GENOMIC MAPPING FOR GRAIN YIELD, STAY GREEN, AND GRAIN QUALITY TRAITS IN SORGHUM

by

SIVAKUMAR SUKUMARAN

B.Sc., Kerala Agricultural University, 2002
M.Sc., Tamil Nadu Agricultural University, 2006

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Agronomy
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2012

Abstract

Knowledge of the genetic bases of grain quality traits will complement plant breeding efforts to improve the end use value of sorghum (*Sorghum bicolor* (L.) Moench). The objective of the first experiment was to assess marker-trait associations for 10 grain quality traits through candidate gene association mapping on a diverse panel of 300 sorghum accessions. The 10 grain quality traits were measured using the single kernel characterization system (SKCS) and near-infrared reflectance spectroscopy (NIRS). The analysis of the accessions through 1,290 genome-wide single nucleotide polymorphisms (SNPs) separated the panel into five subpopulations that corresponded to three major sorghum races (durra, kafir, and caudatum), one intermediate race (guinea-caudatum), and one working group (zerazera/caudatum). Association analysis between 333 SNPs in candidate genes/loci and grain quality traits resulted in eight significant marker-trait associations. A SNP in starch synthase *IIa* (*SSIIa*) gene was associated with kernel hardness (KH) with a likelihood ratio–based R^2 (R_{LR}^2) value of 0.08. SNPs in starch synthase (*SSIIb*) gene ($R_{LR}^2 = 0.10$) and loci *pSB1120* ($R_{LR}^2 = 0.09$) was associated with starch content.

Sorghum is a crop well adapted to the semi arid regions of the world and my harbor genes for drought tolerance. The objective of second experiment was to identify quantitative trait loci (QTLs) for yield potential and drought tolerance. From a cross between Tx436 (food grain type) and 00MN7645 (drought tolerant) 248 recombinant inbred lines (RILs) was developed. Multi-location trials were conducted in 8 environments to evaluate agronomic performance of the RILs under favorable and drought stress conditions. The 248 RILs and their parents were genotyped by genotyping-by-sequencing (GBS). A subset of 800 SNPs was used for linkage map construction and QTL detection. Composite interval mapping identified a major QTLs for grain yield in chromosome 8 and QTL for flowering time in chromosome 9 under favorable conditions. Three major QTLs were detected for grain yield in chromosomes 1, 6, and 8 and two flowering time QTLs on chromosome 1 under drought conditions. Six QTLs were identified for stay green: two on chromosome 4; one each on chromosome 5, 6, 7, and 10 under drought conditions.

GENOMIC MAPPING FOR GRAIN YIELD, STAY GREEN, AND GRAIN QUALITY TRAITS IN SORGHUM

by

SIVAKUMAR SUKUMARAN

B.Sc., Kerala Agricultural University, 2002
M.Sc., Tamil Nadu Agricultural University, 2006

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Agronomy
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2012

Approved by:

Major Professor
Dr. Jianming Yu

Copyright

Abstract

Knowledge of the genetic bases of grain quality traits will complement plant breeding efforts to improve the end use value of sorghum (*Sorghum bicolor* (L.) Moench). The objective of the first experiment was to assess marker-trait associations for 10 grain quality traits through candidate gene association mapping on a diverse panel of 300 sorghum accessions. The 10 grain quality traits were measured using the single kernel characterization system (SKCS) and near-infrared reflectance spectroscopy (NIRS). The analysis of the accessions through 1,290 genome-wide single nucleotide polymorphisms (SNPs) separated the panel into five subpopulations that corresponded to three major sorghum races (durra, kafir, and caudatum), one intermediate race (guinea-caudatum), and one working group (zerazera/caudatum). Association analysis between 333 SNPs in candidate genes/loci and grain quality traits resulted in eight significant marker-trait associations. A SNP in starch synthase *IIa* (*SSIIa*) gene was associated with kernel hardness (KH) with a likelihood ratio–based R^2 (R_{LR}^2) value of 0.08. SNPs in starch synthase (*SSIIb*) gene ($R_{LR}^2 = 0.10$) and loci *pSB1120* ($R_{LR}^2 = 0.09$) was associated with starch content.

Sorghum is a crop well adapted to the semi arid regions of the world and my harbor genes for drought tolerance. The objective of second experiment was to identify quantitative trait loci (QTLs) for yield potential and drought tolerance. From a cross between Tx436 (food grain type) and 00MN7645 (drought tolerant) 248 recombinant inbred lines (RILs) was developed. Multi-location trials were conducted in 8 environments to evaluate agronomic performance of the RILs under favorable and drought stress conditions. The 248 RILs and their parents were genotyped by genotyping-by-sequencing (GBS). A subset of 800 SNPs was used for linkage map construction and QTL detection. Composite interval mapping identified a major QTLs for grain yield in chromosome 8 and QTL for flowering time in chromosome 9 under favorable conditions. Three major QTLs were detected for grain yield in chromosomes 1, 6, and 8 and two flowering time QTLs on chromosome 1 under drought conditions. Six QTLs were identified for stay green: two on chromosome 4; one each on chromosome 5, 6, 7, and 10 under drought conditions.

Table of contents

List of Figures

List of Tables

Acknowledgement

I am grateful to Dr. Jianming Yu for the guidance, patience, and training provided to me and for the unmatched energy and passion in research. I also want to thank my advisory committee Dr. Vara Prasad, Dr. Guihua Bai, and Dr. Harold Trick for their valuable advice during the period. Thanks to Dr. Tesfaye Tesso and Dr. Rob Aiken for their valuable advices on field research. Special mentioned is needed for Dr. Mitch Tuinstra who contributed immensely to the development of two projects. I would also like to acknowledge the Kansas Grain Sorghum Commission and the Centre for Sorghum Improvement (CSI) for the financial support. I would also like to thank our collaborators Dr. Ming Li Wang, Dr. Martha Hamblin, Dr. Stephen Kresovich, Dr. Scott Bean, and Dr. Jeffery Pedersen.

Thanks to all group members in sorghum genetics Lab: Dr. Zhongwei Lin, Dr. Chengsong Zhu, Dr. Xianran Li, Dr. Yuye Wu, Wenwen Xiang, Xin Li, Xia Li, and Guannan Sun for sharing their knowledge, life experiences, research ideas, particularly for the enthusiasm, help, and friendship during our field trips in Kansas and Puerto Rico. I also thank members in the sorghum breeding group Dr. Kellan Kershner, Leah Miller, and Dr. Newton Ochanda.

I also thank Dr. Subbaratnam. Muthukrishnan, Mrs. Asha Muthukrishnan, Ms. Lakshmi Ramaswamy, and Mr. Parangiri Nivas Rajaram for their contributions during my stay at Kansas State University. I also appreciate my beloved wife Mrs. Kalaiyarasi Pidaran and my little daughter for their patience and love. I also thank the almighty for giving the strength to overcome the hurdles and I dedicate this to my parents for their endless support, encouragement, and love.

CHAPTER 1- REVIEW OF LITERATURE: ASSOCIATION MAPPING

Abstract

Association mapping studies in plants continue to contribute not only in detecting the genetic basis of variation in physiological, developmental, and morphological traits (e.g., flowering time, plant height, grain quality, nutrient content, and seed shattering) but also in bringing together researchers to create community resources and genetic platforms. Association mapping is well positioned to exploit the advances in next generation genomic technologies and high-throughput phenotyping. At the same time, researchers in plant genetics and related disciplines need to develop improved genetic designs and computational tools to address the challenges of genetic mapping such as missing heritability, new gene identification, genotyping-by-sequencing, and rare alleles. In this chapter, we describe major progress in understanding population structure, advancements in design and implementation of association mapping, and examples of association mapping in maize, rice, *Arabidopsis*, wheat, barley, soybean, and sorghum. Finally, major opportunities with potential implications in plant genetics are discussed.

Introduction

The plant genetics and breeding research community's priority is to ensure food security for the ever-growing population with limited land and water resources by increasing yield potential and stability of major crops. Genomic tools, plant breeding methods, genetic designs, and biotechnologies need to be integrated to modify the adaptive, agronomic, and economic characteristics of different plants to meet the challenge. Two connected components of this general effort are gene identification and complex trait dissection. While gene identification focuses on individual genes, complex trait dissection emphasizes the genetic contribution and modes of action from many loci that result in phenotypic variation. Linkage mapping and association mapping are the most commonly used methods for dissecting complex traits and identifying genes underlying trait variations in plants, animals, and human genetics (Risch and Merikangas 1996).

Association mapping provides a great platform to exploit genomic technologies and plant germplasm resources simultaneously (Zhu et al. 2008). Compared with the traditional bi-parental linkage analysis, association mapping offers several advantages. Association mapping populations are typically assembled with diverse lines from breeding programs or sampled accessions from germplasm banks. As a result, researchers can initiate genotyping and phenotyping activities with this approach while developing complementary linkage mapping populations. Because association mapping utilizes the historic recombination present in the panel, a higher mapping resolution is expected. The approach is also fast, and it can accommodate a large number of accessions and analyze a higher number of alleles (Myles et al. 2009; Zhu et al. 2008).

Based on the scale of the research, association mapping can be classified into either targeted studies with candidate genes or large-scale genome-wide association studies (GWAS) (Zhu et al. 2008). A combined approach also can be taken in which a specific genomic region is subjected to a high-resolution scan with a large number of markers. The candidate gene approach focuses on targeted traits and genes with known biochemical pathways. In GWAS with high genome coverage of SNP markers, it is possible to identify genes with previously unknown functions. Another integrated approach, Nested Association Mapping (NAM), has been adopted in plants to combine the advantages of linkage analysis and association mapping (Yu et al. 2008). Although earlier association mapping studies focused mainly on candidate genes (Thornsberry et al. 2001) with limited sample sizes and marker numbers, recent advances in genomic technologies have enabled GWAS in many plant species (**Fig. 1**). We encourage interested readers to refer to earlier reviews (Flint-Garcia et al. 2003; Zhu et al. 2008) for specific details of linkage disequilibrium and association mapping. In this chapter, we will focus on major achievements: understanding of the population structure, research strategies, and examples and progress in major crops. We then outline the significant challenges that have broad implications.

Population structure and false positives

Because an association mapping panel is often an assembled population, rather than a random mated or a designed population, the presence of population structure (i.e., unequal genetic relationships among groups of individuals) often can lead to false positive discoveries if this

structure is not adequately accounted for during the analysis of marker-trait association. In general, population structure can arise due to differences in geographical origins, local adaptations, or breeding history of the lines in the panel (Yu et al. 2006).

As is often true in many crops, multiple levels of relatedness frequently exist in association mapping populations as a result of assembling lines and accessions from different geographical regions or breeding programs (*i.e.*, major grouping patterns) as well as different levels of relationship among individuals within individual breeding programs (*i.e.*, pairwise relationship). In such cases, some markers may appear to be significantly associated with the trait of interest from a simple test, but their frequency distributions are in fact correlated with the population structure. Understanding the structure of the population, developing methods to accurately infer the structure, and conducting association analysis with appropriate models are critical to reduce or eliminate these spurious associations or false positives.

Association mapping samples generally fall into five categories based on the population structure and familial relatedness. The population structure is associated with local adaptation or diversifying selection, and the familial relationship is associated with the recent coancestry. The five categories are (1) ideal samples with subtle population structure and familial relatedness, (2) samples with familial relationship, (3) samples with population structure, (4) samples with both population structure and familial relationships, and (5) samples with severe population structure and familial relationships (Zhu and Yu 2009). It is possible to statistically quantify population structure using neutral markers and account for the structure in identifying marker-trait associations (Yu et al. 2006). Several methods have been used to control for population structure in association mapping (**Fig. 2**). These include genomic control (Devlin and Roeder 1999), structured association (Pritchard et al. 2000b), principal component analysis (PCA) (Patterson et al. 2006; Price et al. 2006), unified mixed model (Yu et al. 2006), non-metric multidimensional scaling (nMDS) (Zhu and Yu 2009), and techniques to increase computational speed based on MLM, compressed MLM, and P3D (Zhang et al. 2010b).

Earlier methods of correcting genetic relatedness were based on general linear models (GLM), but recent methods mainly use the mixed linear model (MLM) to account for population structure and familial relatedness. The unified mixed model (Yu et al. 2006) approach for association mapping that accounts for multiple levels of relatedness considers both population structure and pairwise relatedness. In this model, the population structure (Q matrix) estimated

using STRUCTURE software (Pritchard et al. 2000a) is fitted as a fixed effect, and the kinship (K matrix) among individuals estimated using SPAGeDi (Hardy and Vekemans 2002) is incorporated as the variance-covariance structure of the random effects of the individuals in the MLM for association mapping (Yu et al. 2006). This K estimate adjusts the identity by descent (IBD) by adjusting the probability of identity by state (IBS) between two individuals with the average probability of identity by state between random individuals.

Principle component analysis of the genotypic data transforms the variation into a series of orthogonal continuous axes, and these components, typically the first few, can then be used to replace Q to adjust for population structure. EIGENSTRAT is one software package that can be used to infer PCA (Price et al. 2006). Unlike Q, computing of PCA needs no assumptions about the number of groups in a population. In the unified mixed model, the principal components' axes can be fit as a fixed effect with K matrix as the random effect (Zhu and Yu 2009). Alternatively, nMDS can be used. A comparison of nMDS and PCA using the simulated and empirical data from cross- and self-pollinated species showed that models with nMDS resulted in higher power and fewer false positives (Zhu and Yu 2009).

One question emerges as PCA or nMDS is used for population structure control: How many components one must fit into model to account for the population structure, and should this number be the same for all traits, or vary for different traits? A two-stage dimension determination approach was proposed for PCA and nMDS (Zhu and Yu 2009). In general, a model testing should be conducted to select the most appropriate models based on the Bayesian Information Content (BIC), the lower the better. All relevant models are compared under maximum likelihood (ML). With the selected model, all markers can be then tested for marker-trait association. The corresponding Quantitle-Quantile (Q-Q) plot of the selected model has been shown; indeed, different numbers of PCAs are needed to have adequate, but not excessive, control for population.

Using MLM as the core, newer and faster algorithms have been developed to speed up GWAS analysis, particularly when hundreds of thousands of SNPs are tested. Efficient mixed-model association (EMMA) corrects sample structures by accounting for pairwise relatedness between individuals and uses enough markers by modeling phenotype distribution. This method is related to a method developed to simulate a null distribution of variance component test statistics (Crainiceanu and Ruppert 2004). EMMA increases the computational speed and

efficiency of mixed model analysis by enabling statistical tests with single-dimensional optimization. The method also avoids the redundant, computationally expensive matrix operations at iteration and allows converge to the global optimum of the likelihood in variance component estimation with high confidence. This capability was demonstrated in *in silico* whole-genome association mapping of mouse, *Arabidopsis*, and maize datasets. Results from the EMMA method are consistent with published results in reducing false positives and are faster than the previous methods while performing near global optimization (Kang et al. 2010; Kang et al. 2008).

Compressed MLM is another approach that clusters individuals into groups based on kinship estimates, thereby reducing the effective sample size to improve speed in subpopulation determination. Compressed MLM is equivalent to MLM when each group is treated as a single group, and compressed MLM is equal to GLM when all the individuals are in one group. This method is an extension of the pedigree-based sire model (Henderson 1975) with modifications (Zhang et al. 2010b). SSR markers were the predominant molecular markers for determining population structure and familial relatedness, but recent advances in the sequencing technology, higher genome density, lower mutation rate, and ease of detection through high-throughput systems have made single nucleotide polymorphisms (SNPs) the markers of choice. However, with decreasing genotyping costs and increasing numbers of individuals studied, more promising methods are needed to control population structure and the familial relatedness in association mapping.

Nested association Mapping (NAM)

Ideally, an association mapping population can be genotyped once but phenotyped repeatedly for the same sets of traits and new sets of traits; thus, it is advantageous to have a population that can be used by the research community for different purposes. With this in mind, the maize community has developed a nested association mapping (NAM) population to integrate the advantages of linkage analysis and association mapping, with the ultimate goal of dissecting complex traits in maize (Yu et al. 2008). The aims of developing NAM population were to (1) capture maize genetic diversity, (2) exploit the historical recombinations in maize, (3) use a genetic design that can take advantage of next-generation sequencing technologies, (4) generate materials for evaluation of agronomic traits in the field locations of the temperate regions, (5)

develop a population with enough power to detect QTLs and resolve QTLs to the gene level, and (6) provide a community resource that will enable a wide range of community efforts and databases for researchers. A publicly available resource of immortal lines with 26 founders represents the global diversity of maize. A set of 25 diverse inbred lines were crossed to common reference maize inbred line B73, and 5,000 distinct genotypes were created. The 5,000 genotypes are called NAM recombinant inbred lines (NAM RILs).

In essence, NAM is a multiple RIL population derived from crosses between a common founder line and a set of founders. The strategy of NAM is to genotype common-parent-specific (CPS) markers on the founders and progenies but to sequence the founders completely or densely with high-density markers. The genetic information from the CPS is projected from the founders to the progenies after genotyping the founders with high-density markers. Projecting genetic information from the parents to the progenies also reduces genotyping costs. The concept of NAM involves the development of a population. Instead of assembling existing lines to form a population, NAM selects a diverse set of founders. It also provides high allele richness, high statistical power, good mapping resolution, low sensitivity to genetic heterogeneity, low requirement of SNP markers in the progenies, and is amenable to repeated phenotyping (Yu et al. 2008). A series of maize NAM studies has been conducted on genetic map (McMullen et al. 2009), flowering time (Buckler et al. 2009), leaf architecture (Tian et al. 2011), and disease resistance (Kump et al. 2011; Poland et al. 2010). These studies will be reviewed in section 2.4.

Software for association mapping

The most commonly used and frequently updated software for association mapping is TASSEL (Trait Analysis by aSSociation, Evolution and Linkage), which is written in Java and can be used in virtually any operating system (Bradbury et al. 2007). TASSEL implements GLM, MLM, compressed MLM, and P3D approaches for marker-trait association analysis. Other notable functions include evolutionary analysis, computation of LD, imputation of missing data, and data visualization. The program allows calculating and visualizing LD graphically.

Structured association (Pritchard et al. 2000b) as well as the unified mixed model (Yu et al. 2006) were first implemented in TASSEL to reduce the risk of false positives. The Q + K method was implemented in TASSEL as a mixed linear model (MLM) function. TASSEL earlier employed an EM (expectation-maximization) algorithm for MLM analysis. To increase

computing speed and analyze larger datasets, the EMMA algorithm was incorporated into TASSEL. To increase the statistical power, model compression was added.

For even larger datasets, a newer method estimates the population parameters once, then estimates the parameters for test markers were included. Called population parameters previously determined (P3D) (Zhang et al. 2010b), this method is available in the newer version of TASSEL, and compressed estimation of variance components are available in software EMMAX (Kang et al. 2010). EMMA eXpedited (EMMAX) is publically available software that implements a variance component approach for GWAS. EMMAX is built on EMMA's previous approach.

ASREML is a complete package with different modules for mixed model analysis (Gilmour et al. 2002). SAS and R software are generic tools that can be used for association mapping.

GAPIT (Genome Association and Prediction Integrated Tool) is a new tool in the R package that can perform genome-wide association study (Lipka et al. 2012). It integrates the unified mixed model, EMMA, P3D, and compressed MLM with genomic prediction. This software handles large genotypic datasets by subdividing them into multiple files, but the memory requirement remains the same. Genomic predictions are done using a method developed by Zhou et al. (2011; in progress). GAPIT can conduct hierarchical clustering and kinship matrices based on user input and linkage information. The results are produced in the form of Q-Q plot, Manhattan plot, PCA, and association tables.

Computational speed

The unified mixed model method originally developed by Yu et al. (2006) is a widely used technique to correct genetic relatedness in association mapping studies; however, in analyzing genome-wide datasets, solving the mixed model requires a huge amount of computing power. The computing time for solving an MLM increases with the cube of the number of individuals. One approach to reducing computing time is compressed MLM (Zhang et al. 2010b), which decreases the effective sample size of such datasets by clustering individuals into groups. The rationale behind this method has its roots in the sire-model approach (Quass and Pollak, 1996).

A complementary approach to compressed MLM, population parameters previously determined (P3D), has been proposed to reduce computing time by skipping the iteration process

in each individual marker test. In the first step, a base MLM without fitting any marker effect is solved for the variance components. In the second step, an individual marker test with MLM simply uses variance components from the first step without solving the specific mixed model again (Zhang et al. 2010b). This practice has been used in previous mixed model analyses to save computing time (Yu et al. 2005), but the need to reduce the computational burden of MLM is much higher in GWAS. Compressed MLM and P3D, when implemented jointly, significantly reduce computing time and maintain statistical power (Zhang et al. 2010b). These methods are implemented in the software program TASSEL (Bradbury et al. 2007). A different residual analysis approach was also proposed to conduct fast genome-wide pedigree-based association analysis (Aulchenko et al. 2007).

A variance component approach implemented in a publicly available software package, EMMA eXpedited (EMMAX) (Kang et al. 2010), reduces computing time for analyzing large GWAS datasets. First, a pairwise relatedness matrix is computed from high-density markers and used to represent the sample structure. Secondly, the contribution of the sample structure to the phenotype using a variance component model is estimated, resulting in an estimated covariance matrix of phenotypes that models the effect of genetic relatedness on the phenotypes. Thirdly, a generalized least square (GLS) *F*-test (Kariya and Kurata 2004) is applied to each marker to detect associations accounting for the sample structure using the covariance matrix. A study on the welcome trust consortium data (Browning and Browning 2008) found that EMMAX outperforms both PCA (Price et al. 2006) and genomic control (Devlin et al. 2001).

FaST-LMM, a factored spectrally transformed linear mixed model, was recently proposed to further address the computational issues of MLM. FaST-LMM is an algorithm for genome-wide association studies that scales linearly with sample size in both runtime and memory use. With data from 15,000 individuals, FaST-LMM ran an order of magnitude faster than current algorithms; whereas data for 120,000 individuals were analyzed with FaST-LMM in few hours, current algorithms failed. The LMM corrects for confounding by measuring genetic similarity using methods of identify by descent and a realized relationship matrix (RRM), estimated by using a small sample of markers. FaST-LMM can produce results similar to the LMM by reformulating the LMMs with two conditions: (1) SNP used to estimate genetic similarity is less than the number of individuals in the dataset, and (2) the RRM is used to

determine these similarities. This method requires a single spectral decomposition but does not assume variance parameters are same across the SNPs.

Achievements

Association mapping in plants

Association mapping in plants provides a powerful, complementary approach to existing QTL mapping and cloning with bi-parental populations, mutational dissection, and transgenic approaches. It has been widely adopted in almost all major crop species for gene identification, QTL validation, and to understand the genetic basis of complex traits (Zhu et al. 2008). Association mapping also has led to the development of common community resources in important crop species, including maize, rice, sorghum, soybean, and barley. Linkage disequilibrium estimation among a diverse set of accessions within a species typically provides basic knowledge about the potential resolution of association mapping and the marker density requirement. Following these LD studies, population structure analysis of the assembled association mapping is examined in details. The resulting information is then incorporated into either candidate-gene or genome-wide association analysis.

One of the major benefits of association mapping is the diversity captured across many different traits. Unlike specific bi-parental populations in which certain trait differences exist, most of the assembled association mapping panels can be used to study a host of traits so that questions from different angles can be studied, including basic biology, plant architecture, development, agronomic performance, adaptive characteristics, and nutritional value (Atwell et al. 2010; Flint-Garcia et al. 2005).

GWAS in plants

GWAS has become a routine research component in human disease studies and has been carried out with some success in plants. To dissect complex traits through whole-genome association mapping, diverse germplasm panels have been established in *Arabidopsis* (Nordborg et al. 2005), maize (Yu and Buckler 2006), rice (Huang et al. 2010), wheat (Breseghello and Sorrells 2006b), sorghum (Casa et al. 2008), barley (Caldwell et al. 2006), and soybean (Lu et al. 2011) (**Fig. 2**). In this section, we describe some of the experiments that have been successful in detecting marker-trait associations following GWAS strategies in plants (**Table 1**).

Arabidopsis example

Arabidopsis thaliana is a natural organism that exists in a wide range of habitats. *Arabidopsis* HapMap is a resource to study the evolutionary as well as functional genetics of natural populations to resolve complex trait variation due to single genes or individual nucleotides (Clark et al. 2007). Based on the genome analysis, LD decays rapidly in this species, within 50kb (Platt et al. 2010). As a model species with a small genome, *Arabidopsis* has been a frontrunner in association studies. The 1001 Genomes project is sequencing 1001 geographically diverse *Arabidopsis* strains (Weigel and Mott 2009). In the first phase of the project, 80 strains from eight regions of the native species range were sequenced and analyzed (Cao et al. 2011). Another report claims that 471 genomes have been sequenced (unpublished data). Two recent GWAS studies in *Arabidopsis* are reviewed below.

The feasibility of GWAS in plants was demonstrated by studying a sample *Arabidopsis thaliana* global population. The genotyped sample consisted of 95 accessions for which a number of phenotypes were available (Zhao et al. 2007), plus a set of 96 accessions for which flowering traits were available (Atwell et al. 2010). The genotyping chip containing 250,000 SNPs was used to genotype the accessions; thereby, the SNP density was one SNP per 500 bp, which is comparable to studies in humans. The phenotypes studied were related to flowering, plant defense, element concentrations, and developmental traits. This association sample has a highly complex population structure. The mixed model approach performed well in controlling the population structure when compared with other methods commonly used in human genetics. Even though the degree of confounding was different among phenotypes, association analysis effectively identified single genes with known functional polymorphism. Compared with the human GWAS, the sample size was low, but the identification of the gene was possible due to the genetic architecture of the trait. These studies are replicable under controlled conditions that eliminate environmental noise (Atwell et al. 2010).

Another comprehensive study of linkage and association mapping of flowering time was conducted under field and greenhouse conditions (Brachi et al. 2010). The experiment involved phenotyping 20,000 plants over two winters under field conditions. A set of 184 natural accessions from around the world was genotyped with 216,509 SNPs, and 4,366 RILs derived from 13 independent crosses were also examined. More than 60 QTLs with small to medium

effects were identified. The highlight of this research was that linkage mapping, which has a higher power to distinguish true positives from false positives than association mapping, should be integrated with GWAS. Another important finding was that the major genes governing flowering time in greenhouse conditions were not associated with flowering time in field conditions. Instead, a number of genes involved in the regulation of the plant circadian clock were associated (Brachi et al. 2010).

Maize example

The genetic diversity between two different maize inbred lines is roughly equivalent to the diversity between a man and a chimp (Buckler et al. 2006). The maize genome contains about 50,000 genes, and most of the genome comprises repetitive and transposable elements (Schnable et al. 2009). The huge genetic diversity of maize make high-resolution mapping possible, but also requires large numbers of SNPs and systematic analysis (Yu and Buckler 2006). Several maize association mapping populations have been assembled (Camus-Kulandaivelu et al. 2006; Flint-Garcia et al. 2005; Liu et al. 2003; Palaisa et al. 2003; Remington et al. 2001; Yu et al. 2008).

