 83.4% of the chromosome. This phenomenon is in accordance with our previous observations (LIN *et al.* 1995; BOWERS *et al.* 2005; PATERSON *et al.* 2009) that the sorghum pericentromeric heterochromatin is recalcitrant to recombination, with the vast majority of recombination occurring in the distal euchromatin.

2.4.3 Segregation distortion

In the F5 RILs, all chromosomes except chromosome 7 contain regions with segregation distortion significant at the 5% level (Table 2.1). A total of 14 apices (peak genomic regions) of distortion were found, on chromosome 1 near cM 35.8, chromosome 2 near cM 50.6, chromosome 3 near cM 11.2, 35.0, 66.1, and 84.3; chromosome 4 near cM 77.2, chromosome 5 near cM 0.0 and 60.3; chromosome 6 near cM 0.0; chromosome 8 near cM 39.5; chromosome 9 near cM 26.5 and 37.7; and chromosome 10 near cM 88.7. All regions showed enrichment for *S. bicolor* alleles. Other than the chromosome 6 region under selection for day-neutral flowering, the most striking case of segregation distortion was on chromosome 1 – the apex of this distortion was near the locus Xcup24 with a segregation ratio of 154:3 (homozygous *S. bicolor*: *S. propinquum*). This apex was genetically less than 1 cM from the most extreme case found in the F2 population from which these RILs are derived: the locus CSU507 on LG C (BOWERS *et al.* 2003). In a larger set of F2 progeny previously described (LIN *et al.* 1995), we found similarly distorted segregation (203:15) in this region.

We compared the 14 regions of segregation distortion in the F5 RILs to the levels and patterns of segregation found in the F2 population from which these RILs are

derived. Because different DNA markers were used in the two studies, this was done by aligning the F2 and F5 genetic maps to their physical locations on the *S. bicolor* genome (PATERSON *et al.* 2009). A total of 11 regions of segregation distortion were found in the F2 (Table 2.1). Four of the 11 regions of segregation distortion in the F2 population favored the *S. propinquum* alleles, among which, three are no longer distorted in the F5 RILs, and one region near the end of chromosome 2 contains over-representation of *S. bicolor* alleles(!) in the F5 RILs. Those regions with over-representation of *S. bicolor* alleles in the F2 generally also contain such overabundance in the F5 RILs, albeit a few cases lack nearby DNA markers. However, eight regions showing normal segregation in the F2 showed overabundance of the *S. bicolor* allele in the F5 RILs.

2.4.4 Residual heterozygosity

We compared regions of excess/deficiency of residual heterozygosity in the F5 RILs and the F2 population (Table 2.2). In the F2, eight regions show excess and two show deficiency of heterozygotes. All except two of these also show segregation distortion. In the F5, much higher homozygosity makes it difficult to distinguish heterozygote deficiency with statistical significance but 29 regions show excess, 7 (24%) of which also show segregation distortion. In the F2, the regions showing excess are all small (diagnosed by only 1 marker each), however a large region of chromosome 1 shows deficiency of heterozygotes. In the F5, there are 3 large regions showing heterozygote excess in chromosomes 4, 5, and 7 respectively.

2.4.5 Initial QTL mapping

To explore the merit of the RIL population for QTL mapping, we focused on flowering, a trait associated with the tropical origin of *S. propinquum* that had a large confounding effect on many traits in F2 QTL mapping. In the RIL population, near-homozygosity for the *S. bicolor* allele along the salient portion of chromosome 6 reveals that we have largely eliminated genotypes with short-day flowering alleles from *S. propinquum*. A total of three flowering QTLs met a LOD threshold of 2.61 based on 1000 permutation tests, on chromosomes 4, 8, and 9 (Figure 2.2, Table 2.3). The chromosome 9 QTL found here closely overlaps one found in the F2 generation (LIN *et al.* 1995), which also overlaps a QTL found in several other sorghum populations (FELTUS *et al.* 2006; MACE and JORDAN 2011). The chromosome 8 QTL also closely corresponds to one found in the BTx623 x IS3620c cross (BROWN *et al.* 2006), and the *S. propinquum* allele confers early flowering, accounting for the transgressants we observed in F2 and F5. The chromosome 4 QTL is newly discovered in this population, perhaps "unmasked" as a result of removing short-day flowering but is in a region in which flowering QTL have been reported previously (MACE and JORDAN 2011). Indeed, it shows a 'double peak' that may indicate the actions of two nearby genes although we presently infer only a single likelihood interval with statistical confidence. Although a previously-reported QTL on chromosome 2 (LIN *et al.* 1995) did not reach statistical significance here, there was sub-threshold evidence of it (LOD ~1) in the vicinity that it was previously mapped to.

2.5 Discussion

The *S. bicolor* x *S. propinquum* RIL population offers advantages over early-generation populations that promises to shed new light on the genetic, environmental, and physiological/biochemical factors that regulate plant growth and development. Dramatic variation in plant architecture, growth and development, reproduction and life histories of the parental species that segregate among the progeny, together with homozygosity of the RILs and the ability to replicate them across a spectrum of natural and/or controlled conditions, makes this population of high potential importance for the discovery and validation of QTLs. Many of these traits have been measured and will be reported under separate cover.

Advancement of the RIL population in a temperate environment (Lubbock, TX) may improve the ability to resolve QTL for traits that were previously below the significance threshold, also providing a more realistic assessment of variation that is relevant to temperate latitudes. For example, near-homozygosity for the *S. bicolor* allele along the salient portion of chromosome 6 reveals that we have largely eliminated genotypes with photoperiod sensitivity ("short-day flowering"), a trait associated with the tropical origin of *S. propinquum* that had a large confounding effect on many traits in F2-based QTL mapping. Eliminating the profound morpho-physiological alteration associated with short-day flowering permitted us to identify two flowering QTLs that eluded detection in our previous study with 370 F2 plants (LIN *et al.* 1995), one of which accounted for the observation that a few segregants flowered earlier than the early-flowering parent.

The advancement of the RIL population in an environment to which the *S. bicolor* parent was well adapted (indeed bred for), but the *S. propinquum* parent was not, may have had some undesirable consequences as well. All segregation distortions in the F5 generation involved excesses of *S. bicolor* alleles, while the F2 generation showed similar numbers of cases of *S. bicolor* and *S. propinquum* excess. This finding suggests that in addition to the intended removal of short-day flowering, advancement of the population in temperate continental conditions may have caused some inadvertent selection against other traits of *S. propinquum*, a native of southeast Asia that inhabits streamsides and moist places (ANONYMOUS, zipcodezoo.com/Plants/S/Sorghum_propinquum/#footref_2). While these biases favoring *S. bicolor* alleles may impact the ability to map QTLs in a few regions of the genome, the population still exhibits a wide range of morpho-physiological variations, with individual lines more comparable to one another by virtue of the near-absence of *Ma1*.

Benefiting from several additional cycles of recombination beyond our previous F2 population, comparison of this genetic map to the sorghum physical map and sequence highlight the striking bias in distribution of recombination across the sorghum genome. This is a good news-bad news scenario –relatively small amounts of physical DNA per cM may facilitate genomic analyses in the gene-rich portions of the genome, but large blocks of recombination-recalcitrant heterochromatin hinder access to other important genes.

Additional recombination that has accrued in the development of this RIL population may have also improved our ability to resolve apices (peak genomic region) of heterozygote excess, accounting for their greater abundance in F5 than F2 generations,

and occurrence in multiple locations on all chromosomes except the one (chromosome 6) for which about two-thirds of the physical length has been fixed due to selection against the *S. propinquum* short-day flowering allele. A remarkably high 29 apices of heterozygote excess, together with rich genetic and genomic tools for these species, may make this an attractive system in which to further dissect the biology underlying interspecific heterozygote advantage.

2.6 Acknowledgements

We thank members of the Paterson lab and two anonymous reviewers for many valuable contributions, and the US Department of Agriculture (USDA) Agricultural Research Service, USDA *Agriculture and Food Research Initiative* and US Department of Energy – USDA Plant Feedstock Genomics Program for financial support. R.B. gratefully acknowledges financial support as a BOYSCAST Fellow from the Department of Science and Technology, Government of India (SR/BY/L-08/2007). M.K.R. is thankful to the National Agricultural Innovative Project, Indian Council of Agricultural Research for providing financial support for his participation in this work.

Table 2.1. Comparison of regions of segregation distortion between *S. bicolor* (SB) x *S. propinquum* (SP) F5 RIL and F2 populations

	F5				F2			
Chr.	Marker	cM	SB:SP	Location, Mb	Marker	SB:SP	LG	Location, Mb
1	Xcup24	35.8	154:3	14.0	pSB195	203:15	C	14.2
2	CA154181a	50.6	129:24	58.3	pSB101	34:74	B	61.6
2	None near				pSB075	111:35	B	66.1
3	CA199661a	11.2	101:40	0.2	None near		A	
3	CA074959a[a]	35.0	100:23	3.5-6.2		N.D.	A	
3	TC48056a[a]	66.1	86:43	13.8-51.2		N.D.	A	
3	Xcup65	84.3	85:39	55.9		N.D.	A	
3	None near				pSB443b	128:66	A	69.0
4	Xtxp265	77.2	117:37	64.9	None near		F	
4	N.D.				pSB038	34:101	F	14.2
5	mSbCIR329	0.0	134:26	0.2	None near		H	
5	N.D.				pSB064	33:81	H	6.5
5	Xisep1140	60.3	97:52	54.8		N.D.	H	
6	Xgap72	0.0	151:3	41.4	pSB095	104:60	D	50.7
6	None near				pSB428a	93:41	D	38.0
6	None near				pSB643a	65:20	D	4.3
7	N.D.				pSB784	19:50	J	5.9
8	mSbCIR240	39.5	109:49	4.5		N.D.	E	
9	CA142735a	26.5	115:46	31.7		N.D.	G	
9	TC59518b[a]	37.7	105:55	50.2-54.5		N.D.	G	
10	Xcup43	88.7	123:26	59.8	pSB115	124:59	I	60.6

N.D. No distortion (not significantly different from 1:1 segregation)
[a] Physical location not on the corresponding chromosome of the linkage group: apices are estimated by adjacent marker locations.

Table 2.2. Comparison of regions showing over-/under-representations of residual heterozygosity between *S. bicolor* x *S. propinquum* F5 RIL and F2 populations

	F5				F2			
Chr.	marker	cM	H:(SB+SP) [a]	Location, Mb	marker	H:(SB+SP) [a]	LG	Location, Mb
1					pSB102	193:123	C	3.7
1	CA226478a	5.4	21:140	1.8				
1	TC71756a[b]	97.7	24:135	47.7/50.3-73.4	SHO68	106:244[c]	C	46.8
2	Xcup67[b]	12.7	26:133	0.6-2.0				
2	CA296025c	18.2	26:130	2.4				
2	mSbCIR223	37.4	26:131	4.7				
2	Xcup63	51.6	25:136	59.1				
2					pSB101	206:108	B	61.6
2					pSB077	211:103	B	70.0
3	CA152937a	50.2	24:132	7.1				
3	TC48056a[b]	66.1	28:129	13.8-51.2				
3	TC69429a[b]	93.4	21:135	55.9-end				
3					pSB443b	81:194 [c]	A	69.0
4	Xcup61[b]	12.5	36:120	1.5-5.1				
4	Xisep0203	42.5	26:123	10.0				
5	Xtxp065	13.5	23:131	1.9				
5	CA103773a	39.4	25:135	4.8				
5					pSB064	191:114	H	6.5
5	S14_284514_ag17[b]	52.1	41:116	13.5-42.0				
5	CA100232a	55.6	31:130	54.3				
6					pSB643a	178:85	D	4.3
6					pSB140	221:104	D	52.4
6	Xtxp057	26.4	22:134	57.4				
6					pSB487	194:113	D	60.1

	F5				F2			
Chr.	marker	cM	H:(SB+SP)[a]	Location (Mb)	marker	H:(SB+SP) [a]	LG	Location (Mb)
7	Xtxp040	0	28:129	0.9				
7					pSB784	127:69	J	5.9
7	Xtxp278	32.7	26:129	51.1				
7	mSbCIR300	45.6	24:137	58.3				
7	Xisep0829	47.8	31:122	59.4				
8	Xtxp047	29.6	31:126	3.0				
8	CA166256a	51.2	23:137	5.3				
8	Xtxp321	68.1	22:134	50.5				
9	TC50663d	0	20:120	3.0				
9	TC65153a	43.3	23:114	54.5				
9	Xgap206	63.6	21:137	59.2				
10	Xcup49	0	30:130	0.2				
10	CA217392a	26.8	21:132	1.6				
10	CA191677a	81.9	21:139	58.6				

RIL, recombinant inbred line; LG, linkage group.

[a] H:(SB + SP) indicates the ratio of heterozygotes (H) to the sum of parental genotypes (SB + SP).

[b] Physical location not on the corresponding chromosome of the linkage group: apices are estimated by adjacent marker locations.

[c] Deficiency of heterozygotes (all other cases are heterozygote excess).

Table 2.3. Biometric parameters of QTL for days to first flower in the *S. bicolor* x *S. propinquum* RILs

Chr.	LOD	a[a]	R^2	Start (Mb)[b]	End (Mb) [b]	Population/study (if not herein)
2	1	-1.20	0.028	--	--	
2	4.67	-6.70	0.083	61.6	66.1	*S. bicolor x S. propinquum* F2 (LIN *et al.* 1995)
4	3.5	-2.01	0.108	5.1	10.0	
8	2.8	1.64	0.072	49.1	51.9	
8	5.5	[c]	0.134	50.5	55.5	*S. bicolor* BTx623 x IS3620C (BROWN *et al.* 2006)
9	4.2	-2.14	0.114	50.2	54.5	
9	2.53	-10.50	0.042	8.1	57.0	*S. bicolor x S. propinquum* F2 (LIN *et al.* 1995)
9	7.7	--	0.195	[d]	59.1	*S. bicolor* BTx623 x IS3620C (FELTUS *et al.* 2006)

[a] Additive effect, calculated as *S. bicolor* BTx623 – other (*S. propinquum* or IS3620C as appropriate). To match this system, values reported in (LIN *et al.* 1995) were multiplied by -1.

[b] Based on flanking DNA marker locations in the published genome sequence (PATERSON *et al.* 2009b).

[c] Corresponding values not reported.

[d] Only a single nearby marker could be definitively mapped to the genome sequence, span of interval uncertain.

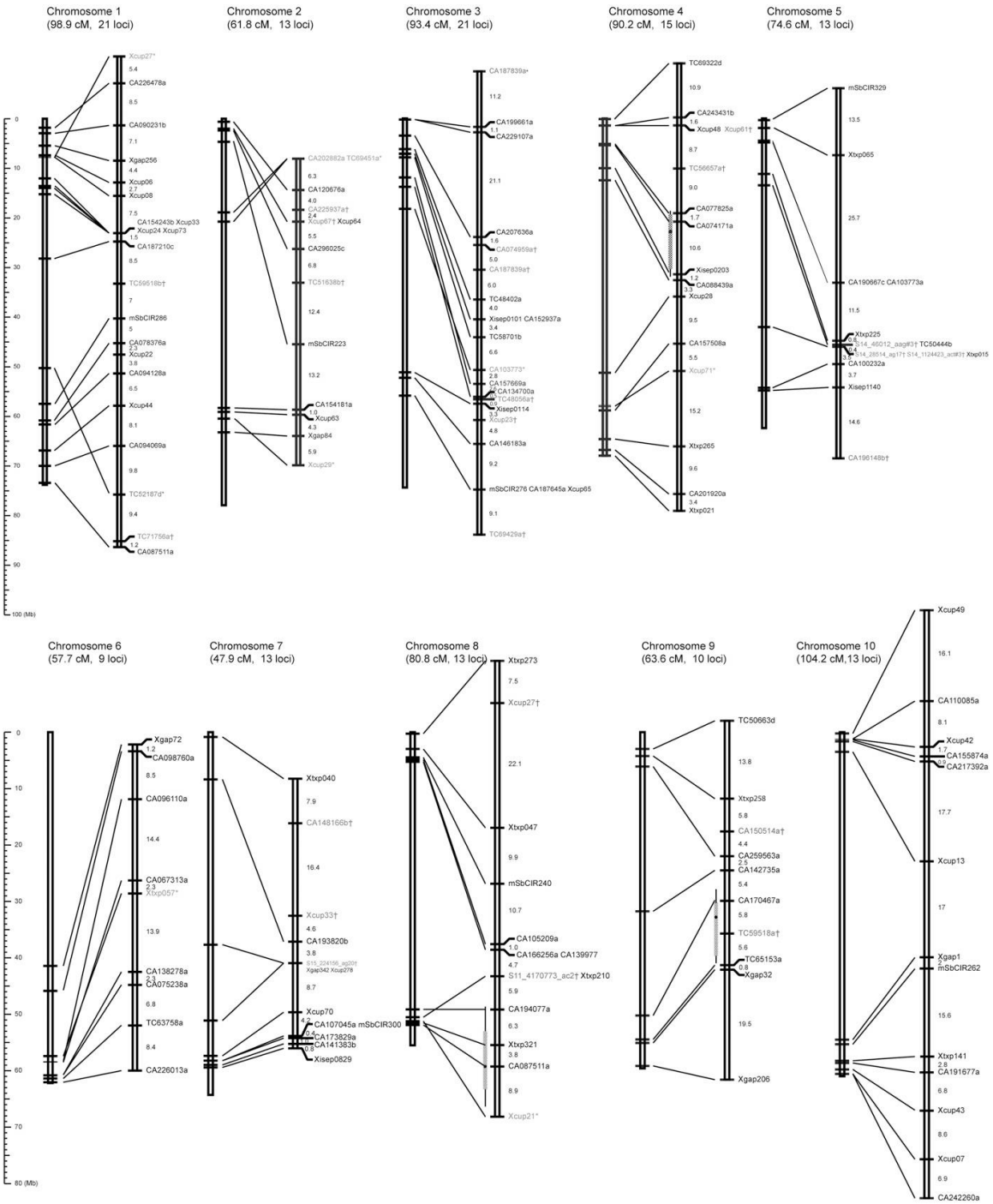


Figure 2.1. A genetic map of the *S. bicolor x S. propinquum* RIL. For each linkage group, genetic distances are shown on the right in Kosambi centimorgans; their corresponding physical chromosomes (from the current genome assembly, as cited) are shown on the left. Markers whose physical positions are unable to be located to their corresponding

chromosomes are indicated by "†"; markers whose genetic orders disagree with their physical positions are indicated by "*". Bar and whisker plots on chromosome 4, 8, and 9 indicate 1- and 2-LOD likelihood intervals for flowering QTL described in the text, with tick marks indicating likelihood peaks.

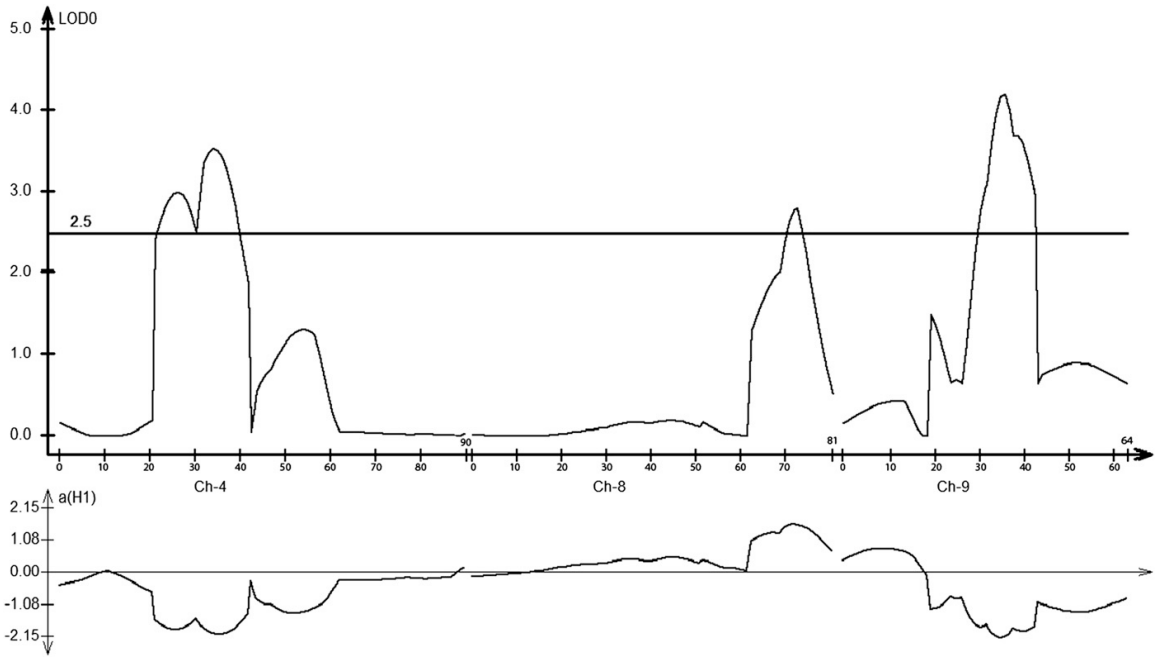


Figure 2.2. QTLs for days to first flower in the *S. bicolor* x *S. propinquum* RILs. Upper plot represents QTL likelihood (LOD scores) at the indicated cM locations on chromosomes 4, 8, and 9. Lower plot indicates additive effect of an allele substitution at the indicated cM locations, calculated based on flowering times associated with *S. bicolor* minus *S. propinquum* alleles (so positive value indicates earliness associated with the late-flowering *S. propinquum* parent).

2.7 References

ALJANABI, S. M., L. FORGET and A. DOOKUN, 1999 An improved and rapid protocol for the isolation of polysaccharide- and polyphenol-free sugarcane DNA. Plant Molecular Biology Reporter **17:** 281-281.

ANONYMOUS, Sorghum propinquum, pp. 600-601 in *Flora of China, Vol. 22*. Missouri Botanical Garden Press.

BOWERS, J. E., C. ABBEY, S. ANDERSON, C. CHANG, X. DRAYE *et al.*, 2003 A high-density genetic recombination map of sequence-tagged sites for sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. Genetics **165:** 367-386.

BOWERS, J. E., M. A. ARIAS, R. ASHER, J. A. AVISE, R. T. BALL *et al.*, 2005 Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. Proceedings of the National Academy of Sciences of the United States of America **102:** 13206-13211.

BROWN, P. J., P. E. KLEIN, E. BORTIRI, C. B. ACHARYA, W. L. ROONEY *et al.*, 2006 Inheritance of inflorescence architecture in sorghum. Theoretical and Applied Genetics **113:** 931-942.

BROWN, S. M., M. S. HOPKINS, S. E. MITCHELL, M. L. SENIOR, T. Y. WANG *et al.*, 1996 Multiple methods for the identification of polymorphic simple sequence repeats (SSRs) in sorghum [Sorghum bicolor (L) Moench]. Theoretical and Applied Genetics **93:** 190-198.

CHITTENDEN, L. M., K. F. SCHERTZ, Y. R. LIN, R. A. WING and A. H. PATERSON, 1994 A Detailed Rflp Map of *Sorghum-Bicolor* X *S-Propinquum*, Suitable for High-

Density Mapping, Suggests Ancestral Duplication of Sorghum Chromosomes or Chromosomal Segments. Theoretical and Applied Genetics **87:** 925-933.

DRAYE, X., Y. R. LIN, X. Y. QIAN, J. E. BOWERS, G. B. BUROW *et al.*, 2001 Toward integration of comparative genetic, physical, diversity, and cytomolecular maps for grasses and grains, using the sorghum genome as a foundation. Plant Physiology **125:** 1325-1341.

FELTUS, F. A., G. E. HART, K. F. SCHERTZ, A. M. CASA, S. KRESOVICH *et al.*, 2006 Alignment of genetic maps and QTLs between inter- and intra-specific sorghum populations. Theor Appl Genet **112:** 1295-1305.

GLOVER, J. D., J. P. REGANOLD, L. W. BELL, J. BOREVITZ, E. C. BRUMMER *et al.*, 2010 Increased Food and Ecosystem Security via Perennial Grains. Science **328:** 1638-1639.

HEATON, E. A., F. G. DOHLEMAN and S. P. LONG, 2008 Meeting US biofuel goals with less land: the potential of Miscanthus. Global Change Biology **14:** 2000-2014.

HU, F. Y., D. Y. TAO, E. SACKS, B. Y. FU, P. XU *et al.*, 2003 Convergent evolution of perenniality in rice and sorghum. Proc Natl Acad Sci U S A **100:** 4050-4054.

JANG, C. S., T. L. KAMPS, H. TANG, J. E. BOWERS, C. LEMKE *et al.*, 2009 Evolutionary fate of rhizome-specific genes in a non-rhizomatous Sorghum genotype. Heredity **102:** 266-273.

KIM, C., D. ZHANG, S. A. AUCKLAND, L. K. RAINVILLE, K. JAKOB *et al.*, 2012 SSR-based genetic maps of *Miscanthus sinensis* and *M. sacchariflorus*, and their comparison to sorghum. Theor Appl Genet **124:** 1325-1338.

KONG, L., J. DONG and G. E. HART, 2000 Characteristics, linkage-map positions, and allelic differentiation of Sorghum bicolor (L.) Moench DNA simple-sequence repeats (SSRs). Theoretical and Applied Genetics **101:** 438-448.

KOSAMBI, D., 1944 The estimation of map distance from recombination values. Annals of Eugenics**:** 172-175.

LANDER, E., P. GREEN, J. ABRAHAMSON, A. BARLOW, M. DALY *et al.*, 1987 MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics **1:** 174-181.

LIN, Y. R., K. F. SCHERTZ and A. H. PATERSON, 1995 Comparative analysis of QTLs affecting plant height and maturity across the Poaceae, in reference to an interspecific sorghum population. Genetics **141:** 391-411.

LIN, Y. R., L. H. ZHU, S. X. REN, J. S. YANG, K. F. SCHERTZ *et al.*, 1999 A Sorghum propinquum BAC library, suitable for cloning genes associated with loss-of-function mutations during crop domestication. Molecular Breeding **5:** 511-520.

MACE, E. S., and D. R. JORDAN, 2011 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. Theoretical and Applied Genetics **123:** 169-191.

MING, R., T. A. DEL MONTE, E. HERNANDEZ, P. H. MOORE, J. E. IRVINE *et al.*, 2002 Comparative analysis of QTLs affecting plant height and flowering among closely-related diploid and polyploid genomes. Genome **45:** 794-803.

MING, R., S.-C. LIU, J. E. IRVINE and A. H. PATERSON, 2001 Comparative QTL Analysis in a Complex Autopolyploid: Candidate Genes for Determinants of Sugar Content in Sugarcane. Genome Research **11:** 2075-2084.

MING, R., S. C. LIU, Y. R. LIN, J. DA SILVA, W. WILSON *et al.*, 1998 Detailed alignment of Saccharum and Sorghum chromosomes: Comparative organization of closely related diploid and polyploid genomes. Genetics **150:** 1663-1682.

PATERSON, A. H., J. E. BOWERS, R. BRUGGMANN, I. DUBCHAK, J. GRIMWOOD *et al.*, 2009 The Sorghum bicolor genome and the diversification of grasses. Nature **457:** 551-556.

PATERSON, A. H., J. E. BOWERS and B. A. CHAPMAN, 2004 Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proceedings of the National Academy of Sciences of the United States of America **101:** 9903-9908.

PATERSON, A. H., Y. R. LIN, Z. K. LI, K. F. SCHERTZ, J. F. DOEBLEY *et al.*, 1995a Convergent Domestication of Cereal Crops by Independent Mutations at Corresponding Genetic-Loci. Science **269:** 1714-1718.

PATERSON, A. H., K. F. SCHERTZ, Y. R. LIN, S. C. LIU and Y. L. CHANG, 1995b The Weediness of Wild Plants - Molecular Analysis of Genes Influencing Dispersal and Persistence of Johnsongrass, Sorghum-Halepense (L) Pers. Proceedings of the National Academy of Sciences of the United States of America **92:** 6127-6131.

PRATT, L. H., C. LIANG, M. SHAH, F. SUN, H. M. WANG *et al.*, 2005 Sorghum expressed sequence tags identify signature genes for drought, pathogenesis, and

skotomorphogenesis from a milestone set of 16,801 unique transcripts. Plant Physiology **139:** 869-884.

RAMU, P., B. KASSAHUN, S. SENTHILVEL, C. A. KUMAR, B. JAYASHREE *et al.*, 2009 Exploiting rice-sorghum synteny for targeted development of EST-SSRs to enrich the sorghum genetic linkage map. Theoretical and Applied Genetics **119:** 1193-1204.

ROONEY, W. L., J. BLUMENTHAL, B. BEAN and J. E. MULLET, 2007 Designing sorghum as a dedicated bioenergy feedstock. Biofuels Bioproducts & Biorefining-Biofpr **1:** 147-157.

SCHLOSS, S. J., S. E. MITCHELL, G. M. WHITE, R. KUKATLA, J. E. BOWERS *et al.*, 2002 Characterization of RFLP probe sequences for gene discovery and SSR development in Sorghum bicolor (L.) Moench. Theoretical and Applied Genetics **105:** 912-920.

SCHNABLE, P. S., D. WARE, R. S. FULTON, J. C. STEIN, F. S. WEI *et al.*, 2009 The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science **326:** 1112-1115.

SWIGONOVA, Z., J. S. LAI, J. X. MA, W. RAMAKRISHNA, M. LLACA *et al.*, 2004 On the tetraploid origin of the maize genome. Comparative and Functional Genomics **5:** 281-284.

TILMAN, D., R. SOCOLOW, J. A. FOLEY, J. HILL, E. LARSON *et al.*, 2009 Beneficial Biofuels-The Food, Energy, and Environment Trilemma. Science **325:** 270-271.

WANG, S., C. J. BASTEN and Z.-B. ZENG, 2011 Windows QTL Cartographer 2.5., pp. Department of Statistics, North Carolina State University, Raleigh, NC. (http://statgen.ncsu.edu/qtlcart/WQTLCart.htm).

CHAPTER 3

IDENTIFICATION AND CHARACTERIZATION OF THE SORGHUM 'DRY' GENE [1]

[1] Huizhe Jin, Hui Guo, Xu Tan, Rosana Compton, Uzay Sezen, Ayten Dizkirici and Andrew H. Paterson. To be submitted to *Molecular Biology and Evolution*.

3.1 Abstract

The ‘dry’ phenotype of sorghum, which reduces plant water content, may offer insights into phenotype evolution and be of practical value in biofuel crops. In *Sorghum bicolor*, the ‘dry’ trait is largely controlled by a single gene difference, with dry, white leaf midribs and dry, pithy stalks being genetically dominant to the alternative ‘juicy’ phenotype. This study characterizes the ‘dry’ trait in sorghum and identifies a probable candidate gene (Sobic.006G145300) by using a combination of linkage mapping and genome-wide association studies (GWAS), followed by targeted re-sequencing. The candidate gene, isoform *Dhr2* of beta-glucosidase (BGLU), is highly expressed in shoots, leaves and panicles, possibly being involved in functional processes such as lignification and cell wall catabolism. The ‘dry’ trait, to date known only in sorghum, appears to have evolved under balancing selection during a recent tandem expansion of the BGLU gene family in sorghum after its divergence from maize. Wider utilization of the ‘dry’ trait may make sorghum a more appealing raw material for biofuel production.

3.2 Introduction

The ‘dry’ phenotype of sorghum is an intriguing trait that has been suggested (SCHERTZ *et al.* 1978) to reduce plant water content. As manifested by white leaf midribs and dry, pithy stalks, the ‘dry’ phenotype is genetically dominant to the alternative ‘juicy’ phenotype, manifested by green midribs and dull, juicy stalks. This single-factor difference between dry and juicy phenotypes in sorghum was first observed in India (HILSON 1916). In general, sorghums with ‘dry’ (D-) genotypes have lower moisture content than those with ‘juicy’ genotypes (dd) (SCHERTZ *et al.* 1978) which is associated

with reduced storage loss from silage (GORDON 1967). Lower moisture content per unit biomass may also make 'dry' sorghums a preferred raw material for biofuel production, reducing energy lost to drying and transportation.

Sorghum is a favorable species to investigate the genetic basis of the 'dry' trait, since in a cultivated x wild sorghum cross one locus explains 100% of the phenotypic variance, with the cultivated sorghum (*S. bicolor*) being green/juicy and the wild sorghum (*S. propinquum*) being dry/white. In contrast, no 'dry/juicy' trait variation has been noted in other cereals to our knowledge.

Previous studies have shown the 'dry' gene (*D*) to co-segregate with an SSR marker Xtxp97 (HART *et al.* 2001), or reside between RFLP marker Xtxs1030 and Xumc34 on sorghum chromosome 6 (XU *et al.* 2000). With the aid of genomic resources for sorghum published in recent years, we can now attempt to identify the 'dry' gene. A high-density genetic map of sorghum (BOWERS *et al.* 2003), high quality sorghum sequence (PATERSON *et al.* 2009) and GBS SNP dataset (MORRIS *et al.* 2013) for a sorghum diversity panel (CASA *et al.* 2008) are now available. *Sorghum propinquum* serves as an ideal cross parent to investigate the 'dry' locus, since the highly polymorphic cross between temperate *S. bicolor* and tropical *S. propinquum* segregates for phenotypically significant variants at many genetic loci.

In this study, we use a combination of linkage and association genetic approaches to investigate the 'dry' gene. Biparental linkage mapping using molecular markers has been widely used in plant breeding with successes in many organisms (TANKSLEY *et al.* 1989; PATERSON 1996). However, biparental linkage mapping is constrained to assess only the alleles that differ between the two parents of a designed cross, and to offer

relatively limited resolution. In contrast, association mapping can improve resolution by taking advantage of high-throughput genotyping strategies and long-term accumulation of historical recombination events, but finds many spurious associations due to genetic relatedness and population structure. Also, the precision available in association mapping can be limited for particular local regions with strong LD in inbreeding organisms such as sorghum (MORRELL *et al.* 2012). Thus, using a combination of linkage and association mapping approaches can be beneficial to gene identification by taking advantage of the relatively high detection power of linkage mapping and the relatively high resolution of association mapping, mitigating constraints of each approach (TANG *et al.* 2013).

In this study we seek to identify the 'dry' gene in sorghum. Main questions addressed in the study include: 1) Determining the precise location of the D gene, integrating information from a *S. bicolor* x *S. propinquum* F2 population and a BTx623 x IS3620c RIL population using linkage mapping, and a sorghum diversity panel of 354 accessions using genome-wide association studies; 2) Characterizing the genetic and physiological basis of the D gene by targeted re-sequencing and functional analysis; 3) Exploring the evolutionary path of the D gene by population genomics and phylogenetic approaches. Other questions such as the relationship between the 'dry' trait and plant moisture content as well as the race-specific pattern and geographic distribution of the 'dry' trait are also discussed.

3.3 Materials and Methods

3.3.1 Genotype

The S. *bicolor* x S. *propinquum* F2 population and molecular methods for linkage mapping are as previously described (CHITTENDEN *et al.* 1994; BOWERS *et al.* 2003), except that the mapping population was expanded to 370 F2 progeny and used 94 approximately evenly-spaced RFLP probes (PATERSON *et al.* 1995). Prefixes of the DNA markers used and their sources are as follows: Sorghum P*st*I genomic DNA: pSB, SHO (A. H. Paterson); Maize cDNA: CSU (E. Coe, M. McMullen).

The genetic map of the *S. bicolor* BTx623 x IS3620c RIL population for QTL mapping was constructed utilizing a total of 399 individuals and 616 genotyping-by-sequencing (GBS) based SNP markers as described (Kong *et al.* unpublished data). It collectively spanned 1,404.8 cM on 10 linkage groups with a 3.8 cM average interval between consecutive markers.

Genotypes for GWAS were generated with GBS (ELSHIRE *et al.* 2011) across a US sorghum association panel (CASA *et al.* 2008), using the BTx623 genome (PATERSON *et al.* 2009) as a reference. A total of 265,487 SNPs from 354 accessions in 27,412 annotated genes were employed (MORRIS *et al.* 2013). For sorghum chromosome 6 (on which the target gene is located), there are a total of 23,224 SNPs with an average >350 SNPs per Mb.

Genotypes for analysis of selective signatures were generated as described (GUO 2015). These included 59 accessions, with a total of 10,482,224 SNPs identified.

3.3.2 Phenotyping.

Since the D gene affecting leaf midribs also influences stem juiciness (HILSON 1916; SWANSON 1931), the appearance of the midribs also indicates the nature of the stalks. In phenotyping the 'dry' trait, we classified leaf midribs for: 1) qualitative variation, including green/juicy midribs (G) and white/dry midribs (W); and 2) modifying variation, including presence/absence of a white stripe within a green midrib or presence/absence of a green cast over a white midrib (Figure 3.1).

Leaf midrib color develops as the plant grows. Sorghum seedlings and young leaves usually have green/juicy midribs. For 'dry' genotypes, the white midrib develops with initially a narrow white stripe emerging at the center of the green midrib and then spreading toward the margins gradually. In mature plants, leaf-midribs with white phenotypes are sometimes covered by green casts. On the other hand, mature leaf midribs with green phenotypes sometimes have a white stripe in the center (Figure 3.1b), making them less distinct from white midribs. In phenotyping the trait, we found that 1) white stripes usually emerge at a much later stage of plant development in green than in white midribs; and 2) margins of green midribs with white stripes are always green, while margins of white midribs are always white. These two observations aided the phenotying process. When analyzing trait data, lines having green midribs with white stripes were treated as G (green/juicy), while lines having white midribs with green casts were treated as W (white/dry).

Phenotypic data for the S. *bicolor* x S. *propinquum* F2 population was scored previously (1993) only for qualitative variation (G versus W). Phenotypic data for the S. *bicolor* BTx623 x IS3620c RIL population was scored in both 2011 and 2012. In 2011,

the ‘dry’ trait was scored twice, once in late June, once in mid August. More than 99% of the genotypes scored as green/juicy in June had developed white stripes within the green midrib by August. The sorghum diversity panel, planted near Watkinsville GA, was scored in 2009 and 2010.

Besides the ‘dry’ trait, the physiological maturity date of a plant was identified and recorded for the sorghum diversity panel based on the coloration of the primary inflorescence. At physiological maturity, fresh weights of leaves, stalks and inflorescences were measured and recorded for the S. *bicolor* BTx623 x IS3620c RIL population and the sorghum diversity panel, with dry weights measured subsequently after drying to stable mass in a tobacco barn. Plant moisture content was measured as the ratio of (fresh weight – dry weight)/fresh weight. Overall BLUP (Best Linear Unbiased Prediction) values of moisture content were used to detect QTLs, calculated for each line using the mixed model: moisture content ~ line + year + line * year + rep(year).

3.3.3 Linkage mapping

Map construction as well as linkage mapping of the ‘dry’ gene was carried out by using Mapmaker version 3.0 (LANDER *et al.* 1987) with the data type ‘F2 intercross’. Linkage groups were built by using two-point linkage analysis with a logarithm of odds ratio (LOD) threshold of 3.00 and a maximum distance of 50 cM. The map distances, cM, were calculated using the Kosambi function (KOSAMBI 1944). To further confirm the position of the ‘dry’ gene in the linkage group, the command ‘ripple’ was used to verify the local maximum likelihood order.

3.3.4 GWAS

Genome-wide association study was carried out using compressed mixed linear models (CMLM) as implemented in the Genomic Association and Prediction Integrated Tool (GAPIT) (LIPKA *et al.* 2012). GAPIT takes into account a genetic marker-based kinship matrix and a population structure term estimated by principal component analysis (ZHANG *et al.* 2010). The model selection feature of GAPIT was used to determine model parameters such as the compression level and optimal number of principal components (ZHU and YU 2009). Log quantile-quantile (QQ) plots were used to monitor systematic sources of spurious association. As the 'Dry' trait was treated as a qualitative binary trait, most associated SNPs identified with the CMLM model were further verified by fitting logistic regression models as implemented in PLINK 1.9 (CHANG *et al.* 2015). Bonferroni-like multiple testing correction (MATTHIES *et al.* 2014; ZHANG *et al.* 2015a) was also used to determine the significance threshold for GWAS. Taking the association test with each SNP as a single test, the traditional Bonferroni method corrects the significance threshold according to the number of independent tests carried out in the study. However, as many closely-linked SNPs are correlated, the total SNP number is much larger than the number of independent tests. Rather, the number of independent tests n can be estimated by using LD-decay information for each chromosome as $n = \sum_i \frac{l_i}{d_i}$, where l_i is the length of chromosome i, and d_i is the extension of LD for chromosome i. The significance threshold for GWAS can then be defined by α/n, where α is the desired significance level of all tests across the genome, and n is the estimated number of independent tests.

3.3.5 Targeted Re-sequencing

Genomic DNA of the diversity panel was extracted from frozen leaf samples by a published method (ALJANABI *et al.* 1999), and concentration measured by a NanoDrop 2000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA). Primer pairs were designed using Primer 3 (UNTERGASSER *et al.* 2012). Polymerase chain reactions (PCR) in 96-well plates were then carried out under standard conditions for all primer pairs (Table S3.1) using 1 μl of Pfu Taq polymerase (0.95 μl Taq and 0.05 μl Pfu) with 3 μl 10 $\times$ Pfu buffer, 3 μl dNTP, 1.8 μl $MgCl_2$, 1.5 μl of each primer, and 20ng of DNA template with a final reaction volume of 30 μl. AmpliTaq Gold 360 GC enhancer was added to some primers, targeting sequences of high GC content, to increase specificity. Thermal cycling was performed with the following program: (1) Preheat at 95°C for 5 min, (2) Denature at 95°C for 30 s, (3) Anneal at 65°C for 1 min (-1°C/cycle), (4) Extend at 72°C for 1 min, (5) 10 cycles of steps (2) to (4), (6) Denature at 95°C for 30 s, (7) Anneal at 55°C for 1 min, (8) Extend at 72°C for 1 min, (9) 43 cycles of steps (6) to (8), (10) Extend at 72°C for 5 min. Agarose gel electrophoresis was performed to separate PCR products, then re-sequencing using BigDye terminators was carried out, with resulting chromatograms examined by Sequencher v5.0.1 (Gene Codes Corporation, MI). Sequences with fewer than 20 bases of 20+ Phred quality scores were discarded. The remaining sequences were trimmed to maximize regions with error rates below 0.1 by CodonCode Aligner software (CodonCode Corporation, Dedham, MA, USA).

Re-sequencing was done first on a carefully chosen sub-panel of 48 sorghum lines and then across the whole diversity panel. The sub-panel (Table 3.1), comprising 24 lines with dry midribs and 24 with juicy midribs, were selected to cover a wide range of

geographic locations including Africa, Asia and North America, and at the same time, to avoid potential population structure. The gene candidates for re-sequencing were those that contained SNPs significantly associated with the phenotype, or within 2 genes immediately flanking significantly associated SNPs that fell into inter-genic regions. Re-sequencing over the sub-panel implied a few genomic fragments of interest, which were re-sequenced over the whole diversity panel for additional verification.

3.3.6 Selective Signatures

The Tajima's D parameter, calculated using libsequence C++ library v1.7.4 (THORNTON 2003), was used to investigate selective signatures in the D gene. The significance threshold was determined by 10,000 repeated simulations of populations under neutral evolution as described (GUO 2015).

3.3.7 Phylogenetic Analysis

The functional domain of a candidate D gene, beta-glucosidase, was used to search for gene family members using Pfam (FINN *et al.* 2016) with manual checking in a number of cereals (*Sorghum bicolor*, *Oryza sativa*, *Brachypodium distachyon*, *Setaria italica*, *Zea mays*), a cereal outgroup (*Musa acuminata*), and a eudicot (*Arabidopsis thaliana*) that is an important botanical model. The retrieved protein sequences were aligned and used to construct neighbor joining trees using Mega version 6.0 (TAMURA *et al.* 2013). Protein sequences shorter than 300 amino acids in length were removed from tree building.

3.3.8 Reference genomes

Gene annotations for *Sorghum bicolor* refer to JGI annotation release Sbi1.4 (PATERSON *et al.* 2009) for linkage and association mapping, and to JGI annotation release Sbi3.1 for genomic studies.

3.4 Results and Discussion

3.4.1 Map construction and initial linkage mapping

The genetic map of the F2 population from a cross of annual *S. bicolor* and perennial *S. propinquum* contains 94 loci on 10 linkage groups corresponding to 10 chromosomes (PATERSON *et al.* 1995). Among the 94 loci, there is an average of 25 (6.7%) missing genotypes per locus, with 95% of the loci having fewer than 18 (4.9%) missing genotypes.

This F2 population is a good base for 'dry' gene mapping, with the two highly polymorphic parents significantly different in phenotype, the *S. bicolor* parent being green/juicy and tropical *S. propinquum* being white/dry. Of the total of 370 F2 progenies studied, 263 showed 'dry' phenotypes with white midribs, and 107 showed 'juicy' phenotypes with green midribs. The result of a chi-squared test (p-value $= 0.0817$) supports the previous observation that the parents differ for a single gene at the significance level of $\alpha = 0.05$, with the null hypothesis being that this F2 population of *S. bicolor* x *S. propinquum* gives a fit that is not significantly different from the monohybrid ratio of 3:1.

The D gene was mapped to linkage group D at 80.9 cM, corresponding to sorghum chromosome 6 of the genome sequence (PATERSON *et al.* 2009), flanked by

markers pSB187 and pSB095 (Figure 3.2). This region has a total recombinational length of 5.3cM and physical length of 1.68Mb, containing 180 annotated genes (Table 3.2). Other studies also reported the D gene to be in this region, co-segregating with an SSR marker Xtxp97 on chromosome 6 (HART *et al.* 2001), and residing between RFLP marker Xtxs1030 and Xumc34 on the same chromosome (XU *et al.* 2000). The marker Xtxp97 was immediately adjacent to marker Xumc34 (BHATTRAMAKKI *et al.* 2000), which co-segregated with marker pSB140 (BOWERS *et al.* 2003). The latter (pSB140), which located 11.2cM from marker pSB187, was also linked to the Dry trait only ~2.45 times less likely than pSB187. On the physical map (PATERSON *et al.* 2009), marker pSB140 was located within the 1.68Mb region between pSB187 and pSB095. The inconsistent order(s) of these three markers in the genetic and corresponding physical map(s) is partly due to segregation distortion of alleles at these loci. With distorted segregation significantly favoring the *S. bicolor* (SB) allele over the *S. propinquum* (SP) alleles at pSB095 (homozygote SB:SP = 104:60) and pSB187 loci (SB:SP = 52:29) but not at pSB140 (SB:SP = 61:43), this mapped region corresponds with regions of segregation distortion found in the F2 and F5 RIL populations (SB:SP = 151:3) derived from this cross (BOWERS *et al.* 2003; KONG *et al.* 2013). Markers in this mapping population were either scored as codominant or, if unable to fit into the map, split into two dominant markers to attempt their mapping separately (BOWERS *et al.* 2003). When examining the ratio of *S. bicolor* to *S. propinquum* alleles at each locus, only individuals with codominant marker scoring were adopted. Taking pSB140 for example, about 60% of the 370 mapping individuals were scored as heterozygotes. This explains why so few lines were informative in exploring the segregation distortion.

Segregation distortion is defined as deviation of the observed segregation ratio of a locus from the expected Mendelian ratio. It can be influenced by a variety of genetic, physiological, and/or environmental factors, with the genetic elements causing segregation distortion being increasingly considered as potent evolutionary forces (SANDLER and NOVITSKI 1957; LIU *et al.* 2010). Given that the mapped region partially overlapped a segregation distortion region on sorghum chromosome 6 (KONG *et al.* 2013), to investigate whether the D gene controlling the 'dry' trait is under selection or not requires more precise mapping of the trait. In a RIL population between BTx623 (green/juicy) and IS3620c (white/dry) with 616 SNP markers (Kong *et al.* unpublished data), the 'dry' trait was mapped to a 1-LOD interval of 4.6cM with a likelihood peak at 57.9cM, corresponding to a 0.89Mb region on the physical map from 51.14Mb to 52.03Mb (Figure 3.3). This reduced the target region mapped in the *S. bicolor* x *S. propinquum* F2 population by about half.

3.4.2 GWAS of the 'Dry' trait

Genome-wide association mapping of the 'Dry' trait was conducted against 265,487 SNPs using trait data collected in 2009 and 2010 respectively, across a sorghum diversity panel (CASA *et al.* 2008). For 2009, there were 164 (49.1%) lines with white midribs and 170 (50.9%) with green midribs. For 2010, there were 199 (58.4%) lines with white midribs and 142 (41.6%) with green midribs. Others among the total of 354 lines were treated as missing data because of germination failure or plant death.

To better control the false positive rate while maintaining sufficient detection power for GWAS, the significance threshold was determined by using a Bonferroni-like approach (see Materials and Methods), which took into account the LD-decay

information for each chromosome (MATTHIES *et al.* 2014). This approach takes into account the fact that sorghum is largely inbreeding, which can result in strong LD patterns with some highly correlated SNPs. Instead of using the total number of SNPs as the denominator in Bonferroni multiple testing corrections, this approach estimates the number of effective (independent) tests n as $n = \sum_i \frac{l_i}{d_i}$, where l_i is the length of chromosome i, and d_i is the extension of LD for chromosome i. Within the sorghum diversity panel, LD decays to background levels ($r^2 < 0.1$) at 250-300 kb, 225-250kb, 200-225kb, 300-400 kb, 200-225 kb, 500-600 kb, 250-300 kb, 175-200 kb, 400-500 kb and 400-500 kb for chromosomes 1 through 10 respectively (MORRIS *et al.* 2013). This results in a significance cutoff within about an order of magnitude of 10^{-5}, to obtain an experiment-wise significance level of 0.05 for all tests.

GWAS (Figure 3.4) indicates 14 SNPs of common variants (minor allele frequency $\geq$ 5%) to be significantly associated with the "dry" trait with a p-value threshold of 1.96×10^{-5}, all of which locate within the 1.68Mb candidate region defined by the *S. bicolor* x S. *propinquum* F2 with no confounding of population stratification (Figure 3.4e, f); all but the first SNP also locate within the refined candidate region of 0.89 Mb based on the BTx623 x IS3620c RILs (Table 3.3). The analysis was performed with compressed mixed linear models (ZHANG *et al.* 2010) implemented in the GAPIT R package (LIPKA *et al.* 2012), with further verification by using logistic models implemented in PLINK 1.9 (CHANG *et al.* 2015). Data from 2009 and 2010 were analyzed separately, accounting for the environment difference.

Logistic models tend to have more power for detection of significant SNPs than CMLM in this case. This finding is not surprising since the 'Dry' trait is binary, while

CMLM models are designed for detection of complex quantitative traits. Also, models with 2010 trait data tend to have a better fit than those with 2009 data. This difference is partly caused by environmental factors, and partly due to the possible higher quality of 2010 trait data, since we were more experienced at phenotyping. To further support the validity of the SNPs detected, we grouped these 14 SNPs into 4 groups (<51.05Mb, 51.60Mb – 51.69MB, 51.70Mb – 51.79Mb, >51.80Mb) based on their locations and mapped them onto the LD plot. A total of 3 groups showed strong linkage disequilibrium (Figure 3.5), the exception being S6_51047142. As spurious associations tend to stand alone, we treated SNP S6_51047142 as a false positive.

3.4.3 Candidate genes showing evidence of effects on the "Dry" trait

For 13 SNPs significantly associated with the 'dry' trait, we selected 9 candidate genes (Table 3.4), which either contained the SNPs or immediately flanked the SNPs in the inter-genic regions. Among these 9 genes, 6 belong to the Glycoside hydrolase family 1, being specifically beta-glucosidase. Others are annotated as CAAX protease self-immunity, DUF640 domain containing protein, and L-allo-threonine aldolase respectively. GO enrichment analysis indicates that the ontology of these 9 genes are over-represented by carbohydrate metabolic process (p-value = 1.45E-04) and hydrolase activity (p-value = 7.63E-08), each of which may be related to the dryness and/or juiciness of the plant.

Exons of these 9 candidate genes nearest to the significant SNPs were re-sequenced in a sub-panel (Table 3.1) of 48 lines (G: 24 lines; W: 24 lines) selected to cover sorghum from a wide range of geographic locations including Africa, Asia and

North America, and to avoid potential population structure. A few selected exon regions that showed at least minor association (P-value > 0.10) with the 'dry' trait in the initial scan, including Sb06g022390 exon 1, Sb06g022450 exon 1 and Sb06g022490 exon 6, were resequenced against the whole panel for additional verification, as the smaller sample size of the sub-panel could compromise the detection power.

Sb06g022390 (version 1.4), or the corresponding Sobic.006G145300 (version 3.1) showed evidence of association with the 'dry' trait at locus S6_51614136 (version 1.4) / S6_50696285 (version 3.1) on exon 1 (Table 3.5), while other candidate genes tested found to have associated loci being synonymous or have no associated loci observed. (Table S3.2). The locus, S6_51614136 of Sb06g022390, resides in a predicted alpha-helical region of the Beta glucosidase protein. Among the cultivated races of *Sorghum bicolor*, there is a non-synonymous codon insertion of TTC (amino acid: L to FL) at this locus for the plants of W phenotype, taking as an example IS3620c which is a representative of race 'guinea' and is substantially diverged from the reference genome of S. *bicolor* BTx623. This insertion was found to be significantly associated with the 'dry' trait across the whole diversity panel, not only for IS3620c. On the other hand, there is a non-synonymous single nucleotide substitution (amino acid: L to F) in Sb06g022390 for S. *propinquum* (W phenotype) (Table 3.5). This suggests that the 'dry' trait may have followed two convergent paths during inter-species and intra-species evolution.

Differences in version 1.4 and version 3.1 *Sorghum bicolor* gene models complicate our study, since targeted re-sequencing data are based on *Sorghum bicolor* v1.4 gene models and predated v3.1, in which some v1.4 genes were altered or merged. The candidate D gene is among the ones most influenced by annotation differences.

Other candidate genes being influenced are Sb06g022440 and Sb06g022480 (Table S3.3). For the D gene, Sobic.006G145300 in v3.1 has almost 13 times the gene length, 7.5 times the protein length, and 12 times the exon number of Sb06g022390 in v1.4. Alignment of the two sequences (Figure S3.1) shows that the Sb06g022390 gene sequence is completely contained (100% match) in Sobic.006G145300, and shares with the latter the same first 63 amino acids of exon 1. The annotation difference of the two starts from the end of v1.4 exon 1, followed by an extra 11 newly annotated exons in v3.1. The D locus that we identified, which is located at 51,614,136bp in v1.4 and 50,696,285bp in v3.1, is within the shared amino acid region on exon 1 and thus is not affected by the updated annotation. However, as our targeted re-sequencing primers were designed based on v1.4 gene models, covering only 1 of 12 Sobic.006G145300 exons, this raises the possibility that additional or other causative mutations of the 'dry' trait may exist in the other 11 exons based on v3.1 gene models. Indeed, 7 more non-synonymous mutations (Table S3.4) were found in exons 4, 5, 7, 9, 11 and 12 of the D gene when aligning the sorghum reference genome to the draft genomes of IS3620c and *S. propinquum*, which would all be possible causative loci of the 'dry' trait. The most recent v3.1 gene model seems more reasonable than v1.4 because the length of the inferred D protein sequence (Sb06g022390: 68 amino acids; Sobic006G145300: 509 amino acids) is more consistent with the beta-glucosidase gene family in sorghum, which generally has a protein length of 500 to 600 amino acids. Thus, the v3.1 gene model Sobic006G145300 is adopted in subsequent studies, unless otherwise noted.

The candidate D gene, Sobic006G145300, which resides on the reverse strand, is implicated in hydrolase activity and carbohydrate metabolic processes (Table 3.6). It

belongs to the beta-glucosidase (BGLU) gene family, which catalyzes the hydrolysis of beta-D-glucosidic bonds with the release of glucose and an aglycone (CICEK and ESEN 1998). As this hydrolysis activity is required for the liberation of many physiologically important compounds, BGLU enzymes function in a number of processes in plant growth, development and defense, such as phytohormone activation (WIESE and GRAMBOW 1986; BRZOBOHATY *et al.* 1993), lignification (DHARMAWARDHANA *et al.* 1995), and cell wall catabolism (SIMOS *et al.* 1994; LEAH *et al.* 1995; GOMEZ-ANDURO *et al.* 2011). Thus, to investigate the potential relationship of the 'dry' trait to other traits in sorghum, such as moisture content (juicy yield), would be of great interest (see below**)**.

It has been previously reported that in sorghum, beta-glucosidase (Dhurrinase) was encoded by a small multigene family, with two isozymes (*Dhr1* and *Dhr2*) isolated and characterized from sorghum seedlings (HOSEL *et al.* 1987; CICEK and ESEN 1998). These two isozymes share ~70% sequence identity with each other and are homologs of the two maize BGLU isozymes: *Glu1* and *Glu2*, which are not at paralogous locations and may have arisen from a single-gene duplication. The *Dhr1* and *Dhr2* protein sequences have 565 and 571 amino acids, respectively. Despite their sequence similarity, they differ in substrate specificity and in activity level within various seedling parts. *Dhr1*, which is strictly specific to the natural substrate dhurrin and substrate analogue sambunigrin, is widely expressed in root, mesocotyl, node and coleoptile tissues. *Dhr2*, has broader substrate specificity, hydrolyzing dhurrin, sambunigrin and certain artificial substrates (4-MUG, PNPG, triglochinin), and is mainly expressed in leaf, node and coleoptile tissues (VERDOUCQ *et al.* 2004; MAHAJAN *et al.* 2015). Previous studies also suggested that since the node includes the shoot apex and primordial leaves, it was likely

that the *Dhr2* found in the node actually came from the primordial leaves (CICEK and ESEN 1998). A *Sorghum bicolor* RNA-seq expression study of 47 experiments in the JGI Plant Gene Atlas Project (Mullet *et al.* unpublished data https://phytozome.jgi.doe.gov/phytomine/aspect.do?name=Expression) indicated comparatively high expression of the D gene in leaves when juvenile or vegetative, in panicles at anthesis, and in shoots. This expression pattern strongly suggests the D gene to be a *Dhr2*, since *Dhr2* is the only isozyme currently reported to be detected in the leaves by immunoblotting (CICEK and ESEN 1998). In green leaf blades of young sorghum seedlings, dhurrin, the natural substrate of beta-glucosidase enzymes, was reported to locate entirely in the epidermal layer, and was suggested to serve as a carbon source for other biosynthetic pathways of physiological significance. *Dhr2* was reported to reside almost exclusively in the mesophyll cells of young sorghum leaves (KOJIMA *et al.* 1979; THAYER and CONN 1981; NGUYEN and BLUM 2004). Phylogenetic analysis (below) suggests that the activity of the D gene is possibly associated with lignification and cell wall catabolism, which may contribute to altered midrib color. Modification in functions associated with lignin structure and content were also suggested previously to be associated with altered cell wall digestibility and plant dry matter (FRITZ *et al.* 1990; THORSTENSSON *et al.* 1992; NGUYEN and BLUM 2004), all providing additional support to our candidate for the D gene.

3.4.4 Selective signatures

To investigate selective signatures around the ‘dry’ gene candidate, we quantified nucleotide variation for each annotated gene of the sorghum reference genome across a

panel of sorghum landraces. As the SNP set used in GWAS of this study (MORRIS *et al.* 2013) contains only ≥ 1 variants within a gene genotyped for 72% of 27,412 annotated genes in the reference genome sequence (ZHANG *et al.* 2015b), there could be a high probability of missing loci within genes. Thus, a population of 59 high coverage re-sequenced sorghum lines (GUO 2015), representing primary taxonomic groups, was adopted to investigate selection. Overall, the Tajima's D value per gene averages -0.3353 for the whole genome and -0.1038 for genes on chromosome 6. The Dry gene candidate Sobic006G145300 has a Tajima's D value of 1.42, which is significantly higher than 0 (See Materials and Methods), indicating that alleles for both phenotypes of the 'Dry' trait are enriched in *S. bicolor* as a result of balancing selection. We further investigated the 1 Mb region around the D gene. This region is not within low heterozygosity regions identified using both quantified genome-wide heterozygosity (MORRIS *et al.* 2013) and a refined ratio of wild sorghum against cultivated types in genetic diversity (ZHANG *et al.* 2015a). In fact, the low heterozygosity regions on sorghum chromosome 6 included one interval ~9.6Mb from the D gene which covered the *Ma1/Dw2* locus, and one interval ~3.3Mb from the D gene which covered a plant height locus and a flowering time locus (BROWN *et al.* 2008; SHIRINGANI *et al.* 2010), with the D gene located in-between. This ordering is consistent with the previous finding that the D gene resides in the middle of two segregation distortion regions. This evidence provide additional support for our finding that the 'dry' trait is under balancing selection, implying that both phenotypes of the 'dry' trait may have been historically equally important or be a natural consequence of adaptation, which is not surprising. Unlike plant height or seed size, the 'dry' trait may not have obvious agronomic importance, but its association with plant moisture content

could lead to adaptation to varied climatic/geographic conditions favoring either of the alleles at the D locus (see below).

3.4.5 Relationship of the 'dry' trait to moisture content in sorghum

The 'dry' trait, phenotyped using the midrib character, determines whether a plant has pithy/dry or dull/juicy stalks. Prior findings have shown that the pithy/dry stalk type is associated with an average of about 1% lower moisture content of the plant overall (SCHERTZ *et al.* 1978). A recent study suggested that the D gene which determined the appearance of the midrib may not influence percent moisture of the plant (FELTUS and VANDENBRINK 2012), however this study used a limited sampling of cultivars harvested at uniform dates rather than each at physiological maturity. Accordingly, there is justification to further investigate the relationship of the dry trait with plant moisture content.

Among 283 lines in the sorghum diversity panel (CASA *et al.* 2008) for which we obtained complete measurements, we tested the relationship between the 'dry' trait and moisture content of the whole plant as well as several parts (leaves, stalks, inflorescences). For each tissue, moisture content is calculated as the ratio of $(\text{fresh weight} - \text{dry weight})/\text{fresh weight}$. A generalized linear regression model is applied, treating plant maturity date as a blocking factor to minimize environmental effects. With windows of 2 weeks, there are altogether 5 blocks grouped (Table 3.7-a). Biases are controlled within each block as well as overall, as the number of samples with 'G' versus 'W' phenotypes does not significantly deviate from 1:1 (Table 3.7-b) based on chi-squared tests.

Genotypes maturing in early November (4th block) have a significantly lower stalk and overall moisture content than the others (Figure 3.6a, b). Likewise, genotypes maturing near the end of the growing season (4th, 5th blocks) showed a decreasing trend in leaf moisture content compared to genotypes maturing earlier (Figure 3.6c). On the other hand, inflorescence moisture contents were little affected by plant maturity dates (Figure 3.6d), as expected since physiological maturity of the inflorescence was the criterion for selecting harvest dates. By fitting the model, we verified that the 'dry' trait is significantly correlated with overall moisture content (p-value: 0.0486). Indeed, 'dry' genotypes had a significant reduction (~3%) in overall moisture content in comparison to 'juicy' genotypes (Table 3.8a). Different plant parts showed different effects of the 'dry' gene (Table 3.8b, c, d), which had a strong positive correlation with leaf moisture content (P-value: 0.0181), but weaker correlation with stalk moisture content (0.0516) and non-significant correlation with inflorescence moisture content (0.5062).

As previous studies concluded that plant moisture content might be a quantitative trait and moderately heritable (MURRAY *et al.* 2008), we mapped stem water content in the BTx623 x IS3620c RIL population (Kong *et al.* unpublished data), detecting a major QTL on chromosome 6 (Figure 3.7) using the BLUP (HENDERSON 1975)(see Materials and Methods) of two years' data. The peak LOD score of 20.2 was located at 59.0cM, with a 1-LOD interval of 7.4cM (1.6Mb), explaining about 20.8% of the phenotypic variation. The refined D candidate gene region is contained within a reported stem moisture QTL qSW6 (HAN *et al.* 2015), spanning 15.0cM (4.5Mb) between SSR marker Ch6-2 and gpsb069 (Figure 3.7b). Using the stem moisture data of year 2012 alone, a second 'putative' QTL of smaller effect was detected on chromosome 1 (Figure 3.7a) in a

region homoeologous to the QTL on chromosome 6, peaking at 60cM with a LOD score of 3.3, and explaining about 3.5% of the phenotypic variation. This QTL eluded detection in 2011 and by the overall BLUP values.

In conclusion, our study indicates that the 'dry' trait is a significant indicator of plant moisture content. Plant moisture content is heritable, but unlike the simply inherited 'dry' trait, it is a quantitative trait with a major QTL co-localized with the 'dry' trait and a small effect QTL at a homoeologous location on chromosome 1.

3.4.6 Race-specific pattern of the 'Dry' trait and geographic variation

Morphological races and geographic origins can lead to patterns of phenotypic variation. Among 354 accessions in a sorghum diversity panel (CASA *et al.* 2008), most members of the sudanense/broomcorn race are of G phenotype (90% in 2009 data, 70% in 2010 data). In a comparison of three sorghum botanical races (caudatum, durra and kafir) using Fisher's exact test, there is a minor significant difference (p-value=0.050), with durra being more enriched for the W phenotype and kafir being more enriched for the G phenotype. No significant difference was found between caudatum and either durra or kafir. This finding is supported by a preliminary study (unpublished data) conducted by Ayten Dizkirici using 47 SSR loci. Now with 265,487 SNPs available across the sorghum genome, we were able to explore phenotypic variation patterns more deeply.