The Maize HapMap project is an excellent resource for plant geneticists to conduct association mapping. The first HapMap in maize (Gore et al. 2009) identified several million polymorphisms (1.4 million SNPs and 200,000 indels) among 27 diverse maize inbred lines and showed that the maize genome is characterized by highly divergent haplotypes. The second maize HapMap (HapMapV2) resulted in the identification of high-quality genotypic data of 50 million SNPs and small indels. These efforts provide the foundation for dissecting the complex trait variation in maize by uniting breeding efforts around the world.

To determine whether standing variation in the regulatory genes in maize contributes to variation in *Balsas teosinte*, association mapping was conducted on 584 *Balsas teosinte* individuals. Forty-eight markers from 9 candidate regulatory genes were tested against 13 traits for plant and inflorescence architecture. Ten associations involving five candidate genes were significantly identified after correcting for multiple testing. The maize homolog of *FLORICAULA* of Antirrhinum *zfl2* was associated with plant height. The maize homolog of *APETALA1* of *Arabidopsis zap1* was associated with inflorescence branching. Five SNPs in the

maize domestication gene, *teosinte branched1*, were significantly associated with either plant or inflorescence architecture (Weber et al. 2007).

To address vitamin A deficiency, the leading cause of blindness , disease, and death from severe infections in children. breeding for increased β-carotene (βC) levels in cereal grains (biofortification) is a realistic approach because β-carotene is a precursor of vitamin A. In the first study, association mapping coupled with linkage analysis, expression analysis, and mutagenesis identified variation in the *lycopene epsilon cyclase* (*lcyE*) gene that accounts for 58% of the variation in the α-carotene versus β-carotene branches of the carotenoid pathway and a threefold difference in provitamin A compounds. The *lcyE* gene was significantly associated with the branching and carotenoid content (Harjes et al. 2008).

In the second study, three association mapping populations were used. SSR and SNP markers were used to estimate the population structure and kinship matrices. GWAS identified a rare variation in the *crtRB1* gene in maize, which increases the β-carotene concentration and conversion in maize kernels (Yan et al. 2010). Results from these studies will facilitate breeding for increased β-carotene levels in cereal grains, thereby addressing the dietary vitamin-A deficiency in the developing world. This is a good example of cross-validating the QTLs using a combination of association mapping and linkage mapping strategies.

This NAM population has been used to elucidate the genetic basis of resistance to southern leaf blight (SLB) disease. SLB resistance was measured on a nine-point scale in three environments. The SLB index values varied among the founder lines, with B73 being the least resistant. The heritability of the SLB index was 87%, indicating potential for accurate mapping. Joint–linkage analysis identified 32 QTLs with small additive effects on SLB resistance. Genome-wide association tests of Maize HapMap were conducted by imputing the founder SNPs onto the NAM RILs; SNPs within and outside the QTLs were found to be associated with the variation for SLB resistance. Limited LD was observed around some SNPs, which indicates that NAM population is good for high-resolution mapping. But half of the QTLs detected by the bi-parental mating studies were not detected in NAM, due to either low frequency or absence of the alleles (Kump et al. 2011).

To gain insight into the genetic architecture of quantitative resistance to plant pathogens, the NAM RILs were evaluated for resistance to northern leaf blight (NLB). Using 1.6 million SNPs, multiple candidate genes related to plant defense were identified. Twenty-nine QTLs were

identified, most of which had multiple alleles. The study concluded that the large amount of variation present in the phenotype could be attributed to a number of loci with small effects (Poland et al. 2010).

Over the years, maize yield in the United States has increased because of reduced planting density and efficient light capture, which has been possible because breeders changed the plant architecture by selecting for small leaf angle and leaf size. One study by Tian et al. focused on the genetic basis of the factors responsible for increased yield in corn. They considered the genetic basis of leaf architecture traits in maize and identified key genes through a GWAS on the NAM population (Tian et al. 2011). This study demonstrated that the genetic architecture of the leaf traits (upper leaf angle, leaf length, and width) are dominated by QTLs with small effects, little epistasis, and environmental interaction or pleiotropy. The study showed that the variation in the liguleless genes has contributed to more upright leaves. For these three leaf traits, 30–36 QTLs were identified.

Flowering time is one of the traits most thoroughly studied by the plant community. It is a complex trait that controls the plant's adaptation to the local conditions. In maize, diversity-based dissection of flowering time is problematic due to tight linkage and population structure. Buckler et al. studied one million plants on the NAM population for flowering time over eight environments and identified numerous QTLs with small effects (Buckler et al. 2009). This study evaluated 5,000 lines plus 500 checks in four environments over two years for flowering time. Days to silking (DS) and days to anthesis (DA) were scored, and anthesis silking interval (ASI) was calculated. The QTLs were mapped on the 25 families separately using composite interval mapping (CIM) and jointly by joint inclusive composite interval mapping (JICIM). JICIM identified twice as many significant QTLs as the individual family analysis. No single QTL with large effects was detected in the study. The NAM founders showed allelic differences for allelic effects. This study showed that for an adaptive trait like flowering time, the genetic architecture of the trait is controlled by small additive genes with few genetic effects or environmental interactions. This finding for maize differs from *Arabidopsis* and rice, where the flowering time is controlled by fewer genes with large effects, epistasis, and environmental interactions.

Structured association analyses have been used to identify pleiotropic genes associated with correlated complex traits. To evaluate the hypothesis that the genes controlling multiple disease resistance (MDR) is present in maize, a mixed model approach for structured association

was extended to multivariate analysis. This analysis of a panel of 253 maize inbred lines identified high positive genetic correlations between resistances to southern leaf blight (SLB), northern leaf blight (NLB), and gray leaf spot (GLS). A glutathione *S*-transferase gene (*GST*) was conferring resistance to the three diseases. These successful examples in maize will encourage GWAS studies in other crops (Wisser et al. 2011).

Rice example

Rice is staple food for half of the world's population, and rice varieties are adapted to varied climatic regions around the globe. Rice is a highly self-fertilizing species with a high-quality reference genome (Sequencing Project International Rice 2005) and phenotyping resources. The genome of domesticated rice contains information that could explain a large amount of the morphological, physiological, and ecological variation present in most of the cultivars throughout out the world (McNally et al. 2009).

Seed shattering is a major trait in the domestication of crop plants. Konishi et al. studied loss of seed shattering through haplotype analysis and association analysis in various rice collections. The study revealed that an SNP highly associated in *japonica* subspecies was the target for artificial selection. QTL analysis revealed that the loss of seed shattering might have occurred independently in *japonica* and *indica* varieties. A QTL of seed shattering in chromosome *1*(*qSH1*) explained 68.6% of the total phenotypic variation in the population. Fine-mapping the *qSH1* gene with 10,388 plants located the natural variation to a 612 bp and identified one SNP. The complementation tests proved that the *qSH1* gene was the homolog of the *RLP* gene in *Arabidopsis*. The researchers also verified the SNP using association analysis of the rice core collections, which indicated that it was highly associated with the degree of seed shattering among the temperate *japonica* rice cultivars. This SNP was a target of artificial selection in rice domestication (Konishi et al. 2006).

Apart from seed shattering, domestication of rice is associated with improvement in grain size, grain number, panicle size, grain quality, plant architecture, and flowering time, but the prime objective in domesticating rice is increasing grain yield. Through fine-mapping, complementation testing, expression analysis, and haplotype testing, researchers found that a deletion in a QTL, *qSW5* (QTL for seed width on chromosome 5) gene resulted in a significant increase in sink size and increased yield of the rice grains (Shomura et al. 2008).

To access the variation present within and between rice cultivars and landraces, the International Rice Functional Genomics Consortium (IRFGC) initiated an SNP discovery project (McNally et al. 2006). With this project, rice was the first crop plant for which a high-quality reference genome sequence from a single variety was produced. Through whole-genome comparisons of the 21 rice genomes including cultivars, landraces, and breeding materials (publically available in www.oryzasnp.org), 160,000 non-redundant SNPs were identified. This provides the foundation for high-resolution genotyping of hundreds or thousands of varieties (Huang et al. 2010).

Rice domestication was a complex process. The deep genetic divergence between the two main varietal groups (*indica* and *japonica*) suggests domestication of rice from two distinct wild populations. GWAS was performed in rice to understand the genome-wide patterns of polymorphism, to characterize population structure, and to infer the introgression history of domesticated Asian rice. The analysis showed that a key gene, *SD1* (*OsGA20* oxidase), determines plant height and was responsible for green revolution (Zhao et al. 2010).

GWAS was performed on 517 landraces of rice with 3.6 million SNPs to understand the genetic basis of diverse varieties in rice. A Rice HapMap was created and GWAS was performed for 14 agronomic traits. The LD decay of indica and japonica were between 123kb and 167kb. The simple as well as the compressed MLM models were used to identify the association signals. On average, the loci identified through GWAS explained ~36 % of the phenotypic variance. The highly significant associations of six loci were close to the previously identified genes. The researchers reported that an approach which integrates the second genome sequencing and GWAS could be used as a powerful complementary strategy to traditional linkage mapping in dissecting complex traits (Huang et al. 2010).

Genome-wide association mapping revealed a rich architecture of complex traits in rice. Numerous common variants influencing physiological, developmental, and morphological traits were identified by a genome-wide association study based on genotyping 44,100 SNP variants across 413 diverse accessions of *O. sativa* collected from 82 countries that were systematically phenotyped for 34 traits. Significant heterogeneity was observed in the genetic architecture associated with subpopulation structure and response to environments. This study was an open-source translational research platform for genome-wide association studies in rice that directly

linked molecular variation in genes and metabolic pathways with the germplasm resources needed to accelerate varietal development and crop improvement (Zhao et al. 2011).

The work of Zhao et al. was followed with another GWAS in a more diverse sample of 950 worldwide varieties that included *indica* and *japonica* subspecies. The researchers identified 32 new loci associated with flowering time and grain-related traits using the compressed MLM approach. The study reveals that an integrated approach following sequencing-based GWAS and functional genome annotation has the potential to reveal more true marker-trait associations (Huang et al. 2012).

Community resources in Wheat, Barley, Soybean, and Sorghum

Wheat is a challenging crop in terms of conducting association mapping owing to hexaploidy, an unfinished genome sequence, and difficulties in sequencing and allocating sequences to the A, B, or D genome. Earlier research in wheat has contributed significantly to our understanding of the potential and strategy of association mapping in crops (Breseghello and Sorrells 2006a, b; Sorrells and Yu 2009). Scaling up studies in wheat is hampered mainly by the lack of a large number of SNPs, but concerted efforts have been made to sequence the wheat genome, which would greatly facilitate association mapping studies and related research. The International Wheat Genome Sequencing Consortium (IWGSC) (http://wheat.pw.usda.gov/PhysicalMapping) was established by plant scientists, plant breeders, and producers to understand the structure and function of the wheat genome (www.wheatgenome.org). Most notably, three custom high-throughput SNP genotyping assays (1,536-, 9,000- and 50,000-SNP) based on Illumina BeadArray and Infinium platforms have been developed (Chao et al. 2010) (E Akhunov, personal communication). A set of 2,994 wheat lines were genotyped with the 9,000-SNP iSelect assay (Cavanagh et al. 2013, manuscript in review).

To understand the genetic diversity, population structure, and linkage disequilibrium in U.S. elite winter wheat, 205 elite breeding lines were analyzed from U.S. winter wheat breeding programs. The accessions were from the Southern and Northern Regional Performance Nurseries, the Regional Germplasm Observation Nursery, the elite hard winter wheat nursery at Oklahoma State University, the Uniform Eastern Soft Red Winter Wheat Nursery, and the Uniform Southern Soft Red Winter Wheat Nursery, plus 22 major cultivars recently released in the hard winter wheat region. Researchers genotyped the 205 elite breeding lines using 245 SSR

markers. Population structure, LD, cluster analysis, and PCA revealed that these collections were highly structured based on their geographical location. The soft and hard winter wheat was separated in the study. The hard winter wheat had more genetic diversity than the soft winter wheat. The LD decay was about 10 cM in the genome (Zhang et al. 2010a).

One of the conclusions from the study was that modern breeding practices maintain reasonable genetic diversity in major U.S. winter wheat breeding gene pools. Also, the presence of higher genetic diversity in hard winter wheat could be used to broaden genetic diversity in soft winter wheat. LD blocks in the genome were identified, but the majority of the genome has lower LD decay, which indicates these could be used for association mapping studies. This study focused on evaluating the germplasm for genetic diversity in the current breeding programs, which will facilitate the use of this information to future cultivar release programs (Zhang et al. 2010a).

Efforts to analyze wheat genetic resources for conducting wheat association studies are progressing. Akhunov et al. performed analysis of nucleotide diversity to construct an SNP database in wheat. They studied about 2,114 genes for nucleotide diversity in *T. aestivum*, *T. dicoccodies* and synthetic 6x wheat. Genetic diversity was similar between A and B genome but was reduced in the D genome. The low variance of the D genome was together with an excess of rare alleles in some genes. The researchers discovered a total of 5,471 SNPs in 1,791 genes. Studying *T. aestivum* and *T .dicoccodies* is a good strategy to develop SNP markers in wheat, where ancestral species are the source of genetic variability. Self pollination and homeologous chromosome pairing could lead to loss of variability in wheat (Akhunov et al. 2010).

Another study on the population structure and genome-wide linkage disequilibrium in wheat used 1536 SNPs (Chao et al. 2010). This study used a panel of 478 spring wheat cultivars from 17 populations across the United States and Mexico, and the population structure analysis identified 9 clusters, indicating that previously inferred populations share a common genetic identity. The assessment of LD and population structure in this assembled panel of diverse lines provides critical information for the development of genetic resources for genome-wide association mapping of agronomically important traits in wheat (Chao *et al.*, 2010).

Barley has a high degree of population sub-structure due to breeding activities. More coordinated community-based approaches are followed in crop plants, and barley is no

exception. Eight founding institutions from six countries have initiated a sequencing project in barley called The International Barley Sequencing Consortium (Schulte et al. 2009). Vernalization requirements for flowering in winter and spring-sown barley is a major subdivision. Major flowering time loci (*VRN-H1* and *VRN-H2*) in barley were identified through association mapping by studying 429 spring and winter barley accession from Europe (Cockram et al. 2008).

Even if the population structure were correlated with the phenotype, by effectively using the statistical methods developed, successful GWAS could be conducted even using low marker density. This assumption was validated by successfully mapping 15 morphological traits in barley. Five hundred barley cultivars were genotyped with 1,536 SNPs. The traits studied were seasonal growth (1H), grain lateral nerve speculation (2H), awn anthocyanin coloration, awn anthocyanin intensity, auricle anthocyanin coloration, auricle anthocyanin intensity, lemma nerve anthocyanin intensity (2H), grain aleurone color (4H), hairiness of leaf sheath (4H), rachille hair type (5H), ear attitude (5H), and grain ventral furrow hair (6H) were significantly associated. By developing a double haploid (DH) population from two of the inbreds in the GWAS panel differing in the anthocyanin pigmentation, the *ANT2* gene on chromosome 2H was fine-mapped and validated by genotyping the population using the 1,536 SNP array (Cockram et al. 2010).

GWA studies have been conducted to understand the genetic basis of domestication in barley, in which domestication has changed the morphological feature of the inflorescence architecture and resulted in two-rowed and six-rowed forms derived from the ancestor two-row wild types. The development of the six-row barley is controlled by a gene, *VRS1,* on chromosome 2H. But the genome-wide scans show that *INTERMEDIUM*-C located on chromosome 4H is an ortholog of the maize domestication gene *TEOSINTE BRANCHED* 1(*TB1*), which acts as the modifier of the of the *VRS1* gene. Ramsay et al. conducted genome-wide association scans of 190 barley cultivars by genotyping them with 2,463 biallelic SNPs. This experiment identified 3 genomic regions on chromosomes 1HL, 2HL, and 4HS associated with the row type. The association on 2H was the *VRS1* gene. The researchers also identified 17 coding mutations in *TB1* correlated with lateral spikelet fertility. The *INT*-C as an ortholog of *ZmTB1* and the confirmation of its involvement in determining both the fertility of the lateral spikelets and of tillering was carried out using a combination of genome-wide association

mapping, and studying conservation of synteny and a collection of well-characterized mutant stocks (Ramsay et al. 2011).

Soybean is an autogamous plant species that exhibits high variation in LD across its genome, indicating that a large number of markers is needed to perform GWAS. The LD pattern in soybean was identified by a study that focused on three genomic regions varying from 336 to 574 kb. The populations used were 26 accessions of the wild ancestor of soybean (*Glycine soja* Seib. et Zucc.), 52 Asian *G. max* landraces, 17 Asian landrace introductions that became the ancestors of North American (N. Am.) cultivars, and 25 elite cultivars from N. Am. In the three cultivated *G. max* groups, LD extended from 90 to 574 kb (Hyten et al. 2007).

Despite high LD in soybean, efforts are ongoing to sequence a number of landraces and cultivars to further understand the genetic structure and to perform association mapping. Seventeen wild and 14 cultivated soybean genomes were sequenced to x5 depth and >90% coverage using the Illumina Genome Analyzer II platform. A comparison of the patterns of genetic variation between wild and cultivated soybean indentified high allele diversity among the wild soybean. Researchers also identified a set of 205,614 tag SNPs useful for LD mapping and linkage analysis. This is a valuable resource for the analysis of wild soybeans and to facilitate future breeding and quantitative trait analysis.

Large-scale SNP discovery has been conducted by deep resequencing of a reduced representation library (Hyten et al. 2010). Researchers then used the generated SNPs to create a high-resolution map that assisted in the assembly of scaffolds from the 8x whole genome shortgun sequences into pseudomolecules corresponding to soybean chromosomes. As in other crops, the release of the soybean genome sequence (Schmutz et al. 2010) would speed up association mapping related research. Concerted efforts are ongoing to develop a large sorghum NAM population with 5,600 lines (B. Diers, personal communication). A set of 40 soybean lines was selected from lines nominated by the soybean community to maximize the genetic diversity based on clustering analysis with 1536 SNPs. The lines were then crossed to IA302, a high-yield line.

Sorghum is a staple food for the people in sub-Saharan Africa. Its C4 photosynthesis, drought resistance, wide adaptation, and high nutritional value hold the promise to alleviate hunger in

Africa. Release of the sorghum genome sequence (Paterson et al. 2009) greatly facilitated research in association mapping. Efforts have been made to assemble information and community resources to conduct association mapping in sorghum and to investigate the LD in sorghum (Casa et al. 2008; Hamblin et al. 2004). Sorghum is an excellent species for association study owing to selfing, low sequence diversity, high LD compared with maize, and availability of a sequenced genome. The variation in sorghum is four-fold less than that of maize (Hamblin et al. 2004).

A recent SNP discovery project through resequencing 8 diverse sorghum accessions successfully identified 283,000 SNPs (Nelson et al. 2011). This study used the restriction site associated DNA (RAD) approach to construct the sequencing library from only genomic DNA fragments whose 5' ends about the recognition site of the selected restriction enzymes, *Pst* or *Bsr*FI. SNP discovery rate of the RAD approach was 10-fold higher than that of a semi-random library (digestion by *Hpa*II and fragment size selection of 200-2000 bp).

A panel of 377 lines of sorghum representing major cultivated lines and lines from the sorghum conversion program (SCP) were assembled and characterized for eight traits. Population structure and linkage disequilibrium were estimated (Casa et al. 2008). A 300-line set (a subset of the 377-line set) has also been characterized for grain quality and genes were identified (unpublished results). A 2000-line sorghum NAM population also has been developed by crossing 10 diverse sorghum lines selected from the sorghum diversity panel with a common parent, Tx430 (Yu et al. 2012).

Candidate gene association mapping has been conducted in sorghum to map the plant height gene. In this study, the sorghum diversity panel was used to effectively characterize the phenotypic effects of the *dw3* mutation and to fine-map a second, epistatic dwarfing QTL on sorghum chromosome 9 (Brown et al. 2008). In addition, sweet sorghum has the potential to become the crop for bio-energy production. Association mapping has been conducted on sweet sorghum for brix and plant height (Murray et al. 2009). Different sweet sorghum collections were also analyzed for population structure and genetic diversity (Ali et al. 2008; Wang et al. 2009a).

Challenges and opportunities

Next-generation sequencing technologies provide new opportunities and challenges to the plant genetics communities (**Fig. 3**). New strategies for high-throughput large-scale phenotyping need

to be developed to match the level of the genotyping/resequencing capacity. Improved bioinformatics, database, statistical methods, and genetic designs are needed for the large-scale data generated from sequencers, fields, growth chambers, greenhouses, and analytical equipment and scanners. Many new areas will be incorporated into association mapping; for example, genotyping strategies will be tested for detecting copy number variation (CNV) (Rogers and Bendich 1987; Springer et al. 2009) and presence-absence variation (PAV) (Springer et al. 2009). Resequencing strategies using RNA-seq (Wang et al. 2009b), exome sequencing (Ng et al. 2009), and genotyping-by-sequencing (GBS) are being optimized.

High-throughput phenotyping is likely to be the most expensive part of plant genetic studies. Phenotypes used in association mapping are being expanded from the traditional, labor-intensive phenotyping to gene expression, and protein/metabolite level. New techniques such as CT-SCAN, near infrared (NIR) spectroscopy, single kernel characterization system for grain quality, global positioning system, and image analysis are all being exploited for faster and more accurate phenotyping. More studies focusing on rare alleles, CR-GWAS (Zhu et al. 2011), multiple alleles, computational speed, data storage, and computational power are also needed. Validation of the results using RNA-seq and other approaches will be important. Methylation in the genome could be used as a marker (Laird 2003) and or as a trait (Lukens and Zhan 2007). Integration of association mapping will certainly be integrated with other areas such as MutMap (Abe et al. 2012), next-generation mapping (NGM) (Schneeberger et al. 2009), genomic selection (Bernardo and Yu 2007a; Meuwissen et al. 2001), and comparative mapping.

Missing heritability

Association mapping strategy is based on the thought that the common phenotypic variation will be caused by common genetic variation, but in most GWA studies, identified significant associations explain little of the phenotypic variation. Most of the studies on flowering time estimate heritability of flowering time as greater than 80%, but none of the studies have identified QTL or SNPs that can explain 80% of the phenotype variation. Consequently, Manolio et al. were correct to call the missing heritability in GWAS the “dark matter” of human genetics (Manolio et al. 2009).

The most important factor that accounts for the missing heritability is the genetic structure of the trait under study. The inability to explain missing heritability also could be

attributed to a number of factors: genotype by environment interaction, a larger number of variants of smaller effects yet to be found; rare variants (possibly with larger effects) that are poorly detected by available genotyping arrays that focus on variants present in 5% or more of the population; structural variants poorly captured by existing arrays; low power to detect gene–gene interactions, inadequate accounting for shared environment among relatives, statistical issues, copy number variation (CNV), and multiple testing issues. The power of GWAS to detect variants of modest effect and low frequency remains lacking due to the low frequency of functional alleles in the mapping population, the low influence of low-frequency alleles on the population, and/or lower detection power of the association mapping strategy. The phenotypic variation caused by numerous small-effect alleles will be difficult to detect compared with a small number of large-effect alleles; this is a challenge for any complex trait dissection studies, including association mapping. Also, while estimating the heritability, whether we are correctly accounting for environmental variances remains a question. Scientists still believe that these are early days to explain missing heritability. The solution is to bridge the gap between the phenotypic variation present in the sample and the variation that can be explained by the marker.

New gene identification

If the complete genome sequence is unavailable while conducting association, newer genes with small effects always will be missed in GWAS. The candidate gene approach is always biased from the beginning, and multiple testing issues and genetic epistasis adds to the problem of lower detection rate of rare alleles. The presently used association mapping methods have limited power to identify new genes. As described earlier, the candidate gene approach uses an SNP discovery panel that limits detection of polymorphism to only a small sample. GWAS methods also focus on variation between a limited number of genotypes due to the sequencing cost. Reduced sequencing costs could change the statistics of new gene identification.

A newly developed and improved approach that combines LD mapping and linkage analysis, NAM also has disadvantages in identification of new genes. First, in maize where the LD decays at 2kb, 1.6 million markers may not be able to capture the complete SNP diversity present in the genome, but this method could be useful in new gene identification in self-fertilizing, low-LD decay species. Secondly, only SNP and INDEL are detected in NAM, and presence absence variations (PAV), and copy number variations (CNV) are not accounted. Third,

a medium number of founders are used in GWAS, which may not capture all the variations available in the species, and fourth, the rare functional alleles cannot be detected because of the lower number of founders used in developing the populations. So identification of new genes through GWAS is still a challenge, which could be addressed by developing new methods that can accommodate higher genome coverage and a higher number of sequenced individuals.

GBS: Genotyping-by-sequencing

Rapidly evolving sequencing and genotyping technologies have fundamentally changed not only the design of specific breeding and selection strategies in crops, but also how the vast amount of available germplasm diversity can be utilized efficiently (Bernardo and Yu 2007b; Heffner et al. 2009; Tester and Langridge 2010). Routine use of GS in plant breeding is becoming possible because of the significantly reduced cost of obtaining molecular marker information, particularly SNPs, thanks to the development of high-throughput technology from DNA extraction, sample preparation, and array-based genotyping technologies as well as cutting-edge GBS technology (Metzker 2010). Current GBS research includes species with a sequenced genome (Huang et al. 2009b), such as rice, maize, and sorghum, and those without, such as wheat and barley (Chutimanitsakun et al. 2011). High-throughput genome sequencing was earlier approached through hybridization-based methods that were laborious, time-consuming, and expensive to design for specific mapping populations. Next-generation sequencing technologies have improved output and made possible sequencing of multiple samples at the same time. Sequencing-based high-throughput genotyping combines the advantages of cost-effectiveness, less time, and dense marker data. A sliding window approach for analyzing the SNPs collectively rather than individually was used on 150 RILs derived from the cross between *indica* and *japonica* rice cultivars. The SNP calling in this method is based on a recombination break point and sliding window. Based on the sequence-based genetic map, a 100-kb region was identified for plant height that is related to a green revolution gene (Huang et al. 2009a). Other approaches attempt to construct the GBS library by reducing the genome complexity through restriction enzymes (REs). Methylation-sensitive REs are used to reduce the genome complexity so lower copy regions are targeted with higher efficiency. This method simplifies the challenges of sequence alignment problems (Elshire et al. 2011).

GBS is an alternative to complex, expensive protocols for sequencing. It has a wide range of applications on breeding, population studies, germplasm characterization, and marker-trait association of diverse germplasms. In the future, plant breeders may be able to do a genomic selection without having prior information about the germplasm. The most important question about GBS is the DNA quality. This technology will accelerate GWAS in crop species and model organisms.

Rare alleles

At present, GWAS is unable to detect rare variants through common SNP markers (Ott et al. 2011). The power to detect an association is a function of allele frequency, and rare alleles have little influence on the population, which renders their detection difficult. The difficulty in detecting the rare alleles is more of a methodological problem than a statistical issue. Rare alleles are supposed to play an important role in the genetics of complex traits, but methodologies to test the interactions at the genome level are lacking due to multiple testing problems. Data mining and multivariate analysis methods have not been not successful in addressing the rare variant issue.

A composite resequencing–based approach could be a solution to the problem of accommodating ever-increasing genomic data. This approach integrates next-generation sequencing, exome sequencing, whole genome sequencing, prediction of biological function of SNPs based on gene prediction, statistical test for rare allele variants, and development of genome databases and gene networks. These approaches could be integrated in rare allele testing and were successfully practiced to identify marker-trait associations in *Arabidopsis* empirical data (Zhu et al. 2011). The method follows the following steps: (1) analyzing gene fragments with statistical methods to identify significant test for gene fragments, (2) testing for genes with common variants to see if prior candidate genes that were significant at 0.001 could be detected by earlier methods, (3) testing rare variants, and (4) using the gene network AraNet to verify the tested genes. Interested readers are directed to Zhu et al. (2011) for more details. When employed on empirical data, these methods showed that a weighted sum test and function-aided sum test were more consistent than the sum test (Zhu et al. 2011). A comparison of results from the three tests could reduce the number of false positives from testing rare alleles.

Conclusion

Association mapping has become one of the major approaches in gene discovery and complex trait dissection in plants. Combined with genetic designs in plant genetics, the development of immortal populations such as NAM showcased the potential of what can be achieved by assimilating knowledge and discovery in other research areas. Concerted efforts are ongoing in almost all major plan species, and we expect great findings in these studies.

Tables and Figures

Table 1. 1 Examples of association mapping study in different crops.

Species	Trait / objective	Population	Marker	Reference
Arabidopsis	107 phenotypes	191 accessions	250,000 SNPs	Atwell et al., 2010
Arabidopsis	Flowering time	184 accessions + 4,366 RILs	216,509 SNPs	Branchi et al., 2010
Zea mays	Leaf angle, leaf length, and width	Maize NAM	1.6 million SNPs	Tian et al., 2011
Zea mays	Southern leaf blight disease	Maize NAM	1.6 million SNPs	Kump et al., 2011
Zea mays	Provitamin A	288 lines	Candidate genes	Harjes et al., 2008
Zea mays	Provitamin A	681 maize germplasm	Candidate genes	Yan et al., 2010
Oryza sativa	14 agronomic traits	517 landraces	3.6 million SNPs	Huang et al., 2010
Oryza sativa	Flowering time and 10 grain-related traits	950 worldwide accessions	4,1 million SNPs	Huang et al., 2012
Triticum aestivum	To understand genetic diversity, population structure, and linkage disequilibrium	205 elite breeding lines	245 SSRs	Zhang et al., 2010
Triticum durum	Drought-adaptive traits and grain yield	189 elite durum	90 SSRs	Maccaferri et al., 2010
Hordeum vulgare	15 morphological traits	500 cultivars + DH populations	1536 SNPs	Cockram et al., 2010
Hordeum vulgare	Domestication traits	190 cultivars	2463 SNPs	Ramsay et al., 2011

Figure 1. 1 Achievements in association mapping methods and future challenges.

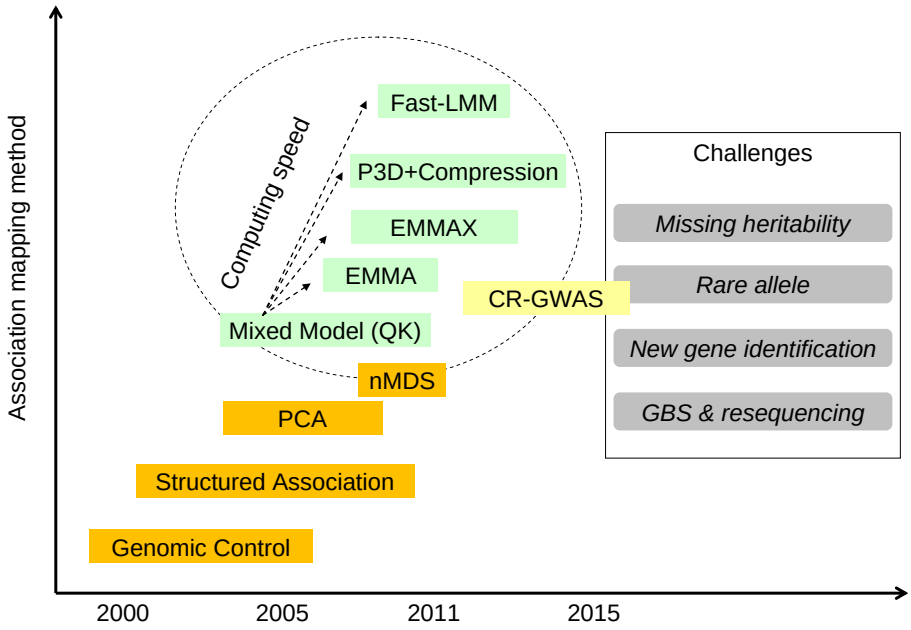


Figure 1. 2. A sample of publications based on population size and scale of markers.

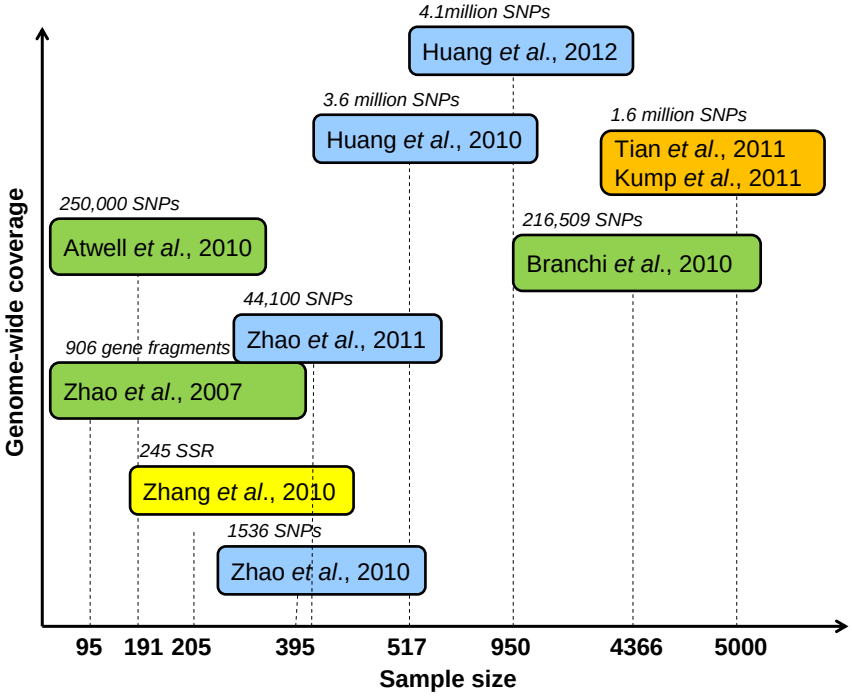
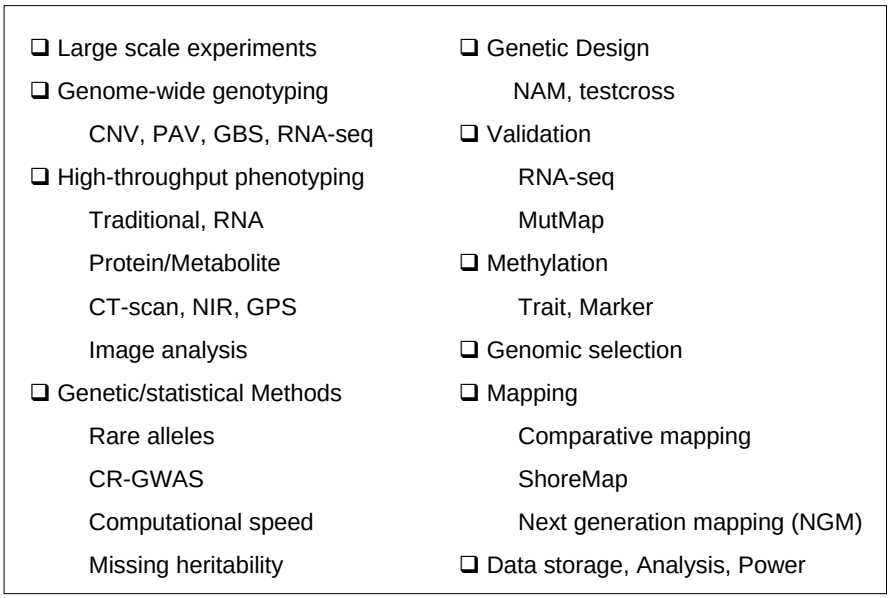


Figure 1. 3. Future opportunities and challenges that are related to association mapping and the general complex trait dissection and selection.



References

Abe A, Kosugi S, Yoshida K, Natsume S, Takagi H, Kanzaki H, Matsumura H, Yoshida K, Mitsuoka C, Tamiru M, Innan H, Cano L, Kamoun S, Terauchi R (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. Nature biotechnology 30:174-178

Akhunov E, Akhunova A, Anderson O, Anderson J, Blake N, Clegg M, Coleman-Derr D, Conley E, Crossman C, Deal K, Dubcovsky J, Gill B, Gu Y, Hadam J, Heo H, Huo N, Lazo G, Luo M-C, Ma Y, Matthews D, McGuire P, Morrell P, Qualset C, Renfro J, Tabanao D, Talbert L, Tian C, Toleno D, Warburton M, You F, Zhang W, Dvorak J (2010) Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. BMC Genomics 11:702

Ali M, Rajewski J, Baenziger P, Gill K, Eskridge K, Dweikat I (2008) Assessment of genetic diversity and relationship among a collection of US sweet sorghum germplasm by SSR markers. Molecular Breeding 21:497-509

Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, de Meaux J, Ecker JR, Faure N, Kniskern JM, Jones JDG, Michael T, Nemri A, Roux F, Salt DE, Tang C, Todesco M, Traw MB, Weigel D, Marjoram P, Borevitz JO, Bergelson J, Nordborg M (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465:627-631

Aulchenko YS, de Koning DJ, Haley C (2007) Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics 177:577-585

Bernardo R, Yu J (2007a) Prospects for genomewide selection for quantitative traits in maize. Crop Science 47:1082-1090

Bernardo R, Yu JM (2007b) Prospects for genomewide selection for quantitative traits in maize. Crop Science 47:1082-1090

Brachi B, Faure N, Horton M, Flahauw E, Vazquez A, Nordborg M, Bergelson J, Cuguen J, Roux F (2010) Linkage and Association Mapping of Arabidopsis thaliana Flowering Time in Nature. PLoS Genet 6:e1000940

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635

Breseghello F, Sorrells ME (2006a) Association analysis as a strategy for improvement of quantitative traits in plants. Crop Science 46:1323-1330

Breseghello F, Sorrells ME (2006b) Association Mapping of Kernel Size and Milling Quality in Wheat (Triticum aestivum L.) Cultivars. Genetics 172:1165-1177

Brown PJ, Rooney WL, Franks C, Kresovich S (2008) Efficient Mapping of Plant Height Quantitative Trait Loci in a Sorghum Association Population With Introgressed Dwarfing Genes. Genetics 180:629-637

Browning B, Browning S (2008) Haplotypic analysis of Wellcome Trust Case Control Consortium data. Human Genetics 123:273-280

Buckler ES, Gaut BS, McMullen MD (2006) Molecular and functional diversity of maize. Current Opinion in Plant Biology 9:172-176

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, Goodman MM, Harjes C, Guill K, Kroon DE, Larsson S, Lepak NK, Li H, Mitchell SE, Pressoir G, Peiffer JA, Rosas MO, Rocheford TR, Romay MC, Romero S, Salvo S, Sanchez Villeda H, da Silva HS, Sun Q, Tian F, Upadyayula N (2009) The genetic architecture of maize flowering time. Science 325:714 - 718

Caldwell KS, Russell J, Langridge P, Powell W (2006) Extreme Population-Dependent Linkage Disequilibrium Detected in an Inbreeding Plant Species, Hordeum vulgare. Genetics 172:557-567

Camus-Kulandaivelu L, Veyrieras JB, Madur D, Combes V, Fourmann M, Barraud S, Dubreuil P, Gouesnard B, Manicacci D, Charcosset A (2006) Maize adaptation to temperate climate: Relationship between population structure and polymorphism in the Dwarf8 gene. Genetics 172:2449-2463

Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Muller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, Weigel D (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43:956-963

Casa AM, Pressoir G, Brown PJ, Mitchell SE, Rooney WL, Tuinstra MR, Franks CD, Kresovich S (2008) Community Resources and Strategies for Association Mapping in Sorghum. Crop Sci 48:30-40

Chao S, Dubcovsky J, Dvorak J, Luo M-C, Baenziger S, Matnyazov R, Clark D, Talbert L, Anderson J, Dreisigacker S, Glover K, Chen J, Campbell K, Bruckner P, Rudd J, Haley S, Carver B, Perry S, Sorrells M, Akhunov E (2010) Population- and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (Triticum aestivum L.). BMC Genomics 11:727

Chutimanitsakun Y, Nipper RW, Cuesta-Marcos A, Cistue L, Corey A, Filichkina T, Johnson EA, Hayes PM (2011) Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. Bmc Genomics 12

Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, Chen H, Frazer KA, Huson DH, SchÃ¶lkopf B, Nordborg M, RÃ¤tsch G, Ecker JR, Weigel D (2007) Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis thaliana. Science 317:338-342

Cockram J, White J, Leigh F, Lea V, Chiapparino E, Laurie D, Mackay I, Powell W, O'Sullivan D (2008) Association mapping of partitioning loci in barley. BMC Genetics 9:16

Cockram J, White J, Zuluaga DL, Smith D, Comadran J, Macaulay M, Luo Z, Kearsey MJ, Werner P, Harrap D, Tapsell C, Liu H, Hedley PE, Stein N, Schulte D, Steuernagel B, Marshall DF, Thomas WTB, Ramsay L, Mackay I, Balding DJ, The AC, Waugh R, O'Sullivan DM (2010) Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. Proceedings of the National Academy of Sciences 107:21611-21616

Crainiceanu CM, Ruppert D (2004) Likelihood ratio tests in linear mixed models with one variance component. Journal of the Royal Statistical Society Series B-Statistical Methodology 66:165-185

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997-1004

Devlin B, Roeder K, Wasserman L (2001) Genomic Control, a New Approach to Genetic-Based Association Studies. Theoretical Population Biology 60:155-166

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6:e19379

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annual review of plant biology 54:357-374

Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J 44:1054 - 1064

Gilmour AR, Gogel BJ, Cullis BR, Welham SJ, Thompson R (2002) ASReml user guide release 1.0. . VSN International Ltd., Hemel Hempstead, UK

Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, Ware DH, Buckler ES (2009) A First-Generation Haplotype Map of Maize. Science 326:1115-1117

Hamblin MT, Mitchell SE, White GM, Gallego W, Kukatla R, Wing RA, Paterson AH, Kresovich S (2004) Comparative population genetics of the panicoid grasses: Sequence polymorphism, linkage disequilibrium and selection in a diverse sample of Sorghum bicolor. Genetics 167:471-483

Hardy OJ, Vekemans X (2002) spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes 2:618-620

Harjes CE, Rocheford TR, Bai L, Brutnell TP, Kandianis CB, Sowinski SG, Stapleton AE, Vallabhaneni R, Williams M, Wurtzel ET, Yan J, Buckler ES (2008) Natural Genetic Variation in Lycopene Epsilon Cyclase Tapped for Maize Biofortification. Science 319:330-333

Heffner EL, Sorrells ME, Jannink JL (2009) Genomic Selection for Crop Improvement. Crop Science 49:1-12

Henderson CR (1975) Comparison of Alternative Sire Evaluation Methods. Journal of Animal Science 41:760-770

Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, Guan J, Fan D, Weng Q, Huang T, Dong G, Sang T, Han B (2009a) High-throughput genotyping by whole-genome resequencing. Genome Research 19:1068-1076

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang Q-F, Li J, Han B (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961-967

Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, Li W, Guo Y, Deng L, Zhu C, Fan D, Lu Y, Weng Q, Liu K, Zhou T, Jing Y, Si L, Dong G, Huang T, Lu T, Feng Q, Qian Q, Li J, Han B (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet 44:32-39

Huang XH, Feng Q, Qian Q, Zhao Q, Wang L, Wang AH, Guan JP, Fan DL, Weng QJ, Huang T, Dong GJ, Sang T, Han B (2009b) High-throughput genotyping by whole-genome resequencing. Genome Research 19:1068-1076

Hyten DL, Cannon SB, Song Q, Weeks N, Fickus EW, Shoemaker RC, Specht JE, Farmer AD, May GD, Cregan PB (2010) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC genomics 11:38

Hyten DL, Choi I-Y, Song Q, Shoemaker RC, Nelson RL, Costa JM, Specht JE, Cregan PB (2007) Highly Variable Patterns of Linkage Disequilibrium in Multiple Soybean Populations. Genetics 175:1937-1944

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-y, Freimer NB, Sabatti C, Eskin E (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42:348-354

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient Control of Population Structure in Model Organism Association Mapping. Genetics 178:1709-1723

Kariya T, Kurata H (2004) Generalized Least Squares Estimators. Generalized Least Squares. John Wiley & Sons, Ltd, pp 25-66

Konishi S, Izawa T, Lin SY, Ebana K, Fukuta Y, Sasaki T, Yano M (2006) An SNP Caused Loss of Seed Shattering During Rice Domestication. Science 312:1392-1396

Kump KL, Bradbury PJ, Wisser RJ, Buckler ES, Belcher AR, Oropeza-Rosas MA, Zwonitzer JC, Kresovich S, McMullen MD, Ware D, Balint-Kurti PJ, Holland JB (2011) Genome-

wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat Genet 43:163-168

Laird PW (2003) The power and the promise of DNA methylation markers. Nat Rev Cancer 3:253-266

Liu KJ, Goodman M, Muse S, Smith JS, Buckler E, Doebley J (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165:2117-2128

Lu H-Y, Liu X-F, Wei S-P, Zhang Y-M (2011) Epistatic Association Mapping in Homozygous Crop Cultivars. PLoS ONE 6:e17773

Lukens LN, Zhan S (2007) The plant genome's methylation status and response to stress: implications for plant improvement. Current Opinion in Plant Biology 10:317-322

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747-753

McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, Brown P, Browne C, Eller M, Guill K, Harjes C, Kroon D, Lepak N, Mitchell SE, Peterson B, Pressoir G, Romero S, Oropeza Rosas M, Salvo S, Yates H, Hanson M, Jones E, Smith S, Glaubitz JC, Goodman M, Ware D (2009) Genetic properties of the maize nested association mapping population. Science 325:737 - 740

McNally KL, Bruskiewich R, Mackill D, Buell CR, Leach JE, Leung H (2006) Sequencing multiple and diverse rice varieties. Connecting whole-genome variation with phenotypes. Plant Physiol 141:26-31

McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, Zeller G, Clark RM, Hoen DR, Bureau TE, Stokowski R, Ballinger DG, Frazer KA, Cox DR, Padhukasahasram B, Bustamante CD, Weigel D, Mackill DJ, Bruskiewich RM, RÃtsch G, Buell CR, Leung H, Leach JE (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. Proceedings of the National Academy of Sciences 106:12273-12278

Metzker ML (2010) Applications of Next-Generation Sequencing Sequencing Technologies - the Next Generation. Nature Reviews Genetics 11:31-46

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157:1819-1829

Murray SC, Rooney WL, Hamblin MT, Mitchell SE, Kresovich S (2009) Sweet Sorghum Genetic Diversity and Association Mapping for Brix and Height. Plant Gen 2:48-62

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES (2009) Association Mapping: Critical Considerations Shift from Genotyping to Experimental Design. The Plant Cell Online 21:2194-2202

Nelson JC, Wang S, Wu Y, Li X, Antony G, White FF, Yu J (2011) Single-nucleotide polymorphism discovery by high-throughput sequencing in sorghum. BMC genomics 12:352

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272-276

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg NA, Shah C, Wall JD, Wang J, Zhao K, Kalbfleisch T, Schulz V, Kreitman M, Bergelson J (2005) The Pattern of Polymorphism in *Arabidopsis thaliana*. PLoS Biol 3:e196

Ott J, Kamatani Y, Lathrop M (2011) Family-based designs for genome-wide association studies. Nat Rev Genet advance online publication

Palaisa KA, Morgante M, Williams M, Rafalski A (2003) Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. Plant Cell 15:1795-1806

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, et al (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457:551-556

Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis. PLoS Genet 2:e190

Platt A, Horton M, Huang YS, Li Y, Anastasio AE, Mulyati NW, Ã…gren J, Bossdorf O, Byers D, Donohue K, Dunning M, Holub EB, Hudson A, Le Corre Vr, Loudet O, Roux F,

Warthmann N, Weigel D, Rivero L, Scholl R, Nordborg M, Bergelson J, Borevitz JO (2010) The Scale of Population Structure in *Arabidopsis thaliana*. PLoS Genet 6:e1000843

Poland JA, Bradbury PJ, Buckler ES, Nelson RJ (2010) Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. Proceedings of the National Academy of Sciences

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904-909

Pritchard JK, Stephens M, Donnelly P (2000a) Inference of Population Structure Using Multilocus Genotype Data. Genetics 155:945-959

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association Mapping in Structured Populations. The American Journal of Human Genetics 67:170-181

Ramsay L, Comadran J, Druka A, Marshall DF, Thomas WTB, Macaulay M, MacKenzie K, Simpson C, Fuller J, Bonar N, Hayes PM, Lundqvist U, Franckowiak JD, Close TJ, Muehlbauer GJ, Waugh R (2011) INTERMEDIUM-C, a modifier of lateral spikelet fertility in barley, is an ortholog of the maize domestication gene TEOSINTE BRANCHED 1. Nat Genet 43:169-172

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doeblay J, Kresovich S, Goodman MM, Buckler ES (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. P Natl Acad Sci USA 98:11479-11484

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516-1517

Rogers SO, Bendich AJ (1987) Ribosomal RNA genes in plants: variability in copy number and in the intergenic spacer. Plant Molecular Biology 9:509-520

Schmutz J, Cannon SB, Schlueter J, Ma J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178-183

Schnable PS, Ware D, Fulton RS, Stein JC, et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112-1115

Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jorgensen J-E, Weigel D, Andersen SU (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat Meth 6:550-551

Schulte D, Close TJ, Graner A, Langridge P, Matsumoto T, Muehlbauer G, Sato K, Schulman AH, Waugh R, Wise RP, Stein N (2009) The International Barley Sequencing Consortium-At the Threshold of Efficient Access to the Barley Genome. Plant Physiol 149:142-147

Sequencing ProjectInternational Rice G (2005) The map-based sequence of the rice genome. Nature 436:793-800

Shomura A, Izawa T, Ebana K, Ebitani T, Kanegae H, Konishi S, Yano M (2008) Deletion in a gene associated with grain size increased yields during rice domestication. Nat Genet 40:1023-1028

Sorrells ME, Yu J (2009) Linkage disequilibrium and association mapping in the Triticeae. In: Feuillet C, J. MG (eds) Genetics and Genomics of the Triticeae, Plant Genetics/Genomics. Springer Verlag, pp 655-684

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Iniguez AL, Barbazuk WB, Jeddeloh JA, Nettleton D, Schnable PS (2009) Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. PLoS Genet 5:e1000734

Tester M, Langridge P (2010) Breeding Technologies to Increase Crop Production in a Changing World. Science 327:818-822

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet 28:286-289

Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat Genet 43:159-162

Wang M, Zhu C, Barkley N, Chen Z, Erpelding J, Murray S, Tuinstra M, Tesso T, Pederson G, Yu J (2009a) Genetic diversity and population structure analysis of accessions in the US historic sweet sorghum collection. TAG Theoretical and Applied Genetics 120:13-23

Wang Z, Gerstein M, Snyder M (2009b) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57-63

Weber A, Clark RM, Vaughn L, de Jesús Sánchez-Gonzalez J, Yu J, Yandell BS, Bradbury P, Doebley J (2007) Major Regulatory Genes in Maize Contribute to Standing Variation in Teosinte (Zea mays ssp. parviglumis). Genetics 177:2349-2359

Weigel D, Mott R (2009) The 1001 Genomes Project for Arabidopsis thaliana. Genome Biology 10

Wisser RJ, Kolkman JM, Patzoldt ME, Holland JB, Yu J, Krakowsky M, Nelson RJ, Balint-Kurti PJ (2011) Multivariate analysis of maize disease resistances suggests a pleiotropic genetic basis and implicates a GST gene. Proceedings of the National Academy of Sciences 108:7339-7344

Yan J, Kandianis CB, Harjes CE, Bai L, Kim E-H, Yang X, Skinner DJ, Fu Z, Mitchell S, Li Q, Fernandez MGS, Zaharieva M, Babu R, Fu Y, Palacios N, Li J, DellaPenna D, Brutnell T, Buckler ES, Warburton ML, Rocheford T (2010) Rare genetic variation at Zea mays crtRB1 increases [beta]-carotene in maize grain. Nat Genet 42:322-327

Yu J, Arbelbide M, Bernardo R (2005) Power of in silico QTL mapping from phenotypic, pedigree, and marker data in a hybrid breeding program. Theor Appl Genet 110:1061-1067

Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Current Opinion in Biotechnology 17:155-160

Yu J, Hamblin MT, Tuinstra MR (2012) Association Genetics Strategies and Resources. In: Paterson A (ed) Genetics and Genomics of the Saccharinae. Springer Verlag (*in press*)

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic Design and Statistical Power of Nested Association Mapping in Maize. Genetics 178:539-551

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203-208

Zhang D, Bai G, Zhu C, Yu J, Carver BF (2010a) Genetic Diversity, Population Structure, and Linkage Disequilibrium in U.S. Elite Winter Wheat. Plant Gen 3:117-127

Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES (2010b) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42:355-360

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M (2007) An Arabidopsis example of association mapping in structured samples. PLoS Genet 3:e4

Zhao K, Tung C-W, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nat Commun 2:467

Zhao K, Wright M, Kimball J, Eizenga G, McClung A, Kovach M, Tyagi W, Ali ML, Tung C-W, Reynolds A, Bustamante CD, McCouch SR (2010) Genomic Diversity and Introgression in O. sativa Reveal the Impact of Domestication and Breeding on the Rice Genome. PLoS ONE 5:e10780

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and Prospects of Association Mapping in Plants. Plant Gen 1:5-20

Zhu C, Li X, Yu J (2011) Integrating Rare-Variant Testing, Function Prediction, and Gene Network in Composite Resequencing-Based Genome-Wide Association Studies (CR-GWAS). G3: Genes, Genomes, Genetics 1:233-243

Zhu C, Yu J (2009) Nonmetric Multidimensional Scaling Corrects for Population Structure in Association Mapping With Different Sample Types. Genetics 182:875-888

CHAPTER 2- ASSOCIATION MAPPING FOR GRAIN QUALITY IN SORGHUM

Abstract

Knowledge of the genetic bases of grain quality traits will complement plant breeding efforts to improve the end use value of sorghum (*Sorghum bicolor* (L.) Moench). Candidate gene association mapping was employed on a diverse panel of 300 sorghum accessions to assess marker-trait associations for 10 grain quality traits, measured using the single kernel characterization system (SKCS) and near-infrared reflectance spectroscopy (NIRS). The analysis of the accessions through 1,290 genome-wide SNPs (single nucleotide polymorphisms) separated the panel into five subpopulations that corresponded to three major sorghum races (durra, kafir, and caudatum), one intermediate race (guinea-caudatum), and one working group (zerazera/caudatum). These subpopulations differed in kernel hardness, acid detergent fiber, and total digestible nutrients. After model testing, association analysis between 333 SNPs in candidate genes/loci and grain quality traits resulted in eight significant marker-trait associations. A SNP in starch synthase *IIa* (*SSIIa*) gene was associated with kernel hardness (KH) with a likelihood ratio–based R^2 (R_{LR}^2) value of 0.08, and a SNP in starch synthase (*SSIIb*) gene was associated with starch content with an R_{LR}^2 value of 0.10, and a SNP in loci *pSB1120* was associated with starch content with an R_{LR}^2 value of 0.09.