The 354 accessions of the population were clustered into 4 subgroups by applying K-means clustering to the first three PCA components (Figure 3.8). Subgroup I contains mostly guinea/caudatum and caudatum races, subgroup II contains mostly race kafir, subgroup III contains mostly race durra, and subgroup IV contains mostly

zerazera/caudatum and caudatum. Guinea/bicolor and sudanense/broomcorn types are mainly clustered in subgroup I, while milo/feterita types are spread among subgroups I, III and IV. Kafir, originating from southern Africa, generally has juicy stalks. Feterita, originating from the Sudan region of Africa; durra, mainly grown in East Africa, Middle East and India; and broomcorn, found in central Africa and Mediterranean region, generally have dry and pithy stalks (MARTIN *et al.* 1931; SWANSON 1931).

As the 'dry' trait is associated with plant moisture content, which may affect adaptation to different environments, we propose that the 'dry' trait may have some race-specific pattern: Feterita, durra, and broomcorn to be enriched by W phenotype, Kafir to be enriched by G phenotype; and be associated with climatic/geographic variation.

Supposing the first part of the hypothesis is true, we would expect subgroup II to have a higher water percentage than subgroup III. The medians rather than the means of stalk and overall plant moisture content were compared for each subgroup (Table 3.9), since the distributions of plant moisture content were largely skewed. Indeed, subgroups I, II and IV had higher overall water percentage as well as stem water percentage than subgroup III. The midrib phenotype (G or W) is significantly associated with the clustering by using Fisher's exact test (p-value = 0.047 for 2009, 0.008 for 2010), suggesting a nonrandom relationship between the Dry trait and population structure. Kafir, durra and caudatum races of sorghum are genetically distinct from each other, consistent with prior studies (ZHANG *et al.* 2015a). With the previous finding that races durra and kafir are significantly different in phenotype enrichment, while neither differs much from race caudatum, we would suggest that races kafir and durra may have

experienced more recent diversification for the ‘dry’ trait with caudatum being an earlier derivative from wild sorghum.

To test the second part of the hypothesis, i.e. that the ‘dry’ trait is associated with climatic/geographic variation, we studied for 2-years’ data the average percentages of individuals with G and W phenotypes respectively for each country (Figure 3.9, Table S3.5) and its geographic variation per race per country (Table 3.10). Complete enrichment for W or G was defined when only W or only G phenotypes were present for at least one year (eg. ‘4/0, 3/0’; ‘0/3, 0/3’). Partial enrichment for W or G was defined as one phenotype (W or G) being enriched for at least one year (eg. ‘9/3, 9/2’; ‘0/4, 2/3’). Heterogeneity was defined as one phenotype being enriched for one year while the alternative was enriched for the other year (eg. ‘0/1, 1/0’). In Africa, most trait variation (partial enrichment and heterogeneity) was aggregated in the northern region under the Mediterranean, arid or semiarid climate where *Sorghum bicolor* originated. Caudatum, guinea/caudatum and guinea/bicolor collected around East and West Africa under arid or tropical dry/wet climates were mostly enriched for the W phenotype, while races collected in central, north-central and south-central Africa under tropical dry/wet, or semiarid climates were enriched for both G and W phenotypes (Figure 3.9b, Table 3.10). Among the 9 races (guinea/bicolor not available) collected from North America under humid subtropical climate, 8 were enriched for the G phenotype, while the only race (caudatum) collected in South America under semiarid or tropical dry climate was completely enriched for the W phenotype. Races collected from West Asia, adjacent to Northeast Africa under arid or semiarid climate, were mostly enriched for the W phenotype, while more G phenotypes emerged to the East, mostly under humid

subtropical climates (Figure 3.9a, Table 3.10). In total, these findings provided support for the hypothesis that the 'dry' trait is associated with climatic/geographic variation. However, limited numbers of plants with trait data per race per region restrained the study of 'dry' trait migration, as there was no data for race sudanese/broomcorn, zerazera/caudatum or guinea/bicolor in Asia, and in most areas there were only less than 3 accessions with trait data available for each race. Thus, larger population data would be needed to more accurately address the association of climatic/geographic variation with the 'dry' trait.

3.4.7 Phylogeny

To investigate the evolutionary history of the D gene, the phylogeny of the whole beta-glucosidase (BGLU) gene family provides a backdrop. In the *Sorghum bicolor* reference genome sequence (PATERSON *et al*. 2009), 39 BGLU paralogs were identified using Pfam (FINN *et al.* 2016) search and subsequent manual curation for verification. In the region around the D gene, there are 8 genes (Sobic.006G145400 to Sobic.006G146200) similar in protein sequence (Figure S3.2) from the same BGLU gene family immediately following the D gene Sobic.006G145300, strongly suggesting at least one tandem expansion of the gene family (Figure 3.10). To further investigate how and when this gene family expansion happened, we identified BGLU genes in 1) *Oryza sativa* (38 homologs), *Setaria italica* (33 homologs), and *Brachypodium distachyon* (36 homologs), each species that like sorghum have not experienced recent whole genome duplication; 2) *Zea mays* (35 homologs) that has experienced a whole genome duplication more recently than its divergence from *Sorghum bicolor*; 3) monocot outgroup *Musa*

acuminata (32 homologs); and 4) eudicot *Arabidopsis thaliana* (48 homologs) which is an important outgroup and botanical model.

The neighbor-joining tree for all 7 species (Figure 3.11a, Figure S3.3) indicates 6 groups of sorghum BGLU genes based on sequence identity at the protein level. A recent phylogenetic study in maize (GOMEZ-ANDURO *et al.* 2011) classified the BGLU gene family into 4 groups with 26 maize BGLU genes, 24 of which were also identified and in same grouping in our study, with the other 2 deleted from the most recent maize gene model at the date of the analysis (B73 RefGen_v3, available at www.maizesequence.org). The 4 groups identified in maize had a perfect one-to-one correspondence to the first 4 groups of our study (Table S3.6). The 2 groups newly identified here each contain a maize BGLU gene that was not previously reported, possibly due to changes in gene number and gene chromosomal coordinates in the most recent maize gene model (B73 RefGen_v3). Our phylogenetic tree has good coverage of all 47 Arabidopsis BGLU genes except one (AT5G16580), which was actually identified in our study but fell short of the alignment requirement for tree construction (299 amino acids compared to the average length of ~500-600 amino acids of other BGLU genes). In all, this phylogenetic tree provides comprehensive information across cereals and serves as a good start to study the evolution of the BGLU gene family in sorghum.

The D gene Sobic.006G145300, together with 8 paralogs, lies in group IV (Figure 3.11b) of the tree, with three Arabidopsis homologs: AT4G21760 (BGLU47), AT1G61820 (BGLU46), AT1G61810 (BGLU45) functionally annotated to be involved in lignin biosynthesis and carbohydrate metabolism. There appear to be three ancient copies of the sorghum BGLU genes around the D gene region: one for

Sobic.006G146000, one for Sobic.006G146100 – 146200, and one for Sobic.006G145300 – 5800. These three ancient copies are likely to have been generated from one even more ancient copy (Figure 3.11b). BGLU genes Sobic.006G145300 – 145600 are clearly recent tandem expansions after the divergence between sorghum and maize at least 11 MYA (SWIGONOVA *et al.* 2004), as they formed a cluster jointly connected with the branch of two maize genes (GRMZM2G012236, GRMZM5G882852). The node between the cluster of 4 sorghum genes and the branch of two maize genes are well supported (bootstrapping value: 100). Sobic.006G145700 – 5800 could be another cluster of sorghum BGLU tandem genes, with the maize copy related most likely to have been lost after divergence.

When addressing the question of how the 'dry' trait evolved, the expansion of gene numbers itself is a potential foundation for evolution of a new trait by invoking sub- or neo-functionalization through duplicate retention. Tandem genes are a major repository for recent genetic innovation in taxa such as sorghum that have not experienced whole genome duplication in an estimated 98 million years (WANG *et al.* 2015). Differentiation of function usually requires a large number of amino acid substitutions. However, a novel function may be achieved through surprisingly few substitutions (GRAUR and LI 2000). For example, lactate dehydrogenase can be changed into malate dehydrogenase by replacing just one of its 317 amino acids (WILKS *et al.* 1988). The 'dry' trait, as discussed earlier, is in association with plant moisture content and thus could be an adaptation to varied climatic/geographic conditions favoring either of the alleles at the D locus. This character, together with the finding that no similar trait has been reported in other cereals, supports our hypothesis that the 'dry' trait evolved

under balancing selection during a recent tandem expansion of the BGLU gene family in sorghum after its divergence from maize.

3.5 Conclusion

In this study, we characterize the ‘dry’ trait in sorghum and identify a probable candidate gene (Sobic.006G145300) by using a combination of linkage mapping and genome-wide association mapping approaches. The ‘dry’ trait may have experienced convergent evolution in wild sorghum (e.g., non-synonymous single nucleotide substitution in *S. propinquum*) and within S. *bicolor* races (e.g., IS3620c non-synonymous codon insertion). The D gene, most likely isoform *Dhr2* of beta-glucosidase, is highly expressed in shoots, leaves and panicles, possibly being involved in lignification and cell wall catabolism.

A relationship between the ‘dry’ trait and plant moisture content is confirmed. Race-specific patterns and geographic variation of the ‘dry’ trait are evident. The ‘dry’ trait, to date known only in sorghum, appears to have evolved under balancing selection, during a recent tandem expansion of the BGLU gene family in sorghum after its divergence from maize.

Table 3.1. Sorghum accessions and countries of origin in the ‘D’ gene sub-panel

Accession	Trait	Origin Country	Accession	Trait	Origin Country
S. propinquum	W		SC42	G	Ethiopia
SC1321	W	Sudan	SC325	G	United States
SC564	W	Uganda	SC63	G	Sudan
Deer-broom	W		SC942	G	United States
SC67	W	Sudan	KELLER	G	
SC757	W	Botswana	SC603	G	Tanzania
Town	W	Botswana	SC630	G	Zambia
SC673	W	Zimbabwe	SC1070	G	Nigeria
WhiteKafir	W	United States	QL3India	G	India
SC326-6	W	United States	SC1319	G	Ethiopia
SC623	W	Congo	Malisor-84-7	G	Mali
SC1246	W	Chad	RCV	G	El Salvador
SC621	W	India	SC833	G	India
SC525	W	Nigeria	SC38	G	Ethiopia
SC1218	W	Sudan	RTx2917	G	United States
SC1439	W	Gambia	SC1057	G	Uganda
SC305	W	Chad	SC124	G	Ethiopia
SC968	W	Zimbabwe	SC309	G	Sudan
SC1063	W	Senegal	RTX2737	G	United States
NTenimissa	W	Mali	SC299	G	Nigeria
SC572	W	China	SC1104	G	Uganda
SC1203	W	Brazil	RIO	G	
El-Mota	W	Niger	SC320	G	Chad
SC333	W	Ethiopia	SC49	G	Sudan

Table 3.2. Locations of markers flanking the ‘Dry’ trait

Marker	LD	cM	Chr.	Mb
pSB095	D	31.2	6	50.74
pSB187	D	36.5	6	52.42

Table 3.3. SNPs with significant association to the ‘dry’ trait (threshold: 1.96E-5). P-values using CMLM and logistics model for 2009 and 2010 are listed.

Chr.	Position	CMLM-2009	CMLM-2010	Logistic-2009	Logistic-2010	Minor allele frequency
6	51047142	5.10E-04	7.43E-06	2.61E-05	7.06E-07	16.08%
6	51614133	3.37E-05	1.16E-07	8.51E-11	3.31E-15	37.80%
6	51614642	4.38E-04	1.59E-05	7.19E-06	9.07E-05	39.82%
6	51659928	5.93E-03	4.50E-06	4.54E-06	2.64E-10	42.48%
6	51659947	5.30E-04	3.94E-06	4.27E-07	2.47E-10	42.77%
6	51673670	1.48E-04	5.09E-06	1.16E-08	1.41E-11	46.31%
6	51711306	1.60E-04	2.90E-06	4.96E-08	7.64E-11	45.72%
6	51715340	1.19E-03	4.45E-06	2.20E-06	2.32E-08	20.94%
6	51715398	4.74E-04	1.86E-06	1.68E-06	1.25E-08	21.24%
6	51834749	3.74E-06	3.93E-06	5.69E-06	4.42E-05	26.05%
6	51834838	2.81E-05	4.12E-06	1.26E-05	3.93E-05	23.19%
6	51834877	2.81E-05	4.12E-06	1.26E-05	3.93E-05	23.19%
6	51839270	1.18E-05	8.79E-06	4.73E-05	2.94E-04	20.93%
6	51839273	1.18E-05	8.79E-06	4.73E-05	2.94E-04	20.93%

Table 3.4. Candidate genes for re-sequencing

Candidate gene	Chr.	Strand	Distance to nearest SNP	Annotation
Sb06g022390	6	-	In gene	Beta glucosidase, Glycosyl hydrolase family 1
Sb06g022440	6	-	In gene	Beta glucosidase, Glycosyl hydrolase family 1
Sb06g022450	6	-	In gene	Beta glucosidase, Glycosyl hydrolase family 1
Sb06g022480	6	-	In gene	Beta glucosidase, Glycosyl hydrolase family 1
Sb06g022490	6	-	In gene	Beta glucosidase, Glycosyl hydrolase family 1
Sb06g022500	6	-	7 kb	Beta glucosidase, Glycosyl hydrolase family 1
Sb06g022600	6	-	41 kb	CAAX protease self-immunity GO: membrane
Sb06g022610	6	+	6 kb	DUF640 domain containing protein
Sb06g022620	6	-	3 kb	L-allo-threonine aldolase, putative, expressed

Table 3.5. Non-synonymous substitution at the ‘dry’ locus. First 20 amino acids of the protein sequence are shown. The non-synonymous changes are marked in red.

Name	**Phenotype**	**Protein segment (first 20 amino acids on exon 1)**
BTx623 (*S. bicolor*)	Green/Juicy	5′---MARLFF-LLLLFPLLCSSVTG---3′
IS3620c (*S. bicolor*)	White/Dry	5′---MARLFFFLLLLFPLLCSSVTG---3′
S. propinquum	White/Dry	5′---MARLFF-FLLLFPLLCSSVTG---3′

Table 3.6. Ontology annotations of Sobic.006G145300

Ontology	Term	Name	Description
GO	GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds	Catalysis of the hydrolysis of any O-glycosyl bond.
GO	GO:0005975	carbohydrate metabolic process	The chemical reactions and pathways involving carbohydrates, any of a group of organic compounds based of the general formula Cx(H2O)y. Includes the formation of carbohydrate derivatives by the addition of a carbohydrate residue to another molecule.
PFAM	PF00232	Glyco_hydro_1	Glycosyl hydrolase family 1
PANTHER	PTHR10353		Glycosyl hydrolase
ENZYME	3.2.1.21	Gentobiase	Beta-glucosidase

Table 3.7. General information for the Dry trait across maturity dates. (a) Maturity dates of blocks. (b) Distribution of 'Dry' phenotypes over blocks.

(a) Block information with dates.

Blocks	Date	Lines	Percent
1	8/29/09	2	0.71
1	9/11/09	12	4.24
2	9/17/09	10	3.53
2	9/18/09	7	2.47
3	10/1/09	36	12.72
3	10/2/09	51	18.02
3	10/6/09	28	9.89
3	10/7/09	21	7.42
3	10/8/09	29	10.25
3	10/9/09	19	6.71
4	11/6/09	33	11.66
4	11/9/09	19	6.71
5	12/11/09	16	5.65

(b) Block information with traits.

Block	Trait		
	G	W	Total
1	9	5	14
2	7	10	17
3	97	87	184
4	27	25	52
5	8	8	16
Total	148	135	283

Table 3.8. Regression models of the ‘dry’ trait versus moisture content (left), and the estimated mean moisture content of ‘dry’ (D) or ‘juicy’ (G) phenotypes (right) considering the impact of maturity date, for the overall plant (a), stalks (b), leaves (c) and inflorescence (d).

(a) Model for overall moisture content and the estimated mean.

Source	DF	F Value	P-value
Dry trait	1	3.92	0.0486

Trait	Overall moisture content
G	58.8%
W	57.0%

(b) Model for stalk moisture content and the estimated mean.

Source	DF	F Value	P-value
Dry trait	1	3.82	0.0516

Trait	Stalk moisture content
G	70.5%
W	69.2%

(c) Model for leaf moisture content and the estimated mean.

Source	DF	F Value	P-value
Dry trait	1	5.66	0.0181

Trait	Leaf moisture content
G	50.1%
W	46.9%

(d) Model for inflorescence moisture content

Source	DF	F Value	P-value
Dry trait	1	0.44	0.5062

Table 3.9. Stalk and overall moisture content for each subgroup

Subgroup	Overall moisture content	Stalk moisture content
I	58.9%	72.6%
II	58.0%	71.4%
III	55.3%	70.4%
IV	56.8%	71.7%

Table 10: Distribution of the 'dry' trait across races and geographic regions. Contents in each cell are in format: 2009W/2009G, 2010W/2010G in plant numbers. For example, for race durra from India there were 19 plants (W:11, G:8) phenotyped in 2009 and 18 (W:10, G:8) in 2010. Geographic locations of a specific race are classified into 5 classes: completely enriched in W (**), partially enriched in W (*), heterogeneity (*†), partially enriched in G (†) and completely enriched in G (††).

Country	**guinea/ caudatum (E. Africa / India)**	**sudanense / broomcorn**	**kafir**	**zerazera/ caudatum**	**durra**	**milo/ feterita**	**guinea/ bicolor**	**guinea/ caudatum (W. Africa)**	**caudatum**	**Region**
Asia										
Pakistan	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	1/0, 1/0**	South Asia
India	4/0, 3/0**	0/0, 0/0	1/0, 1/0**	0/0, 0/0	11/8, 10/8*	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 1/0**	South Asia
Arabia	1/0, 1/0**	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	Western Asia
China	0/0, 0/0	0/0, 0/0	0/1, 1/0*†	0/0, 0/0	0/1, 0/0††	0/0, 0/0	0/0, 0/0	0/0, 0/0	1/1, 2/0*	Central Asia
Japan	0/0, 0/0	0/0, 0/0	0/1, 0/1††	0/0, 0/0	0/0, 0/0	0/1, 1/0*†	0/0, 0/0	0/0, 0/0	1/2, 3/0*†	East Asia
Africa										
West Volta	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/1, 1/1†	0/0, 0/0	West Africa
Burkina Faso	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	1/0, 1/0**	West Africa
Nigeria	0/0, 0/0	0/0, 0/0	0/1, 0/1††	0/1, 1/0*†	0/0, 0/0	1/0, 1/0**	0/0, 0/0	9/3, 9/2*	3/0, 3/0**	West Africa
Niger	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/1, 0/1††	0/1, 1/0*†	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	West Africa
Gambia	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	1/0, 1/0**	0/0, 0/0	0/0, 0/0	West Africa
Senegal	1/0, 1/0**	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	1/0, 1/0**	0/0, 0/0	West Africa
Congo	1/0, 1/0**	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	West central Africa
Zaire	1/0, 1/0**	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	Central Africa

Country	guinea/ caudatum (E. Africa / India)	sudanense / broomcorn	kafir	zerazera/ caudatum	durra	milo/ feterita	guinea/ bicolor	guinea/ caudatum (W. Africa)	caudatum	Region
Zimbabwe	1/0, 1/0**	0/0, 0/0	2/0, 2/0**	0/0, 0/0	0/0, 0/0	0/0, 0/0	1/0, 1/0**	0/1, 0/1††	0/0, 0/0	Southern Africa
Botswana	0/0, 0/0	0/0, 0/0	2/0, 2/0**	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	Southern Africa
Zambia	0/0, 0/0	0/0, 0/0	0/1, 0/1††	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	Southern Africa
South Africa	0/0, 0/0	0/0, 0/0	0/4, 2/3†	0/3, 0/3††	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	3/0, 3/0**	Southern Africa
Malawi	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 1/0**	Southeast Africa
Mosambique	1/0, 1/0**	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	Southeast Africa
Tanzania	1/0, 3/0**	0/0, 0/0	1/1, 1/1*†	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	1/0, 1/0**	East Africa
Ethiopia	0/1, 0/1††	0/0, 0/0	0/0, 0/0	4/6, 3/7†	10/6, 9/7*	0/0, 0/0	4/4, 6/2*	0/0, 0/0	1/0, 1/0**	East Africa
Somalia	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	1/1, 2/0*	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	East Africa
Kenya	0/0, 0/0	0/0, 0/0	1/0, 1/0**	0/0, 0/0	0/0, 0/0	0/0, 0/0	1/0, 1/0**	0/0, 0/0	0/0, 0/0	East Africa
Uganda	2/2, 3/1*	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/1, 0/1††	0/0, 0/0	0/0, 0/0	4/5, 6/5*†	East Africa
Chad	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	1/0, 1/0**	0/0, 0/0	1/0, 1/0**	0/0, 0/0	0/1, 0/1††	Northern central Africa
Mali	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	1/0, 0/0**	1/2, 3/0*†	0/1, 1/0*†	Northeast Africa
Egypt	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	1/0, 1/0**	1/0, 1/0**	0/0, 0/0	Northeast Africa
Sudan	6/1, 6/1*	1/0, 1/0**	0/1, 0/1††	2/6, 5/4*†	0/0, 0/0	1/0, 3/0**	2/3, 4/1*†	1/0, 1/0**	10/6, 11/6*	Northeast Africa

Country	guinea/ caudatum (E. Africa / India)	sudanense / broomcorn	kafir	zerazera/ caudatum	durra	milo/ feterita	guinea/ bicolor	guinea/ caudatum (W. Africa)	caudatum	Region
America										
United States	1/2, 1/2†	0/7, 1/6†	3/14, 4/13†	1/6, 1/6†	0/1, 1/0*†	4/10, 4/11†	0/0, 0/0	0/4, 0/4††	2/4, 1/5†	North America
Central America	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/2, 0/2††	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	Central America
Guatemala	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	1/0, 0/1*†	0/0, 0/0	0/0, 0/0	Central America
Brazil	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	1/0, 1/0**	South America
Venezuela	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	0/0, 0/0	1/0, 1/0**	South America

(a) Qualitative variation

White (W)

Green (G)

(b) Modifying variation

Green with White Stripe

White with Green Cast

Figure 3.1. The classification of midrib color: (a) Qualitative variation of the Dry trait; (b) Modifying variation of the ‘dry’ trait.

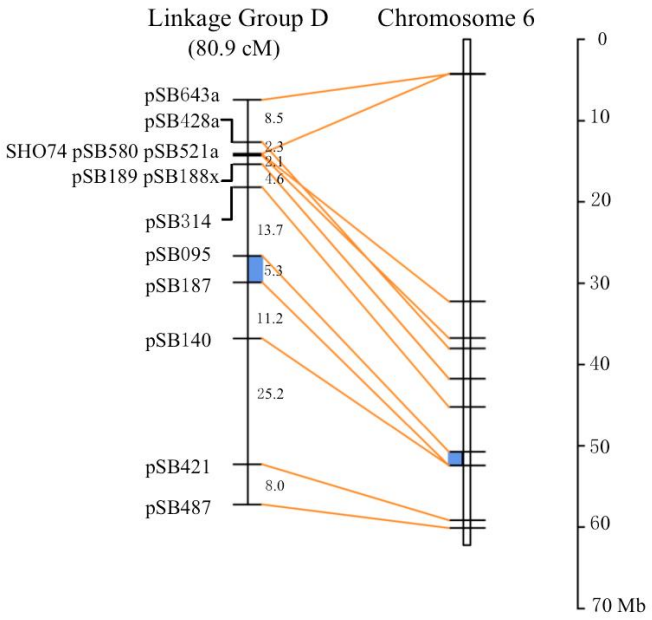


Figure 3.2. Initial mapping of the 'Dry' gene on sorghum chromosome 6. Marker positions on sorghum linkage group D are connected to their physical locations on the chromosome 6 sequence with orange lines. The mapped region of the 'dry' gene is highlighted in blue on both the linkage map and the physical map.

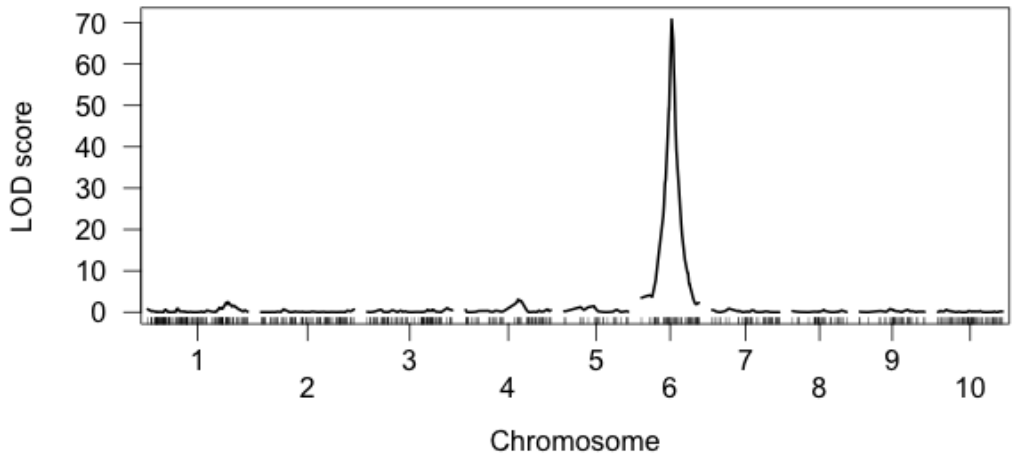
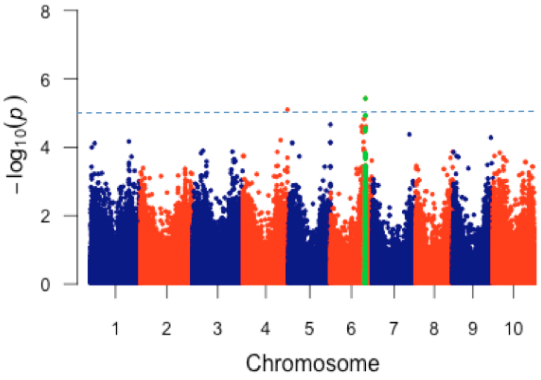
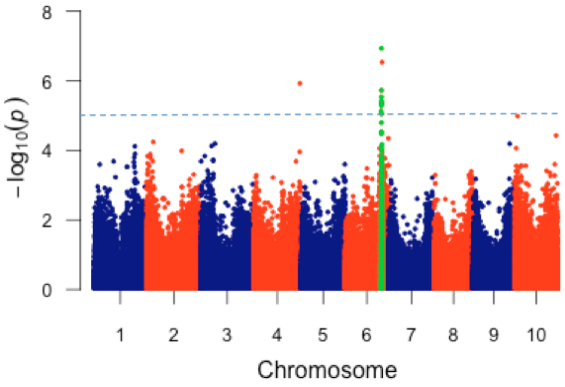


Figure 3.3. Mapping of the ‘dry’ gene in the BTx623 × IS3620c RIL population. The x-axis is genetic distance in cM for each of the 10 chromosomes. The y-axis is the LOD score.

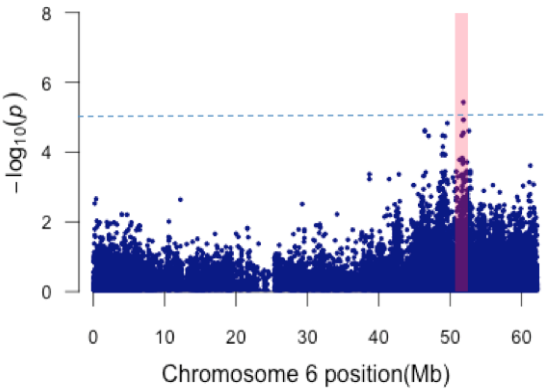
A. Manhattan plot CMLM - 2009



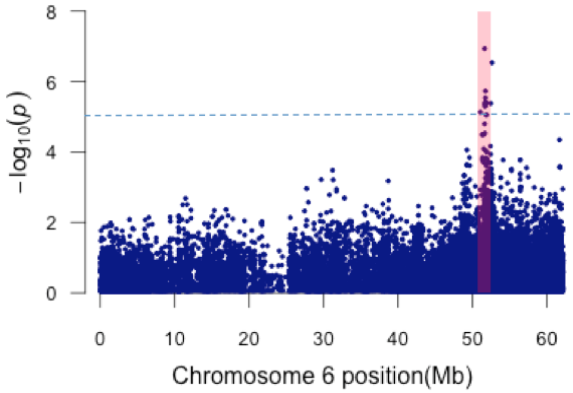
B. Manhattan plot CMLM - 2010



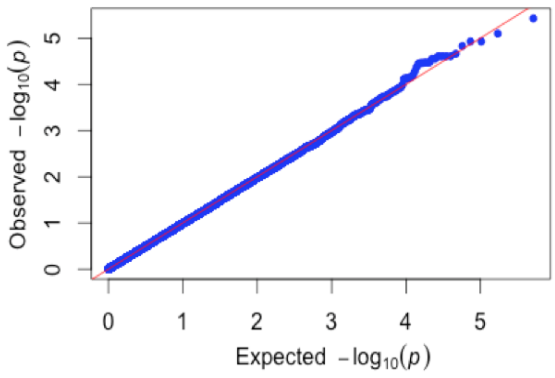
C. Manhattan plot for chr.6 - 2009



D. Manhattan plot for chr.6 - 2010



E. QQ plot - 2009



F. QQ plot - 2010

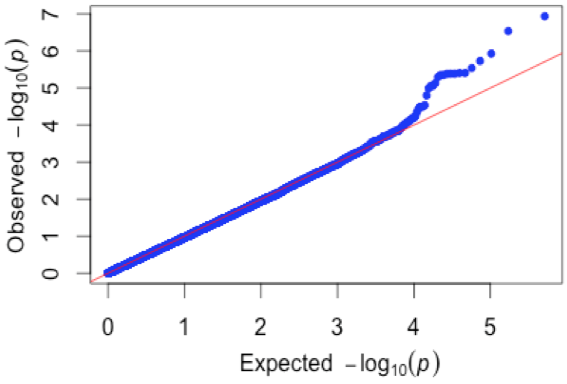


Figure 3.4. GWAS of the ‘dry’ trait with Manhattan plots and Quantile-Quantile plots for 2009 (A, C, E) and 2010 (B, D, F). SNPs within the candidate region previously identified by linkage mapping are highlighted in green (A, B) and shaded in red (C, D). A significance threshold of 0.05 [-$\log_{10}$(p-value)~5] is noted with dashed horizontal lines (A, B, C, D).