Introduction

Sorghum (*Sorghum bicolor* (L.) Moench) is an important cereal crop used as human food in the semi-arid tropics of African and Asian continents by ~500 million people (de Wet, 1978). It is a gluten-free cereal used as whole grain as well as ground flour and it is a source of energy, protein, vitamins, minerals, and nutraceuticals such as antioxidant phenolics and cholesterol-lowering waxes (Taylor et al., 2006). Grain quality varies among different types of sorghum and their cultivated environments. Genetic improvement of grain quality can help sorghum to adapt to varying demands for end-use products.

Grain quality is differentiated by biochemical and physical characteristics in sorghum. Kernel hardness (KH) affects grain mold resistance (Jambunathan et al., 1992), grain storage ability, insect resistance (Bueso et al., 2000), milling behavior (Suhendro et al., 2000), flour particle size, cooking properties (Anglani, 1998; Bettge et al., 2000), and parameters such as adhesion, cooked grain texture, alkali gel stiffness (Cagampang and Kirleis, 1984), porridge quality (Akingbala and Rooney, 1987), and production of high-quality couscous granules (Aboubacar and Hamaker, 1999). Sorghum kernels are round and small in size and vary from about 3–4 mm in diameter. Variation in kernel diameter (KD) exists among cultivars (Wills and Ali, 1983). Large sorghum kernels with corneous endosperm are usually preferred for human consumption and are associated with desirable physical and chemical quality parameters such as high protein concentration, low ash, high milling yields, high water absorbance flour, bright white color, and large particle size (Lee et al., 2002). Small-kernel sorghum that is more likely to be harder and more difficult to mill, is not popular in the grain market (Wills and Ali, 1983). Kernel weight (KW) contributes to grain yield, and its components (kernel moisture content and kernel density) are correlated with milling value. Sorghum grain contains higher levels of acid detergent fiber (ADF) than yellow corn (*Zea mays* L.), and high-tannin varieties contain higher amounts of ADF than non-tannin sorghum (Douglas et al., 1990). Chemical quality parameters such as crude protein (CP), fat (F), phosphorous (P), starch, and total digestible nutrients (TDN) directly influence sorghum nutritional value. Starch content in sorghum kernels affects the consistency of thick porridge, cooked couscous firmness, and rollability (Beta et al., 2001).

Genetic mapping of grain quality traits has been conducted in different cereal crops such as maize (*Zea mays* L.) (Cook et al., 2012; Wilson et al., 2004), rice (*Oryza sativa* L.) (Tian et al., 2009), wheat (*Triticum aestivum* L.) (Bordes et al., 2011; Reif et al., 2011), and sorghum (de

Alencar Figueiredo et al., 2010). Starch is one of the most important grain quality parameters in cereals that provide the basis of subsistence for world population. Four enzymes, Adinosine diphosphate glucose pyrophosphorylase (ADGPP), starch synthase (SS), starch branching enzyme (SBE), and starch debranching enzyme (DBE) catalyze starch biosynthesis in cereals (Preiss and Sivak, 1998). Given that the pathways and enzymes related to grain quality are similar in cereals, it was not unexpected that genes identified through earlier mutational studies are similar to the genes identified by association studies at the population level.

A community resource, sorghum diversity panel was recently established by a collection of sorghum accessions representing all major cultivated races, including lines from the Sorghum Conversion Program (SCP), elite breeding lines, and their progenitors from all around the United States (Casa et al., 2008). The SCP converted tropical lines to photoperiod-insensitive short plants (Stephens et al., 1967). The level of population structure and familial relatedness in this diversity panel was previously assessed using 47 simple sequence repeat (SSR) markers (Casa et al., 2008). Another study analyzed 216 SCP lines using 434 single nucleotide polymorphisms (SNPs) and classified the lines into 4 subpopulations that corresponded closely to four major races. A combined analysis of the breeding lines and lines from the SCP program has not been conducted with a large number of markers; furthermore, genetic mapping studies will complement breeders' efforts to improve grain quality in sorghum. The present research was undertaken to identify marker-trait associations for grain quality traits in sorghum.

Materials and Methods

Plant Germplasm and Phenotypic Characterization

Three hundred lines from the sorghum diversity panel, including 251 SCP lines, and 49 important breeding lines and their progenitors from the U.S. served as the genetic material for this study. The sorghum accessions were planted with randomized complete block design in Manhattan, KS, and West Lafayette, IN, with two replications in 2007 and 2008. Seeds harvested from 10 selfed sorghum heads were analyzed for grain quality using the single kernel characterization system (SKCS) (Martin et al., 1993) and near-infrared reflectance spectroscopy (NIRS) (Pasquini, 2003). SKCS provided data on KH, KD, and KW, and NIRS provided data on ADF, C, CP, F, P, S, and TDN.

Single Kernel Characterization System

The SKCS was the device used to measure physical properties of sorghum kernels such as KH and size characteristics (Bean, 2006; Pedersen, 1996). Seeds of 287 lines from two years and two replications were analyzed through SKCS. Three hundred individual grains were crushed between a serrated rotor and a crescent, and parameters for KH, KD, and KW were estimated and reported. KH was reported as kernel hardness index.

Near-Infrared Reflectance Spectroscopy

NIRS utilizes the near-infrared region of the electromagnetic spectrum (about 800 nm–2500 nm) to determine the concentration of physical and chemical constituents in agricultural materials (Pasquini, 2003). NIRS was used to predict the amount of ADF, CP, F, C, P, S, and TDN in sorghum kernel. A total of 15g of seed were ground in a UDY cyclone mill (UDY Corporation, Fort Collins, CO) with a 1-mm screen, a stainless steel grinding ring, and an aluminum impeller. Two hundred and sixty-nine lines from 2 replications in Manhattan (2007) were scanned using a Foss NIRSystem 6500 monochromator. High R^2 values were obtained for various traits using a validation set of 52 samples. The R^2 for S, CP, F, ADF, and P contents were 0.99, 0.98, 0.91, 0.88, and 0.88, respectively. On the basis of the statistical parameters mentioned above, NIRS was demonstrated to be efficient and accurate in predicting chemical grain quality traits in this panel.

Genotyping and Candidate Genes

Two different genotyping assays were conducted: 1) a genome-wide assay of 1,536 SNPs (Yu, 2011), and 2) candidate gene/loci assay of 384 SNPs (Murray, 2009). The 1,536 SNPs assay was designed to achieve maximum genome coverage. The average distance between SNPs was 400 kb except in the centromere regions. The 384 SNPs assay was developed from SNPs discovered in previously published studies (Hamblin, 2004; Hamblin, 2005; Hamblin, 2006; Hamblin, 2007; Murray, 2009), starch pathways (Hamblin et al., 2007), sucrose pathways (Murray, 2009), and carotenoid pathways (Salas Fernandez et al., 2009). Out of 226 loci represented in the 384 SNPs assay, 39 loci were candidate genes from starch, sucrose, and carotenoid pathways and the remainder were candidate loci distributed across 10 chromosomes. An Illumina GoldenGate assay was used to genotype the samples. Out of the 1,536 SNPs assay, 1,290 SNPs and out of the

384 SNPs assay, 333 SNPs were successful and polymorphic. The program fastPHASE was used to impute missing data (Scheet and Stephens, 2006).

Statistical Analysis

DNA Marker Profile

PowerMarker version 3.25 (Liu and Muse, 2005) was used to calculate Chord distance (Cavalli-Sforza and Edwards, 1967) among accessions. It was also used to compute molecular diversity statistics and to construct the Neighbor Joining (NJ) tree with 100 replications of bootstrapping.

Population Structure Analysis

The program STRUCTURE, version 2.2.3 (Pritchard, 2000), was used to detect population structure and assign individuals to subpopulations. The STRUCTURE program was run 10 times for each subpopulation (*k)* value, ranging from 1–15, using the admixture model with 20,000 replicates for burn-in and 20,000 replicates during analysis. The final subpopulations were determined on the basis of 1) likelihood plot of models, 2) stability of grouping patterns across 10 runs, 3) germplasm information or "breeder's knowledge," 4) cluster analysis (NJ tree), and 5) principal component analysis (PCA). On the basis of this information, we chose $k = 5$ as the optimal grouping. Out of the 10 runs for $k = 5$, the run with the highest likelihood value was selected to assign the posterior membership coefficients (Q) to each accession (**Supplementary Table 1**). A graphical bar plot was then generated with the posterior membership coefficients (**Fig. 1A**), and plots were also plotted for $k = 2$, 3, 4, and 5 for result interpretation.

To validate the genetic structure and to test marker-trait associations, PCA and nonmetric multidimensional scaling (nMDS) were conducted and K matrix was calculated. PCA was conducted to construct a plot of the most significant axes for grouping pattern variation and to obtain axes for further model testing and association mapping (Patterson, 2006; Price, 2006; Zhu, 2009). The combined display of the color-coded subpopulation memberships from STRUCTURE with other analyses, NJ tree (**Fig. 1B**), and PCA (**Fig. 2A**), are shown. Kinship (K) was calculated with SPAGeDi 1.3 (Loiselle et al., 1995; Hardy and Vekemans, 2002).

Model Comparison and Association Analysis

We compared different models to assess the effect of population structure on association mapping of various grain quality traits measured in this diversity panel. Following the previously recommended procedures (Yu, 2006; Zhu, 2009), we tested various mixed models with subpopulation membership percentage (Q), nMDS, and PCA as fixed covariates, and kinship as random effect. The dimension of PCA and nMDS were determined for each trait individually. Among all possible models (the simple model, Q, K, PCA, nMDS (Zhu, 2009), Q+K (Yu, 2006), PCA+K, and nMDS+K), the best fit model was determined for each trait based on the Bayesian Information Criterion (BIC). The selected models were then used to test marker-trait associations between 333 SNPs and 10 grain quality traits. Marker-trait associations were tested in TASSEL (Bradbury, 2007) and were also verified in SAS (SAS institute, 1999). Subsequently, quantile-quantile (Q-Q) plots of the *F-test* statistics for the SNP markers were plotted to assess the adequacy of the best model in controlling type I errors (**Fig. 2.5**). SNPs that passed the threshold of p-value < 10^{-03} were deemed significant if minor allele frequency (MAF) was greater than 5%. In addition, likelihood-ratio-based R^2 was calculated for significant SNPs to provide a general measure for the effect of SNPs in mixed-model association mapping of the traits (Sun et al., 2010). R_{LR}^2 is a generalized form of R^2 in linear regression model that allows comparisons across models with different random and fixed components.

Results

Population Structure and Genetic Diversity

From the SNP data, the STRUCTURE analysis revealed five subpopulations (G1, G2, G3, G4, and G5) that contained 49, 46, 52, 49, and 69 accessions, respectively (**Fig. 1A**). The NJ tree analysis also clustered the data into five branches (B) (**Fig. 1B**). The color-coded branches support the five subpopulation classification. Each subpopulation closely corresponded to durra, kafir, zerazera/caudatum, guinea, and caudatum (**Fig. 1C**). G1 mainly consists of accessions from the race durra (79.6%), G2 comprises kafir (91.3%), G3 consists of the zerazera-caudatum working group (75%), G4 comprises the guinea-caudatum intermediate race (61.2%), and G5 consists of the caudatum race (63.8%). The genetic group guinea-caudatum is the race guinea in traditional classification. We used information from two earlier studies about the sorghum

diversity panel to classify the accessions into genetic groups and races (Brown, 2011; Casa, 2008).

The results from PCA showed that PC1 explains 11.6% variation in the data by separating G1 from G2, G3, G4, and G5, and PC2 explains 6.9% variation in the data by separating G4 from G1, G2, G3, and G5 (**Fig. 2A**). The PCA was color-coded based on the structure results and it generally agrees with STRUCTURE classification of five subpopulations. Even though bicolor is a major race of sorghum, it did not form a specific subpopulation in this diversity panel. We also generated a world map of the accessions based on their sources/origins (**Fig. 2B**). Results from STRUCTURE analysis, NJ tree, and PCA were consistent. Taken together, the sorghum diversity panel was classified into five subpopulations: three main sorghum races (durra, kafir, and caudatum), one intermediate race (guinea-caudatum), and a working group (zerazera-caudatum).

Trait Variation

Data analysis showed a high amount of diversity for grain quality traits. KH, KW, and KD from SKCS showed high consistency across years and environments that were recorded from two locations for two years. The repeatability of KH, KW, and KD were 0.79, 0.84, and 0.78, respectively. The correlation coefficients (r) were calculated for all traits. KH was significantly correlated with all traits except starch. KW and KD were positively correlated (r = 0.91). Protein content and P content were positively correlated (r = 0.75). KD and F were negatively correlated (r= -.24). CP was significantly associated with all traits except KW, KD, and F. Starch content was negatively correlated with ADF (r = -.68), Ca (r = -.31), CP, (r = -.75), F (r = -.25), and P (r = -.68) (**Table 1**).

In addition, KH, ADF, and TDN showed significant differences among the five subpopulations (**Fig. 3**). Caudatum in G5 had the lowest KH (**Fig. 3A**) and TDN (**Fig. 3C**) and the highest ADF (**Fig. 3B**) values. The accessions that formed G3, zerazera-caudatum, had the highest KH and TDN values, followed by guinea-caudatum (G4). Durra (G1) and kafir (G2) accessions had higher KH values than the caudatum (G5), but lower values than zerazera-caudatum (G3) and guinea-caudatum (G4). Caudatum in G5 were significantly different from other subpopulations for these three traits. Other phenotypic traits were not significantly different among the subpopulations.

Marker-Trait Association Analysis

The model comparisons revealed that the mixed model with K matrix was the best model for eight phenotypic traits: KH, KW, KD, ADF, CP, F, starch, and TDN. The simple model was the best model for testing Ca and nMDS2 for P. The intersection of phenotypic data (300 accessions) and genotypic data (265 accessions) yielded combined data set of 200 accessions with both genotypic and phenotypic data. Eight significant marker-trait associations between the SNPs on the candidate genes and grain quality traits were detected after filtering for MAF of 0.05 (**Table 2**). The Q-Q plots for each phenotype showed that the model tested were effective in controlling type I error (**Fig. 2.5**). SNPs associated with grain quality traits were checked for the distribution of alleles among subpopulations (**Fig. 2.6**).

Data analysis revealed three significant SNPs associated with KH: *SB00214.1*, *SB00214.2*, and *SB00116.3* have with *p*-values of 1.84 x 10^{-04}, 1.84 x 10^{-04}, and 7.94 x 10^{-04}, respectively. The consistency of association between significant SNPs and alleles with the trait was checked by plotting the number of alleles in the accessions among five subpopulations (**Fig. 4)**. Accessions with allele *A* in the SNP *SB00214.1* and the accessions with allele *T* in the SNP *SB00214.2* had significantly higher KH values in all subpopulations except in G1 (**Fig. 4**). *SB00214.1* and *SB00214.2* were located in the locus *pSB1700,* and *SB00116.3* was in starch synthase IIa (*SSIIa*) gene. The values of R_{LR}^2 for these SNPs were 8–10%. Except in G2 and G4, accessions with allele *A* in the SNP *SB00116.3* had higher KH values.

Ca, P, and starch had significant SNPs associated with it (**Table 2**). Two significant SNPs were associated with C. *SB00156.1* and *SB00054.1* was significantly associated with Ca; with *p*-values 5.36 x 10^{-04} and 9.04 x 10^{-04}, respectively, each explained 6% of the variation. Except in G4, accessions with allele *A* in the SNP *SB00156.1* had higher Ca content values. In the SNP *SB00054.1*, accessions with allele *C* had higher Ca content values except in G2. *SB00068.1* in candidate loci *pSB0140* in chromosome 6 was significantly associated with P content with a *p*-value of 5.83 x 10^{-04}. This SNP explained 5% of the variation in P. Accessions with allele *G* had higher P content values in all subpopulations, and the allele *A* was fixed in G1, G2, and G3. Two significant SNPs were associated with starch. *SB00115.3* in candidate gene *SSIIb* was associated with a *p*-value of 3.67 x 10^{-04}, and *SB00086.1* in *pSB1120* was associated with a *p*-value of 6.19 x 10^{-04}. The SNP in *SSIIb* explained 10% of the variation and the SNP *SB00086.1* explained 9% of the variation in starch. Accessions with allele *A* in the SNP *SB00115.3* had higher starch

content values in G1 and G5. The accession with allele *C* in the SNP *SB00086.1* had higher starch content values except in G3 (**Fig. 4**).

Discussion

Diversity and Classification in Sorghum

Sorghum is considered to have been domesticated around 5,000–7,000 years ago in the northeastern part of Africa, the present-day Ethiopia (Jennings and Cock, 1977). Earlier efforts to classify sorghum were mainly based on color of grains and glumes, presence/absence of awn, and stem characteristics. The most complete classification of sorghum was in the early part of the last century (Snowden, 1936). In 1972, another classification based on the spikelet characteristics was proposed (Harlan and Dewet, 1972), and cultivated sorghum was mainly classified into five major races. According to that system, bicolor and guinea races have open panicles, kafir and durra races have compact heads, and caudatum spikelets vary in their head type. Broomcorn generally falls in bicolor type, and feterita is considered to be of caudatum type (Harlan and Dewet, 1972). Notably, no barrier between these sorghum races prevents them from crossing and mixing, so a considerable amount of variation within the five races results from admixture that separates them into about 15 mixed races and nearly 70 working groups (Murty, 1967). Sorghum races also vary in their geographical origin and adaptation (**Fig. 2B**). The race bicolor is grown almost everywhere in Africa and does not have a characteristic geographical distribution or ecological adaptation. Guineas have hard seeds and are resistant to insects and mold damage under wet conditions. They are grown in the high rainfall areas of West Africa.

We found that the genetic group guinea-caudatum forms a subgroup with high KH values compared with caudatum (**Fig. 3A**). This genetic group closely corresponds to the race guinea in traditional classification (Brown, 2011). The caudatum race is one of the most important one; almost all modern hybrid sorghums in the U.S. are caudatum or are mixed with caudatum. Caudatum has higher yielding ability, bright seed color, and good seed quality. This race is found mixed with other races of sorghum and the working group zerazera-caudatum had the highest KH characteristics (**Fig. 3A**). The race kafir is found in the southern part of Africa and India. Durra is a drought-tolerant race and is present in India and northern parts of Africa; it is also found mixed with guinea and caudatum.

Earlier efforts to support phenotype-based racial classification in sorghum with genotypic data were successful (Aldrich, 1992; Perumal, 2007; Folkertsma, 2005; Casa, 2008; Brown, 2011), but there is some disagreement between phenotype- and genotype-based racial classifications. One hypothesis is that, in phenotype-based racial classification of sorghum, the traits used (panicle and spikelet characters) are controlled by a limited number of genomic regions but, in genotype-based classification a large number of markers are used to classify sorghum races *i.e.* STRUCTURE classification is based on random markers distributed across the genome that can capture the genomic variation among sorghum races. However, in an earlier study, 216 SCP lines were classified into four genetic groups that closely corresponded to major traditional races (Brown, 2011) except bicolor and bicolor didn't form a separate subpopulation. Our study showed similar patterns; the race bicolor was present in G2–G5 but was not present in the subpopulation G1, which had mostly kafir. Although bicolor is considered the progenitor of all sorghum races (de Wet, 1978), parallel domestication and theories of multiple origin of the sorghum races remain valid. The race kafir might have originated from an early bicolor or a wild race *Sorghum verticilliflorum* (Smith and Frederiksen, 2000). Electrophoresis data from an earlier study suggest that the race kafir was different from the other four major races in protein patterns (Shechter and Wet, 1975).

A recent study of domestication of the shattering gene in cereals reported multiple *Shattering1 (Sh1)* alleles for domesticated races in sorghum, and the *Sh1* allele in Tx623 (an important breeding line) found in kafir and bicolor from south and east Africa is different from the alleles found in guinea and durras, which is different from caudatum (Lin et al., 2012). The race bicolor had multiple alleles of the *sh1* gene and didn't have a dominant *sh1* haplotype that indicates wide distribution of this race. The four major races of sorghum probably have multiple independent domestication events (Lin et al., 2012). However, in this sorghum diversity panel zerazera/caudatum formed a separate subpopulation. Our results indicated that this 300 line sorghum diverse collection was separated into five different subpopulations that closely corresponded to four major traditional races and a working group.

Association Analysis

We followed the unified mixed model approach to account for spurious associations that result from population structure and familial relatedness (Yu, 2006). In deciding the best model to test

marker-trait associations, we compared and tested different models for the best fit to the phenotypic data. Testing a mixed model with the K matrix in SAS is not a straightforward approach and may encounter convergence problems. The best-fit model (the lowest BIC model) was selected for testing markers for each trait; if each phenotype were not tested with the best model, directly fitting both Q and K may overcorrect population structure and familial relatedness for some traits and result in type II error (Zhu and Yu, 2009).

The SNPs that were significant with MAF < 0.05 were not reported and improved methods are needed, to address and identify the true positives (Zhu, 2011). The percentage of variation explained was calculated as R_{LR}^2 that is appropriate for mixed model–based association mapping. After controlling for population structure and admixture, we found eight SNPs on the candidate genes to be significantly associated with the grain quality traits. Candidate gene association mapping approach complements genome-wide association studies (GWAS) and traditional linkage mapping. By using a sufficient number of SNPs coupled with careful selection of candidate genes, this approach can establish the gene-trait relationship at the population level.

Marker-Trait Associations

A SNP on the candidate gene *SSIIa* located on chromosome 10 was associated with KH and explained 8% of the variation in the trait. Earlier studies on kernel hardness in sorghum, wheat, and rice suggest that starch content and the distribution of proteins and lipids on the surface of starch granules are important factors in determining grain hardness (Cagampang and Kirleis, 1984; Chen et al., 2012; Guzman et al., 2012; Morris et al., 1994; Yan et al., 2010). Sorghum grain hardness is related to the vitreousness of the grain and the vitreousness is related to amylose content which is a major component of starch (Cagampang and Kirleis, 1984). The maize homologue of *SSIIa* gene is *sugary2* (*su2*) gene that is in the starch synthesis pathway (Hamblin, 2007). Also, the gelatinization temperature in rice is genetically controlled by the *SSIIa* gene that is related to KH (Yan et al., 2010). So these evidences suggest that the gene *SSIIa* from the starch synthesis pathway plays an important role in KH.

SNPs associated with KH (*SB00214.1* and *SB00214.2*) explained 10% of the variation in the trait and they were located in the locus *pSB1700* on chromosome 3. A bioinformatics analysis of the locus revealed that *pSB1700* locus is similar to *sad1* protein in *Oryza sativa* and SUN4 domain protein in *Zea mays*. The translated nucleotide of *pSB1700* had 49% identity with the

SUN4 protein in maize. These proteins are present in the inner nuclear membrane in a cell and form a link between other proteins, nucleoskeleton, and cytoskeleton that is important in the structure and shape of a cell (Murphy et al., 2010). SNPs *SB00156.1* and *SB00054.1* associated with C content were located in chromosome 3 at 50 cM and 59 cM, respectively. *SB00156.1* was located in a locus, *pSB0289* that was predicted to produce serine/threonine-protein kinase. The function of SNP *SB00054.1*, in locus *PRC1044* in chromosome 6, is not known from NCBI searches. A SNP, *SB00068.1* in *pSB0140* locus in chromosome 6, explained 5% of the variation in P content. A SNP on the candidate locus *pSB1120* on chromosome 3 was significantly associated with S content and explained 9% of the variation in the trait. BLAST searches provided the function of this locus as a gene producing 3-ketoacyl-CoA synthase. These five SNPs-trait associations are novel associations.

A SNP on the starch synthase IIb (*SSIIb*) gene on chromosome 2 was found to be significantly associated with S content and explained 10% of the variation in the trait. The maize homologue of this gene is *SSIIb*. Starch synthase is an enzyme required for starch synthesis in the endosperm of cereals (Fujita et al., 2011). Candidate gene association mapping in maize (Wilson et al., 2004) and rice (Tian et al., 2009) suggest that the starch synthase is an important enzyme in determining starch content and quality in cereals. The genes from starch synthesis pathways form a regulatory network and influence grain quality parameters (Tian et al., 2009). In sweet wheat, in the absence of the granule bound starch synthase II, starch is not formed in its kernels (Shimbata et al., 2011). The findings about the trait differences among different subpopulations and the identified SNPs from the present study can be further exploited in improving grain quality in sorghum and related cereals.

Tables and Figures

Table 2.1. Mean, standard deviation, and correlation of grain quality traits across sorghum accessions. The number of accessions used was 247 for KH, KW, and KD, and 274 for ADF, Ca, CP, F, P, starch, and TDN.

			Correlation (*r*)									
Traits	Mean	*SD*	KH	KW	KD	ADF	Ca	CP	F	P	Starch	TDN
KH	78.15	18.97	-									
KW	24.40	5.20	-.32***	-								
KD	1.70	0.38	-.33***	.91***	-							
ADF	4.81	0.91	-.18**	-.07	-.08	-						
C	0.06	0.01	.23***	-.07	-.12	.21***	-					
CP	13.85	1.67	.18**	.13	.05	.43***	.31***	-				
F	3.25	0.41	.33***	-.22***	-.24***	.05	.52***	-.05	-			
P	0.45	0.05	.23***	.04	-.04	.28***	.33***	.75***	.20***	-		
S	69.26	2.44	-.09	.02	.10	-.68***	-.31***	-.75***	-.25***	-.68***	-	
TDN	84.84	1.40	-.16**	.07	.08	-.99***	-.21***	-.43***	-.05	-.28***	.67***	-

KH = kernel hardness, KW = kernel weight, KD = kernel diameter, ADF = acid detergent fiber, C = calcium, CP = crude protein, F = Fat, P = phosphorous, S = starch, and TDN = total digestible nutrients, *SD* = standard deviation, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.

Table 2.2 Significant SNPs in candidate genes associated with grain quality traits.

SNP	Locus/gene	Ch	Alleles	MAF	Trait	*P*-value (Best model)	*P*-value (Simple model)	R_{LR}^2 (SNP)	R_{LR}^2 (SNP + Best model)	Predicted gene function
SB00214.1	*pSB1700*	3	C/A	0.17	KH	1.84E-04	3.70E-03	0.10	0.53	Hypothetical protein
SB00214.2	*pSB1700*	3	A/T	0.17	KH	1.84E-04	3.70E-03	0.10	0.53	Hypothetical protein
SB00116.3	*SSIIa*	10	A/T	0.28	KH	7.94E-04	3.54E-03	0.08	0.51	Starch synthase IIa
SB00156.1	*pSB0289*	3	G/A	0.22	Ca	5.36E-04	5.36E-04	0.06	0.06	Serine/threonine-protein kinase
SB00054.1	*PRC1044*	3	C/G	0.36	Ca	9.04E-04	9.04E-04	0.06	0.06	Hypothetical protein
SB00068.1	*pSB0140*	6	A/G	0.07	P	5.83E-04	1.04E-03	0.05	0.07	Peptide transporter PTR2
SB00115.3	*SSIIb*	2	A/G	0.17	Starch	3.67E-04	1.18E-03	0.10	0.32	Starch synthase IIb
SB00086.1	*pSB1120*	3	C/A	0.31	Starch	6.19E-04	1.02E-03	0.09	0.31	3-ketoacyl-CoA synthase

Ch = Chromosome, MAF = minor allele frequency, KH = kernel hardness, Ca = calcium, P = phosphorous, R_{LR}^2 = Likelihood ratio-based R^2.

Table 2.3. Information on the sorghum diversity panel: 265 accessions used in the study and their likelihood values, subpopulation membership (Pop), traditional classification, classification by Casa et al. (2008), and origin.

No	Name	Q1	Q2	Q3	Q4	Q5	Pop	Traditional	Casa et al., 2008	Origin
1	01MN1589 B	0.146	0.750	0.040	0.005	0.059	2	bicolor	kafir	USA
2	88V1080	0.048	0.255	0.191	0.336	0.170	4	inbred line	caudatum	USA
3	AJABSIDO	0.229	0.001	0.001	0.065	0.704	5	bicolor	milo-feterita	Sudan
4	AR2002	0.198	0.015	0.172	0.223	0.392	5	bicolor	caudatum	USA
5	BAZ9504	0.132	0.778	0.000	0.086	0.004	2	breeding line	kafir	USA
6	BDLO357	0.009	0.346	0.626	0.009	0.010	3	breeding line	zerazera-caudatum	USA
7	B.OK11	0.000	0.999	0.000	0.000	0.001	2	kafir	kafir	USA
8	B.Tx2752	0.134	0.864	0.001	0.001	0.000	2	breeding line	kafir	USA
9	B.Tx2928	0.003	0.634	0.234	0.050	0.079	2	breeding line	kafir	USA
10	B.Tx3042	0.284	0.715	0.000	0.001	0.000	2	kafir	kafir	USA
11	B.Tx3197	0.000	0.999	0.000	0.000	0.001	2	kafir	kafir	USA
12	B.Tx378	0.026	0.973	0.000	0.000	0.001	2	kafir	kafir	USA
13	B.Tx399	0.248	0.750	0.001	0.001	0.000	2	kafir	kafir	USA
14	B.Tx406	0.324	0.674	0.000	0.001	0.001	2	kafir	kafir	USA
15	B.Tx615	0.000	0.998	0.001	0.000	0.001	2	breeding line	kafir	USA
16	B.Tx623	0.000	0.464	0.534	0.000	0.002	3	kafir	zerazera-caudatum	USA
17	B.Tx623(BMX)	0.001	0.425	0.573	0.000	0.001	3	kafir	zerazera-caudatum	USA
18	B.Tx626	0.002	0.439	0.559	0.001	0.000	3	kafir	zerazera-caudatum	USA
19	B.Tx631	0.001	0.702	0.297	0.000	0.000	2	kafir	kafir	USA
20	B.Tx635	0.002	0.413	0.195	0.037	0.353	5	kafir	kafir	USA
21	B.Tx641	0.120	0.498	0.203	0.096	0.083	2	breeding line	kafir	USA
22	B.Tx642	0.911	0.087	0.000	0.001	0.001	1	durra	durra	USA
23	B.Tx643	0.118	0.439	0.440	0.001	0.002	3	breeding line	kafir	USA
24	B.Tx645	0.080	0.501	0.416	0.002	0.001	2	breeding line	kafir	USA
25	CAPROCK	0.437	0.562	0.000	0.001	0.000	2	breeding line	milo-feterita	USA
26	CHILTEX	0.118	0.466	0.001	0.053	0.362	2	breeding line	milo-feterita	USA
27	COMBINE 7078	0.803	0.193	0.001	0.001	0.002	1	kafir	milo-feterita	USA

28	CS3541	0.003	0.230	0.461	0.001	0.305	3	cultivar	kafir	n/a
29	CSM63	0.005	0.001	0.004	0.979	0.011	4	cultivar	guinea-caudatum	Mali
30	DORADO	0.000	0.000	0.998	0.000	0.002	3	cultivar	zerazera-caudatum	n/a
31	EBA3	0.980	0.000	0.002	0.009	0.009	1	breeding line	durra	n/a
	FETERITA									
32	GISHESH	0.207	0.001	0.001	0.081	0.710	5	cultivar	milo-feterita	n/a
33	HEGARI	0.001	0.000	0.004	0.001	0.994	5	n/a	caudatum	n/a
34	ICSV400	0.000	0.000	0.996	0.000	0.004	3	cultivar	zerazera-caudatum	S. Africa
35	ICSV745	0.001	0.111	0.600	0.091	0.197	3	cultivar	zerazera-caudatum	S. Africa
36	IS 3620C	0.125	0.140	0.002	0.731	0.002	4	breeding line	guinea-bicolor	n/a
37	JOCORO	0.004	0.005	0.854	0.002	0.135	3	durra	zerazera-caudatum	India
38	KARPER	0.152	0.660	0.014	0.154	0.020	2	n/a	kafir	n/a
39	KAT83369	0.000	0.000	0.998	0.000	0.002	3	cultivar	zerazera-caudatum	n/a
40	KS115	0.497	0.001	0.056	0.250	0.196	1	caudatum-bicolor	caudatum	Uganda
41	KUYUMA	0.000	0.000	0.999	0.000	0.001	3	cultivar	zerazera-caudatum	n/a
42	LG70	0.088	0.250	0.442	0.194	0.026	3	cultivar	zerazera-caudatum	USA
43	LIANTANGAI	0.562	0.187	0.001	0.134	0.116	1	breeding line	durra	China
44	M35-1	0.998	0.000	0.001	0.001	0.000	1	cultivar	durra	India
45	MACIA	0.001	0.000	0.938	0.000	0.061	3	c zerazera	zerazera-caudatum	n/a
46	MALISOR 84-7	0.005	0.096	0.891	0.001	0.007	3	cultivar	zerazera-caudatum	Mali
47	MARTIN	0.128	0.871	0.000	0.000	0.001	2	n/a	kafir	n/a
48	MARUPANTSE	0.001	0.998	0.000	0.001	0.000	2	n/a	kafir	Botswana
49	MR732	0.001	0.006	0.886	0.104	0.003	3	cultivar	zerazera-caudatum	Niger
50	N263B	0.036	0.416	0.456	0.091	0.001	3	cultivar	zerazera-caudatum	USA
51	N268B	0.325	0.575	0.073	0.004	0.023	2	n/a	kafir	USA
52	NN 01	0.001	0.001	0.002	0.996	0.000	4	n/a	caudatum	n/a
53	P721	0.018	0.139	0.815	0.024	0.004	3	cultivar	zerazera-caudatum	USA
54	P850029	0.005	0.006	0.986	0.001	0.002	3	breeding line	zerazera-caudatum	USA
55	P898012	0.001	0.000	0.035	0.000	0.964	5	cultivar	caudatum	USA
56	P9517	0.004	0.570	0.292	0.100	0.034	2	n/a	kafir	USA
57	PINKKAFIR	0.001	0.998	0.000	0.000	0.001	2	kafir	kafir	USA
58	PINOLERO 1	0.060	0.095	0.842	0.001	0.002	3	cultivar	zerazera-caudatum	Nicaragua
59	PLAINSMAN	0.416	0.583	0.000	0.001	0.000	2	breeding line	milo-feterita	USA

60	QL3INDIA	0.010	0.983	0.002	0.006	0.000	2	n/a	kafir	India
61	R9188	0.231	0.227	0.321	0.221	0.000	3	cultivar	milo-feterita	n/a
62	REDBINE-60	0.223	0.775	0.001	0.000	0.001	2	breeding line	kafir	USA
63	REDKAFIR	0.001	0.999	0.000	0.000	0.000	2	breeding line	kafir	USA
64	R.TAM2566	0.082	0.000	0.737	0.001	0.180	3	breeding line	zerazera-caudatum	USA
65	R.TAM428	0.042	0.072	0.883	0.001	0.002	3	breeding line	zerazera-caudatum	USA
66	R.Tx2737	0.184	0.186	0.002	0.627	0.001	4	breeding line	guinea-caudatum	USA
67	R.Tx2783	0.182	0.128	0.648	0.017	0.025	3	breeding line	zerazera-caudatum	USA
68	R.Tx2903	0.329	0.231	0.387	0.053	0.000	3	inbred line	milo-feterita	USA
69	R.Tx2907	0.050	0.158	0.267	0.425	0.100	4	caudatum	caudatum	Sudan
70	R.Tx2917	0.365	0.188	0.359	0.087	0.001	3	inbred line	milo-feterita	USA
71	R.Tx430	0.050	0.077	0.222	0.649	0.002	4	breeding line	milo-feterita	USA
72	R.Tx431	0.474	0.149	0.356	0.004	0.017	1	breeding line	milo-feterita	USA
73	R.Tx432	0.001	0.370	0.627	0.001	0.001	3	breeding line	zerazera-caudatum	USA
74	R.Tx433	0.449	0.096	0.451	0.003	0.001	3	caudatum	zerazera-caudatum	Ethiopia
75	R.Tx434	0.399	0.023	0.576	0.001	0.001	3	breeding line	zerazera-caudatum	USA
76	R.Tx435	0.007	0.237	0.205	0.547	0.004	4	breeding line	guinea-caudatum	USA
77	R.Tx436	0.238	0.224	0.317	0.220	0.001	3	inbred line	milo-feterita	USA
78	R.Tx437	0.074	0.073	0.346	0.382	0.125	4	breeding line	milo-feterita	USA
79	SC1017	0.922	0.001	0.001	0.046	0.030	1	durra-bicolor	durra	Ethiopia
80	SC103	0.001	0.048	0.205	0.001	0.745	5	caudatum	caudatum	S. Africa
81	SC1033	0.951	0.002	0.001	0.028	0.018	1	durra-bicolor	durra	Ethiopia
82	SC1038	0.985	0.001	0.001	0.007	0.006	1	durra-bicolor	durra	Ethiopia
83	SC1047	0.998	0.001	0.000	0.001	0.000	1	durra	durra	Ethiopia
84	SC1055	0.012	0.081	0.400	0.003	0.504	5	caudatum	zerazera-caudatum	Sudan
85	SC1056	0.020	0.003	0.440	0.009	0.528	5	other	zerazera-caudatum	Sudan
86	SC1057	0.005	0.195	0.148	0.281	0.371	5	caudatum	milo-feterita	Uganda
87	SC1063	0.028	0.308	0.001	0.663	0.000	4	durra-bicolor	guinea-caudatum	Ethiopia
88	SC1070	0.001	0.474	0.001	0.523	0.001	4	durra	kafir	Ethiopia
89	SC1076	0.005	0.011	0.002	0.886	0.096	4	caudatum-bicolor	guinea-caudatum	Nigeria
90	SC1077	0.005	0.018	0.442	0.007	0.528	5	caudatum	zerazera-caudatum	Nigeria
91	SC108	0.001	0.000	0.998	0.000	0.001	3	caudatum	zerazera-caudatum	Ethiopia
92	SC1080	0.001	0.992	0.000	0.005	0.002	2	kafir	kafir	S. Africa

93	SC1085	0.862	0.001	0.024	0.063	0.050	1	durra	durra	India
94	SC110	0.001	0.000	0.918	0.000	0.081	3	caudatum	zerazera-caudatum	Ethiopia
95	SC1104	0.010	0.130	0.152	0.064	0.644	5	kafir-bicolor	caudatum	Uganda
96	SC1108	0.091	0.427	0.005	0.447	0.030	4	guinea	guinea-caudatum	India
97	SC115	0.031	0.002	0.321	0.104	0.542	5	caudatum-bicolor	guinea-caudatum	Uganda
98	SC1154	0.994	0.000	0.001	0.002	0.003	1	durra-bicolor	durra	Ethiopia
99	SC1155	0.970	0.001	0.001	0.028	0.000	1	durra	durra	Ethiopia
100	SC1158	0.819	0.074	0.001	0.105	0.001	1	durra-bicolor	durra	Ethiopia
101	SC118	0.079	0.000	0.145	0.001	0.775	5	caudatum	caudatum	Sudan
102	SC1201	0.159	0.531	0.045	0.164	0.101	2	guinea-caudatum	kafir	Other
103	SC1203	0.097	0.118	0.161	0.336	0.288	4	other	caudatum	Brazil
104	SC121	0.001	0.002	0.078	0.001	0.918	5	caudatum	caudatum	S. Africa
105	SC1211	0.270	0.184	0.006	0.230	0.310	5	kafir-caudatum	guinea-bicolor	C. America
106	SC1212	0.000	0.001	0.084	0.001	0.914	5	caudatum	caudatum	Venezuela
107	SC1214	0.044	0.083	0.264	0.005	0.604	5	guinea-caudatum	caudatum	Burkina Faso
108	SC124	0.692	0.001	0.001	0.223	0.083	1	durra-bicolor	guinea-bicolor	Ethiopia
109	SC1251	0.003	0.073	0.479	0.008	0.437	3	guinea	zerazera-caudatum	India
110	SC1271	0.002	0.001	0.960	0.000	0.037	3	caudatum	zerazera-caudatum	Ethiopia
111	SC1319	0.001	0.001	0.766	0.001	0.231	3	caudatum	zerazera-caudatum	Ethiopia
112	SC1320	0.001	0.001	0.770	0.001	0.227	3	caudatum	zerazera-caudatum	Ethiopia
113	SC1321	0.216	0.031	0.096	0.176	0.481	5	guinea-caudatum	guinea-caudatum	Sudan
114	SC1328	0.191	0.004	0.130	0.162	0.513	5	caudatum	caudatum	Sudan
115	SC1337	0.001	0.001	0.001	0.996	0.001	4	guinea	guinea-caudatum	Mali
116	SC1345	0.006	0.001	0.171	0.050	0.772	5	caudatum	caudatum	Mali
117	SC135	0.824	0.053	0.003	0.110	0.010	1	durra-bicolor	durra	Ethiopia
118	SC1356	0.106	0.069	0.115	0.145	0.565	5	caudatum	caudatum	Sudan
119	SC1416	0.855	0.001	0.004	0.139	0.001	1	durra-bicolor	durra	Niger
120	SC1424	0.126	0.136	0.008	0.678	0.052	4	kafir-durra	guinea-bicolor	Mali
121	SC1426	0.001	0.000	0.000	0.998	0.001	4	guinea	guinea-caudatum	Mali
122	SC1429	0.001	0.505	0.003	0.423	0.068	2	guinea	guinea-caudatum	Zimbabwe
123	SC144	0.547	0.001	0.084	0.309	0.059	1	durra-caudatum	milo-feterita	Ethiopia
124	SC1465	0.005	0.028	0.461	0.007	0.499	3	other	guinea-bicolor	Sudan

125	SC1476	0.097	0.069	0.132	0.054	0.648	5	other	caudatum	Sudan
126	SC1484	0.997	0.001	0.001	0.001	0.000	1	other	durra	Somalia
127	SC155	0.695	0.083	0.001	0.158	0.063	1	durra-bicolor	durra	Ethiopia
128	SC175	0.070	0.000	0.658	0.001	0.271	3	caudatum	zerazera-caudatum	Ethiopia
129	SC192	0.756	0.151	0.037	0.026	0.030	1	durra	durra	India
130	SC199	0.998	0.001	0.000	0.000	0.001	1	durra	durra	India
131	SC206	0.948	0.010	0.001	0.040	0.001	1	durra	durra	Ethiopia
132	SC209	0.884	0.109	0.001	0.002	0.004	1	durra	durra	India
133	SC21	0.802	0.023	0.001	0.173	0.001	1	kafir-bicolor	guinea-bicolor	Ethiopia
134	SC214	0.946	0.019	0.002	0.032	0.001	1	durra-bicolor	durra	India
135	SC22	0.996	0.001	0.001	0.001	0.001	1	durra	durra	Ethiopia
136	SC224	0.129	0.250	0.002	0.596	0.023	4	bicolor	guinea-bicolor	Ethiopia
137	SC240	0.998	0.001	0.000	0.000	0.001	1	durra	durra	India
138	SC25	0.935	0.059	0.000	0.005	0.001	1	durra	durra	Ethiopia
139	SC265	0.001	0.000	0.001	0.998	0.000	4	guinea	guinea-caudatum	West Volta
140	SC283	0.002	0.443	0.002	0.518	0.035	4	guinea	guinea-caudatum	Tanzania
141	SC295	0.005	0.085	0.001	0.908	0.001	4	guinea	guinea-caudatum	Nigeria
142	SC299	0.002	0.001	0.000	0.996	0.001	4	guinea	guinea-caudatum	Nigeria
143	SC303	0.139	0.126	0.001	0.729	0.005	4	caudatum	guinea-bicolor	Ethiopia
144	SC305	0.059	0.182	0.003	0.644	0.112	4	guinea-caudatum	guinea-bicolor	Chad
145	SC309	0.629	0.004	0.007	0.247	0.113	1	bicolor	guinea-bicolor	Sudan
146	SC319	0.003	0.001	0.260	0.005	0.731	5	caudatum-bicolor	caudatum	Uganda
147	SC320	0.144	0.148	0.006	0.247	0.455	5	kafir	caudatum	Chad
148	SC323	0.002	0.145	0.058	0.474	0.321	4	caudatum	guinea-caudatum	Sudan
149	SC325	0.032	0.351	0.003	0.277	0.337	5	caudatum	guinea-caudatum	USA
150	SC328	0.002	0.003	0.291	0.002	0.702	5	other	caudatum	Uganda
151	SC33	0.953	0.044	0.000	0.001	0.002	1	durra	durra	Ethiopia
152	SC333	0.099	0.006	0.001	0.155	0.739	5	durra	caudatum	Ethiopia
153	SC334	0.029	0.023	0.107	0.095	0.746	5	caudatum	guinea-caudatum	Sudan
154	SC348	0.027	0.001	0.001	0.971	0.000	4	caudatum	guinea-caudatum	Nigeria
155	SC35	0.998	0.000	0.000	0.001	0.001	1	durra	durra	Ethiopia
156	SC373	0.007	0.000	0.016	0.928	0.049	4	caudatum	guinea-caudatum	Nigeria
157	SC38	0.999	0.000	0.000	0.000	0.001	1	durra	durra	Ethiopia

158	SC391	0.001	0.001	0.000	0.997	0.001	4	caudatum	guinea-caudatum	Egypt
159	SC396	0.002	0.000	0.001	0.995	0.002	4	caudatum	guinea-caudatum	Nigeria
160	SC399	0.126	0.012	0.002	0.803	0.057	4	caudatum	guinea-caudatum	Nigeria
161	SC411	0.160	0.069	0.109	0.649	0.013	4	caudatum-bicolor	guinea-caudatum	Sudan
162	SC418	0.001	0.122	0.126	0.216	0.535	5	kafir-caudatum	guinea-caudatum	Tanzania
163	SC42	0.011	0.117	0.178	0.292	0.402	5	kafir	guinea-caudatum	S. Africa
164	SC420	0.023	0.001	0.148	0.120	0.708	5	kafir-caudatum	caudatum	Sudan
165	SC423	0.046	0.002	0.413	0.003	0.536	5	caudatum	zerazera-caudatum	Sudan
166	SC425	0.145	0.043	0.002	0.177	0.633	5	durra-caudatum	guinea-bicolor	Sudan
167	SC441	0.758	0.128	0.005	0.084	0.025	1	durra	durra	India
168	SC449	0.268	0.319	0.002	0.356	0.055	4	kafir-bicolor	guinea-caudatum	Uganda
169	SC465	0.003	0.368	0.002	0.502	0.125	4	guinea	guinea-caudatum	Arabia
170	SC467	0.807	0.189	0.001	0.002	0.001	1	durra-bicolor	durra	India
171	SC473	0.880	0.109	0.001	0.009	0.001	1	durra	durra	India
172	SC480	0.998	0.001	0.000	0.000	0.001	1	durra	durra	India
173	SC489	0.937	0.061	0.001	0.001	0.000	1	durra	durra	India
174	SC49	0.019	0.103	0.074	0.153	0.651	5	guinea-caudatum	caudatum	Sudan
175	SC498	0.998	0.000	0.000	0.000	0.002	1	durra	durra	India
176	SC500	0.998	0.001	0.000	0.001	0.000	1	durra	durra	India
177	SC502	0.217	0.048	0.237	0.044	0.454	5	durra-caudatum	caudatum	Sudan
178	SC51	0.297	0.021	0.032	0.211	0.439	5	caudatum	caudatum	Sudan
179	SC53	0.332	0.058	0.005	0.033	0.572	5	durra-caudatum	milo-feterita	Sudan
180	SC532	0.049	0.069	0.001	0.881	0.000	4	guinea	guinea-caudatum	West Volta
181	SC55	0.216	0.001	0.001	0.151	0.631	5	caudatum	milo-feterita	Sudan
182	SC557	0.141	0.144	0.036	0.423	0.256	4	durra	guinea-caudatum	Ethiopia
183	SC56	0.147	0.029	0.186	0.067	0.571	5	caudatum	guinea-caudatum	Sudan
184	SC562	0.163	0.001	0.003	0.147	0.686	5	caudatum	caudatum	Sudan
185	SC563	0.174	0.079	0.079	0.353	0.315	4	caudatum	caudatum	Nigeria
186	SC564	0.109	0.014	0.002	0.137	0.738	5	caudatum	caudatum	Uganda
187	SC566	0.002	0.005	0.011	0.972	0.010	4	caudatum	guinea-caudatum	Nigeria
188	SC566-14	0.002	0.005	0.004	0.979	0.010	4	caudatum	guinea-caudatum	Nigeria
189	SC569	0.003	0.000	0.001	0.995	0.001	4	caudatum	guinea-caudatum	Nigeria
190	SC57	0.218	0.001	0.046	0.103	0.632	5	caudatum	caudatum	Uganda

191	SC572	0.187	0.001	0.133	0.005	0.674	5	kafir-caudatum	caudatum	China
192	SC574	0.009	0.050	0.276	0.062	0.603	5	caudatum	caudatum	Pakistan
193	SC58	0.143	0.034	0.187	0.093	0.543	5	caudatum	caudatum	Sudan
194	SC587	0.998	0.001	0.000	0.001	0.000	1	durra	durra	India
195	SC59	0.012	0.249	0.010	0.364	0.365	5	caudatum-bicolor	guinea-caudatum	Sudan
196	SC598	0.786	0.116	0.059	0.036	0.003	1	other	milo-feterita	Uganda
197	SC599	0.013	0.311	0.330	0.124	0.222	3	caudatum	caudatum	USA
198	SC60	0.002	0.001	0.309	0.001	0.687	5	caudatum	caudatum	Sudan
199	SC603	0.029	0.786	0.000	0.183	0.002	2	guinea	kafir	Tanzania
200	SC605	0.120	0.144	0.004	0.704	0.028	4	guinea	guinea-bicolor	Kenya
201	SC606	0.159	0.305	0.097	0.093	0.346	5	guinea-bicolor	caudatum	China
202	SC609	0.061	0.741	0.001	0.196	0.001	2	bicolor	kafir	China
203	SC621	0.718	0.162	0.001	0.108	0.011	1	bicolor	durra	India
204	SC623	0.132	0.391	0.019	0.312	0.146	4	durra-bicolor	guinea-caudatum	Congo
205	SC627	0.023	0.928	0.014	0.027	0.008	2	caudatum	kafir	Nigeria
206	SC628	0.002	0.997	0.000	0.001	0.000	2	kafir	kafir	S. Africa
207	SC63	0.074	0.038	0.015	0.135	0.738	5	caudatum	guinea-caudatum	Sudan
208	SC630	0.001	0.998	0.000	0.000	0.001	2	kafir	kafir	Zambia
209	SC637	0.010	0.033	0.117	0.002	0.838	5	kafir-caudatum	caudatum	Uganda
210	SC639	0.005	0.281	0.056	0.004	0.654	5	kafir-caudatum	caudatum	India
211	SC64	0.011	0.001	0.545	0.009	0.434	3	kafir-caudatum	caudatum	Sudan
212	SC641	0.001	0.001	0.147	0.000	0.851	5	kafir-caudatum	caudatum	Uganda
213	SC645	0.001	0.083	0.142	0.001	0.773	5	kafir-caudatum	caudatum	Uganda
214	SC648	0.015	0.930	0.001	0.052	0.002	2	kafir-caudatum	kafir	S. Africa
215	SC655	0.003	0.036	0.144	0.016	0.801	5	caudatum	caudatum	Sudan
216	SC659	0.001	0.963	0.001	0.030	0.005	2	guinea-kafir	kafir	USA
217	SC663	0.001	0.997	0.000	0.001	0.001	2	guinea-kafir	kafir	USA
218	SC671	0.007	0.558	0.002	0.365	0.068	2	kafir-caudatum	kafir	Kenya
219	SC672	0.005	0.865	0.001	0.074	0.055	2	kafir-caudatum	kafir	Zimbabwe
220	SC673	0.001	0.997	0.000	0.001	0.001	2	kafir-caudatum	kafir	Zimbabwe
221	SC679	0.002	0.034	0.006	0.801	0.157	4	guinea-caudatum	caudatum	Sudan
222	SC701	0.066	0.005	0.248	0.017	0.664	5	caudatum	zerazera-caudatum	Sudan
223	SC702	0.237	0.013	0.073	0.015	0.662	5	caudatum	zerazera-caudatum	Sudan

224	SC704	0.002	0.006	0.003	0.001	0.988	5	caudatum	caudatum	Japan
225	SC708	0.012	0.006	0.218	0.004	0.760	5	caudatum	caudatum	Uganda
226	SC720	0.001	0.069	0.273	0.001	0.656	5	caudatum	caudatum	Kenya
227	SC725	0.060	0.059	0.192	0.028	0.661	5	caudatum	caudatum	Japan
228	SC734	0.014	0.982	0.001	0.002	0.001	2	caudatum	kafir	Sudan
229	SC738	0.110	0.110	0.146	0.069	0.565	5	caudatum	caudatum	Sudan
230	SC748	0.032	0.099	0.465	0.002	0.402	3	guinea-caudatum	zerazera-caudatum	Sudan
231	SC749	0.223	0.382	0.001	0.002	0.392	5	caudatum-bicolor	milo-feterita	Japan
232	SC757	0.003	0.969	0.001	0.005	0.022	2	kafir	kafir	Botswana
233	SC760	0.043	0.024	0.181	0.005	0.747	5	kafir-caudatum	caudatum	Sudan
234	SC782	0.061	0.779	0.000	0.159	0.001	2	caudatum	kafir	India
235	SC790	0.186	0.111	0.355	0.002	0.346	3	caudatum	bicolor	Sudan
236	SC798	0.001	0.007	0.542	0.002	0.448	3	durra	caudatum	Ethiopia
237	SC803	0.004	0.001	0.480	0.006	0.509	3	caudatum	zerazera-caudatum	Sudan
238	SC805	0.005	0.001	0.801	0.003	0.190	3	caudatum	caudatum	Uganda
239	SC833	0.934	0.060	0.002	0.003	0.001	1	durra	durra	India
240	SC855	0.547	0.067	0.003	0.281	0.102	1	durra	guinea-bicolor	Egypt
241	SC91	0.003	0.010	0.002	0.984	0.001	4	other	guinea-caudatum	Zimbabwe
242	SC910	0.618	0.122	0.012	0.245	0.003	1	Guinea-durra	durra	India
243	SC929	0.887	0.053	0.001	0.001	0.058	1	caudatum	durra	S. Africa
244	SC937	0.358	0.244	0.007	0.389	0.002	4	bicolor	sudanense-broomcorn	Sudan
245	SC941	0.215	0.260	0.006	0.287	0.232	4	bicolor	sudanense-broomcorn	Sudan
246	SC942	0.374	0.176	0.006	0.439	0.005	4	bicolor	sudanense-broomcorn	Sudan
247	SC949	0.143	0.662	0.001	0.159	0.035	2	guinea	kafir	Nigeria
248	SC964	0.001	0.000	0.292	0.000	0.707	5	caudatum	caudatum	Uganda
249	SC968	0.632	0.053	0.002	0.260	0.053	1	durra-bicolor	guinea-bicolor	Zimbabwe
250	SC970	0.006	0.002	0.163	0.099	0.730	5	other	caudatum	Uganda
251	SC971	0.003	0.313	0.003	0.532	0.149	4	kafir-durra	guinea-caudatum	USA
252	SC979	0.066	0.017	0.735	0.001	0.181	3	caudatum	zerazera-caudatum	Ethiopia
253	SC982	0.001	0.005	0.786	0.010	0.198	3	caudatum	zerazera-caudatum	Ethiopia

254	SC984	0.001	0.000	0.824	0.001	0.174	3	caudatum	zerazera-caudatum	Ethiopia
255	SC991	0.059	0.319	0.007	0.496	0.119	4	bicolor	guinea-caudatum	Uganda
256	SEGAOLANE	0.002	0.919	0.001	0.065	0.013	2	n/a	kafir	n/a
257	SEPON82	0.001	0.001	0.812	0.002	0.184	3	guinea	zerazera-caudatum	Senegal
258	SOBERANO	0.001	0.000	0.997	0.000	0.002	3	cultivar	zerazera-caudatum	C. America
259	SRN39	0.148	0.002	0.183	0.047	0.620	5	cultivar	caudatum	n/a
260	SURENO	0.001	0.000	0.712	0.000	0.287	3	cultivar	zerazera-caudatum	C. America
261	TX2741	0.136	0.150	0.394	0.318	0.002	3	inbred line	milo-feterita	USA
262	TX2882	0.154	0.039	0.319	0.463	0.025	4	breeding line	milo-feterita	USA
263	TX2891	0.118	0.101	0.281	0.003	0.497	5	breeding line	caudatum	USA
264	TX2911	0.137	0.312	0.193	0.004	0.354	5	cultivar	kafir	USA
265	WHITEKAFIR	0.001	0.997	0.000	0.001	0.001	2	breeding line	kafir	USA

n\a = no information available

Figure 2. 1. Diversity analysis of the sorghum accessions: (A) STRUCTURE results inferring five subpopulations. (G). (B) Neighbor-joining tree: branches (B1–B5) color-coded based on STRUCTURE results. (C) Distribution of races among five subpopulations. D: durra, K: kafir, ZC: zerazera-caudatum, B: bicolor, GB: guinea-bicolor, MF: milo-feterita, SB: sudanese-broomcorn, C: caudatum, and GC: guinea-caudatum.