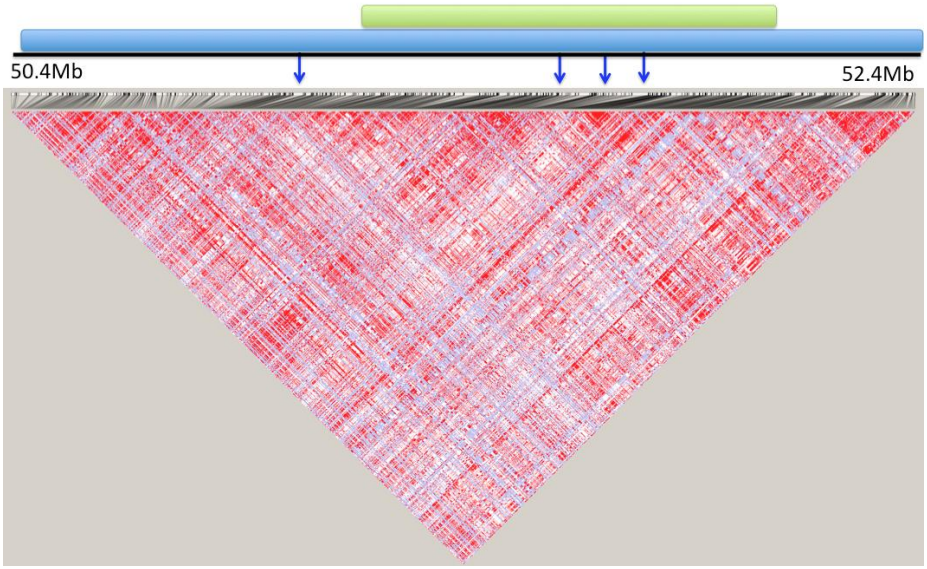
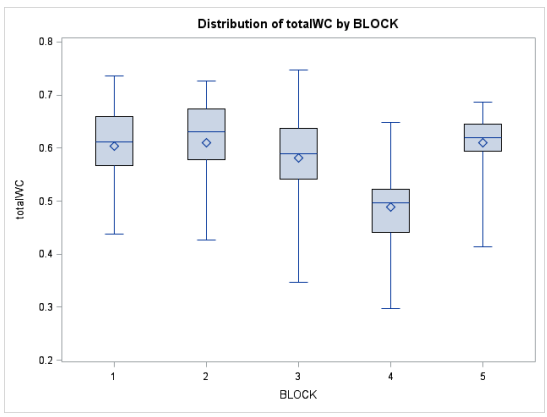
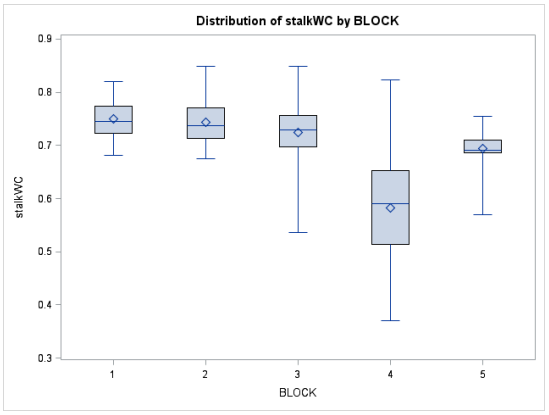


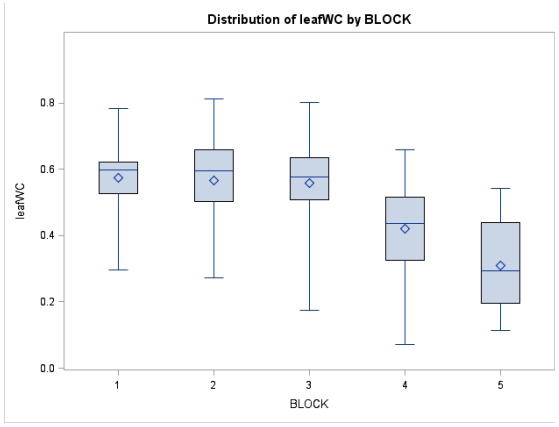
Figure 3.5. LD blocks with significant SNPs aligned to sorghum chromosome 6. Pre-defined D gene regions are marked with bars in blue (*S. bicolor* × *S. propinquum* F2 population) and green (BTx623 × IS3620c RILs population). Significant SNPs are specified by arrows.



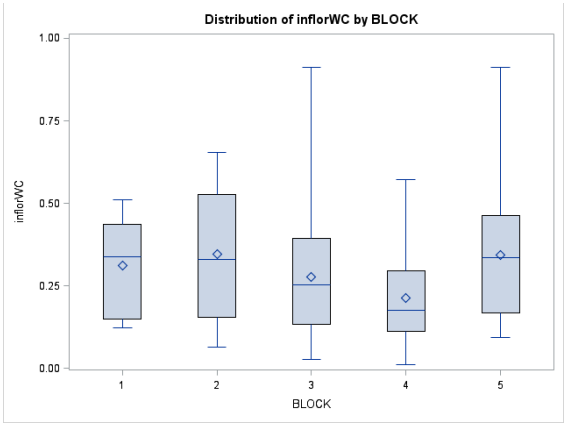
(a)



(b)



(c)



(d)

Figure 3.6. The distribution of overall plant (a), stalk (b), leaf (c), and inflorescence (d) moisture content of the sorghum genotypes over the growing season

a) QTLs affecting stem moisture content in the BTx623 x IS3620c RILs.

Chr.	Peak (cM)	LOD	Additive[a]	R^2 (%)	Start (Mb)[b]	End (Mb)[b]	Source
6	59.0	20.2	-0.0127	20.859	51.1	52.7	BLUP
1	60.0	3.3	-0.0110	3.482	18.4	26.2	2012

[a] Additive effect calculates as IS3620c – BTx623.

[b] Based on DNA marker locations flanking 1-LOD interval in the published genome sequence (PATERSON *et al.* 2009).

b) Relationship of the D gene region to the stem moisture QTL on chromosome 6.

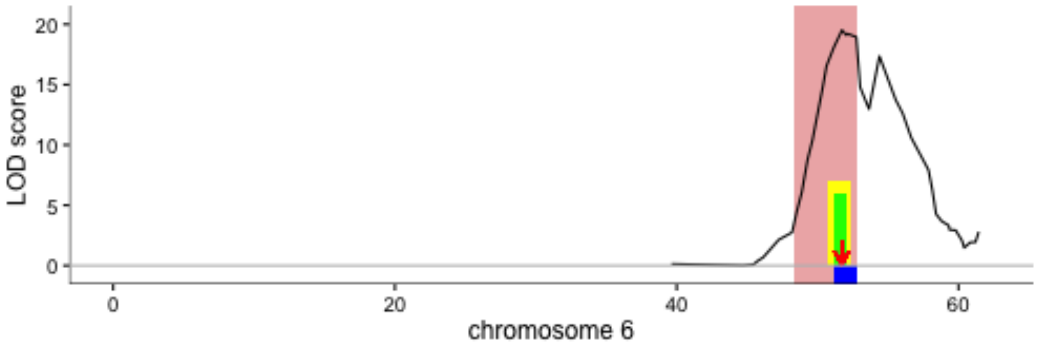


Figure 3.7. Stem moisture QTL mapping. a) QTLs affecting stem moisture content in the BTx623 x IS3620c RILs. b) Relationship of the D gene region to a stem moisture QTL on chromosome 6. The x-axis is the physical distance in Mb. The 1-LOD likelihood interval for the QTL mapped on chr.6 is highlighted in blue with the likelihood peak indicated by a red arrow. qSW6 (HAN *et al.* 2015), a previously identified stem moisture QTL, is highlighted in pink. The mapped D gene regions are highlighted in yellow (for the *S. bicolor* × *S. propinquum* F2 population) and green (for the BTx623 × IS3620c RIL population).

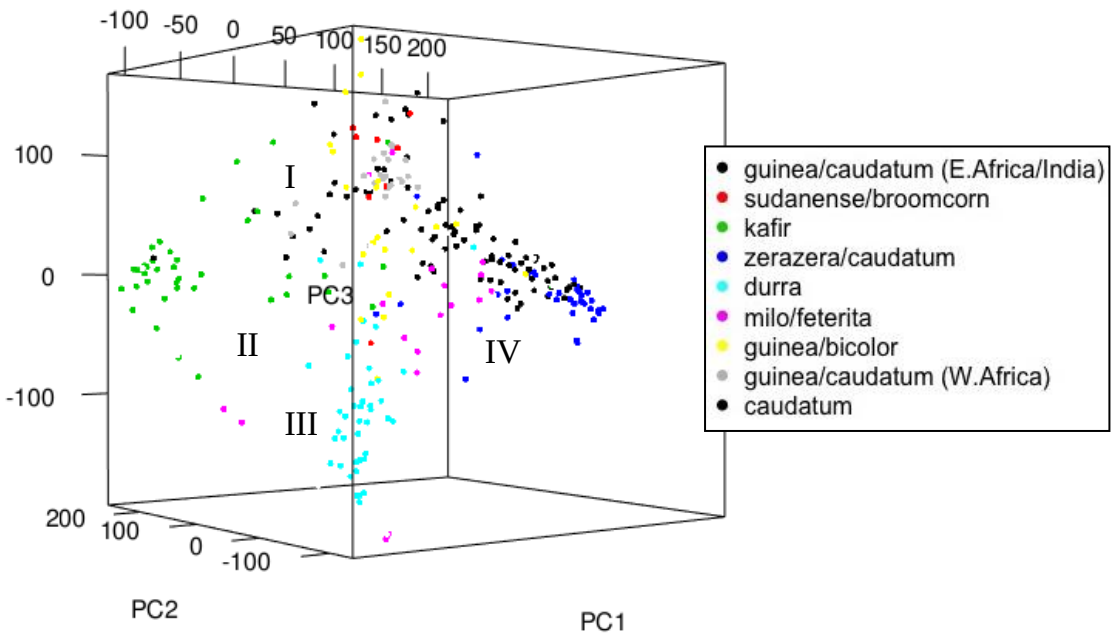
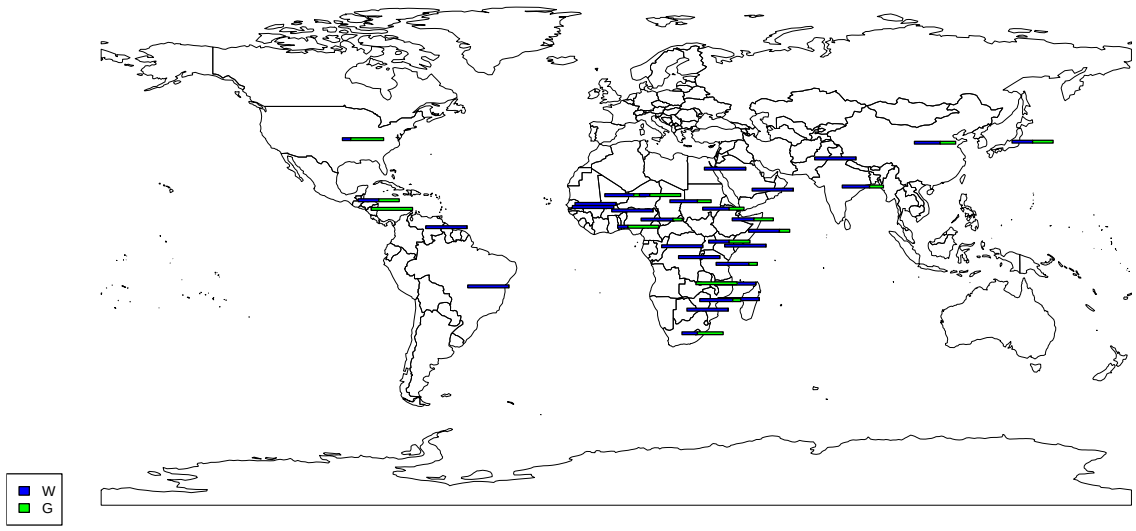


Figure 3.8. K-means clustering using the first three PCA components. Boundaries of 4 subgroups were marked by colored lines.

a)



b)

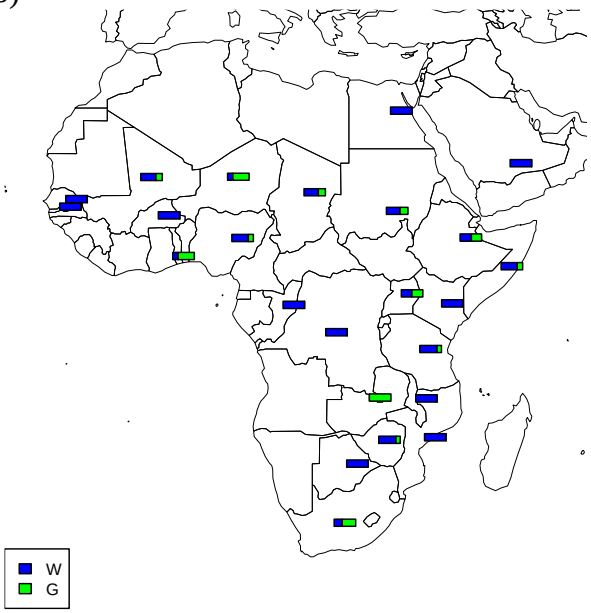


Figure 3.9. Geographic distribution of the ‘dry’ trait based on the sorghum diversity panel. Bar charts represent the percentages of individuals (Table S3.5) with phenotype G (green) and W (blue) at countries in Africa (b) and across the world (a).

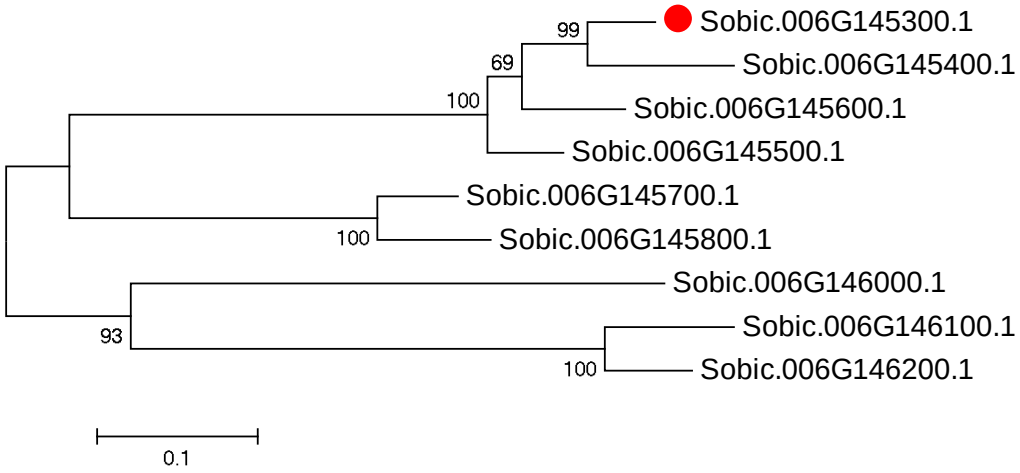
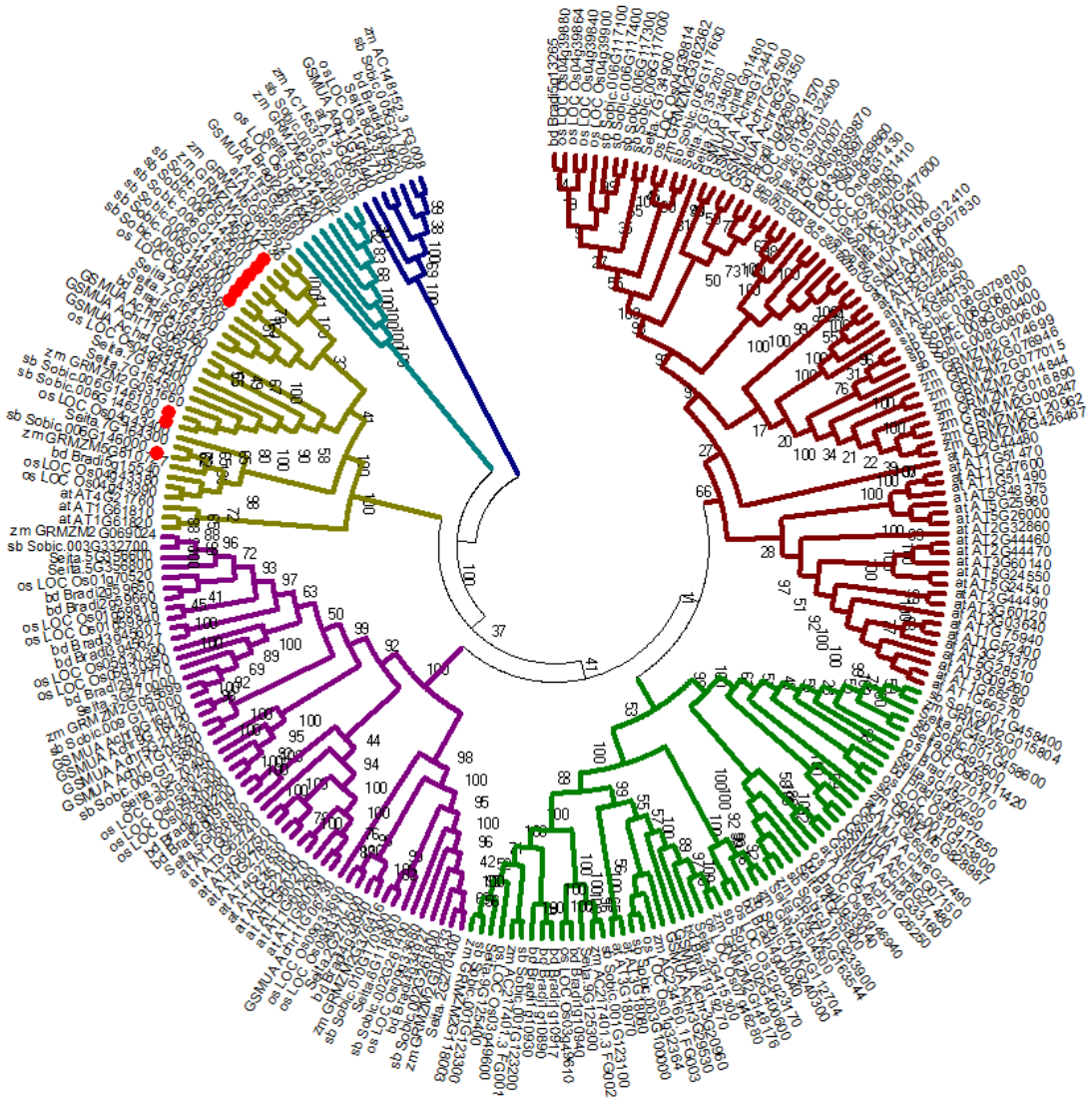
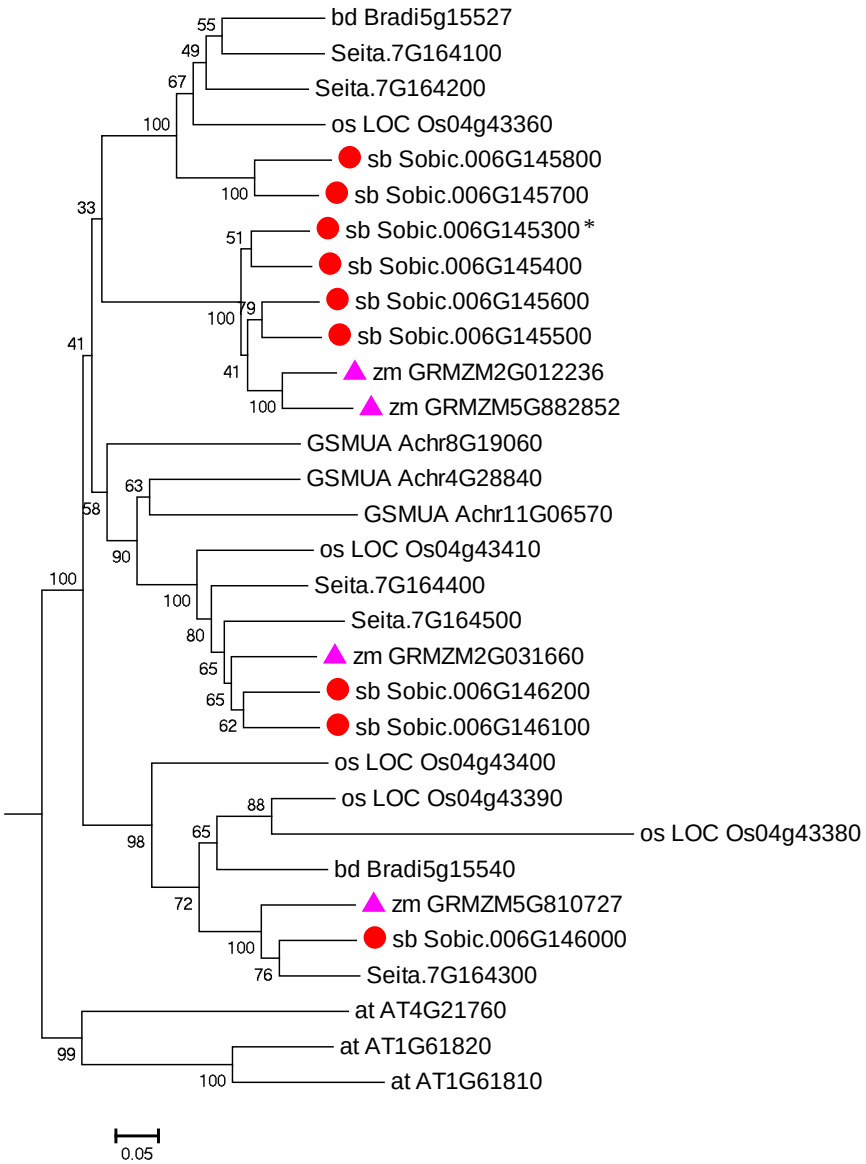


Figure 3.10. Dendogram of 9 immediately adjacent sorghum beta-glucosidase genes on chromosome 6. The D gene Sobic.006G145300 is marked with a red circle.

a)





b)

Figure 3.11. Phylogenetic tree of the beta-glucosidase gene family. a) The complete tree with 6 groups identified, in maroon (group I), green (II), purple (III), olive (IV), aqua (V), and blue (VI) respectively. The D gene (*) is identified together with its immediate 8 paralogs around the region, by red circles. b) The branch for group IV of the tree where the D gene resides. Sorghum BGLU genes are marked in red circles, while maize BGLU genes are marked in fuchsia triangles.

Supplemental Table S3.1. Primer pairs of selected exons for targeted re-sequencing

Reseq gene	Forward primer	Reverse Primer
Sb06g022390	TAGAGATACGCTAGGGATGGAC	GGGAAACGGTAGAAGGAGAATG
Sb06g022440	TGCAGTGCATTGTTGGTAGTG	AAAGAGGAATGTGCAGTGTGG
Sb06g022450	TCGATCTTTCAAAGCCTAGCTC	GACTGGCACCGTATCAGTTT
	ATCTGCTTTCCAAGTAGCCAAG	TCCACCATATCGCTTCTCTAGT
Sb06g022480	TTAAACGGAGCGAGTAAGACC	ATGAAGCACTTCTCGGATGC
	TAGCCTGGTTGCTTTCAGTG	AGTTGACACCCAAGGAATGC
Sb06g022490	TAGCGTGACAAGGTGATGAC	AAGAAAGAGGGCAAGACGAC
	AGGCGAGCACCTTTAAATACG	CTGCGTTCAGCAAGTAGTACAG
	CTTCAGACATACGAAGGGAACG	TCGACAGGTTGAAGTTTCGTC
Sb06g022500	ATGTTCCGAGTTTCCCTTCC	TCAATGGTACGGTACAGGTTTG
Sb06g022600	TTCTTCACCTGTCGTTCCAG	TCGCAGACATGGACAGATTG
Sb06g022610	CCACTGCGAAAGAGTAAGAAGC	GCACCTTAGTCTTGCCGAAC
	CTCACAGCTACACTGCTTTCTC	GCAAACACTTGCTCACTCCTAC
Sb06g022620	ACTGTGTGCAGTCTTGGAAC	ACCCTTCCACAAACATCGAG

Supplemental Table S3.2. Significant association variants characterized in targeted re-sequencing. Candidate gene-exons which contained the significant SNPs or immediately flanked the SNPs in introns or inter-genic regions were sequenced.

Gene identifier	Locus	Chr.	Exon | Intron	Alleles	Syn | Non-syn	Note
Sb06g022390	51614136	6	Exon	G/GGAA, G/A	Non-syn	Linked to SNPs found in GWAS
Sb06g022440	51659928	6	Intron	T/C	N/A	From GWAS
Sb06g022440	51659947	6	Intron	C/G	N/A	From GWAS
Sb06g022450	51673670	6	Intron	C/T	N/A	From GWAS
Sb06g022450	51674775	6	Exon	A/G	Syn	Linked to SNPs found in GWAS
Sb06g022480	51711306	6	Intron	T/C	N/A	From GWAS
Sb06g022490	51715340	6	Exon	T/G	Non-syn	Found in GWAS but not in *S. propinquum*
Sb06g022490	51715351	6	Exon	A/G	Syn	Linked to SNPs found in GWAS
Sb06g022490	51715398	6	Exon	G/T	Syn	From GWAS
Sb06g022500						No associated locus
Sb06g022600						No associated locus
Sb06g022610						No associated locus
Sb06g022620						No associated locus

Syn: synonymous mutation. Non-syn: non-synonymous mutation

Supplemental Table S3.3. Synonym for gene model v1.4 in gene model v3.1 at the selected candidate gene region. Genes in v1.4 that has no synonym in v3.1 are either altered or merged.

Gene model v3.1	Gene model v1.4
Sobic.006G145300.1	Sb06g022385
Sobic.006G145500.1	Sb06g022410
Sobic.006G145700.1	Sb06g022450
Sobic.006G145800.1	Sb06g022460
Sobic.006G146000.1	Sb06g022490
Sobic.006G146100.1	Sb06g022500
Sobic.006G146200.1	Sb06g022510
Sobic.006G146300.1	Sb06g022520
Sobic.006G146500.1	Sb06g022540
Sobic.006G146700.1	Sb06g022550
Sobic.006G146800.1	Sb06g022560
Sobic.006G146900.1	Sb06g022570
Sobic.006G147101.1	Sb06g022585
Sobic.006G147200.1	Sb06g022590
Sobic.006G147300.1	Sb06g022600
Sobic.006G147400.1	Sb04g036340
Sobic.006G147450.1	Sb04g036336
Sobic.006G147500.1	Sb06g022610
Sobic.006G147600.1	Sb06g022620

Supplemental Table S3.4. Possible causative loci of the ‘dry’ trait

Chr.	Exon	Location	Strand	Ref	Alt	Ref_aa	Alt_aa
6	1*	51614136	-	C	T/+TTC	L	F
6	4	51606932	-	G	A	G	S
6	5	51606549	-	C	A	Q	K
6	7	51605990	-	A	C	N	H
6	9	51605590	-	A	G	H	R
6	11	51605005	-	A	G	S	G
6	12	51604789	-	C	G	T	S
6	12	51604763	-	G	A	A	T

*Causative locus reported by targeted re-sequencing.

Supplemental Table S3.5. Distribution of individuals with W and G phenotypes in each country based on the Sorghum diversity panel

Origin Country	**No. Of Individuals** (W09/G09, W10/G10)	**Average Percentage**	
		W	**G**
Arabia	1/0, 1/0	100.0%	0.0%
Botswana	2/0, 2/0	100.0%	0.0%
Brazil	1/0, 1/0	100.0%	0.0%
Burkina Faso	1/0, 1/0	100.0%	0.0%
Central America	0/2, 0/2	0.0%	100.0%
Chad	2/1, 2/1	66.7%	33.3%
China	1/3, 3/0	62.5%	37.5%
Congo	1/0, 1/0	100.0%	0.0%
Egypt	2/0, 2/0	100.0%	0.0%
Ethiopia	19/17, 19/17	52.8%	47.2%
Gambia	1/0, 1/0	100.0%	0.0%
Guatemala	1/0, 0/1	50.0%	50.0%
India	16/8, 15/8	65.9%	34.1%
Japan	1/4, 4/1	50.0%	50.0%
Kenya	2/0, 2/0	100.0%	0.0%
Malawi	0/0, 1/0	100.0%	0.0%
Mali	2/3, 4/0	70.0%	30.0%
Mosambique	1/0, 1/0	100.0%	0.0%
Niger	0/2, 1/1	25.0%	75.0%
Nigeria	13/5, 14/3	77.3%	22.7%
Pakistan	1/0, 1/0	100.0%	0.0%
Senegal	2/0, 2/0	100.0%	0.0%
Somalia	1/1, 2/0	75.0%	25.0%
South Africa	3/7, 5/6	37.7%	62.3%
Sudan	23/17, 31/13	64.0%	36.0%
Tanzania	3/1, 5/1	79.2%	20.8%
Uganda	6/8, 9/7	49.6%	50.4%
United States	11/48, 13/47	20.2%	79.8%
Venezuela	1/0, 1/0	100.0%	0.0%
West Volta	0/1, 1/1	25.0%	75.0%
Zaire	1/0, 1/0	100.0%	0.0%
Zambia	0/1, 0/1	0.0%	100.0%
Zimbabwe	4/1, 4/1	80.0%	20.0%

Supplemental Table S3.6. Phylogeny of beta-glucosidase gene family

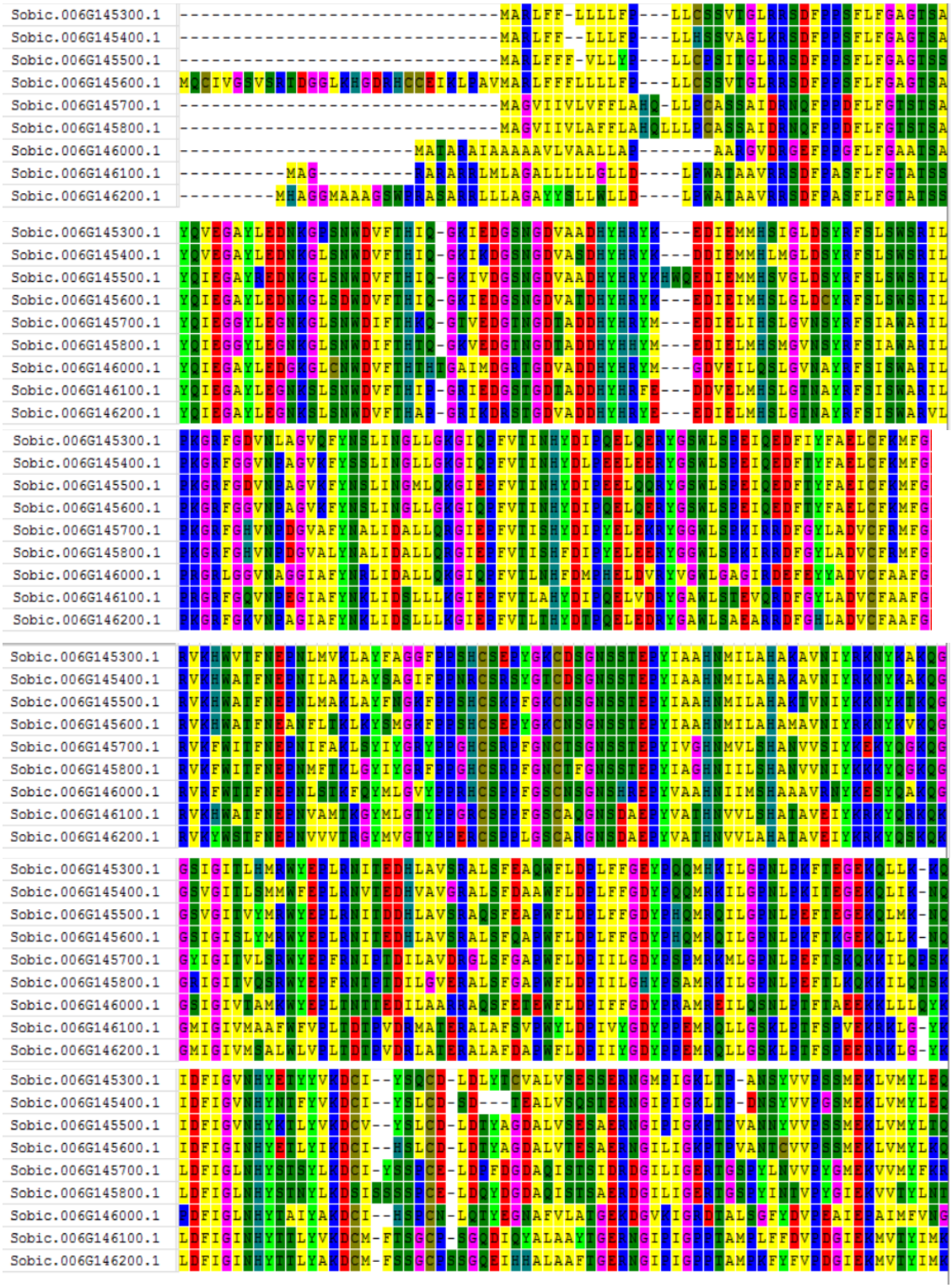
Group	*Sorgum bicolor*	*Zea mays*	*Brachypodium distachyon*	*Oryza sativa*	*Setaria italic*	*Musa acuminate*	*Arabidopsis thaliana*
I	Sobic.002G247600	GRMZM2G008247	Bradi1g42690	LOC_Os04g39814	Seita.2G254000	GSMUA_Ach	AT5G44640
	Sobic.006G117000	GRMZM2G014844	Bradi3g39997	LOC_Os04g39840	Seita.2G254100	r4G01460	AT5G42260
	Sobic.006G117100	GRMZM2G016890	Bradi3g40007	LOC_Os04g39864	Seita.4G134400	GSMUA_Ach	AT2G25630
	Sobic.006G117300	GRMZM2G076946	Bradi5g13265	LOC_Os04g39880	Seita.4G139700	r6G12410	AT2G44450
	Sobic.006G117400	GRMZM2G077015		LOC_Os04g39900	Seita.7G134800	GSMUA_Ach	AT3G60130
	Sobic.006G117600	GRMZM2G120962		LOC_Os06g21570	Seita.7G134900	r7G20500	AT2G44480
	Sobic.008G079800	GRMZM2G174699		LOC_Os08g39860	Seita.7G135200	GSMUA_Ach	AT1G51470
	Sobic.008G080100	GRMZM2G362362		LOC_Os08g39870		r8G24350	AT1G47600
	Sobic.008G080400	GRMZM2G426467		LOC_Os09g31410		GSMUA_Ach	AT1G51490
	Sobic.008G080600			LOC_Os09g31430		r9G07830	AT5G48375
	Sobic.010G132400					GSMUA_Ach	AT5G25980
						r9G12440	AT5G26000
							AT2G32860
							AT2G44460
							AT2G44470
							AT3G60140
							AT5G24550
							AT5G24540
							AT2G44490
							AT3G60120
							AT3G03640
							AT1G75940
							AT1G52400
							AT3G21370
							AT5G28510
							AT3G09260
							AT1G66280
							AT1G66270

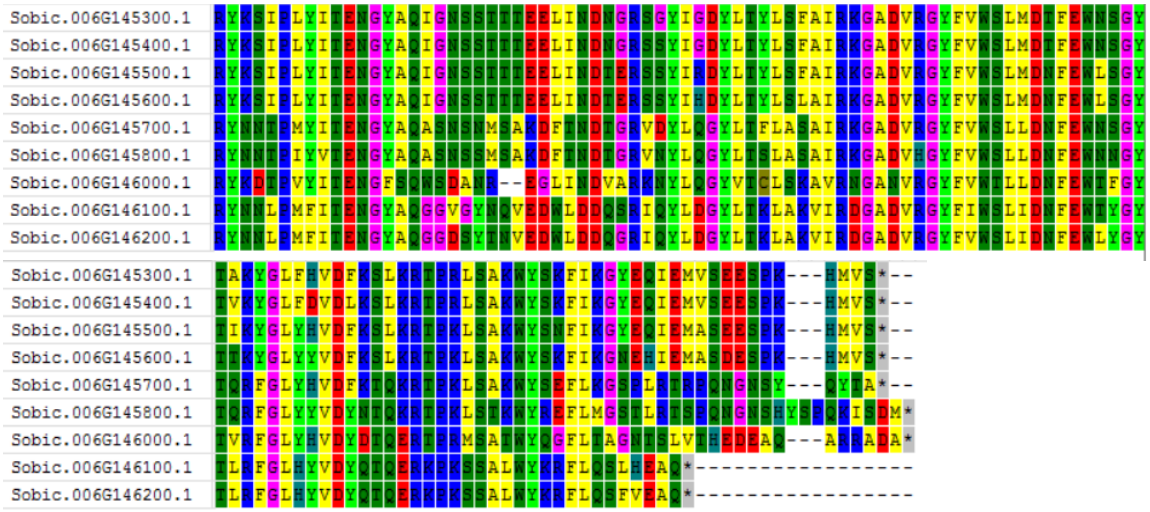
II	Sobic.003G332700 Sobic.009G114000 Sobic.009G113800 Sobic.010G170300 Sobic.002G261400 Sobic.002G261600	GRMZM2G069024 GRMZM2G055699 GRMZM2G376416 GRMZM2G108133	Bradi2g59660 Bradi2g59650 Bradi3g45627 Bradi3g45607 Bradi2g27770 Bradi2g09187 Bradi2g09200 Bradi4g34940 Bradi4g34927	LOC_Os01g70520 LOC_Os01g59840 LOC_Os01g59819 LOC_Os05g30390 LOC_Os05g30350 LOC_Os05g30250 LOC_Os05g30280 LOC_Os09g33680 LOC_Os09g33710 LOC_Os05g30300 LOC_Os09g33690	Seita.5G356600 Seita.5G356800 Seita.3G270400 Seita.3G270000 Seita.5G056800 Seita.2G270500 Seita.6G118900	GSMUA_Achr11G05590 GSMUA_Achr5G21420 GSMUA_Achr9G16420 GSMUA_Achr9G16470 GSMUA_Achr11G06730	AT4G27830 AT4G27820 AT3G62750 AT3G62740 AT1G02850 AT4G22100 AT1G45191 AT1G60270 AT1G60090 AT1G60260
III	Sobic.001G458400 Sobic.001G123300 Sobic.001G123200 Sobic.001G123100 Sobic.003G100000 Sobic.002G400600 Sobic.010G240300 Sobic.010G233900 Sobic.001G153800 Sobic.001G458600	GRMZM2G015804 GRMZM5G828987 GRMZM2G163544 GRMZM2G112704 GRMZM2G148176 GRMZM2G118003 AC217401.3_FG001 AC217401.3_FG002 AC234160.1_FG003	Bradi1g10930 Bradi1g10917 Bradi1g10890 Bradi1g10940 Bradi1g19270 Bradi4g08040 Bradi4g08040 Bradi1g33040 Bradi3g00650 Bradi1g70170	LOC_Os03g49600 LOC_Os03g49610 LOC_Os01g32364 LOC_Os07g46280 LOC_Os12g23170 LOC_Os06g46940 LOC_Os10g17650 LOC_Os03g11420	Seita.9G125400 Seita.9G125300 Seita.2G415300 Seita.3G304500 Seita.4G255800 Seita.9G492700 Seita.9G492600 Seita.9G492500	GSMUA_Achr3G29530 GSMUA_Achr3G20960 GSMUA_Achr8G33160 GSMUA_Achr1G26250 GSMUA_Achr9G07150 GSMUA_Achr6G27480 GSMUA_Achr6G27480 GSMUA_Achr6G27490	AT3G18080 AT3G18070 AT5G54570 AT1G26560

IV	Sobic.006G145300 Sobic.006G145400 Sobic.006G145500 Sobic.006G145600 Sobic.006G145700 Sobic.006G145800 Sobic.006G146000 Sobic.006G146100 Sobic.006G146200	GRMZM2G012236 GRMZM2G031660 GRMZM5G810727 GRMZM5G882852	Bradi5g15527 Bradi5g15540	LOC_Os04g43360 LOC_Os04g43380 LOC_Os04g43390 LOC_Os04g43400 LOC_Os04g43410	Seita.7G164100 Seita.7G164200 Seita.7G164300 Seita.7G164400 Seita.7G164500	GSMUA_Achr4G28840 GSMUA_Achr8G19060 GSMUA_Achr11G06570	AT1G61810 AT1G61820 AT4G21760
V	Sobic.003G389000	GRMZM2G044092 AC155376.2_FG005	Bradi2g57640	LOC_Os01g67220	Seita.5G414600	GSMUA_Achr3G21660	AT5G36890
VI	Sobic.005G217000	AC148152.3_FG008	Bradi4g09920	LOC_Os11g45710	Seita.8G237500	GSMUA_Achr1G18440	AT3G06510

```
Sb06g022390      MARLFFLLLLFPLLCSSVTGLRRSDFPPSFLFGAGTSAYQVEGA
Sobic.006G145300 MARLFFLLLLFPLLCSSVTGLRRSDFPPSFLFGAGTSAYQVEGA

Sb06g022390      YLEDNKGPSNWDVFTHIQGAFHWS*-------------------
Sobic.006G145300 YLEDNKGPSNWDVFTHIQG-----KIEDGSNGDVAADHYHRYKE

Sb06g022390      ---------------------------------------------
Sobic.006G145300 DIEMMHSIGLDSYRFSLSWSRILPKGRFGDVNLAGVQFYNSLIN

Sb06g022390      ---------------------------------------------
Sobic.006G145300 GLLGKGIQPFVTINHYDIPQELQERYGSWLSPEIQEDFIYFAEL

Sb06g022390      ---------------------------------------------
Sobic.006G145300 CFKMFGDRVKHWVTFNEPNLMVKLAYFAGGFPPSHCSEPYGKCD

Sb06g022390      ---------------------------------------------
Sobic.006G145300 SGNSSTEPYIAAHNMILAHAKAVNIYRKNYKAKQGGSIGITLHM

Sb06g022390      ---------------------------------------------
Sobic.006G145300 RWYEPLRNITEDHLAVSRALSFEAQWFLDPLFFGEYPQQMHKIL

Sb06g022390      ---------------------------------------------
Sobic.006G145300 GPNLPKFTEGEKQLLKKQIDFIGVNHYETYYVKDCIYSQCDLDL

Sb06g022390      ---------------------------------------------
Sobic.006G145300 YTCVALVSESSERNGMPIGKLTPANSYVVPSSMEKLVMYLEQRY

Sb06g022390      ---------------------------------------------
Sobic.006G145300 KSIPLYITENGYAQIGNSSTTTEELINDNGRSGYIGDYLTYLSF

Sb06g022390      ---------------------------------------------
Sobic.006G145300 AIRKGADVRGYFVWSLMDTFEWNSGYTAKYGLFHVDFKSLKRTP

Sb06g022390      --------------------------------
Sobic.006G145300 RLSAKWYSKFIKGYEQIEMVSEESPKHMVS*
```

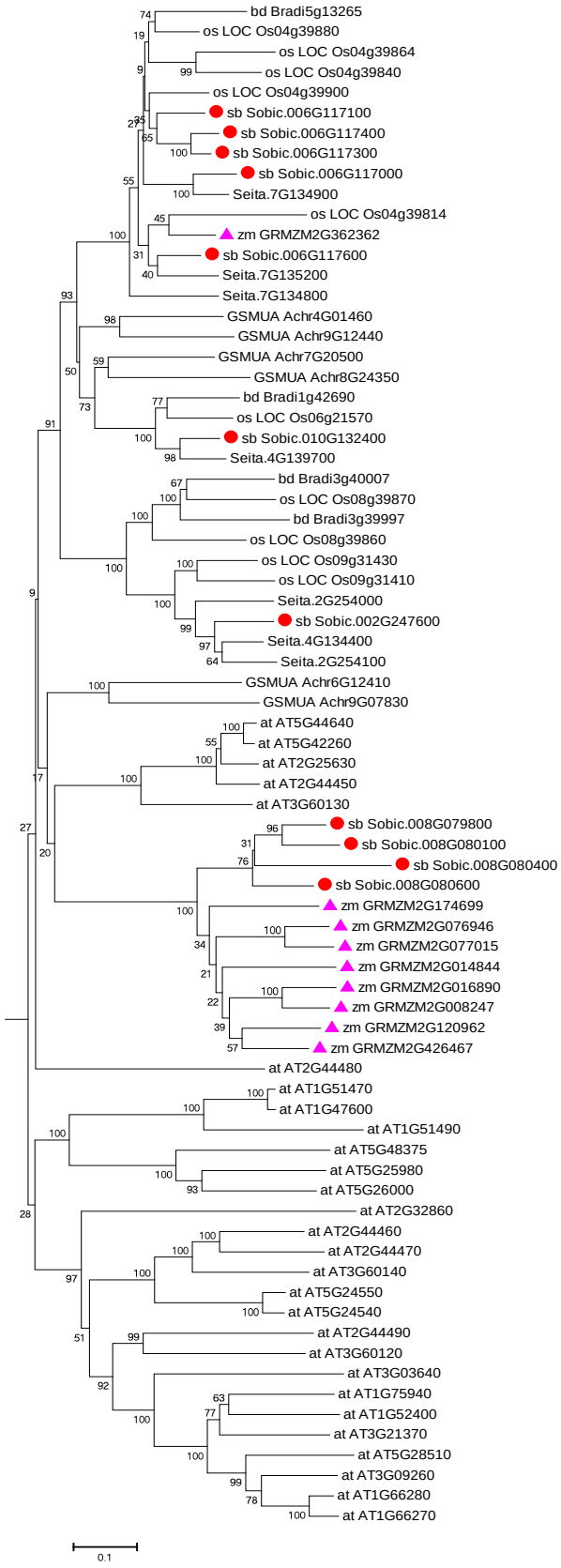
Supplemental Figure S3.1. Comparison of the D protein sequence in *Sorghum bicolor* genome annotation version 1.4 and version 3.1.


Sobic.006G145300.1
Sobic.006G145400.1
Sobic.006G145500.1
Sobic.006G145600.1
Sobic.006G145700.1
Sobic.006G145800.1
Sobic.006G146000.1
Sobic.006G146100.1
Sobic.006G146200.1

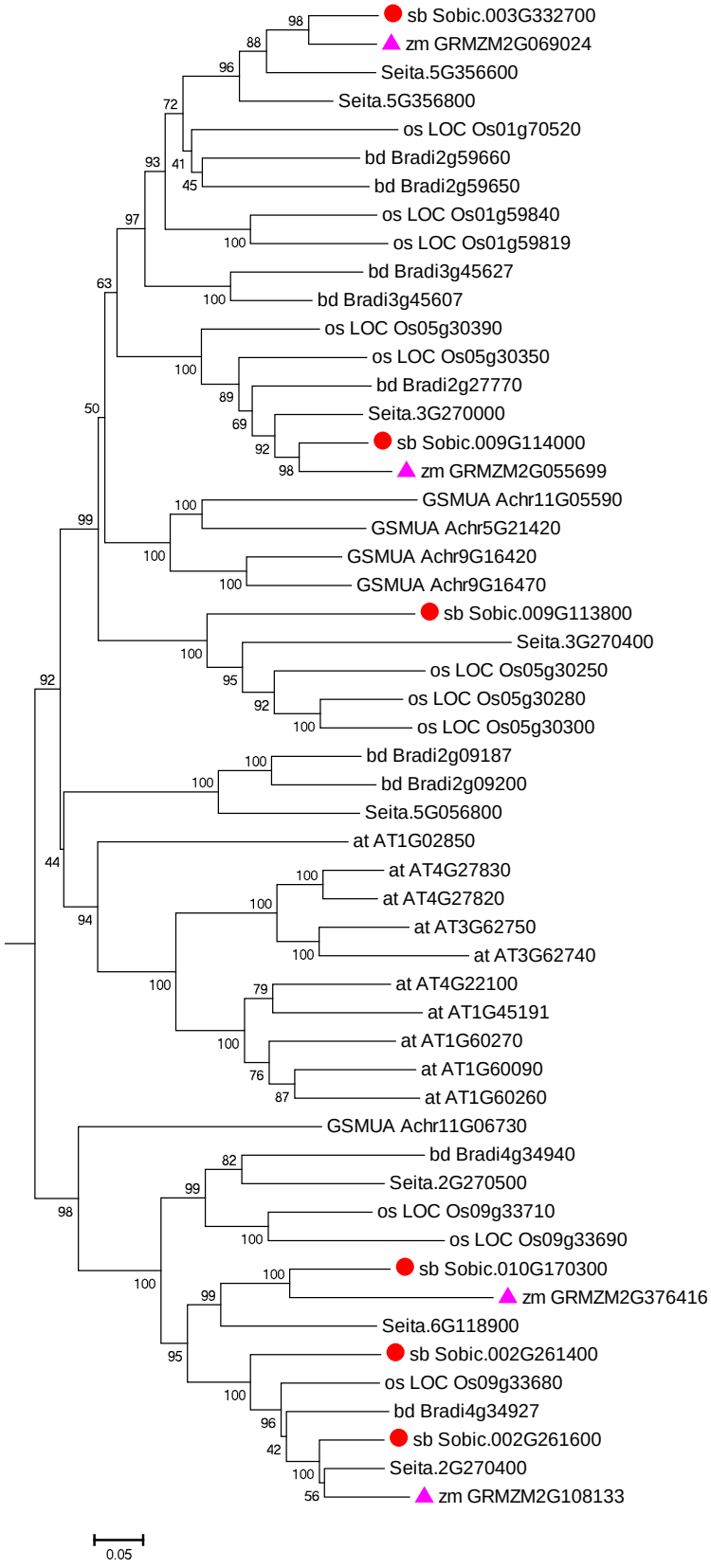


Supplemental Figure S3.2. Alignment of 9 beta-glucosidase proteins at the D candidate gene regions. The different colors highlight the varied chemical properties of amino acids.

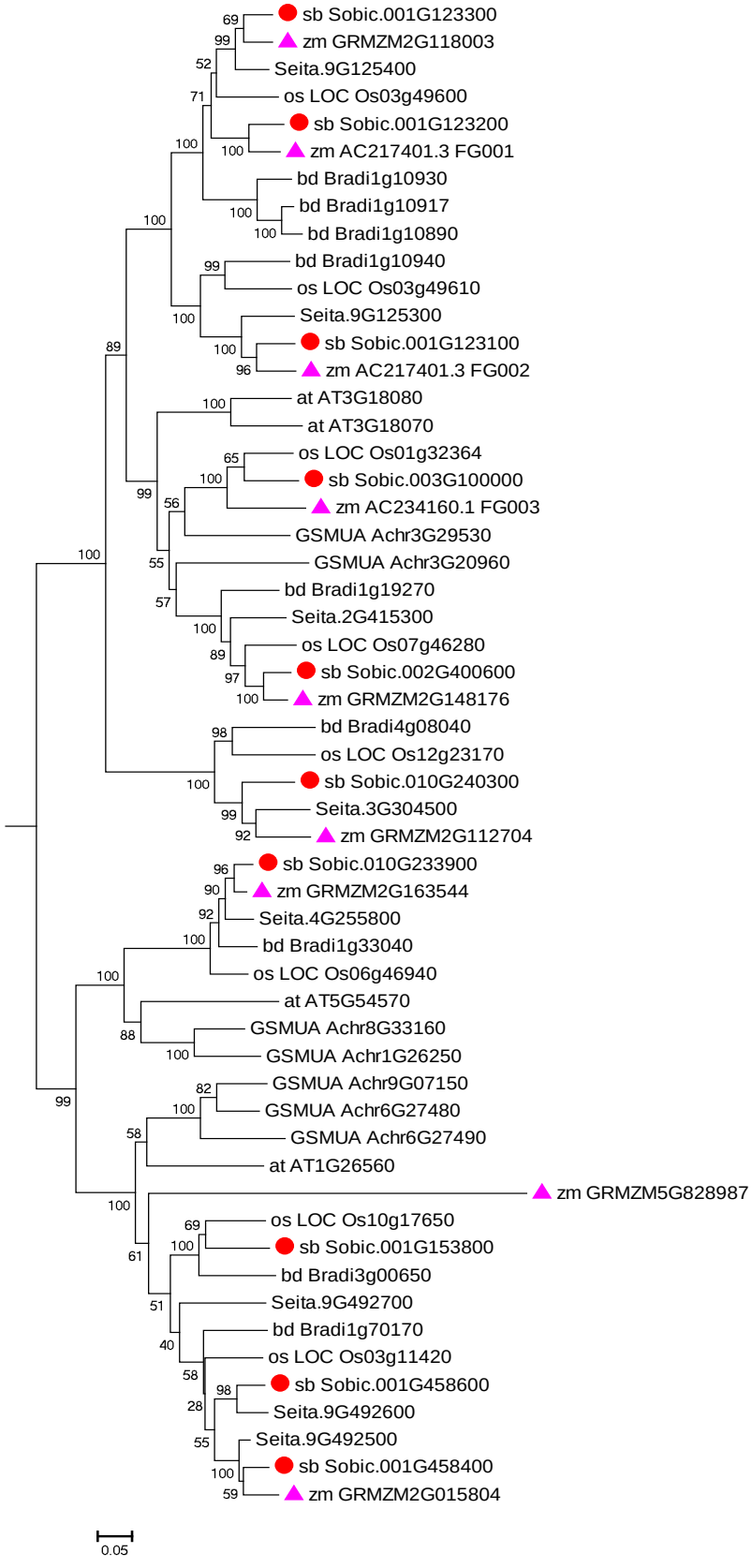
a)


bd Bradi5g13265
os LOC Os04g39880
os LOC Os04g39864
os LOC Os04g39840
os LOC Os04g39900
sb Sobic.006G117100
sb Sobic.006G117400
sb Sobic.006G117300
sb Sobic.006G117000
Seita.7G134900
os LOC Os04g39814
zm GRMZM2G362362
sb Sobic.006G117600
Seita.7G135200
Seita.7G134800
GSMUA Achr4G01460
GSMUA Achr9G12440
GSMUA Achr7G20500
GSMUA Achr8G24350
bd Bradi1g42690
os LOC Os06g21570
sb Sobic.010G132400
Seita.4G139700
bd Bradi3g40007
os LOC Os08g39870
bd Bradi3g39997
os LOC Os08g39860
os LOC Os09g31430
os LOC Os09g31410
Seita.2G254000
sb Sobic.002G247600
Seita.4G134400
Seita.2G254100
GSMUA Achr6G12410
GSMUA Achr9G07830
at AT5G44640
at AT5G42260
at AT2G25630
at AT2G44450
at AT3G60130
sb Sobic.008G079800
sb Sobic.008G080100
sb Sobic.008G080400
sb Sobic.008G080600
zm GRMZM2G174699
zm GRMZM2G076946
zm GRMZM2G077015
zm GRMZM2G014844
zm GRMZM2G016890
zm GRMZM2G008247
zm GRMZM2G120962
zm GRMZM2G426467
at AT2G44480
at AT1G51470
at AT1G47600
at AT1G51490
at AT5G48375
at AT5G25980
at AT5G26000
at AT2G32860
at AT2G44460
at AT2G44470
at AT3G60140
at AT5G24550
at AT5G24540
at AT2G44490
at AT3G60120
at AT3G03640
at AT1G75940
at AT1G52400
at AT3G21370
at AT5G28510
at AT3G09260
at AT1G66280
at AT1G66270
0.1

b)


sb Sobic.003G332700
zm GRMZM2G069024
Seita.5G356600
Seita.5G356800
os LOC Os01g70520
bd Bradi2g59660
bd Bradi2g59650
os LOC Os01g59840
os LOC Os01g59819
bd Bradi3g45627
bd Bradi3g45607
os LOC Os05g30390
os LOC Os05g30350
bd Bradi2g27770
Seita.3G270000
sb Sobic.009G114000
zm GRMZM2G055699
GSMUA Achr11G05590
GSMUA Achr5G21420
GSMUA Achr9G16420
GSMUA Achr9G16470
sb Sobic.009G113800
Seita.3G270400
os LOC Os05g30250
os LOC Os05g30280
os LOC Os05g30300
bd Bradi2g09187
bd Bradi2g09200
Seita.5G056800
at AT1G02850
at AT4G27830
at AT4G27820
at AT3G62750
at AT3G62740
at AT4G22100
at AT1G45191
at AT1G60270
at AT1G60090
at AT1G60260
GSMUA Achr11G06730
bd Bradi4g34940
Seita.2G270500
os LOC Os09g33710
os LOC Os09g33690
sb Sobic.010G170300
zm GRMZM2G376416
Seita.6G118900
sb Sobic.002G261400
os LOC Os09g33680
bd Bradi4g34927
sb Sobic.002G261600
Seita.2G270400
zm GRMZM2G108133
0.05

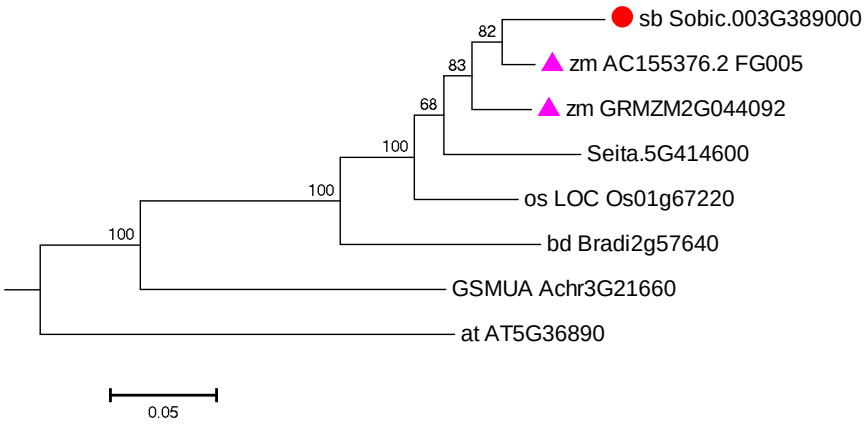
c)


sb Sobic.001G123300
zm GRMZM2G118003
Seita.9G125400
os LOC Os03g49600
sb Sobic.001G123200
zm AC217401.3 FG001
bd Bradi1g10930
bd Bradi1g10917
bd Bradi1g10890
bd Bradi1g10940
os LOC Os03g49610
Seita.9G125300
sb Sobic.001G123100
zm AC217401.3 FG002
at AT3G18080
at AT3G18070
os LOC Os01g32364
sb Sobic.003G100000
zm AC234160.1 FG003
GSMUA Achr3G29530
GSMUA Achr3G20960
bd Bradi1g19270
Seita.2G415300
os LOC Os07g46280
sb Sobic.002G400600
zm GRMZM2G148176
bd Bradi4g08040
os LOC Os12g23170
sb Sobic.010G240300
Seita.3G304500
zm GRMZM2G112704
sb Sobic.010G233900
zm GRMZM2G163544
Seita.4G255800
bd Bradi1g33040
os LOC Os06g46940
at AT5G54570
GSMUA Achr8G33160
GSMUA Achr1G26250
GSMUA Achr9G07150
GSMUA Achr6G27480
GSMUA Achr6G27490
at AT1G26560
zm GRMZM5G828987
os LOC Os10g17650
sb Sobic.001G153800
bd Bradi3g00650
Seita.9G492700
bd Bradi1g70170
os LOC Os03g11420
sb Sobic.001G458600
Seita.9G492600
Seita.9G492500
sb Sobic.001G458400
zm GRMZM2G015804
0.05

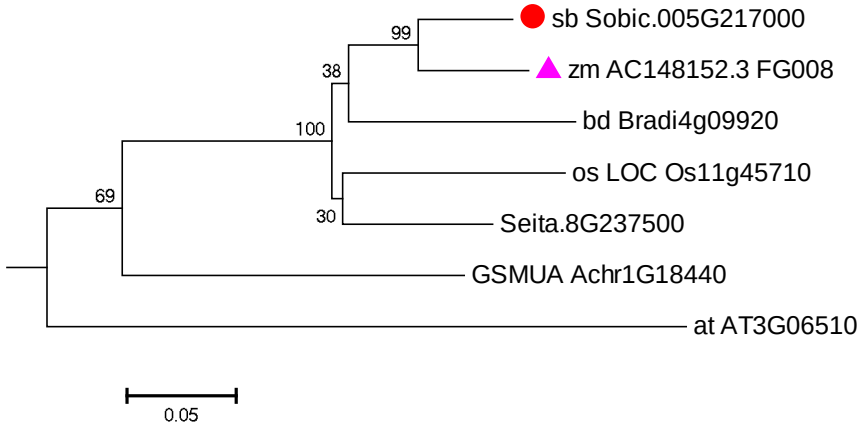
d)

bd Bradi5g15527
Seita.7G164100
Seita.7G164200
os LOC Os04g43360
sb Sobic.006G145800
sb Sobic.006G145700
sb Sobic.006G145300
sb Sobic.006G145400
sb Sobic.006G145600
sb Sobic.006G145500
zm GRMZM2G012236
zm GRMZM5G882852
GSMUA Achr8G19060
GSMUA Achr4G28840
GSMUA Achr11G06570
os LOC Os04g43410
Seita.7G164400
Seita.7G164500
zm GRMZM2G031660
sb Sobic.006G146200
sb Sobic.006G146100
os LOC Os04g43400
os LOC Os04g43390
os LOC Os04g43380
bd Bradi5g15540
zm GRMZM5G810727
sb Sobic.006G146000
Seita.7G164300
at AT4G21760
at AT1G61820
at AT1G61810
0.05

e)



f)



Supplemental Figure S3.3. Six branches of the tree corresponding to group I to VI (a to f). Sorghum BGLU genes are marked in green circles, while maize BGLU genes are marked in fuchsia triangles.

3.6 References

ALJANABI, S. M., L. FORGET and A. DOOKUN, 1999 An improved and rapid protocol for the isolation of polysaccharide- and polyphenol-free sugarcane DNA. Plant Molecular Biology Reporter **17:** 281-281.

BHATTRAMAKKI, D., J. M. DONG, A. K. CHHABRA and G. E. HART, 2000 An integrated SSR and RFLP linkage map of Sorghum bicolor (L.) Moench. Genome **43:** 988-1002.

BOWERS, J. E., C. ABBEY, S. ANDERSON, C. CHANG, X. DRAYE *et al.*, 2003 A high-density genetic recombination map of sequence-tagged sites for sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. Genetics **165:** 367-386.

BROWN, P. J., W. L. ROONEY, C. FRANKS and S. KRESOVICH, 2008 Efficient mapping of plant height quantitative trait loci in a sorghum association population with introgressed dwarfing genes. Genetics **180:** 629-637.

BRZOBOHATY, B., I. MOORE, P. KRISTOFFERSEN, L. BAKO, N. CAMPOS *et al.*, 1993 Release of Active Cytokinin by a Beta-Glucosidase Localized to the Maize Root-Meristem. Science **262:** 1051-1054.

CASA, A. M., G. PRESSOIR, P. J. BROWN, S. E. MITCHELL, W. L. ROONEY *et al.*, 2008 Community resources and strategies for association mapping in sorghum. Crop Science **48:** 30-40.

CHANG, C. C., C. C. CHOW, L. C. TELLIER, S. VATTIKUTI, S. M. PURCELL *et al.*, 2015 Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience **4:** 7.

CHITTENDEN, L. M., K. F. SCHERTZ, Y. R. LIN, R. A. WING and A. H. PATERSON, 1994 A detailed RFLP map of Sorghum bicolor x S. propinquum, suitable for high-density mapping, suggests ancestral duplication of Sorghum chromosomes or chromosomal segments. Theor Appl Genet **87:** 925-933.

CICEK, M., and A. ESEN, 1998 Structure and expression of a dhurrinase (beta-glucosidase) from sorghum. Plant Physiology **116:** 1469-1478.

DHARMAWARDHANA, D. P., B. E. ELLIS and J. E. CARLSON, 1995 A Beta-Glucosidase from Lodgepole Pine Xylem Specific for the Lignin Precursor Coniferin. Plant Physiology **107:** 331-339.

ELSHIRE, R. J., J. C. GLAUBITZ, Q. SUN, J. A. POLAND, K. KAWAMOTO *et al.*, 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One **6:** e19379.

FELTUS, F. A., and J. P. VANDENBRINK, 2012 Bioenergy grass feedstock: current options and prospects for trait improvement using emerging genetic, genomic, and systems biology toolkits. Biotechnology for Biofuels **5**.

FINN, R. D., P. COGGILL, R. Y. EBERHARDT, S. R. EDDY, J. MISTRY *et al.*, 2016 The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research **44:** D279-D285.

FRITZ, J. O., K. J. MOORE and E. H. JASTER, 1990 Digestion Kinetics and Cell-Wall Composition of Brown Midrib Sorghum X Sudangrass Morphological Components. Crop Science **30:** 213-219.

GOMEZ-ANDURO, G., E. A. CENICEROS-OJEDA, L. E. CASADOS-VAZQUEZ, C. BENCIVENNI, A. SIERRA-BELTRAN *et al.*, 2011 Genome-wide analysis of the beta-

glucosidase gene family in maize (Zea mays L. var B73). Plant Molecular Biology **77:** 159-183.

GORDON, C. H., 1967 Storage Losses in Silage as Affected by Moisture Content and Structure. Journal of Dairy Science **50:** 397-&.

GRAUR, D., and W.-H. LI, 2000 *Fundamentals of molecular evolution*. Sinauer, Sunderland, Ma.

GUO, H., 2015 The Evolution of Crop Genetic Diveristy (Unpublished doctoral dissertation), pp., University of Georgia, Athens GA.

HAN, Y. C., P. LV, S. L. HOU, S. Y. LI, G. S. JI *et al.*, 2015 Combining Next Generation Sequencing with Bulked Segregant Analysis to Fine Map a Stem Moisture Locus in Sorghum (Sorghum bicolor L. Moench). Plos One **10**.

HART, G. E., K. F. SCHERTZ, Y. PENG and N. H. SYED, 2001 Genetic mapping of Sorghum bicolor (L.) Moench QTLs that control variation in tillering and other morphological characters. Theoretical and Applied Genetics **103:** 1232-1242.

HENDERSON, C. R., 1975 Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics **31:** 423-447.

HILSON, G. R., 1916 A NOTE ON THE INHERITANCE OF CERTAIN STEM CHARACTERS IN SORGHUM. Agr. Jour. India **11:** 150-155.

HOSEL, W., I. TOBER, S. H. EKLUND and E. E. CONN, 1987 Characterization of Beta-Glucosidases with High Specificity for the Cyanogenic Glucoside Dhurrin in Sorghum-Bicolor (L) Moench Seedlings. Archives of Biochemistry and Biophysics **252:** 152-162.

KOJIMA, M., J. E. POULTON, S. S. THAYER and E. E. CONN, 1979 Tissue Distributions of Dhurrin and of Enzymes Involved in Its Metabolism in Leaves of Sorghum-Bicolor. Plant Physiology **63:** 1022-1028.