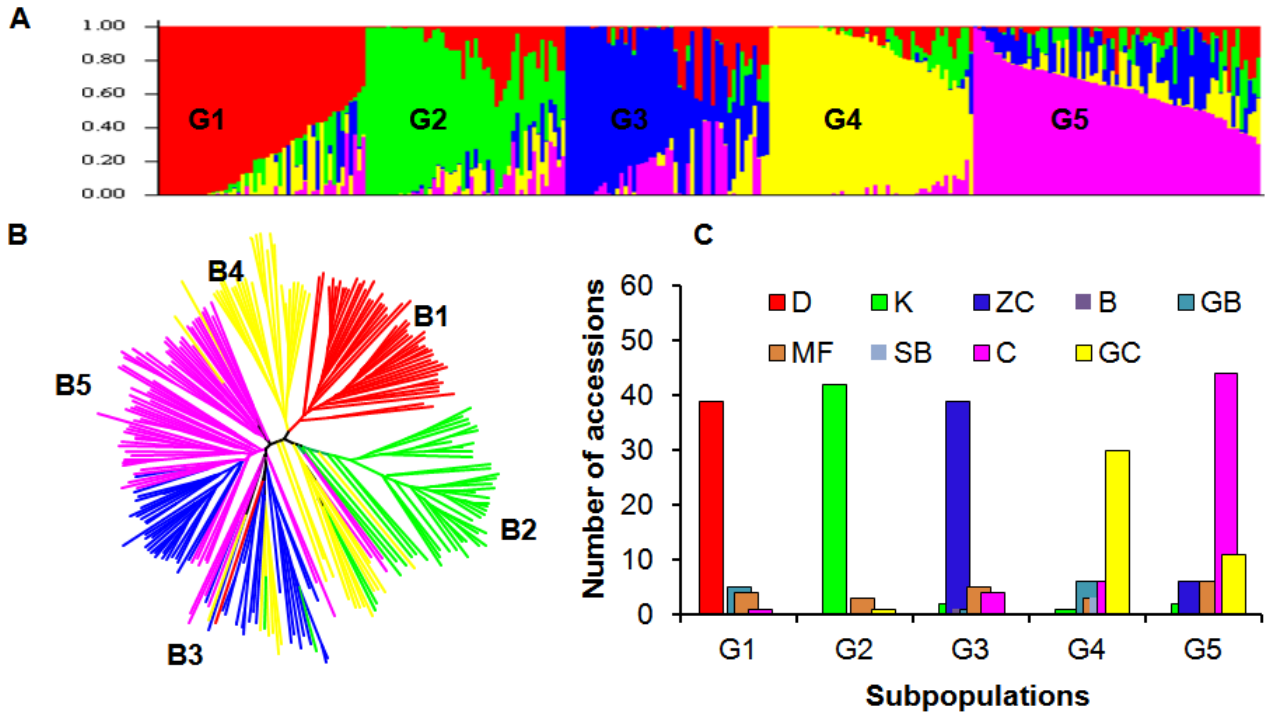


Figure 2.2. General congruence among principal component analysis (PCA), STRUCTURE classification, and race classification. (A) PCA and STRUCTURE classification. Each color represents a subpopulation based on STRUCTURE results. (B) Distribution of 265 sorghum accessions based on its collection site/origin. Red triangle = G1, green triangle = G2, blue circles = G3, yellow boxes = G4, and pink boxes = G5.

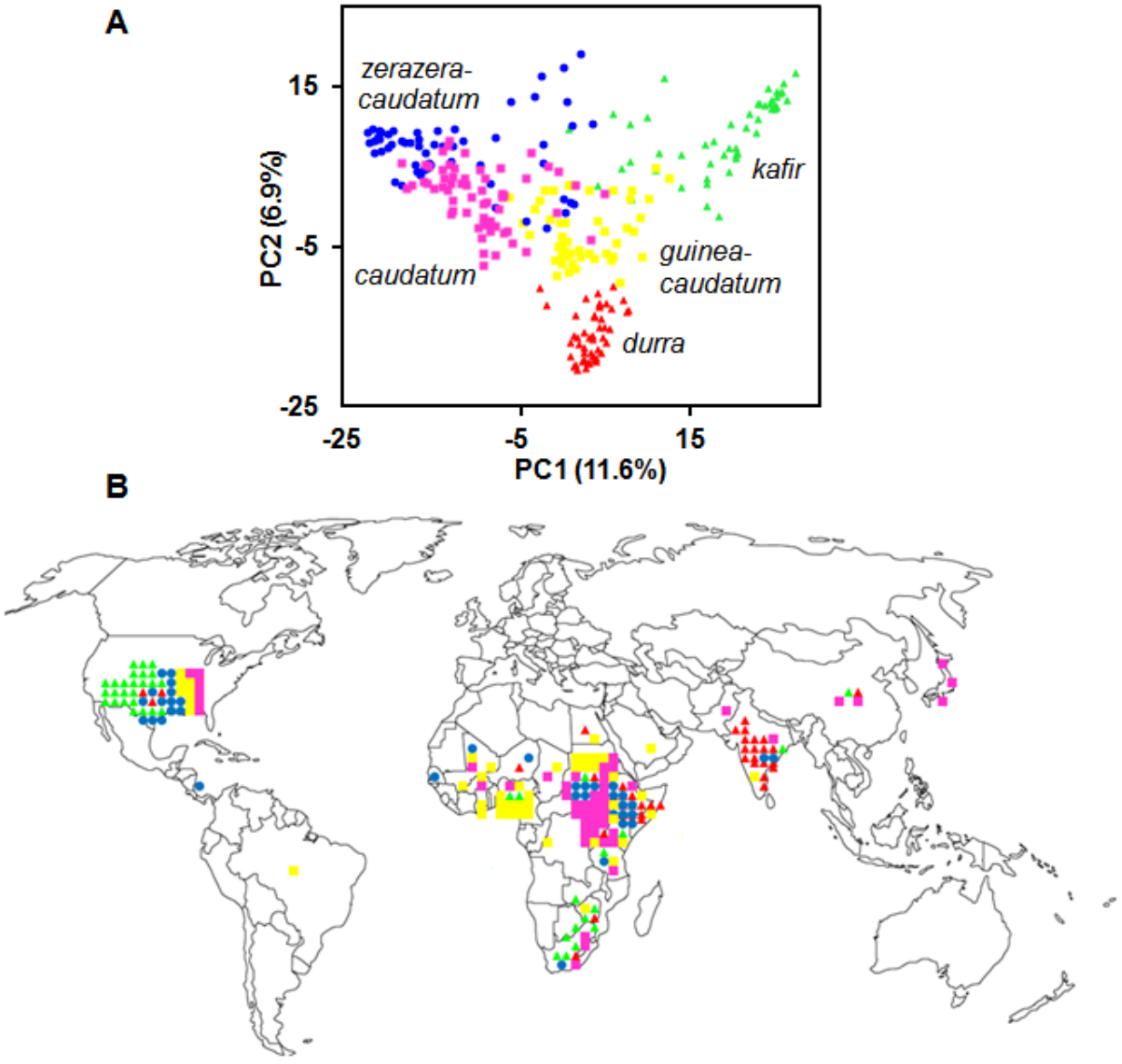


Figure 2. 3. Significant variations in grain quality traits found among different subpopulations of sorghum diversity panel. (a) KH: kernel hardness. (b) ADF: acid detergent fiber. (c) TDN: total digestible nutrients. The error bar represents the standard error. G1, G2, G3, G4, and G5 consist of 49, 46, 52, 49, and 69 accessions, respectively.

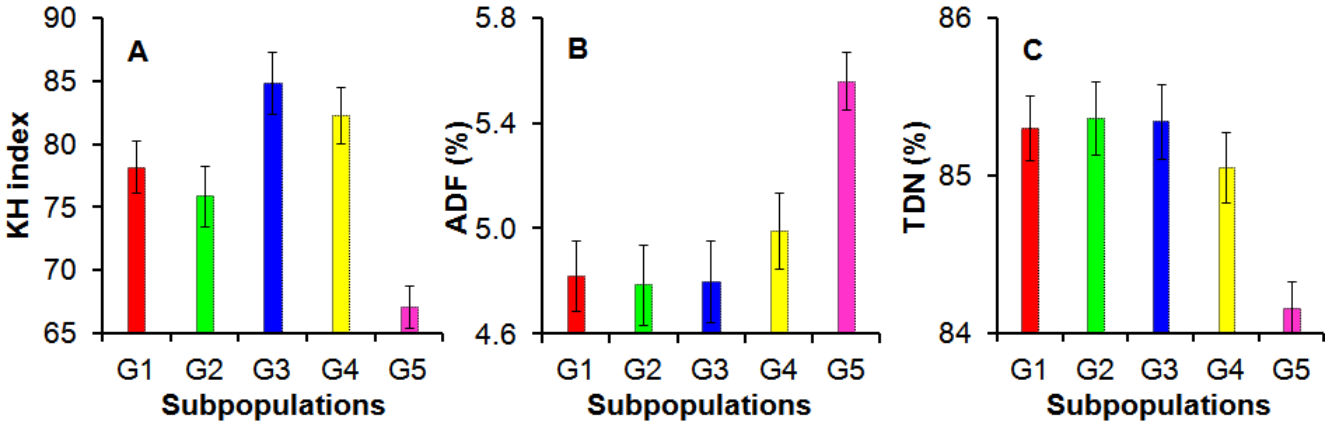


Figure 2.4. Consistency of SNP alleles are shown across five subpopulations. Each bar represents mean value of accessions with significant SNP allele. Error bar represents the standard error.

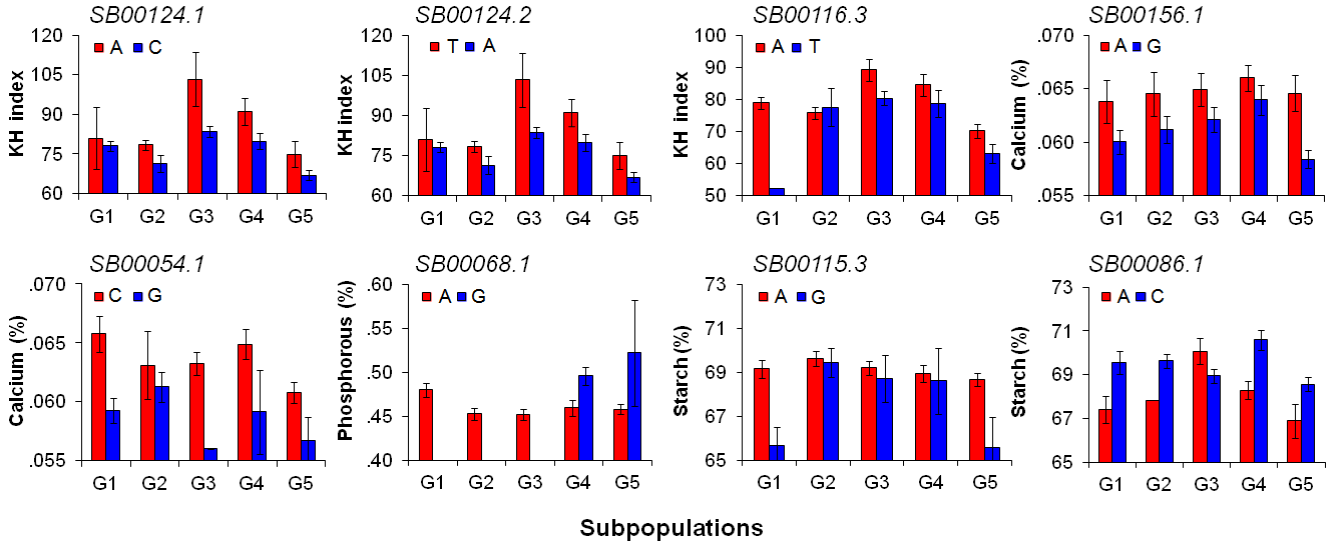


Figure 2.5. Quantile-quantile plots of the 10 grain quality traits with 1,523 SNP markers. The Q-Q plots showed the control of type I error by the selected models. KH = kernel hardness, KW = kernel weight, KD = kernel diameter, ADF = acid detergent fiber, Ca = calcium, CP = crude protein, F = Fat, P = phosphorous, S = starch, and TDN = total digestible nutrients.

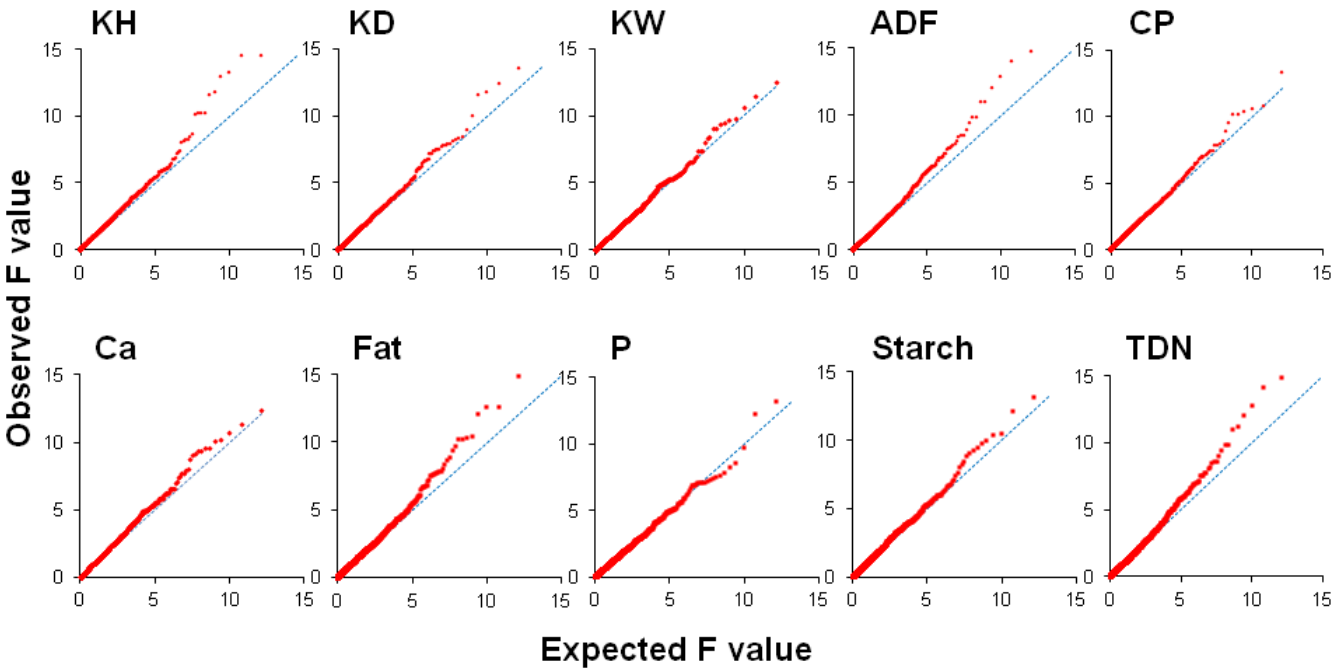
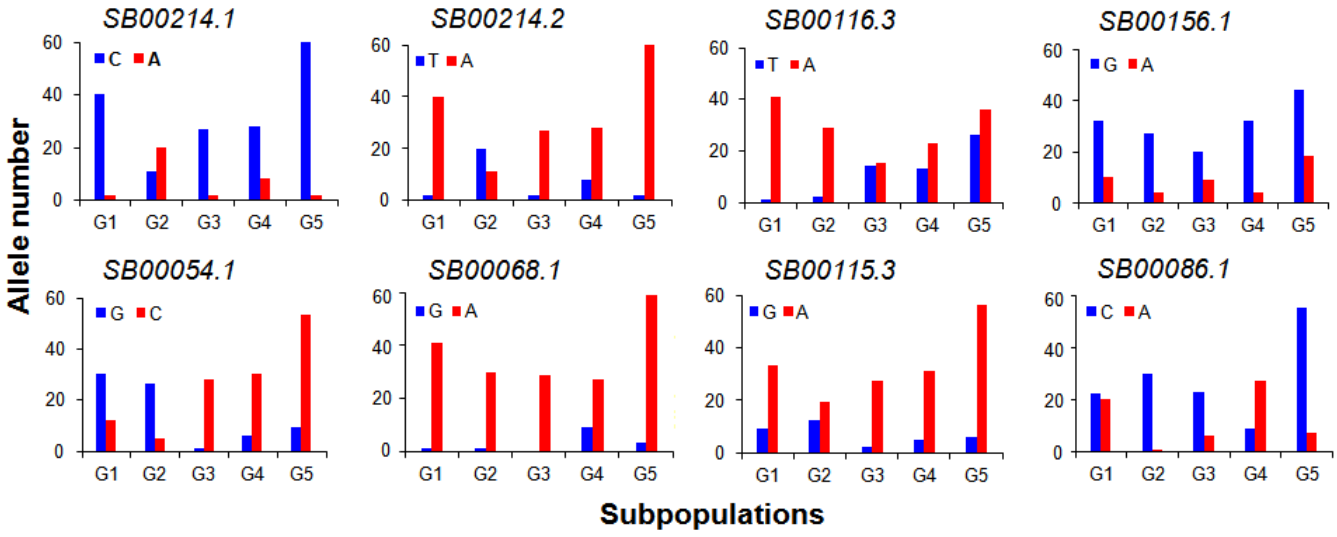


Figure 2.6 Variation in the frequency of significant SNPs associated with grain quality traits across the five subpopulations. Each bar represents the number of alleles for each SNP in the subpopulation.



References

Aboubacar, A., and B.R. Hamaker. 1999. Physicochemical properties of flours that relate to sorghum couscous quality. Cereal Chem. 76:308-313.

Akingbala, J.O., and L.W. Rooney. 1987. Paste Properties of Sorghum Flour and Starches. J. Food Proc. Pres. 11:13-24.

Aldrich, P.R., and J. Doebley. 1992. Restriction fragment Variation in the Nuclear and Chloroplast Genomes of Cultivated and Wild *Sorghum bicolor*. Theor. Appl. Genet. 85:293-302.

Anglani, C. 1998. Sorghum endosperm texture- a review. Plant Food Hum. Nutri. 52:67-76.

Bean, S.R., O.K. Chung, M.R. Tuinstra, J.F. Pedersen, and J. Erpelding. 2006. Evaluation of the Single Kernel Characterization System (SKCS) for Measurement of Sorghum Grain Attributes. Cereal Chem. 83:108-113.

Beta, T., H. Corke, L.W. Rooney, and J.R.N. Taylor. 2001. Starch properties as affected by sorghum grain chemistry. J. Sci. Food Agri. 81:245-251.

Bettge, A.D., M.J. Giroux, and C.F. Morris. 2000. Susceptibility of waxy starch granules to mechanical damage. Cereal Chem. 77:750-753.

Bordes, J., C. Ravel, J. Le Gouis, A. Lapierre, G. Charmet, and F. Balfourier. 2011. Use of a global wheat core collection for association analysis of flour and dough quality traits. J. Cereal Sci. 54:137-147.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635.

Brown, P.J., S. Myles, and S. Kresovich. 2011. Genetic Support for Phenotype-based Racial Classification in Sorghum. Crop Sci. 51:224-230.

Bueso, F.J., R.D. Waniska, W.L. Rooney, and F.P. Bejosano. 2000. Activity of antifungal proteins against mold in sorghum caryopses in the field. J. Agri. Food Chem. 48:810-816.

Cagampang, G.B., and A.W. Kirleis. 1984. Relationship of Sorghum Grain Hardness to Selected Physical and Chemical Measurements of Grain Quality. Cereal Chem. 61:100-105.

Casa, A.M., G. Pressoir, P.J. Brown, S.E. Mitchell, W.L. Rooney, M.R. Tuinstra, C.D. Franks, and S. Kresovich. 2008. Community resources and strategies for association mapping in

sorghum. Crop Sci. 48:30-40.

Chen F., Zhang F.Y., Xia X.C., Dong Z.D., and D.Q.Cui. 2012. Distribution of puroindoline alleles in bread wheat cultivars of the Yellow and Huai valley of China and discovery of a novel puroindoline a allele without PINA protein. Mol. Breed. 29:371-378.

Cook, J.P., M.D. McMullen, J.B. Holland, F. Tian, P. Bradbury, J. Ross-Ibarra, E.S. Buckler, and S.A. Flint-Garcia. 2012. Genetic Architecture of Maize Kernel Composition in the Nested Association Mapping and Inbred Association Panels. Plant Physiol. 158:824-834.

de Alencar Figueiredo, L., B. Sine, J. Chantereau, C. Mestres, G. Fliedel, J.F. Rami, J.C. Glaszmann, M. Deu, and B. Courtois. 2010. Variability of grain quality in sorghum: association with polymorphism in *Sh2*, *Bt2*, *SssI*, *Ae1*, *Wx* and *O2*. Theor. Appl. Genet. 121:1171-1185.

de Wet, J.M.J. 1978. Systematics and Evolution of Sorghum Sect. Sorghum (Gramineae). Amer. J. Bot. 65:477-484.

Douglas, J.H., T.W. Sullivan, P.L. Bond, and F.J. Struwe. 1990. Nutrient Composition and Metabolizable Energy Values of Selected Grain Sorghum Varieties and Yellow Corn. Poult. Sci. 69:1147-1155.

Folkertsma, R.T., H.F.W. Rattunde, S. Chandra, G.S. Raju, and C.T. Hash. 2005. The pattern of genetic diversity of Guinea-race *Sorghum bicolor* (L.) Moench landraces as revealed with SSR markers. Theor. Appl. Genet. 111:399-409.

Fujita, N., R. Satoh, A. Hayashi, M. Kodama, R. Itoh, S. Aihara, and Y. Nakamura. 2011. Starch biosynthesis in rice endosperm requires the presence of either starch synthase I or IIIa. J. Exp. Bot. 62:4819-4831.

Guzman C., Caballero L., Martin M.A., and J.B. Alvarez. 2012. Molecular characterization and diversity of the Pina and Pinb genes in cultivated and wild diploid wheat. Mol. Breed. 30:69-78.

Hamblin, M.T., M.G.S. Fernandez, M.R. Tunistra, W.L. Rooney, and S. Kresovich. 2007. Sequence variation at candidate loci in the starch metabolism pathway in sorghum: Prospects for linkage disequilibrium mapping. Crop Sci. 47:S125-S134.

Hamblin, M.T., M.G.S. Fernandez, A.M. Casa, S.E. Mitchell, A.H. Paterson, and S. Kresovich. 2005. Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass *Sorghum bicolor*. Genetics 171:1247-

1256.

Hamblin, M.T., A.M. Casa, H. Sun, S.C. Murray, A.H. Paterson, C.F. Aquadro, and S. Kresovich. 2006. Challenges of detecting directional selection after a bottleneck: Lessons from *Sorghum bicolor*. Genetics 173:953-964.

Hamblin, M.T., S.E. Mitchell, G.M. White, J. Gallego, R. Kukatla, R.A. Wing, A.H. Paterson, and S. Kresovich. 2004. Comparative Population Genetics of the Panicoid Grasses: Sequence Polymorphism, Linkage Disequilibrium and Selection in a Diverse Sample of *Sorghum bicolor*. Genetics 167:471-483.

Harlan, J.R., and J.M.J. Dewet. 1972. Simplified Classification of Cultivated Sorghum. Crop Sci. 12:172.

Jennings, P., and J. Cock. 1977. Centres of origin of crops and their productivity. Econ. Bot. 31:51-54.

Lee, W.J., J.F. Pedersen, and D.R. Shelton. 2002. Relationship of Sorghum kernel size to physiochemical, milling, pasting, and cooking properties. Food Res. Int. 35:643-649.