KONG, W., H. JIN, C. D. FRANKS, C. KIM, R. BANDOPADHYAY *et al.*, 2013 Genetic analysis of recombinant inbred lines for Sorghum bicolor x Sorghum propinquum. G3 (Bethesda) **3:** 101-108.

KONG, W., C. KIM, H. GUO and A. H. PATERSON, 2016 Genetic mapping in recombinant inbred lines of Sorghum bicolor BTx623 × IS3620C, pp., Unpublished manuscript.

KOSAMBI, D. D., 1944 The estimation of map distance from recombination values. Annu. Eugen. **12:** 172-175.

LANDER, E. S., P. GREEN, J. ABRAHAMSON, A. BARLOW, M. J. DALY *et al.*, 1987 MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics **1:** 174-181.

LEAH, R., J. KIGEL, I. SVENDSEN and J. MUNDY, 1995 Biochemical and Molecular Characterization of a Barley Seed Beta-Glucosidase. Journal of Biological Chemistry **270:** 15789-15797.

LIPKA, A. E., F. TIAN, Q. S. WANG, J. PEIFFER, M. LI *et al.*, 2012 GAPIT: genome association and prediction integrated tool. Bioinformatics **28:** 2397-2399.

LIU, X., L. GUO, J. YOU, X. LIU, Y. HE *et al.*, 2010 Progress of Segregation Distortion in Genetic Mapping of Plants. Research Journal of Agronomy **4:** 78-83.

MAHAJAN, C., K. PATEL, B. M. KHAN and S. S. RAWAL, 2015 In silico ligand binding studies of cyanogenic beta-glucosidase, dhurrinase-2 from Sorghum bicolor. Journal of Molecular Modeling **21**.

MARTIN, J. H., J. A. HARRIS and I. D. JONES, 1931 Freezing-point depression and specific conductivity of sorghum tissue fluids. Journal of Agricultural Research **42:** 0057-0069.

MATTHIES, I. E., M. MALOSETTI, M. S. RODER and F. VAN EEUWIJK, 2014 Genome-wide association mapping for kernel and malting quality traits using historical European barley records. PLoS One **9:** e110046.

MORRELL, P. L., E. S. BUCKLER and J. ROSS-IBARRA, 2012 Crop genomics: advances and applications. Nature Reviews Genetics **13:** 85-96.

MORRIS, G. P., P. RAMU, S. P. DESHPANDE, C. T. HASH, T. SHAH *et al.*, 2013 Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci U S A **110:** 453-458.

MURRAY, S. C., A. SHARMA, W. L. ROONEY, P. E. KLEIN, J. E. MULLET *et al.*, 2008 Genetic Improvement of Sorghum as a Biofuel Feedstock: I. QTL for Stem Sugar and Grain Nonstructural Carbohydrates. Crop Science **48:** 2165-2179.

NGUYEN, H. T., and A. BLUM, 2004 Physiology and biotechnology integration for plant breeding, pp. xiii, 628 p. in *Books in soils, plants, and the environment*. Marcel Dekker,, New York.

PATERSON, A. H., 1996 *Genome mapping in plants*. Academic Press R.G. Landes Co.

PATERSON, A. H., J. E. BOWERS, R. BRUGGMANN, I. DUBCHAK, J. GRIMWOOD *et al.*, 2009 The Sorghum bicolor genome and the diversification of grasses. Nature **457:** 551-556.

PATERSON, A. H., K. F. SCHERTZ, Y. R. LIN, S. C. LIU and Y. L. CHANG, 1995 The Weediness of Wild Plants - Molecular Analysis of Genes Influencing Dispersal and Persistence of Johnsongrass, Sorghum-Halepense (L) Pers. Proceedings of the National Academy of Sciences of the United States of America **92:** 6127-6131.

SANDLER, L., and E. NOVITSKI, 1957 Meiotic Drive as an Evolutionary Force. American Naturalist **91:** 105-110.

SCHERTZ, K. F., J. A. VIERA and J. W. JOHNSON, 1978 Sorghum Stover Digestibility as Affected by Juiciness. Crop Science **18:** 456-458.

SHIRINGANI, A. L., M. FRISCH and W. FRIEDT, 2010 Genetic mapping of QTLs for sugar-related traits in a RIL population of Sorghum bicolor L. Moench. Theoretical and Applied Genetics **121:** 323-336.

SIMOS, G., C. A. PANAGIOTIDIS, A. SKOUMBAS, D. CHOLI, C. OUZOUNIS *et al.*, 1994 Barley Beta-Glucosidase - Expression during Seed-Germination and Maturation and Partial Amino-Acid-Sequences. Biochimica Et Biophysica Acta-General Subjects **1199:** 52-58.

SWANSON, A. F., AND PARKER, JOHN H., 1931 INHERITANCE OF SMUT RESISTANCE AND JUICINESS OF STALK IN THE SORGHUM CROSS, RED AMBER X FETERiTA. Jour. Hered **22:** 51-56.

SWIGONOVA, Z., J. S. LAI, J. X. MA, W. RAMAKRISHNA, V. LLACA *et al.*, 2004 Close split of sorghum and maize genome progenitors. Genome Research **14:** 1916-1923.

TAMURA, K., G. STECHER, D. PETERSON, A. FILIPSKI and S. KUMAR, 2013 MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Molecular Biology and Evolution **30:** 2725-2729.

TANG, H. B., H. E. CUEVAS, S. DAS, U. U. SEZEN, C. B. ZHOU *et al.*, 2013 Seed shattering in a wild sorghum is conferred by a locus unrelated to domestication. Proceedings of the National Academy of Sciences of the United States of America **110:** 15824-15829.

TANKSLEY, S. D., N. D. YOUNG, A. H. PATERSON and M. W. BONIERBALE, 1989 Rflp Mapping in Plant-Breeding - New Tools for an Old Science. Bio-Technology **7:** 257-264.

THAYER, S. S., and E. E. CONN, 1981 Subcellular-Localization of Dhurrin Beta-Glucosidase and Hydroxynitrile Lyase in the Mesophyll-Cells of Sorghum Leaf Blades. Plant Physiology **67:** 617-622.

THORNTON, K., 2003 libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics **19:** 2325-2327.

THORSTENSSON, E. M. G., D. R. BUXTON and J. H. CHERNEY, 1992 Apparent Inhibition to Digestion by Lignin in Normal and Brown Midrib Stems. Journal of the Science of Food and Agriculture **59:** 183-188.

UNTERGASSER, A., I. CUTCUTACHE, T. KORESSAAR, J. YE, B. C. FAIRCLOTH *et al.*, 2012 Primer3-new capabilities and interfaces. Nucleic Acids Research **40**.

VERDOUCQ, L., J. MORINIERE, D. R. BEVAN, A. ESEN, A. VASELLA *et al.*, 2004 Structural determinants of substrate specificity in family 1 beta-glucosidases - Novel insights from the crystal structure of sorghum dhurrinase-1, a plant beta-

glucosidase with strict specificity, in complex with its natural substrate. Journal of Biological Chemistry **279:** 31796-31803.

WANG, X. Y., J. P. WANG, D. C. JIN, H. GUO, T. H. LEE *et al.*, 2015 Genome Alignment Spanning Major Poaceae Lineages Reveals Heterogeneous Evolutionary Rates and Alters Inferred Dates for Key Evolutionary Events. Molecular Plant **8:** 885-898.

WIESE, G., and H. J. GRAMBOW, 1986 Indole-3-Methanol-Beta-D-Glucoside and Indole-3-Carboxylic Acid-Beta-D-Glucoside Are Products of Indole-3-Acetic-Acid Degradation in Wheat Leaf Segments. Phytochemistry **25:** 2451-2455.

WILKS, H. M., K. W. HART, R. FEENEY, C. R. DUNN, H. MUIRHEAD *et al.*, 1988 A Specific, Highly-Active Malate-Dehydrogenase by Redesign of a Lactate-Dehydrogenase Framework. Science **242:** 1541-1544.

XU, W., P. K. SUBUDHI, O. R. CRASTA, D. T. ROSENOW, J. E. MULLET *et al.*, 2000 Molecular mapping of QTLs conferring stay-green in grain sorghum (Sorghum bicolor L. Moench). Genome **43:** 461-469.

ZHANG, D., W. KONG, J. ROBERTSON, V. H. GOFF, E. EPPS *et al.*, 2015a Genetic analysis of inflorescence and plant height components in sorghum (Panicoidae) and comparative genetics with rice (Oryzoidae). BMC Plant Biol **15:** 107.

ZHANG, D., J. LI, R. O. COMPTON, J. ROBERTSON, V. H. GOFF *et al.*, 2015b Comparative Genetics of Seed Size Traits in Divergent Cereal Lineages Represented by Sorghum (Panicoidae) and Rice (Oryzoidae). G3 (Bethesda) **5:** 1117-1128.

ZHANG, Z., E. ERSOZ, C. Q. LAI, R. J. TODHUNTER, H. K. TIWARI *et al.*, 2010 Mixed linear model approach adapted for genome-wide association studies. Nat Genet **42:** 355-360.

ZHU, C., and J. YU, 2009 Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. Genetics **182:** 875-888.

CHAPTER 4

GENETIC ANALYSIS OF STEM DIAMETERS AND WATER CONTENTS TO IMPROVE BIOENERGY TRAITS IN SORGHUM [1]

[1] Huizhe Jin, Wenqian Kong and Andrew H. Paterson. To be submitted to *Theoretical and Applied Genetics*.

4.1 Abstract

This study identified quantitative trait loci (QTLs) for three stem diameters and stem/leaf water contents of sorghum in a *S. bicolor* BTx623 x IS3620c recombinant inbred line population of 399 genotypes, and validated in a sorghum diversity panel of 354 accessions using genome-wide association studies (GWAS). The observations of both co-localized and non-overlapping loci affecting stem diameter traits suggest that stem widths at different heights (base, middle, rachis) may share some common genetic control, but also have some distinct genetic influences. Co-localizations of stem diameter and water content traits with a number of other biomass traits, including plant height, flowering time, branching, stem volume and the 'dry' trait, suggest that their inheritance may be linked functionally (pleiotropy) or physically (linkage disequilibrium). Exploration of genomic regions responsible for these bioenergy traits can serve as a foundation for future positional cloning of causal genes, and contribute to genetic improvement of sorghum as a feedstock for biofuel production.

4.2 Introduction

Stem diameter, conferring thicker stems and reduced lodging (KASHIWAGI *et al.* 2008); and water content, are important traits for biofuel production. Given the advantages of sustainability and environmental friendliness, plant-based biofuel production has been explored in the United States since at least 1978, and has been envisioned by some to displace up to 30% of current U.S. petroleum consumption (PATZEK and PIMENTEL 2005). The US Energy Information Administration projects a 48% increase in energy consumption between 2012 and 2040 (EIA 2016). With

increasing energy demand and diminishing petroleum reserves, energy sources derived from renewable plant-based biomass are suggested to be the only direct substitute for fossil fuels available on a significant scale and with less pollution (TOMES *et al.* 2011).

Plants having potential as biofuel feedstocks should require limited inputs (irrigation, fertilizer, pesticides), produce high yields of biomass, and be convertible to bio-based products in a relatively efficient and economical manner (VERMERRIS 2011). Leading candidates include but are not limited to 'conventional' biofuel food cereals and oilseed food crops, e.g. corn (*Zea mays*) starch and sugarcane (*Saccharum officinarum*) for ethanol, soybean (*Glycine max*) for biodiesel; early 'advanced' biofuel non-food annual crops, e.g. sorghum (*Sorghum bicolor*) starch and sugar for ethanol, camelina (*Camelina sativa*) for biodiesel; biofuel perennial grasses and short-rotation forests, e.g. Miscanthus (*Miscanthus* spp.), poplar (*Populus* spp.) for lignocellulosic ethanol; and aquatic plants, e.g. algae for biomass-derived biodiesel (YOUNG 1980; CHAUMONT 1993; TOMES *et al.* 2011; PATERSON 2013).

Being the fifth most widely grown cereal crop in the world, sorghum was proposed as a potential biofuel feedstock over 30 years ago (BURTON 1986). A C4 crop that converts energy more efficiently into biomass than C3 plants, sorghum also has many advantages that make it an attractive biofuel feedstock, with lower need for fertilizers and pesticides than many crops, high water-use efficiency, drought tolerance, and wide adaptability to a variety of climates and soil conditions (SABALLOS 2008). In addition to its resilient nature, sorghum is appealing for biofuel production also for its short life cycle (4 months on average), abundant genetic diversity, and a history of improvement of lignocellulose, sugar and starch yields (ROONEY *et al.* 2007). Efficient

harvesting, storage, and transportation methodologies are in place, with various energy conversion options being explored.

To maximize biomass yield per geographic unit and production system, genetic improvement needs to target bioenergy traits such as stem diameter and plant water content. Stem diameter, thicker being preferred for biomass yield, is itself a measurement to monitor plant growth. In sugarcane, Milligan et al. (MILLIGAN *et al.* 1990) reported a high correlation between stem weight and stem diameter (r = 0.70). In sorghum, Murray et al. (MURRAY *et al.* 2008) reported a weak but significant correlation between sugar in dry stems and stem diameter (r = 0.11). In sorghum, Shiringani et al (SHIRINGANI *et al.* 2010) reported 13 stem diameter QTLs on chromosomes 1, 2, 3, 4, 6, 7, 8, and 9 in a *S. bicolor* SS79 x M71 RIL population, using 157 AFLP, SSR, and EST-SSR markers. However, previous studies mostly measured stem diameter only once at the base, e.g., 5 cm above the ground in Miscanthus (ATIENZA *et al.* 2003), 20 cm above the ground in sorghum (SHIRINGANI *et al.* 2010), or 40 cm above the ground in rice (KASHIWAGI and ISHIMARU 2004). By simultaneously measuring stem diameters at the base, midpoint and rachis, our study provides a more comprehensive investigation of QTLs affecting stem architecture.

Plant water content, with less being preferred for biomass yield and storage (GORDON 1967), is a useful indicator of plant moisture level and is associated with a variety of physiological mechanisms. Previous studies have indicated that this trait is plastic in nature and affected by many genes (MURRAY *et al.* 2008; HAN *et al.* 2015). Some QTLs have been reported in a number of sorghum mapping populations and with different measurement statistics (SHIRINGANI *et al.* 2010; GUAN *et al.* 2011; LV *et al.*

2013; HAN *et al.* 2015). For example, Guan et al. (GUAN *et al.* 2011) reported 6 QTLs for juice weight on chromosome 1, 4, 7 and 9 in a Shihong137 (grain sorghum) x L-Tian (sweet sorghum) cross, with the trait measured as wet weight of stems and leaves within 24 hours of harvesting. Shiringani et al. (SHIRINGANI *et al.* 2010) reported a single QTL for relative juice weight on chromosome 2 in a M71 (grain sorghum) x SS79 (sweet sorghum) RIL population, with the trait measured as the weight difference between wet and dry stems. One common drawback of these studies is that juiciness was not generally segregating well in mapping populations. For example, in Shiringani et al.'s (SHIRINGANI *et al.* 2010) study, both parents of the RILs were juicy -- although transgressive segregation was observed, the power and accuracy of QTL detections could be reduced (LI *et al.* 2005). Water content, the percent difference between wet and dry biomass weight, may be a better estimate of plant moisture level than juice weight or relative juice weight, as it normalizes the moisture level data to a scale of 0 to 1. Consistent with this definition, Han et al. (HAN *et al.* 2015) reported a major QTL (qSW6) for stem water content on chromosome 6, accounting for 34.7-56.9% of the phenotypic variation at different internodes. However, there is still a lack of knowledge about the genetic basis of leaf water content and the interaction between water contents of stems and leaves.

With the availability of genomic resources including genetic maps (BOWERS *et al.* 2003; Kong *et al.* unpublished data), a high quality genome sequence (PATERSON *et al.* 2009) and GBS SNP data (MORRIS *et al.* 2013) for a sorghum diversity panel (CASA *et al.* 2008), the genetic basis of stem diameter and water content traits can be elucidated. A combination of linkage and association genetic approaches was adopted here. Biparental linkage mapping of quantitative trait loci (QTLs) using molecular markers has been

widely used in plant breeding. Analysis of overlapping QTLs can suggest the genetic basis of relationships among traits. Genome-wide association studies (GWAS) can provide support to QTL detection by taking advantage of high-throughput genotyping strategies and long-term accumulation of historical recombination events. The lower false-positive rate of QTL mapping and higher resolution of GWAS complement each other, providing valuable information for trait enhancement while mitigating constraints of each approach to accelerate the process of gene mapping and identification (TANG *et al.* 2013).

In this study, we report QTLs for three stem diameters and stem/leaf water contents in a *S. bicolor* BTx623 x IS3620c RIL population. These QTLs were further examined using GWAS in a sorghum diversity panel (CASA *et al.* 2008). The relationships of these traits to other important bioenergy traits, including plant height, flowering time, branching, and the 'dry' midrib trait are discussed. Exploration of genomic regions responsible for stem diameter and water content traits in this study offers a means to validate the positions and effects of previously discovered QTLs, and can serve as a foundation for future positional cloning of causal genes. Genomic regions identified here can also contribute to genetic improvement of biofuel sorghum, further enhancing its merit as a means to capture and store solar energy.

4.3 Materials and Methods

4.3.1 Genotypes

The genetic map of the *S. bicolor* BTx623 x IS3620c RIL population used for QTL mapping, as previously described (Kong *et al.* unpublished data), was constructed

utilizing 399 individuals and 616 genotyping-by-sequencing (GBS) based SNP markers. It collectively spanned 1,404.8 cM on 10 linkage groups with a 3.8 cM average interval between consecutive markers.

The genotypes for genome-wide association study (GWAS) were generated for a US sorghum diversity panel (CASA *et al.* 2008), including a total of 265,487 SNPs in 27,412 annotated genes across 354 sorghum accessions (MORRIS *et al.* 2013).

4.3.2 Phenotypes

Phenotypic data for QTL mapping was measured for the *S. bicolor* BTx623 x IS3620c RIL population in 2011 and 2012. Phenotypic data for GWAS was measured for the sorghum diversity panel in 2009 and 2010. Both populations were grown at the University of Georgia Plant Science Farm near Watkinsville GA. Three stem diameters (base, middle, rachis) and four plant weights (wet stem weight, dry stem weight, wet leaf weight, dry leaf weight) were recorded. The stem diameters were measured using calibrated digital calipers. Fresh weights of leaves and stalks were measured and recorded at physiological maturity, with dry weights measured subsequently after drying to stable mass in a tobacco barn. The water contents of stems and leaves were defined as: (wet weight – dry weight) / wet weight * 100% respectively. Overall BLUP (Best Linear Unbiased Prediction) values used to detect QTLs were calculated for each line using the mixed model: moisture content ~ line + year + line * year + rep(year) by R software (RCORETEAM 2015).

4.3.3 Heritability

Broad-sense heritability for stem diameter traits and water content traits were calculated based on the impact of genotype (G), environment (E) and genotype by environment interaction (G X E), using the lme4() library of R version 3.2.3 (BATES *et al.* 2015). Different years (2009 and 2010) were treated as different environments. The mixed model utilized is: Trait ~ Line + Year + Line * Year + Rep(Year), where Line, Year and their interactions are all considered random factors. Variance components used to calculate heritability were determined by the restricted maximum-likelihood (REML) method, with their significance estimates tested by model comparison with likelihood ratio tests (LONGIN and WURSCHUM 2014). Broad-sense heritability was then calculated as: $H = var_{Line}/(var_{Line} + \frac{var_{Line\times Year}}{E} + \frac{var_{Residual}}{ER})$, in which E is the number of environments and R is the number of replications (KONG *et al.* 2014).

4.3.4 QTL mapping

QTLs were mapped by interval mapping using the R/qtl Package (BROMAN *et al.* 2003), with a LOD (logarithm of the odds ratio) threshold of 2.5. Significant QTLs detected were then selected and fitted into multiple-QTL models (ARENDS *et al.* 2010). Physical locations of QTLs were delineated by anchoring to the reference genome the two flanking markers nearest to the 1-LOD interval boundaries that have alignment information as described (ZHANG *et al.* 2013), based on colinearity between genetic and physical marker positions.

QTL nomenclature is as described (MCCOUCH *et al.* 1997), starting with a lowercase ‘q’ followed by abbreviations of trait names in capital letters, then the year (if

not for overall BLUP values, optional), chromosome number, and a decimal numeric identifier to differentiate multiple QTLs on the same chromosome.

4.3.5 GWAS

Genome-wide association mapping (GWAS) was conducted using a compressed Mixed Linear Model (cMLM), which took into account a genetic marker-based kinship matrix and a principal component-based population structure term (ZHANG *et al.* 2010). The model selection feature of Genomic Association and Prediction Integrated Tool (GAPIT) (LIPKA *et al.* 2012) was used to determine the compression level and the optimal number of principal components (ZHU and YU 2009). To ensure the quality of GWAS, log quantile-quantile (QQ) plots were used to monitor systematic sources of spurious associations. To determine the significance threshold of GWAS, a Bonferroni-like multiple testing correction (MATTHIES *et al.* 2014; ZHANG *et al.* 2015) was used rather than the traditional Bonferroni method. As not all SNPs were mutually independent, the assumption of the traditional Bonferroni method was violated, which may result in reduced power in detecting true associations. To balance an acceptable false positive rate with sufficient detection power, we integrated LD-information for each chromosome in determining the significance threshold as $\alpha/(\sum_i \frac{l_i}{d_i})$ (See Chapter 2 Materials and Methods), where l_i is the length of chromosome i, d_i is the extension of LD for chromosome i, and α is the desired significance level of all tests across the genome. As a result, a P-value of 1.96×10^{-5} was used as the significance threshold for GWAS.

4.3.6 Reference Genomes

Sorghum bicolor gene annotations refer to JGI annotation release Sbi1.4 (PATERSON *et al.* 2009).

4.4 Results and Discussion

4.4.1 Phenotypic distribution

Descriptive statistics for the three stem width variables: basal stem diameter (BD), middle stem diameter (MD), and rachis diameter (RD); four weight variables: wet stem weight (WSW), dry stem weight (DSW), wet leaf weight (WLW), and dry leaf weight (DLW); and two derived water-content variables: stem water content (SWC), leaf water content (LWC), of the sorghum diversity panel (CASA *et al.* 2008) and BTx623 x IS3620c RILs (Kong *et al.* unpublished data) are shown in Table 4.1. For both populations, basal stem diameter tends to be largest, followed by middle stem (MD) and rachis (RD). Plant water content is normally higher with less variation in stems (SWC) than leaves (LWC). Differences in means and ranges between years reflect environmental effects.

Correlation matrices of these 9 traits in the sorghum diversity panel and BTx623 x IS3620c RILs are in Table 4.2, with correlation coefficients ranging from -0.30 to 0.95. For many traits, correlation coefficients were similar between the two populations. Significant correlations were determined respectively among three stem variables: BD, MD, RD (Table 4.2, correlation in 2009/2010: $r_{BD:MD}$ = 0.60/0.57, $r_{BD:RD}$ = 0.56/0.68, $r_{MD:RD}$ = 0.72/0.72, all p-values < 0.0001); four weight variables: WSW, DSW, WLW, DLW (Table 4.2, $r_{DSW:WSW}$ = 0.92/0.95, $r_{DSW:DLW}$ = 0.74/0.85, $r_{DSW:WLW}$ = 0.58/0.75, r-

$r_{WSW:DLW}$ = 0.80/0.85, $r_{WSW:WLW}$ = 0.66/0.80, $r_{DLW:WLW}$ = 0.93/0.93, all p-values < 0.0001), and two water-content variables: SWC, LWC (Table 4.2, $r_{SWC:LWC}$ = 0.21/0.48, p-value < 0.0001), indicating possible overlapping QTLs in each set of traits. Positive correlations between stem diameter (BD, MD, RD) and biomass yield (DSW, DLW) are verified. However, the correlation of water-content variables to the other 7 traits varied to some extent between the two populations. There were positive correlation of SWC with BD, RD; LWC with WLW, DLW; and negative correlation of SWC with DSW in both populations. However, significant correlations of SWC with MD, WSW, WLW (Table 4.2, $r_{SWC:MD}$ = 0.06, p-value = 0.0171; $r_{SWC:WSW}$ = 0.12, $r_{SWC:WLW}$ = 0.17, p-value < 0.0001), and of LWC with BD, WSW, DSW (Table 4.2, $r_{LWC:BD}$ = 0.11, $r_{LWC:WSW}$ = 0.24, $r_{LWC:DSW}$ = 0.14, p-value < 0.0001) were only observed in the BTx623 x IS3620c RILs, while significant correlation of LWC with MD (Table 4.2, $r_{LWC:MD}$ = -0.09, p-value = 0.0138) was only observed in the sorghum diversity panel. The significant correlation of LWC and RD was slightly negative (Table 4.2, $r_{LWC:RD}$ = -0.09, p-value = 0.0200) in the sorghum diversity panel but positive (Table 4.2, $r_{LWC:RD}$ = 0.13, p-value < 0.0001) in the BTx623 x IS3620c RILs. In all, among the three stem diameter variables, basal diameter is consistently positively correlated with water contents in stems and leaves. From a biofuel perspective, however, exceptional genotypes in which basal stem diameter is associated with reduced water content would be preferred.

4.4.2 Heritability

Heritability estimates (Table 4.3) for stem diameter and water content traits were similar in the two populations, ranging from 0.38 for stem/leaf water content to 0.70 for middle stem diameter in the sorghum diversity panel, and from 0.25 for leaf water

content to 0.68 for basal stem diameter in the BTx623 x IS3620c RILs. The genotypic effects are statistically significant (p-value < 0.001) for all traits except SWC in the sorghum diversity panel (Table 4.3a), where only the environmental effect is significant (p-value < 0.001). This is caused by large variance in residuals, which could be due to stem wet and dry weights being highly variable among years (Table 4.1a). For stem diameter variables (BD, MD, RD), genotype by environment interactions are statistically significant (p-value < 0.001) in both populations. Their heritabilities were moderate-high, ranging from 0.55 to 0.70, similar to what was reported in *S. bicolor* M71 x SS79 RILs (h = 0.60) (SHIRINGANI *et al.* 2010). For water content variables, the environmental effects are statistically significant (p-value < 0.001) in both populations. Their heritabilities are moderate-low, ranging from 0.25 to 0.39, indicating more plasticity than stem diameter traits. In consideration of different levels of plasticity among traits, we adopt both overall BLUP values (see Materials and Methods) across environments (years) and single year values of the traits in QTL analyses.

4.4.3 QTLs for stem diameter traits

A total of 6 QTLs affecting basal stem diameter (on chromosomes 1, 3, 6, 7, 8), 6 QTLs affecting middle stem diameter (on chromosomes 1, 6, 7, 8), and 5 QTLs affecting rachis stem diameter (on chromosomes 1, 6, 8) were detected based on the overall BLUP values, jointly explaining 28.9%, 26.0% and 20.0% of phenotypic variation in full models respectively (Table 4.4). Additional environment-specific QTLs detected (Table S4.1) that do not overlap with those from overall BLUP values include: 2 for BD (on chromosome 4), 3 for MD (on chromosomes 1, 3, 4), and 1 for RD (on chromosome 3),

reflecting genotype x environment interactions as stated above (see section "Heritability").

Two overlapping QTL regions for BD, MD and RD (Table 4.5a, Figure 4.1) based on overall BLUP values were found on chromosomes 6 and 8. By taking into account QTLs based on single year values, a third overlapping QTL region for BD, MD and RD was detected on chromosome 3. In addition, 4 overlapping QTL regions for BD and MD were observed on chromosomes 1, 4, and 7 (Table 4.5b); 2 overlapping QTL regions for BD and RD were observed on chromosomes 1 and 6 (Table 4.5c); and 1 overlapping QTL region for MD and RD was observed on chromosome 1 (Table 4.5d). These genomic regions influencing 2 or more stem diameter traits support our expectation that stem widths at different stem heights (base, middle, rachis) may share some common genetic control. On the other hand, 1 non-overlapping QTL for BD was found on chromosome 4 (qBD12_4.2) for 2012 data. Likewise, 1 non-overlapping QTL for MD was found on chromosome 6 (qMD6.1) based on overall BLUP values. Both give evidence that stem widths at different heights may also have some degree of distinct genetic control.

In a previous study, Shiringani et al. (SHIRINGANI *et al.* 2010) reported 13 QTLs for stem diameter on chromosomes 1, 2, 3, 4, 6, 7, 8, and 9 in a *S. bicolor* SS79 x M71 RIL population, with the trait measured 20 cm above the ground. We determined the physical locations (Table S4.2) of most of these QTLs by aligning markers flanking the support intervals to the reference genome (PATERSON *et al.* 2009). Many of these QTLs were covered in our study, except the ones on chromosomes 2, 8 and 9 and a second QTL on chromosome 3. Most of those QTLs (except the one on chromosome 8) that eluded

detection in our QTL mapping were found by GWAS (below), which is a good complement to biparental linkage mapping. Besides, Brown et al. (BROWN *et al.* 2006) reported a QTL for rachis diameter in a smaller *S. bicolor* BTx623 x IS3620c RIL population of 119 lines. Located on chromosome 6 (3Mb to 49 Mb), the QTL was also detected in our study (qRD6.1). Curiously, qRD6.1 for RD and the co-localized qBD6.1 for BD have effects in the opposite direction of what would be expected based on parental phenotypes (Table 4.4): the BTx623 allele is associated with reduced stem diameter, even though BTx623 has much thicker stems than IS3620c. This opposite effect was also observed by Brown et al.

By investigating several stem diameters distributed over the plant rather than a single measurement, our study provides more information than prior studies about patterns of genetic control. For example, the 2 overlapping QTL regions for BD, MD and RD (Table 4.5a) on chromosomes 6 and 8; 3 overlapping QTL regions for BD and MD (Table 4.5b) on chromosomes 1 and 4; and 1 overlapping QTL region for MD and RD (Table 4.5d) on chromosome 1 that were discovered in our work, eluded detection in previous studies.

4.4.4 GWAS for stem diameter traits

Genome-wide association studies were conducted using 265,487 SNPs for the sorghum diversity panel (CASA *et al.* 2008) and trait data collected in 2009 and 2010 respectively, accounting for environmental differences. The significance threshold for GWAS was determined by using a Bonferroni-like approach (Materials and Methods), which took into account LD-decay information for each chromosome (MATTHIES *et al.* 2014), to better control the false positive rate while maintaining sufficient detection

power. This results in a threshold of 1.96×10^{-5} to achieve an experiment-wide significance level of 0.05 for all tests.

A total of 33 SNPs (14 for BD, 10 for MD, 9 for RD) (Figures 4.1-4.2, Table S4.3) of common variants (minor allele frequency $\geq$ 5%) were significantly associated with stem diameter traits, with no confounding by population stratification (Figure S4.1). False positives were removed based on marker locations and linkage disequilibrium, since spurious associations tend to stand alone (i.e. with no nearby SNPs showing association). Few significant associations were shared between the two years, a finding for which environmental factors (Table 4.3) might be a large element of the reason (MANOLIO *et al.* 2009). Even given 100% accurate information about phenotype and genotype, some associations may still be unrepeatable due to interaction between genotype and environment (ROSS-IBARRA *et al.* 2007).

To investigate the validity of the SNPs detected, we grouped them based on their locations and compared them to the QTLs (Figure 4.1, Table 4.6). Significant SNPs (on chromosomes 1, 3, 4, 6) for BD, MD and RD in association with QTLs were listed in Table 4.6. Other significant SNPs (on chromosomes 2, 3, 9) that were not covered by our linkage mapping results were found to be in association with some of the QTLs reported by Shiringani (SHIRINGANI *et al.* 2010). Among these, the ones on chromosome 9 suggested a fourth overlapping genomic region (54.1Mb to 59.4Mb) for BD, MD and RD; while associations on chromosome 2 (peak at 62.8Mb) and 3 (peak at 3.8Mb) were detected in year 2010 only, suggesting environment-specific loci for BD. This provides further support to our hypothesis stated earlier that stem widths at different heights (base,

middle, rachis), share some common genetic control but also have some distinct genetic influences.

4.4.5 QTLs for water content traits

A total of 2 QTLs affecting stem water content on chromosome 1 (only from 2012 data) and 6 (from BLUP), and 1 QTL affecting leaf water content on chromosome 1 (only from 2012 data) were detected, explaining 3.5%, 20.9% and 3.6% of phenotypic variation respectively (Table 4.7, Figure 4.1). There were no QTLs detected for LWC based on overall BLUP values and single year values from 2011. This could result from the plasticity of this trait. With a heritability of only 0.25 in the mapping population, LWC is significantly affected by environment and genotype x environment interactions (Table 4.3b). Hence, it is not surprising that the only overlapping QTL region detected for SWC and LWC, which locates on chromosome 1, was from 2012 data (Table 4.8, Figure 4.1). This genomic region also overlapped with a QTL for BD (qBD1.1) on chromosome 1 (Table 4.8). There is another overlapping QTL region detected for SWC (qSWC6.1) and RD (qRD6.2) on chromosome 6 (Table 4.8). These two QTL clusters, which may be explained by inter-relationship between traits, were supported by significant correlations in the mapping population among BD, SWC, and LWC (Table 4.2b), and between RD and SWC (Table 4.2b). They could be either due to pleotropic effects of single genes, or to proximal locations of different genes related to stem width and water content. In comparison to prior studies, qSWC6.1 (Table 4.7) detected on chromosome 6 was found to be contained in a previously report stem moisture locus qWC6 (HAN *et al.* 2015) in a Shihong137 (grain sorghum) x L-Tian (sweet sorghum) cross, while qSWC12_1.1, which

contained regions homoeologous to qSWC6.1 (Table S4.4), eluded detection in the previous report.

4.4.6 GWAS for water content traits

A total of 24 SNPs (20 for SWC, 4 for LWC) (Figure 4.1, 4.3, Table S4.5) of common variants (minor allele frequency ≥ 5%) were significantly associated with water content traits, with no confounding of population stratification (Figure S4.2). As with the analysis for stem diameter traits, these 24 SNPs were grouped based on their locations and compared to the QTLs (Figure 4.1, Table 4.9). Significant SNPs (on chromosomes 1, 6) in association with QTLs (qSWC6.1, qSWC12_1.1) for SWC were listed in Table 4.9. There are no significant SNPs associated with the QTL (qLWC12_1.1) for LWC, possibly due to the plasticity of the trait, since it is largely affected by environment (Table 4.3a). Other SNPs (Figure 4.1, Table S4.5) detected suggest three additional environment-specific loci for SWC on chromosomes 1 (peak at 59.2Mb), 6 (38.7Mb) and 9 (57.2Mb), and two for LWC on chromosome 6 (5.7Mb, 50.7Mb respectively). Thus, the GWAS result validated many QTLs, and indicated putative loci that eluded detection from linkage mapping in the study population. In addition, the improved resolution of GWAS can aid in targeted re-sequencing, e.g., by re-sequencing genes surrounding the peak of associations as to find causative loci. Together with other resources beyond the QTL meta-data and GWAS comparative data that we have used here, e.g., examining expression profiles in particular tissues, one can envision a practical path to the final determination of candidate genes.

4.4.7 Relationships among traits

Co-localizations of loci affecting stem diameter and water content traits with other biofuel-related traits were observed. On chromosome 6, there is a cluster of associations (39.6Mb to 47.2Mb) for basal stem diameter (qBD6.1, association peak at S6_44581098), middle stem diameter (association peak at S6_45935408), rachis diameter (qRD6.1, association peak at S6_45935408), plant height *Dw2* (MORRIS *et al.* 2013), flowering time *Ma6* (BRADY 2006) and primary branching number (BROWN *et al.* 2006; KONG *et al.* 2014). On chromosome 9, there is a cluster of associations (54.1Mb to 59.4Mb) for basal stem diameter (association peak at S9_56928114), middle stem diameter (association peak at S9_57240634), rachis diameter (association peak at S9_54137819), plant height and flowering time (ZHANG *et al.* 2015).

Another bioenergy trait of merit to look at is stem volume (SV). Assuming the stem is conical, stem volume can be calculated as: $SV = \frac{1}{24}\pi * (BD^2 + BD * MD + MD^2) * PH + \frac{1}{24}\pi * (MD^2 + MD * RD + RD^2) * PH$. QTLs (Table 4.10) detected reflect environmental effects, with only one common region (4.9Mb to 58.1Mb on Chr. 6) observed in 2 years. Obviously from the formula, stem volume must be positively correlated with stem diameters. Thus, it is not surprising that all QTLs detected for SV are found to locate at overlapping QTL regions of at least 2 stem diameter traits.

With regard to water content traits, there is co-localization of stem water content qSWC6.1 and the D gene for the 'dry' trait (see Chapter 2). As the 'dry' trait is also significantly associated with leaf water content (see Chapter 2), we were surprised to see no QTL detected for LWC on chromosome 6. However, GWAS identifies a locus (peak at 50.7Mb) very close to the D gene (<1Mb), complementing QTL mapping, especially

for traits of high plasticity. These co-localizations of bioenergy traits indicate that their inheritance may be linked functionally (pleiotropy) or physically (linkage disequilibrium).

4.5 Conclusion

This study identified the genetic basis for three stem diameter variables and two water content variables in sorghum by using a combination of linkage mapping and genome-wide association mapping approaches. QTLs for basal stem diameter (on chromosomes 1, 3, 4, 6, 7, 8), middle stem diameter (on chromosomes 1, 3, 4, 6, 7, 8), rachis diameter (on chromosomes 1, 3, 6, 8), stem water content (on chromosomes 1, 6), and leaf water content (on chromosome 1) were reported in a BTx623 x IS3620c RIL population and verified using GWAS in a sorghum diversity panel (CASA *et al.* 2008). GWAS, a complement to linkage mapping, also suggested several additional putative loci for each of the traits (BD: on Chr. 2, 3, 9; MD: on Chr. 9; RD: on Chr. 9, SWC: on Chr. 1, 6, 9; LWC: on Chr. 6). The observations of both co-localized and non-overlapping loci affecting stem diameter traits suggest that stem widths at different heights (base, middle, rachis) may share some common genetic control, but also have some distinct genetic influences. Besides, co-localizations of stem diameter and water content traits with a number of other bioenergy traits, including plant height, flowering time, branching, stem volume and the 'dry' trait, suggest that their inheritance may be functionally (pleiotropy) or physically (linkage disequilibrium) linked.

Increasing knowledge of the genetic control of stem diameter and water content traits, and in identification of corresponding genes and their functions, may lead to tools

and strategies for either enhancing or suppressing these traits, supporting advances toward greater use of plant-based biomass in biofuel production.

Table 4.1. Descriptive statistics for stem and water content traits in a sorghum diversity panel and BTx623 x IS3620c RILs

a) Sorghum diversity panel

Trait	Unit	2009				2010			
		MEAN	SD	Min	Max	MEAN	SD	Min	Max
BD	mm	14.39	3.76	7.50	47.70	19.29	4.47	6.40	48.70
MD	mm	9.30	2.59	3.90	18.65	11.93	3.26	5.50	26.70
RD	mm	7.25	1.71	3.05	12.05	8.82	2.07	2.50	17.20
WSW	g	103.68	91.18	13.65	731.05	239.71	236.38	22.50	2174.50
DSW	g	34.66	48.08	3.90	501.95	59.91	64.93	6.50	547.50
WLW	g	68.54	42.29	6.65	256.85	89.24	69.30	8.00	534.00
DLW	g	28.87	15.17	5.80	93.10	42.10	27.75	6.00	239.00
SWC	%	69.82	9.65	17.41	93.64	75.57	5.36	42.62	91.60
LWC	%	53.56	13.45	7.26	84.77	46.44	15.19	3.23	75.63

b) BTx623 x IS3620c RILs

Trait	Unit	2011				2012			
		MEAN	SD	Min	Max	MEAN	SD	Min	Max
BD	mm	14.82	3.05	4.10	24.95	15.24	3.29	7.43	26.66
MD	mm	9.55	2.48	2.80	19.30	9.26	2.03	4.46	16.69
RD	mm	7.52	1.51	3.10	12.30	7.65	1.48	3.75	15.19
WSW	g	185.28	148.67	15.50	1279.00	233.12	172.67	26.00	1196.50
DSW	g	59.28	45.60	5.50	371.00	61.24	50.97	7.00	476.00
WLW	g	106.12	77.00	7.00	471.50	149.56	99.06	9.50	717.50
DLW	g	49.68	29.84	6.00	184.00	52.98	33.80	4.50	230.50
SWC	%	67.32	4.89	48.85	81.58	74.07	4.36	34.97	85.79
LWC	%	47.85	13.13	4.95	75.57	62.67	8.23	25.93	87.21

BD basal stem diameter, MD middle stem diameter, RD rachis diameter, DSW dry stem weight, WSW wet stem weight, DLW dry leaf weight, WLW wet leaf weight, SWC stem water content, LWC leaf water content.

Table 4.2. Correlation coefficients of stem and water content traits

a) Sorghum diversity panel

	BD	MD	RD	WSW	DSW	WLW	DLW	SWC	LWC
BD	1								
MD	0.60***	1							
RD	0.56***	0.72***	1						
WSW	0.43***	0.53***	0.27***	1					
DSW	0.37***	0.50***	0.23***	0.92***	1				
WLW	0.46***	0.36***	0.26***	0.66***	0.58***	1			
DLW	0.52***	0.46***	0.34***	0.80***	0.74***	0.93***	1		
SWC	0.17***	0.05	0.14***	-0.03	-0.30***	0.04	-0.05	1	
LWC	0.01	-0.09*	-0.09*	0.00	-0.06	0.47***	0.20***	0.21***	1

b) BTx623 x IS3620c RILs

	BD	MD	RD	WSW	DSW	WLW	DLW	SWC	LWC
BD	1								
MD	0.57***	1							
RD	0.68***	0.72***	1						
WSW	0.25***	0.39***	0.22***	1					
DSW	0.22***	0.38***	0.18***	0.95***	1				
WLW	0.35***	0.27***	0.23***	0.80***	0.75***	1			
DLW	0.36***	0.31***	0.22***	0.85***	0.85***	0.93***	1		
SWC	0.15***	0.06*	0.19***	0.12***	-0.10***	0.17***	0.01	1	
LWC	0.11***	0.05	0.13***	0.24***	0.14***	0.47***	0.24***	0.48***	1

BD basal stem diameter, MD middle stem diameter, RD rachis diameter, DSW dry stem weight, WSW wet stem weight, DLW dry leaf weight, WLW wet leaf weight, SWC stem water content, LWC leaf water content.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.3. Heritability and variance components for stem diameter and water content traits, based on genotype, year, and genotype by year interaction percentages

a) Sorghum diversity panel

Traits	Genotype (%)	Year (%)	Genotype x year (%)	Residual (%)	Heritability
BD	21.70***	39.23***	7.02***	32.05	0.65
MD	32.15***	24.37***	11.44***	32.04	0.70
RD	25.71***	20.26***	18.63***	35.41	0.59
SWC	18.48	21.68***	< 0.01	59.84	0.38
LWC	21.24***	10.31***	38.58	29.87	0.38

b) BTx623 x IS3620c RILs

Traits	Genotype (%)	Year (%)	Genotype x year (%)	Residual (%)	Heritability
BD	38.18***	0.45	9.00***	52.37	0.68
MD	28.37***	0.60*	20.19***	50.84	0.55
RD	29.20***	0.15	16.70***	53.95	0.57
SWC	9.50***	45.95***	14.73***	29.81	0.39
LWC	6.32***	45.19***	28.09***	20.40	0.25

BD basal stem diameter, MD middle stem diameter, RD rachis diameter, SWC stem water content, LWC leaf water content.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.4. QTLs affecting stem diameter traits using overall BLUP values in BTx623 x IS3620c RILs

Trait	QTL name	Chr.	Peak (cM)	LOD	Additive[a]	R2 (%)	Start (Mb)[b]	End (Mb)[b]
BD	qBD1.1	1	67.0	3.54	0.82	3.01	1.9	59.4
	qBD3.1	3	113.8	4.83	-0.76	4.14	61.4	63.4
	qBD6.1	6	32.0	5.97	0.85	5.15	39.6	47.2
	qBD6.2	6	93.0	8.18	-1.00	7.16	56.6	59.3
	qBD7.1	7	97.0	4.41	-0.72	3.77	58.4	60.1
	qBD8.1	8	83.0	6.79	-0.92	5.88	51.8	52.8
MD	qMD1.1	1	96.5	3.85	0.42	3.43	50.0	53.7
	qMD1.2	1	124.9	5.25	-0.48	4.67	57.2	59.4
	qMD6.1	6	51.0	2.76	-0.35	2.44	48.5	50.6
	qMD6.2	6	82.0	2.54	-0.31	2.27	55.5	59.3
	qMD7.1	7	87.0	8.76	0.60	8.07	57.7	59.5
	qMD8.1	8	83.3	5.61	-0.48	5.01	51.8	52.8
RD	qRD1.1	1	25.7	3.97	-0.36	4.03	7.1	8.1
	qRD1.2	1	160.5	7.57	-0.36	7.02	66.7	68.2
	qRD6.1	6	4.0	3.12	0.26	3.24	0.0	47.2
	qRD6.2	6	69.0	3.78	-0.26	3.82	51.7	57.9
	qRD8.1	8	84.0	2.77	-0.24	3.09	50.2	54.4

BD basal stem diameter, MD middle stem diameter, RD rachis diameter

[a] Additive effects calculated as IS3620c – BTx623.

[b] Based on DNA marker locations flanking 1-LOD intervals in the published genome sequence (PATERSON *et al.* 2009).

Table 4.5. Overlapping QTL regions among different measures of stem diameter

a) BD, MD and RD

Chr.	BD	MD	RD	Start (Mb)	End (Mb)
3	qBD3.1	qMD12_3.1*	qRD12_3.1*	61.5	63.4
6	qBD6.2	qMD6.2	qRD6.2	56.6	57.9
8	qBD8.1	qMD8.1	qRD8.1	51.8	52.8

b) BD and MD

Chr.	BD	MD	Start (Mb)	End (Mb)
1	qBD1.1	qMD1.1	50.0	53.7
1	qBD1.1	qMD1.2	57.2	59.4
4	qBD12_4.1*	qMD12_4.1*	4.3	6.9
7	qBD7.1	qMD7.1	58.4	59.5

c) BD and RD

Chr.	BD	RD	Start (Mb)	End (Mb)
1	qBD1.1	qRD1.1	7.1	8.1
6	qBD6.1	qRD6.1	39.6	47.2

d) MD and RD

Chr.	MD	RD	Start (Mb)	End (Mb)
1	qMD11_1.1*	qRD1.2	66.7	68.2

BD basal stem diameter, MD middle stem diameter, RD rachis diameter
*QTLs based on single year values

Table 4.6. Stem diameter QTLs with corresponding significant SNPs

Trait	QTL	Chr.	Significant SNPs	Distance*
BD	qBD1.1	1	S1_22691388	Within†
BD	qBD3.1	3	S3_59642849	1.8Mb
BD	qBD12_4.1	4	S4_5581387, S4_5581390, S4_5774418	Within†
BD	qBD6.1	6	S6_44581098	Within†
MD	qMD6.1	6	S6_42736415, S6_46077506, S6_46194160, S6_46217845, S6_46217893, S6_46217966, S6_48234025	0.3Mb
RD	qRD1.2	1	S1_66715300	Within†
RD	qRD12_3.1	3	S3_60554780	1Mb
RD	qRD6.1	6	S6_45935408	Within†

BD basal stem diameter, MD middle stem diameter, RD rachis diameter
*Distance between QTL to the nearest significant SNP outside the QTL interval.
†Significant SNPs are contained in the QTL interval.

Table 4.7. QTLs affecting water content traits using overall BLUP and single year values in BTx623 x IS3620c RILs

Trait	Source	QTL name	Chr.	Peak (cM)	LOD	Additive[a]	R2 (%)	Start (Mb)[b]	End (Mb)[b]
SWC	BLUP	qSWC6.1	6	59.0	20.24	-0.0127	20.86	51.1	52.7
	2012	qSWC12_1.1[c]	1	60.0	3.28	-0.0110	3.48	18.4	26.2
LWC	2012	qLWC12_1.1[c]	1	70.0	3.01	-0.0207	3.55	19.8	72.4

SWC stem water content, LWC leaf water content

[a] Additive effects calculated as IS3620c – BTx623.

[b] Based on DNA marker locations flanking 1-LOD interval in the published genome sequence (PATERSON *et al.* 2009).

[C] QTLs based on single year values that do no overlap with those from overall BLUP values.

Table 4.8. Overlapping QTL regions between water content and stem diameter traits

Chr.	SWC	LWC	BD	RD	Start (Mb)	End (Mb)
1	qSWC12_1.1*	qLWC12_1.1*	qBD1.1		19.8	26.2
6	qSWC6.1			qRD6.2	51.7	52.7

BD basal stem diameter, RD rachis diameter, SWC stem water content, LWC leaf water content
*QTLs based on single year values

Table 4.9. Water content QTLs with corresponding significant SNPs

Trait	QTL	Chr.	Significant SNPs	Distance*
SWC	qSWC12_1.1	1	S1_25989043	Within†
SWC	qSWC6.1	6	S6_49958866, S6_49958867, S6_51042333	60kb

SWC stem water content, LWC leaf water content
*Distance between QTL to the nearest significant SNP outside the QTL interval.
†Significant SNPs are contained in the QTL interval.

Table 4.10. QTLs affecting stem volume trait in BTx623 x IS3620c RILs

Source	QTL name	Chr.	Peak (cM)	LOD	Additive[a]	R2 (%)	Start (Mb)[b]	End (Mb)[b]
2011	qSV11_6.1	6	24.60	3.11	-810.80	3.21	1.5	59.4
	qSV11_7.1	7	88.00	7.02	1275.90	7.42	57.7	59.5
	qSV11_8.1	8	101.60	3.11	-818.00	3.22	51.8	54.4
2012	qSV12_1.1	1	96.50	2.79	776.30	2.71	1.9	54.5
	qSV12_1.2	1	126.00	6.72	-1178.50	6.68	57.2	59.1
	qSV12_3.1	3	118.00	4.15	-891.10	4.06	60.9	64.2
	qSV12_4.1	4	26.00	4.11	872.70	4.02	4.3	6.9
	qSV12_6.1	6	82.00	3.56	-813.70	3.48	4.9	58.1

[a] Additive effects calculated as IS3620c – BTx623.
[b] Based on DNA marker locations flanking 1-LOD interval in the published genome sequence (PATERSON *et al.* 2009).

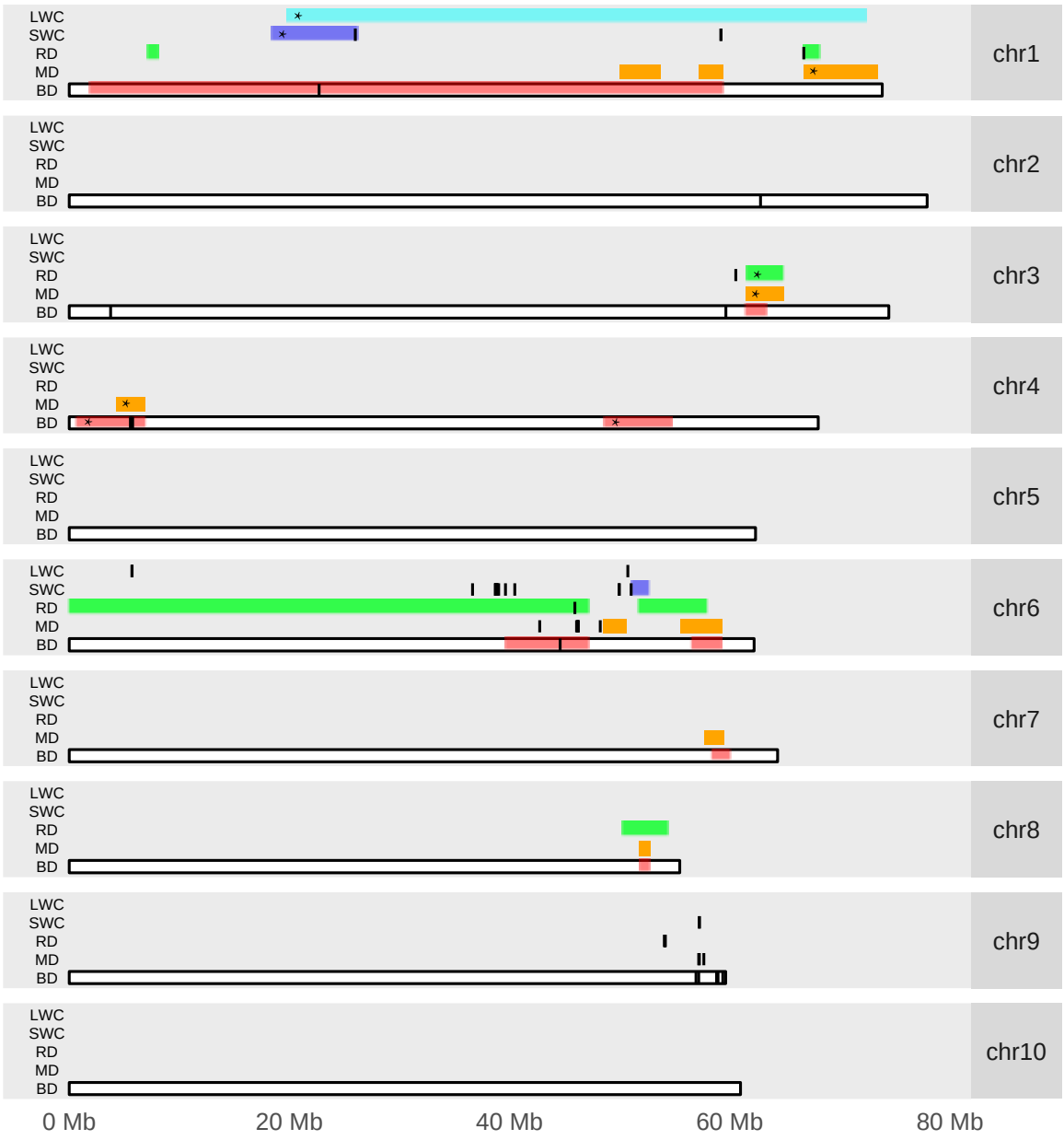
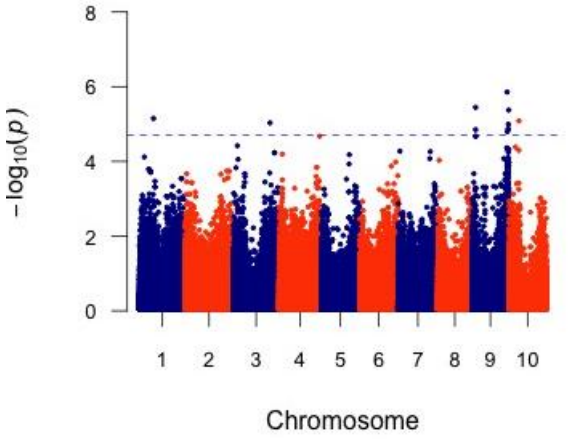


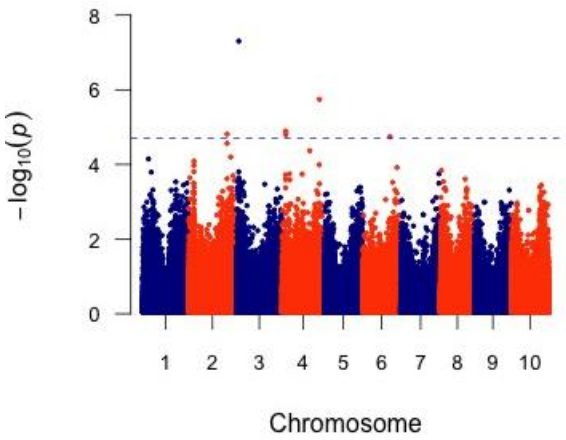
Figure 4.1. QTLs and significant SNPs for stem diameter and water content traits along the published *S. bicolor* reference genome (PATERSON *et al.* 2009). 1-LOD intervals of QTLs in BTx623 x IS3620c RILs for BD (red), MD (orange), RD (green), SWC (purple) and LWC (blue) are shown. QTL intervals derived using single year values are marked with asterisks. Significant SNPs in the sorghum diversity panel at threshold of 1.96×10^{-5}

are shown as solid lines. BD basal stem diameter, MD middle stem diameter, RD rachis diameter, SWC stem water content, LWC leaf water content

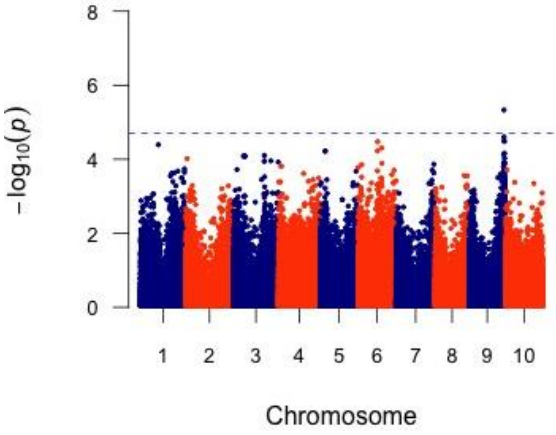
A. Basal diameter (BD) - 2009



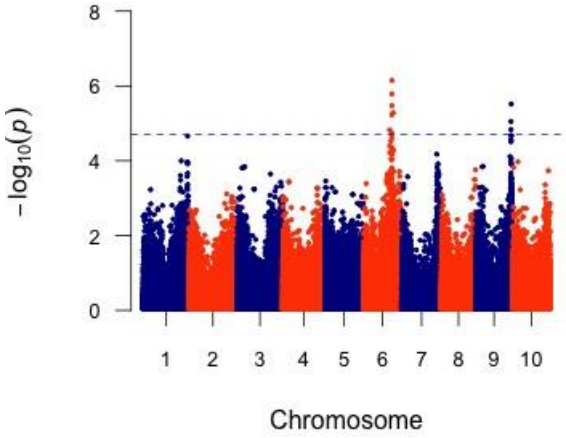
B. Basal diameter (BD) - 2010



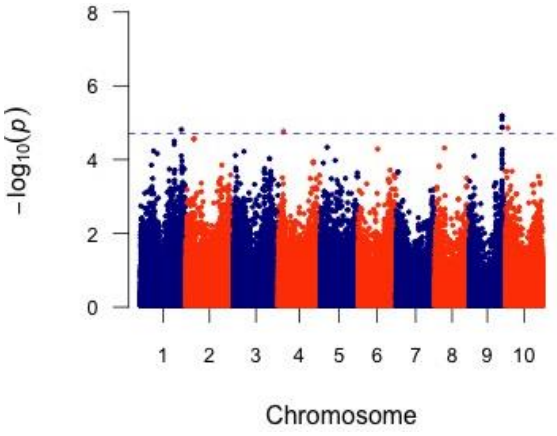
C. Middle diameter (MD) - 2009



D. Middle diameter (MD) - 2010



E. Rachis diameter (RD) - 2009



F. Rachis diameter (RD) - 2010

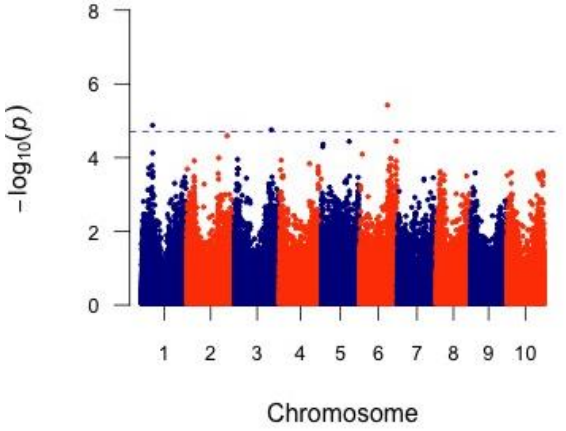


Figure 4.2. Manhattan plots for stem diameter traits, including basal (A, B), middle (C, D) and rachis (E, F) diameter, in the sorghum diversity panel. A significance threshold of 0.05 [$-\log_{10}$(p-value)$\sim$4.7] is noted with dashed horizontal lines.

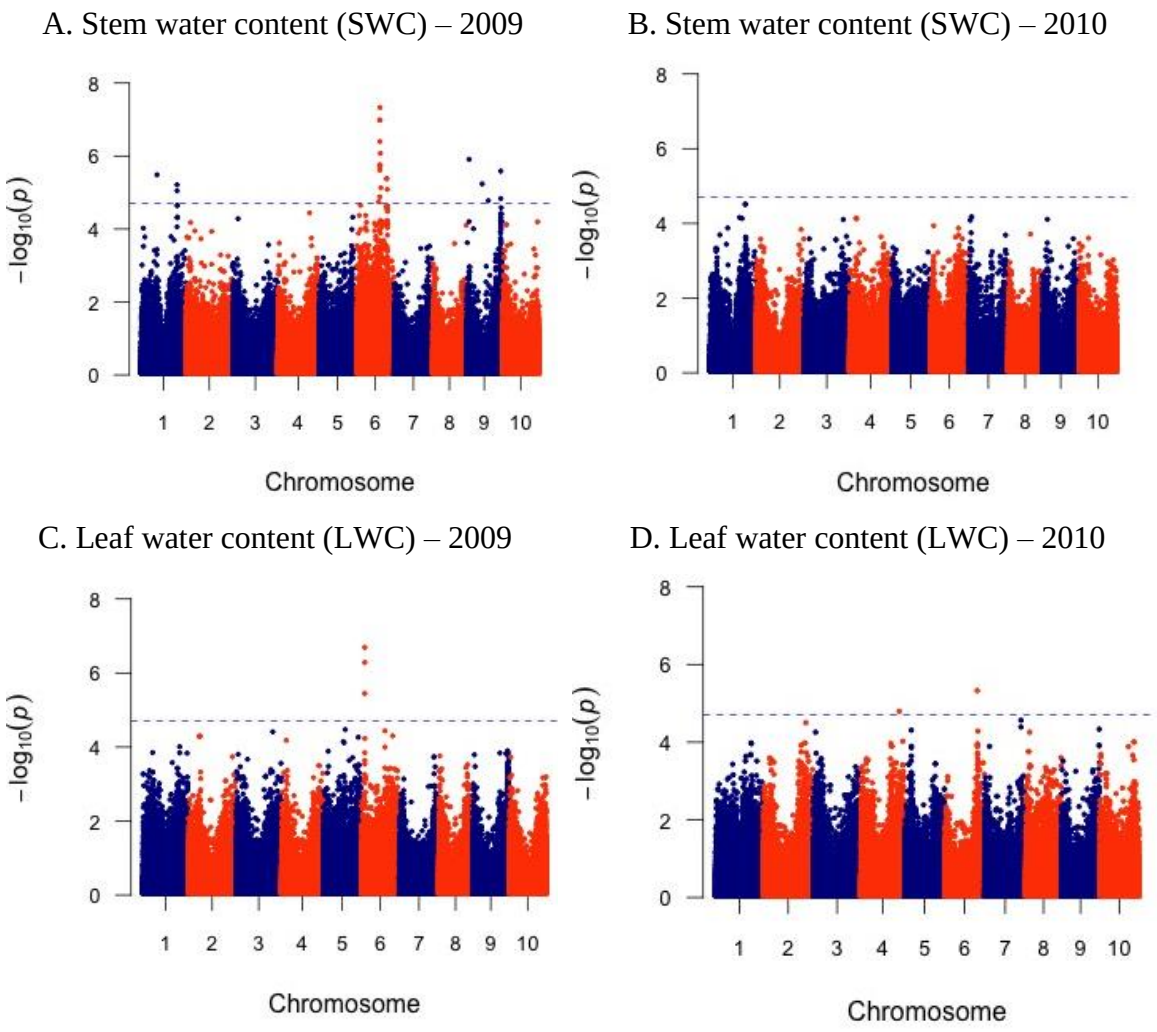


Figure 4.3. Manhattan plots for water content traits, including stem water content (A, B) and leaf water content (C, D), in the sorghum diversity panel. A significance threshold of 0.05 [$-\log_{10}$(p-value)~4.7] is noted with dashed horizontal lines.