Lin, Z., X. Li, L.M. Shannon, C.-T. Yeh, M.L. Wang, G. Bai, Z. Peng, J. Li, H.N. Trick, T.E. Clemente, J. Doebley, P.S. Schnable, M.R. Tuinstra, T.T. Tesso, F. White, and J. Yu. 2012. Parallel domestication of the Shattering1 genes in cereals. Nat. Genet. 44:720-724.

Liu, K., and S.V. Muse. 2005. PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128 - 2129.

Martin, C.R., R. Rousser, and D.L. Brabec. 1993. Development of a Single-Kernel Wheat Characterization System. Trans. ASAE 36:1399-1404.

Morris, C.F, G.A. Greenblatt, A.D. Bettge, and H.I. Malkawi. 1994. Isolation and characterization of multiple forms of friabilin. J. Cereal Sci. 20:167–174.

Murphy, S.P., Simmons, C.R., Bass, H.W. 2010. Structure and expression of the maize (Zea mays L.) SUN-domain protein gene family: evidence for the existence of two divergent classes of SUN proteins in plants. BMC Plant Biol. 10: 269.

Murray, S.C., W.L. Rooney, M.T. Hamblin, S.E. Mitchell, and S. Kresovich. 2009. Sweet Sorghum Genetic Diversity and Association Mapping for Brix and Height. Plant Genome 2:48-62.

Murty, B.R. 1967. Catalogue of World Collection of Sorghum and Pennisetum - Acknowledgement. Indian J. Genet. Pl. Breed. 27:R5.

Pasquini, C. 2003. Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. J. Braz. Chem. Soc. 14:198-219.

Patterson, N., A.L. Price, and D. Reich. 2006. Population structure and eigenanalysis. PLoS Genet. 2:e190.

Pedersen, J.F., C.R. Martin, F.C. Felker, and J.L. Steele. 1996. Application of the single kernel wheat characterization technology to sorghum grain. Cereal Chem. 73:421-423.

Perumal, P., P. Krishnaramanujam, M.A. Menz, S. Katile, J. Dahlberg, C.W. Magill, and W.L. Rooney. 2007. Genetic diversity among sorghum races and working groups based on AFLPs and SSRs. Crop Sci. 47:1375-1383.

Price, A.L. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904-909.

Pritchard, J.K., M. Stephens, N.A. Rosenberg, and P. Donnelly. 2000. Association Mapping in Structured Populations. Am. J. Hum. Genet. 67:170-181.

Rami, J.F., P. Dufour, G. Trouche, G. Fliedel, C. Mestres, F. Davrieux, P. Blanchard, and P. Hamon. 1998. Quantitative trait loci for grain quality, productivity, morphological and agronomical traits in sorghum (*Sorghum bicolor* L. Moench). Theor. Appl. Genet. 97:605-616.

Reif, J.C., M. Gowda, H.P. Maurer, C.F.H. Longin, V. Korzun, E. Ebmeyer, R. Bothe, C. Pietsch, and T. Wurschum. 2011. Association mapping for quality traits in soft winter wheat. Theor. Appl. Genet. 122:961-970.

Sabadin, P.K., M. Malosetti, M.P. Boer, F.D. Tardin, F.G. Santos, C.T. Guimaraes, R.L. Gomide, C.L.T. Andrade, P.E.P. Albuquerque, F.F. Caniato, M. Mollinari, G.R.A. Margarido, B.F. Oliveira, R.E. Schaffert, A.A.F. Garcia, F.A. van Eeuwijk, and J.V. Magalhaes. 2012. Studying the genetic basis of drought tolerance in sorghum by managed stress trials and adjustments for phenological and plant height differences. Theor. Appl. Genet. 124:1389-1402.

Salas Fernandez, M., I. Kapran, S. Souley, M. Abdou, I. Maiga, C. Acharya, M. Hamblin, and S. Kresovich. 2009. Collection and characterization of yellow endosperm sorghums from West Africa for biofortification. Genet. Resources Crop Evol. 56:991-1000.

SAS institute. 1999. SAS/STAT user's guide version 8. Cary.

Scheet, P., and M. Stephens. 2006. A Fast and Flexible Statistical Model for Large-Scale

Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. Am. J. Hum. Genet. 78:629-644.

Shechter, Y., and J.M.J.D. Wet. 1975. Comparative Electrophoresis and Isozyme Analysis of Seed Proteins from Cultivated Races of Sorghum. Am. J. Bot. 62:254-261.

Shimbata, T., T. Inokuma, A. Sunohara, P. Vrinten, M. Saito, T. Takiya, and T. Nakamura. 2011. High Levels of Sugars and Fructan in Mature Seed of Sweet Wheat Lacking GBSSI and SSIIa Enzymes. J. Agri. Food Chem. 59:4794-4800.

Smith, C.W., and R.A. Frederiksen. 2000. Sorghum: origin, history, technology and production John Wiley and Sons, New York.

Snowden, J.D. 1936.. The cultivated races of sorghum. Adlard, London, UK.Stephens, J.C., F.R. Miller, and D.T. Rosenow. 1967. Conversion of Alien Sorghums to Early Combine Genotypes. Crop Sci. 7:396-396.

Suhendro, E.L., C.F. Kunetz, C.M. McDonough, L.W. Rooney, and R.D. Waniska. 2000. Cooking characteristics and quality of noodles from food sorghum. Cereal Chem. 77:96-100.

Sun, G., C. Zhu, M.H. Kramer, S.S. Yang, W. Song, H.P. Piepho, and J. Yu. 2010. Variation explained in mixed-model association mapping. Heredity 105:333-340.

Taylor, J.R.N., T.J. Schober, and S.R. Bean. 2006. Novel food and non-food uses for sorghum and millets. J. Cereal Sci. 44:252-271.

Tian, Z., Q. Qian, Q. Liu, M. Yan, X. Liu, C. Yan, G. Liu, Z. Gao, S. Tang, D. Zeng, Y. Wang, J. Yu, M. Gu, and J. Li. 2009. Allelic diversities in rice starch biosynthesis lead to a diverse array of rice eating and cooking qualities. Proc. Natl. Acad. Sci. USA 106:21760-21765.

Wills, R.B.H., and M.R. Ali. 1983a. Effect of Grain-Size on Degree of Milling, Color and Cooking Time of Sorghum. J. Food Sci. 48:650-651.

Wills, R.B.H., and M.R. Ali. 1983b. Effect of Grain-Size on Dehulling of Sorghum. Cereal Chem. 60:12-14.

Wilson, L.M., S.R. Whitt, A.M. Ibanez, T.R. Rocheford, M.M. Goodman, and E.S. Buckler. 2004. Dissection of maize kernel composition and starch production by candidate gene association. Plant Cell 16:2719-2733.

Yan, H.-B., X.-X. Pan, H.-W. Jiang, and G.-J. Wu. 2009. Comparison of the starch synthesis genes between maize and rice: copies, chromosome location and expression divergence.

Theor. Appl. Genet. 119:815-825.

Yan L., Peng X., Yafang S., Gan Z., Thanwanit T., and B. Jinsong. 2010. Development of new markers to genotype the functional SNPs of SSIIa, a gene responsible for gelatinization temperature of rice starch. J. Cereal Sci. 52:438-443.

Yu, J., and E.S. Buckler. 2006. Genetic association mapping and genome organization of maize. Curr. Opin. Biotechnol. 17:155 - 160.

Yu, J., MT Hamblin, and a.M. Tuinstra. 2011. Association Genetics Strategies and Resources, in AH Paterson (ed.), Genetics and Genomics of Saccharinae, Springer-Science, New York, USA.

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203-208.

Zhu, C., X. Li, and J. Yu. 2011. Integrating rare-variant testing, function prediction, and gene network in composite resequencing-based genome-wide association studies (CRGWAS). G3. 1:233–243.

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 3- LITERATURE REVIEW: DROUGHT TOLERANCE

Introduction

Grain sorghum (*Sorghum bicolor* (L.) Moench) is the fifth most important grain crop grown across about 99 countries in the world. It is a predominantly self pollinating crop native to Sub-Saharan Africa and has been cultivated for centuries as a staple in much of Africa and Asia. It is an important source of food, feed, and bio-fuel especially in the semi-arid tropics due to its wide adaptation to the harsh, drought prone environments. It has gained considerable importance as a fodder and feed crop in the last decade in US. Sorghum is a tropical C_4 grass with a relatively compact genome of 736MB (Brown et al. 2006) and it has been sequenced .

Drought is a major constraint in sorghum production worldwide and is considered as the most important cause of yield reduction in crop plants. Even though sorghum is well adapted to the hot and dry environments in Africa and Asia, sorghum cultivation in US is checked by frequent drought and high temperature. More than 80% of the sorghum hybrids in USA are grown in non-irrigated conditions and are prone to drought stress thereby reduced yield. Sorghum is regarded as a model for studying drought tolerance among the grass species, particularly due to its synteny with genome of maize. When sorghum hybrids are exposed to drought, rapid and premature leaf senescence occurs which could lead to charcoal rot, stalk lodging, and significant yield losses.

So breeding for drought tolerance in sorghum is an important objective. But conventional methods to evaluate drought tolerance under natural field conditions are difficult due to unpredictable moisture stress and large genotype by environment interactions. QTL mapping is the most popular traditional method to detect complex traits in plants. So the focus has changed to study QTLs for drought tolerance and genome mapping for drought tolerance (Tuberosa 2003). However, the plant responses to drought stress are complex and consistent QTL for grain yield and drought tolerance are still lacking.

Drought tolerance physiology

Plants that are tolerant to drought have evolved certain adaptive mechanisms with different degree of tolerance. It is determined by genetic plasticity that allows them to survive the adverse conditions or change specific growth habits that avoid stress conditions. Drought stress leads to cellular dehydration, which causes osmotic stress and removal of water from the cytoplasm into the extracellular space resulting in a reduction of the cytosolic and vacuolar volumes. Another consequence is the production of reactive oxygen species which negatively affect cellular structures and metabolism. Early response to drought involves decrease in photosynthesis and rising levels of plant hormone Abscissic acid (Bartels and Sunkar 2005).

Plant's resistance to drought has been divided in to three different types: drought escape, drought avoidance, and drought tolerance strategies. These strategies are not mutually exclusive and plants may combine a range of response types. Plants that escape drought exhibit a high degree of developmental plasticity and complete their life cycle before physiological water deficit occurs. This is more common with the plants that grow in arid environments. Drought avoidance is a mechanism for avoiding lower water status in tissues during drought by maintaining cell turgor pressure and cell volume either through aggressive water uptake with an extensive root system or through reduction of water loss from transpiration and other non-stomatal pathways (Chaves et al. 2003). Plants can overcome drought conditions by avoiding tissue dehydration, while maintaining tissue water potential as high as possible, or by tolerating low tissue water potential. There are two characters related to drought avoidance (i) minimizing water loss (ii) maximizing water uptake. Water loss is minimized by closing stomata by reducing light absorbance through rolled leaves, a dense trichome layer increasing reflectance, or steep leaf angles or by decreasing canopy leaf area through reduced growth and shedding of older leaves. Water uptake is maximized by adjusting the allocation pattern, increasing the root uptake (Jackson et al. 2000). The epi-cuticular wax and cuticle structure determines the hydraulic permeability of the leaf surface and sorghum has high epi-cuticular resistance which prevents water loss when stomata are closed.

Plants that are drought tolerant maintain metabolic activities even at low water potential. Osmotic adjustment and antioxidant capacity highly influence drought tolerance. When plants experience water stress, the accumulation of compatible solutes in the cells lowers osmotic potential in the cell that helps to maintain the turgor of both shoots and roots that is osmotic

adjustment. Improved reproductive success also includes better partitioning of assimilates to developing fruits (Gebbing *et al*., 1999). This is associated with plant's ability to mobilize the reserves for fruit production, this ability is increased in drought tolerant plants (Yang et al. 2001).

Plant's adjustments to the low resources in arid environments include altered leaf structure, increased proportions of assimilates allocation to roots, slow growth rates, and slow organ turnover rates (Poorter.H and Nagel. 2000). Whereas short lived organs like leaves can be discarded in response to stress, long lived organs must optimize their resource gain. Tolerance to low tissue water potential involves osmotic adjustment, more rigid cell walls or smaller cells. (Wilson *et al*., 1980). Stomatal closure together with leaf growth inhibition is the earliest responses to drought protecting the plants from earliest water loss, which might result in cell dehydration. Stomatal closure is likely to be mediated by chemical signals travelling from the dehydration roots to the shoots (Chaves et al. 2003).

Drought tolerance pathways

Drought is a complex trait that has more than one specific pathway involved in the expression of the trait. The regulatory circuits include stress sensors, signaling pathways comprising a network of protein-protein interactions, transcription factors, promoters, proteins, and other metabolites. Each gene produces a polypeptide to enhance or restrict a biochemical action in a pathway. Differential stress tolerances are attributed to differences in plant reactivity in terms of stress perception, signal transduction, appropriate gene expression programs, and other novel metabolic pathways.

At least four signal transduction pathways involved in plant responses to osmotic stress are investigated by researchers; two are ABA dependent (I and II) and two are ABA independent (I and II) (Shinozaki and Yamaguchi-Shinozaki., 2007). Until now two ABA dependent pathways are known to mediate gene expression in plants during osmotic stress. The distinction is largely based on *cis*- elements that exist in the promoters of ABA inducible genes. The ABA dependent pathways are thought to mediate the gene expression through an *ABRE* element and b-*ZIP* transcription factors (Busk and Pages, 1998). ABA dependent pathway I require protein synthesis to activate transcription factors *MYC/MYB* or *b-ZIP* which bind to DNA regions other than the *ABREs*. ABA-dependent pathway II activates *b-ZIP* (Hollung et al. 1997; Nakagawa et

al. 1996), a transcription factor that turns on gene expression through binding to *ABREs*. The main ABA biosynthetic pathway starts from carotinoids C-40. Biochemical studies have indicated that 9 cis-epoxycarotenoid dioxygenase (*NCED*) is a key enzyme in ABA biosynthesis. The induction of the *NCED* gene by drought stress has been reported in maize, tomato, *Arabidopsis*, bean, and cowpea (Bartels *et al*., 2007)

Comparative analysis of *Arabidopsis* with rice (*Oryza sativa*) showed that among the 73 genes identified as stress inducible in rice, 51 genes has shown similar responses to stress tolerance in *Arabidopsis* (Rabbani et al. 2003). The investigation of dehydration induced genes in *Arabidopsis* has revealed that there exist ABA independent signal transduction pathways genes which do not require ABA for their expression under drought conditions. ABA independent genes have a conserved dehydration responsive element (*DRE*, with *TACCGACAT* consensus) in their promoters that is involved in gene regulation by interaction with an ABA independent signaling cascade (Knight and Knight, 2001). Some of the proteins involved in the stress adaptation are late embryogenesis abundant proteins (*LEA*). Accumulation of *LEA* proteins during embryogenesis correlates with increased levels of ABA and with acquisition of desiccation tolerance (Bartels *et al*., 2007).

Through the microarray analysis the products of drought inducible genes identified can be divided into two groups. The first group includes chaperons, LEA proteins, Osmotin, anti freeze proteins, mRNA binding proteins, key enzymes for osmolyte biosynthesis, water channel proteins, sugar and proline transporters, detoxification enzymes, and various proteases. The second group is regulatory proteins, protein factors involved in signal transduction and stress responsive gene expression. There is different protein molecules involved in the signaling pathway for drought response. CDKs are a family of proteins each with a positive regulatory subunit termed a cyclin and the catalytic subunit CDK (den Boer and Murray, 2000). The decrease in cell division in response to water stress is characterized by lower CDK activity which is correlated to tyrosine phosphorylation.

Protein phosphorylation is one of the major mechanisms for controlling cellular functions in response to external signals. (Wang et al. 1998) reported that ABA induces expression of an inhibitor of *CDK*. These mechanisms could be responsible for the ABA dependent cell cycle arrest during osmotic stress in plants. *NtC7*, a gene is that is responsive protein kinases (*MAPK*) are common signaling modules in eukaryotic cells, including plants. In Arabidopsis *AtMEKK1*

and *AtMPK3* are activated by dehydration. *AtMPK4* and *AtMPK6* are post transcriptionally activated by osmotic stress. *SNF1/AMP* activated protein kinases are expressed in response to dehydration or ABA. salicylic acid, nitric oxide, phosphatidic acid, inositol 1,4,5-triphosphate are some of the stress signaling molecules identified (Bartels and Sunkar, 2005).

Genetics of drought tolerance: QTL mapping studies

Classical breeding approaches revealed that stress tolerance is governed by many genes that make the genetic selection of these traits through traditional breeding a difficult task. In sorghum, two types of drought responses is known. In pre-flowering drought tolerance and post-flowering drought tolerance that is controlled by two different genetic mechanisms. Pre-flowering in sorghum refers to the stage from panicle differentiation to flowering. Drought stress during this stage affects panicle size, number of grains in a panicle, and grain yield. Post-flowering refers to the grain development stage (GS-3). Drought at this stage is characterized by rapid premature death which leads to charcoal rot, stalk lodging, and yield loss. (Sanchez et al. 2002). There were several studies to map the QTLs for stay green in sorghum (Crasta et al. 1999; Harris et al. 2006; Haussmann et al. 2002; Kebede et al. 2001; Sanchez et al. 2002; Subudhi et al. 2000; Tuinstra et al. 1997; Xu et al. 2000)

In an earlier study 98 recombinant inbred lines from a cross of B35 (pre-flowering drought susceptible and post-flowering tolerant) and Tx7078 (pre-flowering tolerant and post-flowering susceptible) was evaluated for post flowering drought tolerance in Kansas. This population was characterized under drought and non-drought conditions for the traits associated with post-flowering drought tolerance and related components of grain development. They identified 13 QTLs associated with post-flowering drought tolerance. Two QTLs were identified with major effects on yield and stay green under post flowering drought. QTLs for stay green were identified in seven genomic regions; two QTLs were on linkage groups F and I, which were also associated with yield under post flowering drought tolerance. Three other QTLs on linkage groups B and G were associated with stay green and together these QTLs explained 53% of the variability for the stay green trait. Genetic studies of stay green have generally indicated a complex pattern of inheritance, both dominant and recessive expression have been reported Tuinstra *et al*. (1996).

In another study, QTLs controlling premature leaf senescence and their association under post flowering drought tolerance were evaluated. A recombinant inbred population derived from a cross of B35 and Tx430 was the genetic material for the experiment. They developed a linkage map of sorghum using 142 restriction fragment length polymorphism (RFLP) markers. These populations along with parents were evaluated in four different environments for post-flowering drought tolerance and maturity. Using simple interval mapping they identified seven stay green QTLs and two maturity QTLs. Out of the seven QTLs, three major stay-green QTLs (SGA, SGD, and SGG) explained 42% of the phenotypic variability and four minor QTLs (SGB, SGI.1, SGI.2, and SGJ) explained an additional 25% of the variability of stay green ratings. Composite interval mapping with an additional analysis of the QTL by environment interaction confirmed the above results and the identified QTLs explained about 90% of the genetic variability for stay green. Six out of the seven stay green QTLs identified were independent of the QTLs influencing grain maturity (Crasta et al. 1999).

Four stay-green QTLs, located on three linkage groups were identified that control the stay-green and chlorophyll content in sorghum by using a restriction fragment length polymorphism (RFLP) map. This study used a RIL population. The QTLs, *Stg1* and *Stg2* were on linkage group A, and the others, *Stg3* and *Stg4* were on linkage groups D and J, respectively. QTLs, *Stg1* and *Stg2*, explaining 13–20% and 20–30% of the phenotypic variability, respectively, were consistently identified at different locations in two years. Three QTLs for chlorophyll content (*Chl1, Chl2, and Chl3*), explaining 25–30% of the phenotypic variability were also identified under post-flowering drought stress. All coincided with the three stay-green QTL regions (*Stg1, Stg2,* and *Stg3*) accounting for 46% of the phenotypic variation. The *Stg1* and *Stg2* regions also contain the genes for key photosynthetic enzymes, heat shock proteins, and the stress hormone Abscissic acid (ABA). that the study concluded that the linkage group A is important for drought- and heat-stress tolerance and yield improvement in sorghum (Xu et al. 2000).

A recombinant inbred line (RILs) mapping population that was developed from a cross of B35 and Tx7000 was evaluated over two locations for two years to study the consistency of stay green QTLs in sorghum (Subudhi et al., 2001). They reconfirmed all the QTLs identified for stay green traits by Xu et al., (2000). The four stay green QTLs (*Stg1*, *Stg2*, *Stg3*, and *Stg4*) were consistent in different environments. They concluded that the QTL, *Stg2* was the most important

QTL explaining maximum amount of phenotypic variation and targeting the Stg-2 region for gene discovery will help the researchers to identify the basics of stay green phenomenon in cereals (Subudhi et al., 2000).

Another study focused on post flowering drought tolerance, pre-flowering drought tolerance, and lodging tolerance using a F_7, recombinant inbred line (RIL) population derived from a cross between SC56 and Tx7000 (Kebede et al. 2001). Composite interval mapping detected nine QTLs located over seven linkage groups. This study showed that the three QTLs identified on linkage groups A, G, J were consistent in different environments. Comparative analysis showed that two stay green QTLs identified in sorghum corresponded to stay green QTLs identified in maize. Cross species comparative analysis of genomic regions for drought tolerance traits gave an indication that there may be some orthologous conserved regions for drought tolerance in rice, maize and sorghum which can be targeted for intensive investigation in future in improve drought tolerance in cereals (Kebede et al. 2001).

Using recombinant inbred lines (RILs) and Near Inbred lines (NILs) developed from a cross of B35 and Tx7000, four genomic regions associated with the stay green trait in sorghum were identified. These four major stay-green QTLs were consistently identified in all field trials and accounted for 53.5% of the phenotypic variance (Sanchez et al. 2002). In another study, using NIL derived from BTx642 and RTx7000, alleles that contribute to the stay green trait were mapped to the four major QTLs, *Stg1- Stg4* in sorghum (Harris et al. 2006).

Table 3.1 Studies that identified QTLs for drought tolerance in sorghum

No	Lines	QTLs identified	Article
1.	BTx642 × RTx7000 (NIL)	QTLs *Stg1, Stg2, Stg3, and Stg4*	Harris et al., 2006
2.	SC56 × Tx7000 (RIL)	QTLs on A,G, and J	Kebede et al., 2001
3.	B35 × Tx7000 (NIL)	Four major stay green QTL were identified.	Sanchez et al., 2002
3.	B35 × Tx430 (RIL)	*Stg1* and *Stg2* on A, *Stg3* on D, *Stg4* on J.	Sanchez et al., 2002
4.	QL39 × QL41(RIL)	Five stay green QTL identified	Sanchez et al., 2000
5.	Tx7078 × B35 (RIL)	13 regions of the genome associated with post flowering drought tolerance were identified. Two QTL identified	Tuinstra et al., 1997
6.	B35 × Tx7000 (RIL)	Four QTLs were identified	Subudhi et al.,(2000)
7.	IS9830 × E36-1 and N13 × E36-1	Three stay green QTLs identified in chromosomes A, E and G.	Haussmann et al., (2002)

References

Bartels D, Sunkar R (2005) Drought and Salt Tolerance in Plants. Critical Reviews in Plant Sciences 24:23 - 58

Chaves MM, maroco jP, Pereira JS (2003) Understanding plant responses to drought from genes to the whole plant Functional Plant Biology 30:239-264

Crasta OR, W. W. Xu, D. T. Rosenow, J. Mullet, H. T. Nguyen (1999) Mapping of post-¯owering drought resistance traits in grain sorghum:association between QTLs in¯uencing premature senescence and maturity. Mol Gen Genet 262:578-588

Crasta OR, Xu WW, Rosenow DT, Mullet J, Nguyen HT (1999) Mapping of post-flowering drought resistance traits in grain sorghum: association between QTLs influencing premature senescence and maturity. Molecular and General Genetics MGG 262:579-588

Harris K, Subudhi PK, Borrell A, Jordan D, Rosenow D, Nguyen H, Klein P, Klein R, Mullet J (2006) Sorghum stay-green QTL individually reduce post-flowering drought-induced leaf senescence. Jounal of Experimental Botany:erl225

Haussmann BIG, Mahalakshmi V, Reddy BVS, Seetharama N, Hash CT, Geiger HH (2002) QTL mapping of stay-green in two sorghum recombinant inbred populations. Theor Appl Genet 106:133-142

Haussmann BIG, V. Mahalakshmi, B. V. S. Reddy · N. Seetharama · C. T. Hash, H. H. Geiger (2002) QTL mapping of stay-green in two sorghum recombinant inbred populations. Theor Appl Genet 106:133-142

Hollung K, Espelund M, Schou K, Jakobsen KS (1997) Developmental, stress and ABA modulation of mRNA levels for bZip transcription factors and Vp1 in barley embryos and embryo-derived suspension cultures. Plant Molecular Biology 35:561-571

Jackson RB, Sperry JS, Dawson TE (2000) Root water uptake and transport: using physiological processes in global predictions. Trends in Plant Science 5:482-488

Kebede H, Subudhi PK, Rosenow DT, Nguyen HT (2001) Quantitative trait loci influencing drought tolerance in grain sorghum (Sorghum bicolor L. Moench). Theor Appl Genet 103:266-276

Nakagawa H, Ohmiya K, Hattori T (1996) A rice bZIP protein, designated OSBZ8, is rapidly induced by abscisic acid. The Plant Journal 9:217-227

Nguyen TTT, Klueva N, Chamareck V, Aarti A, Magpantay G, Millena ACM, Pathan MS, Nguyen HT (2004) Saturation mapping of QTL regions and identification of putative candidate genes for drought tolerance in rice. Molecular Genetics and Genomics 272:35-46

Poorter.H, Nagel. O (2000) The role of biomass allocation in the growth response of plants to different levels of light, CO2, nutrients and water: a quantitative review. Aust J Plant Physiol 27:595-607

Rabbani MA, Maruyama K, Abe H, Khan MA, Katsura K, Ito Y, Yoshiwara K, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2003) Monitoring Expression Profiles of Rice Genes under Cold, Drought, and High-Salinity Stresses and Abscisic Acid Application Using cDNA Microarray and RNA Gel-Blot Analyses. Plant Physiol 133:1755-1767

Sanchez AC, P.K. Subudhi,D.T. Rosenow, and H.T. Nguyen (2002) Mapping QTLs associated with drought resistance in sorghum (Sorghum bicolor L. Moench). Plant Mol Bio 48:713-726

Sanchez AC, Subudhi PK, Rosenow DT, Nguyen HT (2002) Mapping QTLs associated with drought resistance in sorghum (Sorghum bicolor L. Moench). Plant Molecular Biology 48:713-726

Shinozaki K, Yamaguchi-Shinozaki K (2007) Gene networks involved in drought stress response and tolerance. Nat Rev Genet 58:221-227

Subudhi PK, Rosenow DT, Nguyen HT (2000) Quantitative trait loci for the stay green trait in sorghum (Sorghum bicolor L. Moench): consistency across genetic backgrounds and environments. Theor Appl Genet 101:733-741

Tuberosa R, S.Grillo,R.P.Ellis. (2003) Unravelling the genetics basis of drought tolerance in crops. Springer

Tuinstra MR, Edwin M. Grote, Peter B. Goldsbrough and Gebisa Ejeta (1997) Genetic analysis of post-flowering drought tolerance and components of grain development in Sorghum bicolor (L.) Moench. Molecular Breeding 3:439-448

Tuinstra MR, Grote EM, Goldsbrough PB, Ejeta G (1997) Genetic analysis of post-flowering drought tolerance and components of grain development in Sorghum bicolor (L.) Moench. Molecular Breeding 3:439-448

Vinod MS, Sharma N, Manjunatha K, Kanbar A, Prakash NB, Shashidhar HE (2006) Candidate genes for drought rolerance and improved productivity in rice (*Oryza sativa* L.). Indian Academy of Sciences 31(1):69-74

Xu-sheng W, J. Z, L. M, R. B (2005) Identification of candidate genes for drought stress tolerance in rice by the integration of a genetic (QTL) map with the rice genome physical map*. J Zhejiang Univ SCI 6B(5):382-388

Xu WW, Subudhi PK, Crasta OR, Rosenow DT, Mullet JE, Nguyen HT (2000) Molecular mapping of QTLs conferring stay-green in grain sorghum (Sorghum bicolor L. Moench). Genome 43:461-469

CHAPTER 4-QTL MAPPING OF YIELD POTENTIAL AND DROUGHT TOLERANCE IN SORGHUM

Abstract

Drought is the most important cause of yield reduction in crop plants. Sorghum is a crop well adapted to semi arid regions of USA and may harbor genes for drought tolerance. The objective of this experiment was to identify quantitative trait loci (QTLs) for yield potential and drought tolerance in sorghum. A population of 248 recombinant inbred lines (RILs) was developed from a cross between Tx436 (food grain type) and 00MN7645 (drought tolerant). Multi-location trials were conducted in 8 environments across Kansas to evaluate agronomic performance of the RILs under favorable and drought stress conditions. The 248 RILs and their parents were genotyped by genotyping-by-sequencing (GBS) and 8000 SNPs were identified. A subset of 800 SNPs was used for linkage map construction and QTL detection. Composite interval mapping identified a major QTL for grain yield in chromosome 8 and another QTL for flowering time in chromosome 9 under favorable conditions. Three major QTLs were detected for grain yield under drought conditions in chromosome 1, 6, and 8. We identified two flowering time QTLs on chromosome 1 under drought conditions. Six QTLs were identified for stay green: two on chromosome 4; one each on chromosome 5, 6, 7, and 10 under drought conditions.