Supplemental Table S4.1. QTLs affecting stem diameter traits using single year values in BTx623 x IS3620c RILs. Environment (year)-specific QTLs that elude detection by using overall BLUP values are marked in bold

Trait	Year	QTL name	Chr.	Peak (cM)	LOD	Additive[a]	R2 (%)	Start (Mb)[b]	End (Mb)[b]
BD	2011	qBD11_1.1	1	14.0	2.56	-0.5431	2.547	1.0	43.2
	2011	qBD11_6.1	6	33.0	3.50	0.5879	3.503	42.4	47.2
	2011	qBD11_6.2	6	98.0	5.18	-0.7210	5.234	58.8	59.8
	2011	qBD11_8.1	8	83.3	6.05	-0.7877	6.151	51.8	52.8
	2012	qBD12_1.1	1	128.0	4.09	-0.6200	3.547	7.4	59.4
	2012	qBD12_3.1	3	115.0	3.57	-0.5911	3.093	61.4	64.9
	2012	**qBD12_4.1**	4	16.4	3.30	0.5592	2.851	0.7	6.9
	2012	**qBD12_4.2**	4	94.5	4.05	-0.6208	3.502	48.6	54.8
	2012	qBD12_6.1	6	27.0	5.29	0.7180	4.617	39.6	46.2
	2012	qBD12_6.2	6	93.0	9.18	-0.9433	8.180	58.1	59.3
	2012	qBD12_7.1	7	98.0	6.60	-0.7959	5.800	58.4	60.1
MD	2011	**qMD11_1.1**	1	181.8	2.53	-0.3869	2.378	66.7	73.4
	2011	qMD11_6.1	6	98.1	2.88	-0.4000	2.675	48.9	59.8
	2011	qMD11_7.1	7	89.0	11.52	0.8300	11.349	57.7	59.5
	2011	qMD11_8.1	8	101.6	4.60	-0.5109	4.391	51.8	55.5
	2012	qMD12_1.1	1	25.0	2.97	-0.4372	2.930	5.8	53.7
	2012	qMD12_1.2	1	127.0	4.41	-0.4191	4.384	57.2	68.2
	2012	**qMD12_3.1**	3	117.0	3.69	-0.3878	3.648	61.4	64.9
	2012	**qMD12_4.1**	4	25.8	3.17	0.3493	3.131	4.3	6.9
	2012	qMD12_6.1	6	82.0	4.35	-0.4211	4.323	48.9	57.9
RD	2011	qRD11_1.1	1	0.0	2.60	-0.2557	2.730	0.0	13.6
	2011	qRD11_1.2	1	161.0	3.17	-0.2757	3.338	66.7	71.8
	2011	qRD11_8.1	8	84.0	4.44	-0.3329	4.720	51.8	54.4
	2012	qRD12_1.1	1	160.5	5.42	-0.3505	5.603	66.7	68.2
	2012	**qRD12_3.1**	3	116.6	3.86	-0.3054	3.954	61.5	64.9
	2012	qRD12_6.1	6	1.0	3.11	0.3313	3.168	0.0	47.2
	2012	qRD12_6.2	6	60.6	3.06	-0.2653	3.115	51.1	58.1

BD basal stem diameter, MD middle stem diameter, RD rachis diameter

[a] Additive effects calculated as IS3620c – BTx623.

[b] Based on DNA marker locations flanking 1-LOD interval in the published genome sequence (PATERSON *et al.* 2009).

Supplemental Table S4.2. QTLs in Shiringani et al.'s study (SHIRINGANI *et al.* 2010)

QTL names	Chr.	Linked Marker	Support Interval (cM)	Left (Mb)*	Right (Mb)*
Q.Shiringani2010.STEMDIA.1-1	1	E43M53-500	124-132	0.0	23.8
Q.Shiringani2010.STEMDIA.2-1	2	Xtxp56	78-90	61.6	63.4
Q.Shiringani2010.STEMDIA.3-1	3	Xtxp285	2-16	61.8	67.8
Q.Shiringani2010.STEMDIA.3-2	3	Xtxp31	74-90	0.0	55.4
Q.Shiringani2010.STEMDIA.3-3	3	E32M55-215	98-108	0.0	55.4
Q.Shiringani2010.STEMDIA.3-4	3	Xtxp9	134-150	0.0	55.4
Q.Shiringani2010.STEMDIA.4-1	4	E35M49-230	52-54	48.8	51.2
Q.Shiringani2010.STEMDIA.6-1	6	Xtxp97	64-74	0.0	51.2
Q.Shiringani2010.STEMDIA.7-1	7	E45M57-120	12-20	ND	ND
Q.Shiringani2010.STEMDIA.7-2	7	E31M61-420	28-34	ND	ND
Q.Shiringani2010.STEMDIA.7-3	7	E31M59-202	76-88	ND	ND
Q.Shiringani2010.STEMDIA.8-1	8	E32M47-455	104-126	0.0	48.2
Q.Shiringani2010.STEMDIA.9-1	9	E43M41-210	72-86	ND	ND

*Based on DNA marker locations flanking support intervals in the published genome sequence (PATERSON *et al.* 2009).

ND: physical locations of QTL boundaries not determined because lack of marker sequence information.

Supplemental Table S4.3. SNPs with significant association to stem diameter traits (threshold: 1.96E-5)

Trait	Environment	Chr.	Position	P-value	Minor allele frequency
BD	2009	1	22691388	7.07E-06	9.7%
BD	2010	2	62790559	1.54E-05	5.1%
BD	2010	3	3759453	4.97E-08	28.8%
BD	2009	3	59642849	9.31E-06	6.6%
BD	2010	4	5581387	1.30E-05	36.5%
BD	2010	4	5581390	1.30E-05	36.5%
BD	2010	4	5774418	1.61E-05	7.8%
BD	2010	6	44581098	1.80E-05	8.0%
BD	2009	9	56928114	1.39E-06	22.2%
BD	2009	9	57185867	1.60E-05	23.3%
BD	2009	9	58810386	1.41E-05	16.5%
BD	2009	9	58810387	1.34E-05	18.6%
BD	2009	9	58939522	1.01E-05	39.8%
BD	2009	9	59369385	4.23E-06	45.6%
MD	2010	6	42736415	1.51E-05	32.2%
MD	2010	6	46077506	1.83E-05	50.0%
MD	2010	6	46194160	1.64E-06	32.0%
MD	2010	6	46217845	3.35E-06	29.7%
MD	2010	6	46217893	5.93E-06	28.8%
MD	2010	6	46217966	7.12E-07	28.1%
MD	2010	6	48234025	5.22E-06	45.8%
MD	2009, 2010	9	57185867	4.67E-06	23.4%
MD	2010	9	57240634	1.47E-05	37.2%
MD	2010	9	57641496	3.04E-06	23.6%
RD	2009	1	66715300	1.54E-05	10.7%
RD	2010	3	60554780	1.76E-05	11.7%
RD	2010	6	45935408	3.76E-06	8.1%
RD	2009	9	54084773	1.34E-05	41.3%
RD	2009	9	54084785	1.34E-05	41.3%
RD	2009	9	54084970	8.12E-06	43.2%
RD	2009	9	54137788	6.55E-06	39.1%
RD	2009	9	54137799	6.55E-06	39.1%
RD	2009	9	54137819	6.55E-06	39.1%

BD basal stem diameter, MD middle stem diameter, RD rachis diameter

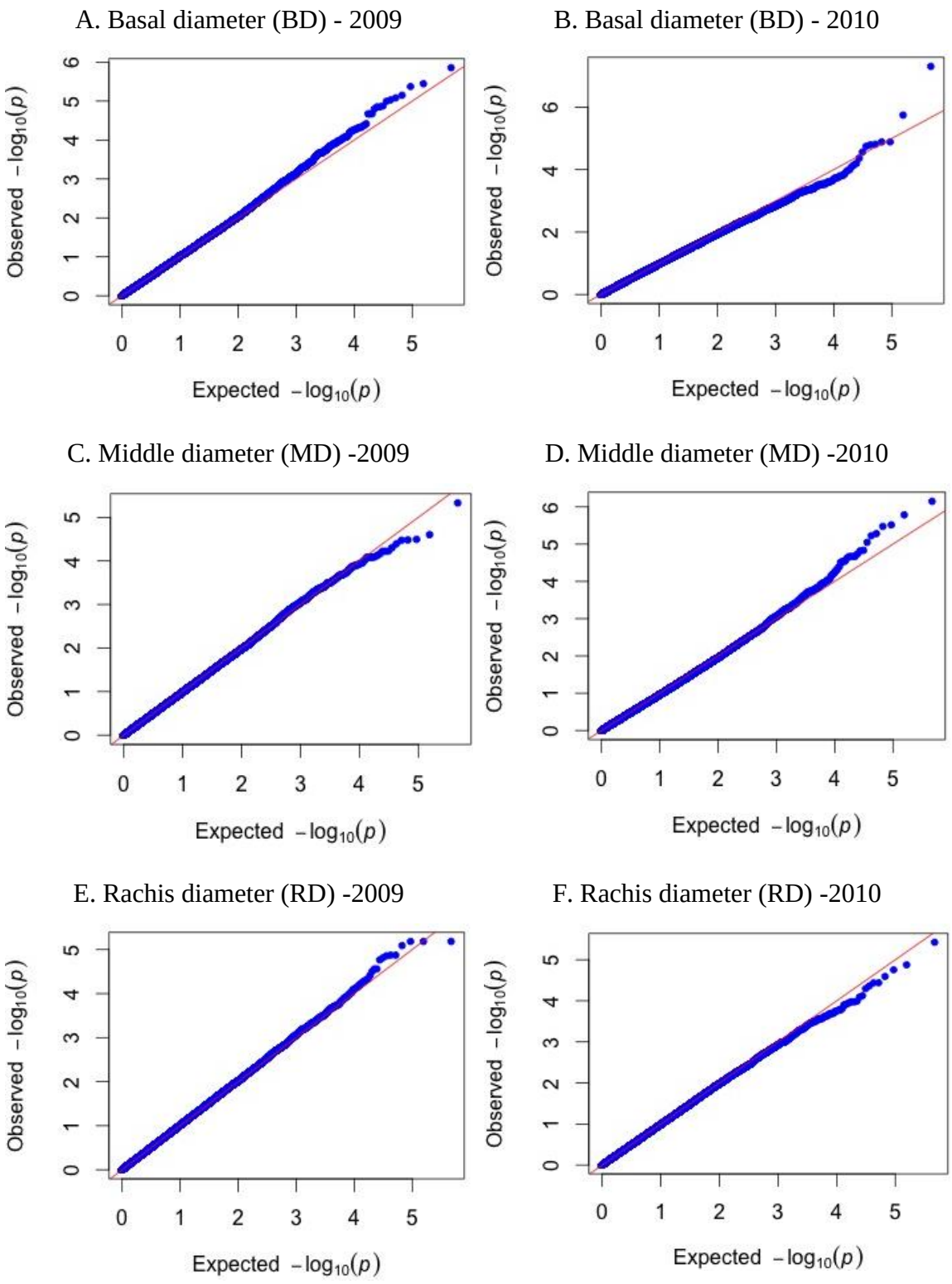
Supplemental Table S4.4. Synteny block contained in qSWC12_1.1 and qSWC6.1 (Score=481.0, E-value=2e-27), which locates at 20.0Mb - 22.3Mb on chromosome 1 and 51.0Mb - 52.1Mb on chromosome 6

Locus 1	Locus 2
Sb01g019000	Sb06g022930
Sb01g019100	Sb06g022820
Sb01g019130	Sb06g022810
Sb01g019150	Sb06g022790
Sb01g019170	Sb06g022780
Sb01g019270	Sb06g022660
Sb01g019290	Sb06g022610
Sb01g019390	Sb06g022560
Sb01g019590	Sb06g022320
Sb01g019830	Sb06g022220
Sb01g020000	Sb06g021990
Sb01g020190	Sb06g021850

Supplemental Table S4.5. SNPs with significant association to water content traits (threshold: 1.96E-5)

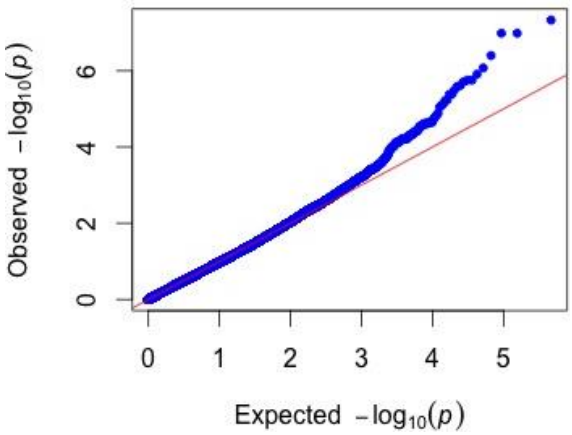
Trait	Environment	Chr.	Position	P-value	Minor allele frequency
SWC	2009	1	25989043	3.26E-06	17.9%
SWC	2009	1	59191971	6.12E-06	7.2%
SWC	2009	1	59200440	8.97E-06	6.1%
SWC	2009	6	36628658	1.80E-05	29.7%
SWC	2009	6	38708238	1.32E-05	10.7%
SWC	2009	6	38732363	1.04E-07	11.8%
SWC	2009	6	38732407	1.04E-07	11.8%
SWC	2009	6	38738938	1.76E-06	11.6%
SWC	2009	6	38823536	4.00E-07	11.5%
SWC	2009	6	38823793	2.44E-06	11.9%
SWC	2009	6	38836159	1.94E-06	13.5%
SWC	2009	6	38869751	1.76E-06	12.6%
SWC	2009	6	39007151	4.68E-08	12.9%
SWC	2009	6	39631053	8.43E-07	8.7%
SWC	2009	6	40464974	7.41E-06	49.1%
SWC	2009	6	49958866	4.16E-06	14.1%
SWC	2009	6	49958867	4.16E-06	14.1%
SWC	2009	6	51042333	8.19E-06	9.5%
SWC	2009	9	57236791	1.47E-05	36.8%
SWC	2009	9	57240454	2.60E-06	33.2%
LWC	2009	6	5707120	3.57E-06	16.0%
LWC	2009	6	5707143	5.19E-07	20.7%
LWC	2009	6	5713788	2.04E-07	18.7%
LWC	2010	6	50745257	4.70E-06	7.2%

SWC stem water content, LWC leaf water content

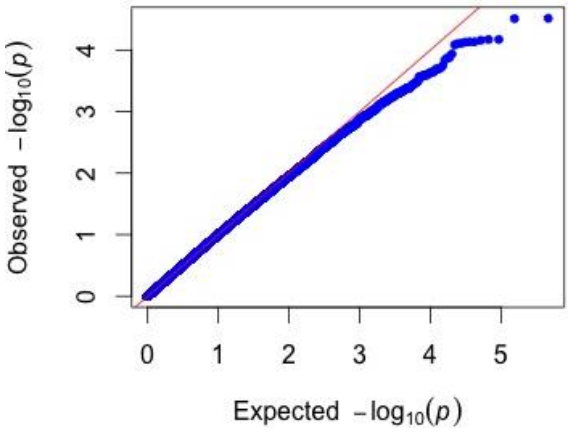


Supplemental Figure S4.1. QQ-plots for stem diameter traits of the sorghum diversity panel in 2009 and 2010.

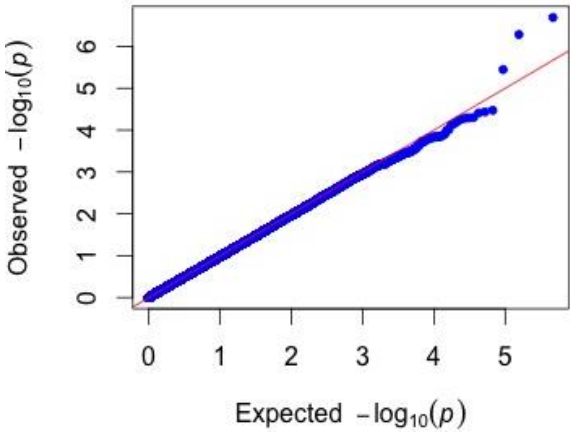
A. Stem water content (SWC) - 2009



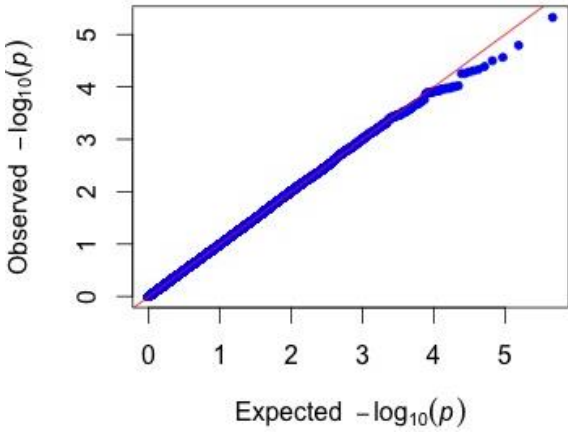
B. Stem water content (SWC) - 2010



C. Leaf water content (LWC) - 2009



D. Leaf water content (LWC) - 2010



Supplemental Figure S4.2. QQ-plots for water content traits of the sorghum diversity panel in 2009 and 2010.

4.6 References

ARENDS, D., P. PRINS, R. C. JANSEN and K. W. BROMAN, 2010 R/qtl: high-throughput multiple QTL mapping. Bioinformatics **26:** 2990-2992.

ATIENZA, S. G., Z. SATOVIC, K. K. PETERSEN, O. DOLSTRA and A. MARTIN, 2003 Identification of QTLs influencing agronomic traits in Miscanthus sinensis Anderss. I. Total height, flag-leaf height and stem diameter. Theoretical and Applied Genetics **107:** 123-129.

BATES, D., M. MACHLER, B. M. BOLKER and S. C. WALKER, 2015 Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software **67:** 1-48.

BOWERS, J. E., C. ABBEY, S. ANDERSON, C. CHANG, X. DRAYE *et al.*, 2003 A high-density genetic recombination map of sequence-tagged sites for sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. Genetics **165:** 367-386.

BRADY, J. A., 2006 Sorghum Ma5 and Ma6 Maturity Genes. Ph.D. dissertation. Texas A&M University.

BROMAN, K. W., H. WU, S. SEN and G. A. CHURCHILL, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics **19:** 889-890.

BROWN, P. J., P. E. KLEIN, E. BORTIRI, C. B. ACHARYA, W. L. ROONEY *et al.*, 2006 Inheritance of inflorescence architecture in sorghum. Theor Appl Genet **113:** 931-942.

BURTON, G. W., 1986 Biomass production from herbaceous plant, pp. xiv, 667 p. in *Biomass energy development*, edited by W. H. SMITH. Plenum Press, New York.

CASA, A. M., G. PRESSOIR, P. J. BROWN, S. E. MITCHELL, W. L. ROONEY *et al.*, 2008 Community resources and strategies for association mapping in sorghum. Crop Science **48:** 30-40.

CHAUMONT, D., 1993 Biotechnology of Algal Biomass Production - a Review of Systems for Outdoor Mass-Culture. Journal of Applied Phycology **5:** 593-604.

EIA, 2016 International Energy Outlook 2016, pp.

GORDON, C. H., 1967 Storage Losses in Silage as Affected by Moisture Content and Structure. Journal of Dairy Science **50:** 397-&.

GUAN, Y. A., H. L. WANG, L. QIN, H. W. ZHANG, Y. B. YANG *et al.*, 2011 QTL mapping of bio-energy related traits in Sorghum. Euphytica **182:** 431-440.

HAN, Y. C., P. LV, S. L. HOU, S. Y. LI, G. S. JI *et al.*, 2015 Combining Next Generation Sequencing with Bulked Segregant Analysis to Fine Map a Stem Moisture Locus in Sorghum (Sorghum bicolor L. Moench). Plos One **10**.

KASHIWAGI, T., and K. ISHIMARU, 2004 Identification and functional analysis of a locus for improvement of lodging resistance in rice. Plant Physiology **134:** 676-683.

KASHIWAGI, T., E. TOGAWA, N. HIROTSU and K. ISHIMARU, 2008 Improvement of lodging resistance with QTLs for stem diameter in rice (Oryza sativa L.). Theor Appl Genet **117:** 749-757.

KONG, W., C. KIM, H. GUO and A. H. PATERSON, 2016 Genetic mapping in recombinant inbred lines of Sorghum bicolor BTx623 × IS3620C, pp., Unpublished manuscript.

KONG, W. Q., H. GUO, V. H. GOFF, T. H. LEE, C. KIM *et al.*, 2014 Genetic analysis of vegetative branching in sorghum. Theoretical and Applied Genetics **127:** 2387-2403.

LI, R. H., M. A. LYONS, H. WITTENBURG, B. PAIGEN and G. A. CHURCHILL, 2005 Combining data from multiple inbred line crosses improves the power and resolution of quantitative trait loci mapping. Genetics **169:** 1699-1709.

LIPKA, A. E., F. TIAN, Q. S. WANG, J. PEIFFER, M. LI *et al.*, 2012 GAPIT: genome association and prediction integrated tool. Bioinformatics **28:** 2397-2399.

LONGIN, C. F. H., and T. WURSCHUM, 2014 Genetic variability, heritability and correlation among agronomic and disease resistance traits in a diversity panel and elite breeding material of spelt wheat. Plant Breeding **133:** 459-464.

LV, P., G. S. JI, Y. C. HAN, S. L. HOU, S. Y. LI *et al.*, 2013 Association analysis of sugar yield-related traits in sorghum [Sorghum bicolor (L.)]. Euphytica **193:** 419-431.

MANOLIO, T. A., F. S. COLLINS, N. J. COX, D. B. GOLDSTEIN, L. A. HINDORFF *et al.*, 2009 Finding the missing heritability of complex diseases. Nature **461:** 747-753.

MATTHIES, I. E., M. MALOSETTI, M. S. RODER and F. VAN EEUWIJK, 2014 Genome-wide association mapping for kernel and malting quality traits using historical European barley records. PLoS One **9:** e110046.

MCCOUCH, S. R., Y. G. CHO, M. YANO, E. PAUL, M. BLINSTRUB *et al.*, 1997 Report on QTL nomenclature. Rice Genet Newslett **14:** 11-13.

MILLIGAN, S. B., K. A. GRAVOIS, K. P. BISCHOFF and F. A. MARTIN, 1990 Crop Effects on Genetic-Relationships among Sugarcane Traits. Crop Science **30:** 927-931.

MORRIS, G. P., P. RAMU, S. P. DESHPANDE, C. T. HASH, T. SHAH *et al.*, 2013 Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci U S A **110:** 453-458.

MURRAY, S. C., A. SHARMA, W. L. ROONEY, P. E. KLEIN, J. E. MULLET *et al.*, 2008 Genetic Improvement of Sorghum as a Biofuel Feedstock: I. QTL for Stem Sugar and Grain Nonstructural Carbohydrates. Crop Science **48:** 2165-2179.

PATERSON, A. H., 2013 *Genomics of the Saccharinae*. Springer, New York.

PATERSON, A. H., J. E. BOWERS, R. BRUGGMANN, I. DUBCHAK, J. GRIMWOOD *et al.*, 2009 The Sorghum bicolor genome and the diversification of grasses. Nature **457:** 551-556.

PATZEK, T. W., and D. PIMENTEL, 2005 Thermodynamics of energy production from biomass. Critical Reviews in Plant Sciences **24:** 327-364.

RCORETEAM, 2015 R: A language and environment for statistical computing., pp. in *R Foundation for Statistical Computing, Vienna, Austria. URL https://*http://www.R-project.org/.

ROONEY, W. L., J. BLUMENTHAL, B. BEAN and J. E. MULLET, 2007 Designing sorghum as a dedicated bioenergy feedstock. Biofuels Bioproducts & Biorefining-Biofpr **1:** 147-157.

ROSS-IBARRA, J., P. L. MORRELL and B. S. GAUT, 2007 Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proceedings of the National Academy of Sciences of the United States of America **104:** 8641-8648.

SABALLOS, A., 2008 Development and utilization of sorghum as a bioenergy crop., pp. xxi, 449 p. in *Genetic improvement of bioenergy crops*, edited by W. VERMERRIS. Springer, New York.

SHIRINGANI, A. L., M. FRISCH and W. FRIEDT, 2010 Genetic mapping of QTLs for sugar-related traits in a RIL population of Sorghum bicolor L. Moench. Theoretical and Applied Genetics **121:** 323-336.

TANG, H. B., H. E. CUEVAS, S. DAS, U. U. SEZEN, C. B. ZHOU *et al.*, 2013 Seed shattering in a wild sorghum is conferred by a locus unrelated to domestication. Proceedings of the National Academy of Sciences of the United States of America **110:** 15824-15829.

TOMES, D. T., P. LAKSHMANAN and D. SONGSTAD, 2011 *Biofuels : global impact on renewable energy, production agriculture and technological advancements*. Springer, New York.

VERMERRIS, W., 2011 Survey of Genomics Approaches to Improve Bioenergy Traits in Maize, Sorghum and SugarcaneFree Access. Journal of Integrative Plant Biology **53:** 105-119.

YOUNG, H. E., 1980 Forest Biomass as a Renewable Source of Energy - Inventory, Productivity and Availability. Pulp & Paper-Canada **81:** 21-&.

ZHANG, D., H. GUO, C. KIM, T. H. LEE, J. P. LI *et al.*, 2013 CSGRqtl, a Comparative Quantitative Trait Locus Database for Saccharinae Grasses. Plant Physiology **161:** 594-599.

ZHANG, D., W. KONG, J. ROBERTSON, V. H. GOFF, E. EPPS *et al.*, 2015 Genetic analysis of inflorescence and plant height components in sorghum (Panicoidae) and comparative genetics with rice (Oryzoidae). BMC Plant Biol **15:** 107.

ZHANG, Z., E. ERSOZ, C. Q. LAI, R. J. TODHUNTER, H. K. TIWARI *et al.*, 2010 Mixed linear model approach adapted for genome-wide association studies. Nat Genet **42:** 355-360.

ZHU, C., and J. YU, 2009 Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. Genetics **182:** 875-888.

CHAPTER 5

CONCLUSIONS

The primary intent of the work presented here is to improve understanding of several biomass-related traits in sorghum, including the 'dry' trait, stem architecture and water content components. It also illustrates how GWAS may be used to complement conventional linkage mapping in exploring the genetic basis of simple and complex traits.

A genetic map, representing the distribution of a set of linked loci within a genome, is fundamental for linkage mapping. In Chapter 2, we constructed a genetic map derived from an annual *S. bicolor* x perennial *S. propinquum* recombinant inbred line (RIL) population of 161 F5 genotypes, collectively spanning 773.1 cM. It contains 135 markers segregating for 141 loci on 10 linkage groups corresponding to the 10 sorghum chromosomes. The average interval between consecutive loci is 5.48 cM, ranging from 0.0 cM between co-segregating markers to 25.7 cM in the largest gap (on chromosome 5).

S. bicolor x S. propinquum is thought to be the widest euploid cross that can be made with the cultigen (*S. bicolor*) by conventional means, and interspecific populations from these species offer opportunities to genetically dissect a wide range of traits related to plant domestication and crop productivity (CHITTENDEN *et al.* 1994; LIN *et al.* 1995; PATERSON *et al.* 1995a; PATERSON *et al.* 1995b; LIN *et al.* 1999; HU *et al.* 2003; FELTUS *et al.* 2006). The RIL population of our study was advanced in an environment to which

the *S. bicolor* parent was well adapted (indeed bred for), but the *S. propinquum* parent was not. The removal of short-day flowering on chromosome 6 resulting from advancement in this environment improved the ability to resolve QTL for traits that were previously below the significance threshold. For example, eliminating the profound morpho-physiological alteration associated with short-day flowering permitted us to identify two flowering QTLs that eluded detection in a previous study of 370 F2 plants from this cross (LIN *et al.* 1995). The present map also provides a more realistic assessment of variation that is relevant to temperate latitudes. Together with homozygosity of the RILs and the ability to replicate them across a spectrum of natural and/or controlled conditions, this genetic map is of high potential importance for the discovery and validation of QTLs.

With the availability of genomic resources including genetic maps (BOWERS *et al.* 2003; Kong *et al.* unpublished data), a high quality genome sequence (PATERSON *et al.* 2009) and GBS SNP data (MORRIS *et al.* 2013) for a sorghum diversity panel (CASA *et al.* 2008), the genetic control of biomass traits of interest can be elucidated.

In Chapter 3, we characterized the 'dry' trait in sorghum and identified a probable candidate for the D gene (Sobic.006G145300) on chromosome 6. The candidate D gene, isoform *Dhr2* of beta-glucosidase (BGLU), is highly expressed in shoots, leaves and panicles, possibly being involved in processes such as lignification and cell wall catabolism. A relationship between the 'dry' trait and plant moisture content is confirmed. Race-specific patterns and geographic variation of the 'dry' trait are evident. The 'dry' trait, to date known only in sorghum, appears to have evolved under balancing

selection, during a recent tandem expansion of the BGLU gene family in sorghum after its divergence from maize.

In Chapter 4, QTLs for basal stem diameter (on chromosomes 1, 3, 4, 6, 7, 8), middle stem diameter (on chromosomes 1, 3, 4, 6, 7, 8), rachis diameter (on chromosomes 1, 3, 6, 8), stem water content (on chromosomes 1, 6), and leaf water content (on chromosome 1) were reported and verified using GWAS. GWAS, a complement to linkage mapping, also suggested several additional putative loci for each of the traits (BD: on Chr. 2, 3, 9; MD: on Chr. 9; RD: on Chr. 9, SWC: on Chr. 1, 6, 9; LWC: on Chr. 6). The observations of both co-localized and non-overlapping loci affecting stem diameter traits suggest that stem widths at different heights (base, middle, rachis) may share some common genetic control, but also have some distinct genetic influences. Co-localizations of stem diameter and water content traits with a number of other biomass traits, including plant height, flowering time, branching, stem volume and the 'dry' trait, suggest that their inheritance may be functionally (pleiotropy) or physically (linkage disequilibrium) linked.

In conclusion, these analyses have provided a comprehensive description of genes and genomic regions responsible for a list of biomass-related traits as well as their interactions. For future studies, the genetic map of the *S. bicolor* x *S. propinquum* RIL population can offer advantages over early-generation populations, shedding new light on genetic, environmental, and physiological/biochemical factors that regulate plant growth and development. Identification and characterization of the D gene can aid in trait improvement. Exploration of genomic regions responsible for stem diameter and water content traits can serve as a foundation for future positional cloning of causal genes. In

all, the genetic map and genomic regions for a list of biomass-related traits reported in these studies can contribute to genetic improvement of sorghum as a feedstock for production of biofuel, further enhancing its merit as a means to capture and store solar energy.

5.1 References

BOWERS, J. E., C. ABBEY, S. ANDERSON, C. CHANG, X. DRAYE *et al.*, 2003 A high-density genetic recombination map of sequence-tagged sites for sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. Genetics **165:** 367-386.

CASA, A. M., G. PRESSOIR, P. J. BROWN, S. E. MITCHELL, W. L. ROONEY *et al.*, 2008 Community resources and strategies for association mapping in sorghum. Crop Science **48:** 30-40.

CHITTENDEN, L. M., K. F. SCHERTZ, Y. R. LIN, R. A. WING and A. H. PATERSON, 1994 A Detailed Rflp Map of Sorghum-Bicolor X S-Propinquum, Suitable for High-Density Mapping, Suggests Ancestral Duplication of Sorghum Chromosomes or Chromosomal Segments. Theoretical and Applied Genetics **87:** 925-933.

FELTUS, F. A., G. E. HART, K. F. SCHERTZ, A. M. CASA, S. KRESOVICH *et al.*, 2006 Alignment of genetic maps and QTLs between inter- and intra-specific sorghum populations. Theor Appl Genet **112:** 1295-1305.

HU, F. Y., D. Y. TAO, E. SACKS, B. Y. FU, P. XU *et al.*, 2003 Convergent evolution of perenniality in rice and sorghum. Proc Natl Acad Sci U S A **100:** 4050-4054.

KONG, W., C. KIM, H. GUO and A. H. PATERSON, 2016 Genetic mapping in recombinant inbred lines of Sorghum bicolor BTx623 × IS3620C, pp., Unpublished manuscript.

LIN, Y. R., K. F. SCHERTZ and A. H. PATERSON, 1995 Comparative analysis of QTLs affecting plant height and maturity across the Poaceae, in reference to an interspecific sorghum population. Genetics **141:** 391-411.

LIN, Y. R., L. H. ZHU, S. X. REN, J. S. YANG, K. F. SCHERTZ *et al.*, 1999 A Sorghum propinquum BAC library, suitable for cloning genes associated with loss-of-function mutations during crop domestication. Molecular Breeding **5:** 511-520.

MORRIS, G. P., P. RAMU, S. P. DESHPANDE, C. T. HASH, T. SHAH *et al.*, 2013 Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci U S A **110:** 453-458.

PATERSON, A. H., J. E. BOWERS, R. BRUGGMANN, I. DUBCHAK, J. GRIMWOOD *et al.*, 2009 The Sorghum bicolor genome and the diversification of grasses. Nature **457:** 551-556.

PATERSON, A. H., Y. R. LIN, Z. K. LI, K. F. SCHERTZ, J. F. DOEBLEY *et al.*, 1995a Convergent Domestication of Cereal Crops by Independent Mutations at Corresponding Genetic-Loci. Science **269:** 1714-1718.

PATERSON, A. H., K. F. SCHERTZ, Y. R. LIN, S. C. LIU and Y. L. CHANG, 1995b The Weediness of Wild Plants - Molecular Analysis of Genes Influencing Dispersal and Persistence of Johnsongrass, Sorghum-Halepense (L) Pers. Proceedings of the National Academy of Sciences of the United States of America **92:** 6127-6131.