Introduction

Sorghum (*Sorghum bicolor* L. Moench) is one of the most drought tolerant crops in the world. It is a crop that contributes to food security in the hot and water limited parts of Africa. In the United States, it is mainly grown as a feed stock and for ethanol production. But most of the commercial sorghum grown in the United States is prone to drought stress. Drought is a major environmental factor limiting crop productivity (Bartels and Sunkar, 2005). Drought affects molecular, physiological, and morphological mechanisms of the plant, resulting in yield losses higher than all other causes combined (Farooq et al., 2009; Passioura, 2006). From physiological and agronomic perspectives, drought tolerance is a loosely defined trait related to water use efficiency; but when considered from the perspective of gene discovery, drought tolerance is a complex trait due to the possible number of genes involved, their interactions among each other, and interactions with the environmental conditions (Blum, 2011). Hence, identifying the genetic causes of drought tolerance is challenging through traditional breeding methodologies and quantitative trait loci (QTL) based approaches can contribute to better understanding of the crop performance under stress conditions (Collins et al., 2008).

QTL mapping studies for drought tolerance in sorghum have identified four major stay green QTLs (*Stg1-Stg4*) using recombinant inbred lines (RILs) and near iso-genic lines (NILs) (Crasta et al., 1999; Jordan et al., 2012; Kebede et al., 2001; Mace and Jordan, 2011; Sanchez et al., 2002; Subudhi et al., 1999; Xu et al., 2000). Stay green is a physiological mechanism that is an incidence of delayed or inoperative foliar senescence in maize, sorghum, oats, rice, wheat, and other plant species (Thomas and Smart, 1993). Stay-green is a desirable character in sorghum for higher grain yield under water limiting conditions compared to non-stay green varieties (Jordan et al., 2012). Stay-green genotypes shows resistance to charcoal rot (Garud et al., 2002) and shoot lodging (Jordan et al., 2012). In sorghum, the expression of stay-green is a physiologically complex process depending upon the onset and duration of drought. Sorghum genotypes vary in their initiation and rate of senescence and field studies suggest that this trait is highly expressed under drought stress (vanOosterom et al., 1996).

Most of the QTL mapping experiments for drought tolerance have used unimproved accessions (BT×642 (formerly known as B35), SC56, E36-1, and Tx7000) as parents for making mapping populations. In the present study a RIL population was developed from Tx436, a food quality grain sorghum line and 00MN7645 (R45), an improved pollinator line with stay green

and lodging resistance. The pedigree of 00MN7645 (SC35//B35/80060) is heavily based on SC35 and B35. But 80060, a parental line from from Department of Plant Industry, Queensland makes 00MN7645 a new type of stay green line. The objective of the present study was to identify genomic regions associated with grain yield and stay green in this RIL population of sorghum under favorable and drought stress conditions.

Materials and Methods

Genetic material

A recombinant inbred line population (RIL) of sorghum containing 248 RILs was developed from a cross between Tx436 and 00MN7645 through single seed descent (SSD) method (**Fig 4.22**). Tx436 is a widely used, 3-dwarf (*$dw_1Dw_2dw_3dw_4$*), non stay green, tan-plant (*ppQQ*), pollinator line used in U.S. hybrid seed production. The panicle is semi-open with erect branches at maturity. It is also resistant to anthracnose (*Colletotrichum graminicola*), fusarium head blight (*Fusarium spp*.), leaf blight (*Exserohilium turticium*), and downy mildew (*Peronosclerospora sorghi*) and tolerant to head smut (*Sporisorium holci-sorghi*). This pollinator line has necessary characteristics to produce food-quality grain hybrids with good yield potential and resistance to prevalent biotic stresses. 00MN7645 is a new type of stay green, drought tolerant, pollinator line with red caryopsis, and outstanding yield potential released in 2003. Hybrids produced using 00MN7645 have higher yield and lodging resistance than many check hybrids. The RIL population was crossed to an A-line, ATx3042, a common tester used in sorghum hybrid breeding programs. The resulting population was called recombinant inbred line testcross (RILTC) population. In 2008 and 2009, 188 RILTCs were used for phenotyping under favorable conditions and in 2011, 248 RILTC were used for phenotyping under drought conditions.

Phenotyping for grain yield and drought tolerance

The RILTC populations were phenotyped for grain yield and drought tolerance in a randomized complete block design (RBD) with two replication under favorable and stressed environments: Ottawa (2008 and 2009), Manhattan (2008, 2009, and 2011), Hesston (2008 and 2009), and Hays (2011). One hundred and eighty eight RILTC along with the parents were phenotyped in 2008 and 2009 across 6 environments. These six environments received >500 inches of rainfall in the crop growth period so they were considered as favorable environments. In 2011, 248 RILTC

along with the parents were phenotyped in 2 locations (Manhattan and Hays). In 2011, the rainfall was less than 300 inches and the two locations were considered as drought environments (**Fig. 4.23 and Fig. 4.24**).

The phenotyping was done for grain yield, flowering time (days to 50% anthesis), and stay green measurements. Stay green measurements included relative chlorophyll content, leaf fluorescence (F_v/F_m), and green leaf area visual rating (GLAVS). Minolta SPAD-502 meter was used to measure relative chlorophyll content from 5 tagged plants in a row. These measurements were taken 10, 20, 30, and 40 days after flowering on the flag leaf of the plants in 2008 and 2009. Measurements were taken at the middle of the flag leaf 1 cm from the edge of the leaf lamina. In 2011, chlorophyll content was measured on 2^{nd} and 4^{th} leaves, one at the base of the leaf, one at the middle of the leaf 1 cm from the edge, and one at the tip of the leaf. The average of these measurements was used for the analysis.

Leaf fluorescence was measured using OPTI-Science OS 30p-chlorophyll fluorometer on the tagged plants. The plants were dark adapted for 20 minutes before the measurement using OS 30p-chlorophyll Fluorometer. F_v/F_m indicates maximum quantum yield of photosystem II that is an indicator of tolerance to drought (Li et al., 2006). In addition, green leaf area was visually scored in 2011 on a scale of 1-5, where 1 = completely dead plant and 5 = completely green plant. This score was obtained based on number of green leaves on the plant, size of the leaf, and dry area in the leaf at maturity. These plots were harvested using combine harvester and grain moisture content was recorded for grain yield calculations.

Genotyping, molecular markers, and map construction

DNA extraction of the 248 RILs and their parents were done using a modified CTAB method (Murray and Thompson, 1980). Genotyping was done by genotyping by sequencing (GBS) that identified 8000 segregating SNPs between the parents Tx436 and 00MN7645 (Elshire et al., 2011). GBS is an alternative to complex, expensive protocols for genotyping. It reduces the genome complexity by using restriction enzymes and is a low cost highly reproducible method that can reach genomic regions inaccessible to other sequence capture approaches.

From the 8,000 markers SNPs were selected based on its distribution in the genome, missing value, and chi-square test for segregation distortion to make linkage maps (**Fig.4.6**). The resulting markers were at least 10kb apart based on the physical distance between the markers.

Eight hundred SNPs were used to construct linkage map using JoinMap® (Van Ooijen and Voorrips, 2001). Linkage map was constructed for each chromosome and physical distance was also considered to join the linkage groups together into a same chromosome (**Fig.4.7**). Within each linkage group, the best order of markers was determined according to the physical position of the markers and the genetic distances (cM) were calculated using the Kosambi (1944) function. A plot of physical position of the markers with genetic distance was reported (**Fig.4.8**).

Statistical analysis

The phenotypic data obtained from 2008 and 2009 were combined to do the analysis of variance using PROC GLM. Genotype and genotype x environment interaction were considered as random effects and other effects were fixed effects. Also heritability estimates of the traits were done using PROC MIXED in SAS 9.1. (SAS Institute Inc., 2012). Entry-mean based heritability (h^2) was estimated for each trait using the formula:

$$h^2 = \sigma_g^2 / (\sigma_g^2 + (\sigma_{gl}^2/r) + (\sigma_e^2/rl)),$$

where σ_g^2, σ_{ge}^2, and σ_e^2 are estimated variance components for genotype, genotype by environment, and error, respectively; r is the number of replications; and l is the number of environments. The correlation coefficient between all traits was estimated. In 2011, the drought stress was drastically different for each location resulting in low heritability estimates so the two locations were analyzed separately to detect QTLs.

QTL detection

Composite interval mapping (CIM) was employed in QTL cartographer to detect QTLs in favorable and drought stressed environments. A stepwise regression model 6 was used to select cofactors for CIM. A 10 cM scan window was used for analysis and the LOD statistic was computed at a walk speed of 1 cM. QTLs were detected with an LOD threshold of 2.5 and percentage of variation explained (PVE) by the QTL is reported.

Results

Agronomic performance of RILs and their parents in 2008 and 2009

Analysis of variance showed that genotypes were significantly different for all the traits studied (**Table 4.1**). In 2008 and 2009, the heritability was highest for flowering time (h^2 = 0.78) followed by grain yield (h^2 = 0.65) and chlorophyll content (h^2 = 0.30). Chlorophyll fluorescence (h^2 = 0.22) had the lowest heritability estimates among the traits studied.

In 2008 and 2009, grain yield was positively correlated with chlorophyll content (r = 0.16) and chlorophyll fluorescence (r = 0.18) negatively correlated with flowering time (r = -0.40). Flowering time was positively correlated with chlorophyll content (r =0.12) and chlorophyll fluorescence (r = 0.19) (**Table 4.3**). 00MN7645 flowered earlier than Tx436 in all the locations (**Fig. 4.1 A**). The grain yield was higher for Tx436 than 00MN7645 (**Fig. 4.1 B**). Chlorophyll content readings were not different between the parents (**Fig. 4.4**).

Agronomic performance of 248 RILs and their parents in 2011

The RILs showed significant differences for grain yield, flowering time, chlorophyll content, and stay green visual ratings. However, a combined analysis of the data from two locations showed lower heritability estimates for all the traits (**Table 4.2)**

High correlation between the stay green ratings and the SPAD meter readings were observed at Hays (r = 0.65) and Manhattan (r = 0.49) (**Fig. 4.5**). In Manhattan 2011, flowering time was positively correlated with SPAD (r = 0.16) and GLAVS (r = 0.10). Grain yield per plant and GLAVS was also positively correlated (r = 0.12).

The parents of RILs performed differently for grain yield at Hays and Manhattan in 2011. Tx436 had higher yields in Hays compared to 00MN7645 (**Fig. 4.3**) and 00MN7645 had higher yield than Tx436 in Manhattan (**Fig 4.2**). Tx436 flowered late in both locations compared to 00MN7645 (**Fig. 4.2 and Fig. 4.3**). Stay green ratings were lower for Tx436 at both locations and the differences between them were higher at Hays than Manhattan (**Fig.4.4**).

Linkage map

A total of 800 SNPs were used in the final map construction using JoinMap. The total length of the linkage groups was 2149.3cM; Ch1 (269.3cM), Ch2 (231.9cM), Ch3 (239.1cM), Ch4 (217.0cM), Ch5 (185.7cM), Ch6 (233.3cM), Ch7 (224.8cM), Ch8 (191.8cM), Ch9 (247.6cM),

and Ch10 (108.8cM). The average interval between the loci was 2.68cM. The linkage groups were compared with the physical map of each chromosome (**Fig. 4.9**). The lower recombination rate of the centromere regions is shown with lower slope and chromosome arms had higher recombinations.

QTLs identified under favorable conditions

Grain yield and flowering time QTLs

A major QTL was detected for grain yield on chromosome 8 with an LOD score of 4.5 and a phenotypic variation explained (PVE) value of 13.2 (**Fig. 4.10**). A major flowering time QTL was detected on chromosome 9 (**Fig. 4.11**) with an LOD of 5.0 and a PVE value of 12.8 (**Table4.4**).

QTLs identified under drought conditions

Grain yield and flowering time QTLs

Three major QTLs were identified for grain yield: chromosome 1 (LOD = 3.3), chromosome 6 (LOD = 2.6), and chromosome 8 (LOD = 3.6). Their explained phenotypic variation was 5.4, 5.4, and 7.2, respectively. QTL on chromosome 1 was located between 183.07cM and 203.07cM (**Fig. 4.17**). QTL on chromosome 6 was between 40.7cM and 58.3cM (**Fig. 4.18**) and QTL on chromosome 8 was between 103.0cM and 106cM (**Fig.4.12**).

Two QTLs were detected on chromosome 1 for flowering time under drought conditions. A QTL was located between 35.4cM and 52.9cM with an LOD score of 3.0 and a PVE value of 4.9 (**Fig. 4.13**). Another QTL was located on the same chromosome between 207.3cM and 228.2cM with an LOD score of 3.1 and a PVE value of 5.4 (**Fig. 4.19**) (**Table. 4.5**).

Stay green QTLs

We identified six QTLs for stay green under drought conditions. A QTL was detected on chromosome 5, between 104.2cM and 122.4cM (LOD = 2.6, PVE = 4.3) for chlorophyll content from Manhattan 2011 data (**Fig. 4.20**). Another QTL was detected on chromosome 6 between 19.6cM and 31.0cM (LOD = 3.0, PVE = 5.9) from the Manhattan 2011 data. For chlorophyll fluorescence three QTLs were identified, on chromosomes 4, 7, and 10. The QTL on chromosome 4 located between 103.2cM and 134.75cM explained 4.51 of the variation in

chlorophyll fluorescence (**Fig.4.15**). Another QTL on chromosome 7 (LOD = 4.0) between 187.6cM and 204.2cM explained 9.44 of the variation in the trait (**Fig. 4.21**) and a QTL on chromosome 10 (LOD = 2.95) between 52.5cM and 68.0cM explained 4.9 of the variation in the trait (**Fig. 4.16**). A QTL was identified for visual rating on chromosome 4 between 103.2cM and 134.75cM (LOD = 2.8) that explained 4.51 variation in the trait (**Fig.4.14**).

Discussion

Drought stress is specific to the crop, growth period of the crop, and environment conditions. In the present study grain yield and flowering time was studied under favorable and drought stressed conditions. These agronomic traits showed high heritability under favorable conditions but under drought conditions the heritability was lower in the RILTC population. The parents didn't differ for stay green measurements under favorable conditions but under drought conditions stay green measurements showed differences (**Fig. 4.4**). Stay green was positively correlated with grain yield under favorable conditions indicating that stay green genotypes yield higher also under favorable conditions. The QTLs detected for grain yield didn't coincide with stay green QTLs detected under drought conditions indicating possibility of different genetic mechanisms for grain yield and stay green. QTLs for grain yield were detected in chromosomes 1, 6, and 8 under drought conditions. Under favorable conditions the QTL for grain yield was identified on chromosome 9. This indicates that different mechanisms determine grain yield under favorable and drought stress conditions.

Earlier studies have reported QTLs for stay green in chromosomes 2, 3, 4, 6, 8, and 10 (Crasta et al., 1999; Kebede et al., 2001; Sanchez et al., 2002; Subudhi et al., 2000; Subudhi et al., 1999; Xu et al., 2000). In our study, we detected. QTLs for stay green under drought stressed environments in chromosomes 4, 5, 6, 7, and 10. The QTLs detected on chromosome 4 for GLAVS (122.8cM - 129.3cM) and chlorophyll fluorescence (103.2cM - 134.75cM) overlapped. These QTLs overlapped with a QTL (112.0cM) detected on chromosome 4 for stay green from an earlier study (Sabadin et al., 2012). QTL detected for chlorophyll content on chromosome 5 (104.2cM - 122.4cM) was near the QTL for stay green (98.3cM) from the earlier study and QTL for chlorophyll content overlapped with a QTL detected on chromosome 6 at 30cM (Sabadin et al., 2012). Transgressive segregation was present for stay green similar to earlier studies that mapped QTLs for drought tolerance in sorghum (Crasta et al., 1999; Kebede et al., 2001).

In our study, we detected QTLs for flowering time in chromosome 8 under favorable conditions and two QTLs on chromosome 1 under drought stress conditions. Flowering time and grain yield was negatively correlated under favorable conditions indicating that early flowering genotypes yielded higher. Several flowering genes have been identified in sorghum Ma_1, Ma_2, Ma_3, Ma_4, Ma_5, and Ma_6 (Childs et al., 1997). It is a key trait for adaptation of the plant to its environmental conditions. We identified two QTLs for flowering time on chromosome 1 (35.4cM - 52.9cM and 207.3cM - 228.2cM) under drought conditions. Earlier studies using photoperiod sensitive and photoperiod insensitive lines of sorghum has identified several QTLs for flowering time (El Mannai et al., 2012). Our experimental material RILTC didn't show drastic variation in flowering time because of the common genetic background created by crossing RILs with ATx3042 an early flowering line.

Most of the QTL mapping studies for drought tolerance used RILs or NILs for phenotyping. In our study, RILs in a hybrid background (RILTC) was phenotyped and RIL was genotyped. RILs are normally homozygous lines with some possible heterozygosity, but RILTC are RILs crossed to a commonly used A-line tester. In other words, the phenotyping was done on potential hybrids, and we found several lines that are higher yielding than the checks Pioneer 84G62, Tx3042/Tx2737, and CS1114/R45. The idea of phenotyping RILTC instead of RILs was to reduce the confounding effects of flowering time on grain yield and stay green measurements.

Traits that are governed by single genes or fewer genes have low genotyping by environment interaction and higher stability across environment and thus higher heritability (eg. flowering time and resistance to diseases). But complex traits that are governed by multiple genes and multiple QTLs interact between them and the environment. Hence, the expression of these traits varies across environments and show lower heritability estimates. The potential traits for selecting drought tolerance should be positively associated with grain yield and must have higher heritability estimates than the grain yield itself. These trait measurements should be nondestructive, rapid, accurate, and inexpensive (Tuberosa, 2012). Previous studies also reported lower heritability estimates for stay green traits using hand held instruments. These low heritability traits chlorophyll content and chlorophyll fluorescence are point measurements and doesn't capture the dynamics of whole plant response to drought stress. To improve phenotyping accuracy and efficiency for drought tolerance, integrated platforms that have high-throughput phenotyping ability using near-infrared (NIR) spectroscopy on agricultural harvesters, canopy

spectral reflectance, infra red thermography, magnetic resonance imaging, position emission tomography, and nuclear magnetic resonance are needed (Mir et al., 2012). These precision phenotyping tools with advancements in genotyping through next generation sequencing technologies and genetic designs can increase the speed of gene discovery for drought tolerance in plants.

Conclusions

QTLs were detected for grain yield, flowering time, and stay green in this present study. QTLs for stay green were detected under drought conditions only, and they overlapped with some of the earlier identified QTLs for stay green. To dissect drought tolerance into gene level, large scale coordinated approaches that include integrated phenotyping platforms, high density genotyping, and genetic designs are needed.

Tables and Figures

Table 4.1. Analysis of variance for grain yield, flowering time, chlorophyll content, and chlorophyll fluorescence from Hesston, Manhattan, and Ottawa in 2008 and 2009.

		Mean squares			
Source of Variation	**df**	**Grain yield**	**Flowering time**	**Chlorophyll content**	**Chlorophyll fluorescence**
Environment	5	933.08	8423.00	3444.48	0.106059
Replication (Environment)	6	20.39	75.14	456.47	0.030241
Genotype	189	5.86***	19.14***	24.84**	0.002237**
Genotype × Environment	942	1.52	3.71	17.62	0.001828
Error	1043	1.34	3.27	17.52	0.001709

***,Significant at $\alpha = 0.001$; **,significant at $\alpha = 0.01$, and *,significant at $\alpha = 0.05$.

Table 4.2. Analysis of variance for grain yield, flowering time, chlorophyll content, stay green rating, and chlorophyll fluorescence in the year 2011.

Source of variation	df	Mean squares				
		Grain yield	Flowering time	Chlorophyll content	Stay green rating	Chlorophyll fluorescence
Environment	1	194.69	42314.6	854.82	1.54	0.0279
Rep(Environment)	2	54.85	29.41	5979.13	5.06	0.1259
Genotype	244	1.64***	16.03***	193.26*	0.82*	0.0155
Genotype × Environment	244	1.45*	11.40**	191.52*	0.73*	0.0158
Error	473	1.38	6.12	154.21	0.59	0.0169

***, Significant at $\alpha = 0.001$: **, significant at $\alpha = 0.01$; and *, significant at $\alpha = 0.05$.

Table 4.3 Correlation between grain yield, flowering time, chlorophyll content, and chlorophyll fluorescence under favorable conditions across 188 recombinant inbred lines and their parents.

Traits	Grain yield	Flowering time	Chlorophyll content	Chlorophyll fluorescence
Grain yield		-0.40***	0.16*	0.18*
Flowering time		-	0.12*	0.19***
Chlorophyll content			-	0.20**

***,Significant at $\alpha = 0.001$; **,significant at $\alpha = 0.01$, and *,significant at $\alpha = 0.05$.

Table 4.4. QTLs identified for grain yield and flowering time through composite interval mapping using 800 SNPs. One hundred and eighty eight RILs were phenotyped under favorable conditions in 2008 and 2009.

Trait	Chr	Position (cM)	Closest marker(s)	LOD	PVE	Additive effect	Contributing parent
Grain yield	8	14.3 - 33.2	S8_5490811 - S8_9689857	4.5	13.2	0.012	00MN7645
Flowering time	9	6.8 - 27.3	S9_561031 - S9_1547975	5.0	12.8	0.50	Tx436

Chr = Chromosome, cM = centimorgan (kosambi), LOD = logarithm of odds, and PVE = phenotypic variance explained

Table 4.5. QTLs identified through composite interval mapping using 800 SNPs on 248 RILs under drought conditions in 2011.

Trait	Environment	Chr	Position (cM)	Closest marker(s)	LOD	PVE	Additive effect	Contributing parent
Grain yield	Hays	8	103 - 106	S8_51181394 - S8_51372548	3.6	7.2	0.11	Tx436
	Manhattan	1	183.7 - 203.7	S1_57773312 - S1_60383946	3.3	5.4	0.04	00MN7645
		6	40.7 - 58.3	S6_1980781 - S6_2631628	2.6	5.7	0.03	00MN7645
Flowering time	Hays	1	35.4 - 52.9	S1_1965838 - S1_4067572	3.0	4.9	0.68	00MN7645
	Manhattan	1	207.3 - 228.2	S1_61406226 - S1_65253574	3.1	5.4	0.56	Tx436
				Stay green				
Visual rating	Hays	4	122.8 - 129.3	S4_62557400 - S4_62775392	2.7	4.7	0.15	00MN7645
Chlorophyll content	Hays	5	104.2 - 122.4	S5_10950864 - S5_19539291	2.6	4.3	1.68	Tx436
	Manhattan	6	19.6 - 31.0	S6_1279361 - S6_1486757	3.0	5.9	1.17	Tx436
Chlorophyll fluorescence	Hays	4	103.2 - 134.75	S4_13602506 - S4_62725094	2.8	4.5	0.12	00MN7645
		10	52.5 - 68.0	S10_5960530 - S10_9250692	2.9	4.9	0.12	Tx436
	Manhattan	7	187.6 - 204.2	S7_58198679 - S7_62799938	4.0	9.4	0.04	Tx436

Chr = Chromosome, cM = centimorgan (Kosambi), LOD = logarithm of odds, and PVE = phenotypic variance explained

Figure 4.1 Histogram of flowering time and grain yield of 188 RILs and their parents in 2008 and 2009. (A) Flowering time and (B) Grain yield per plant.

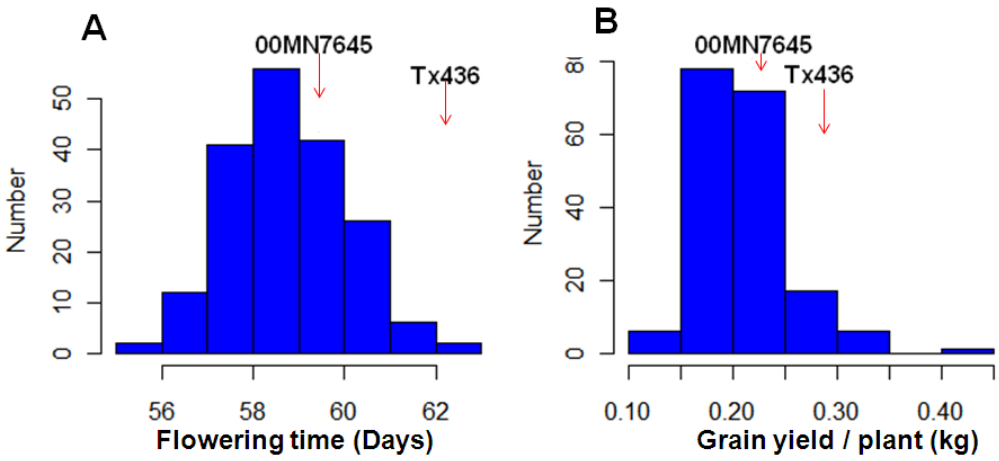


Figure 4.2 Histogram of flowering time and grain yield of 248 RILs and their parents in Manhattan 2011. (A) Flowering time and (B) grain yield per plant.

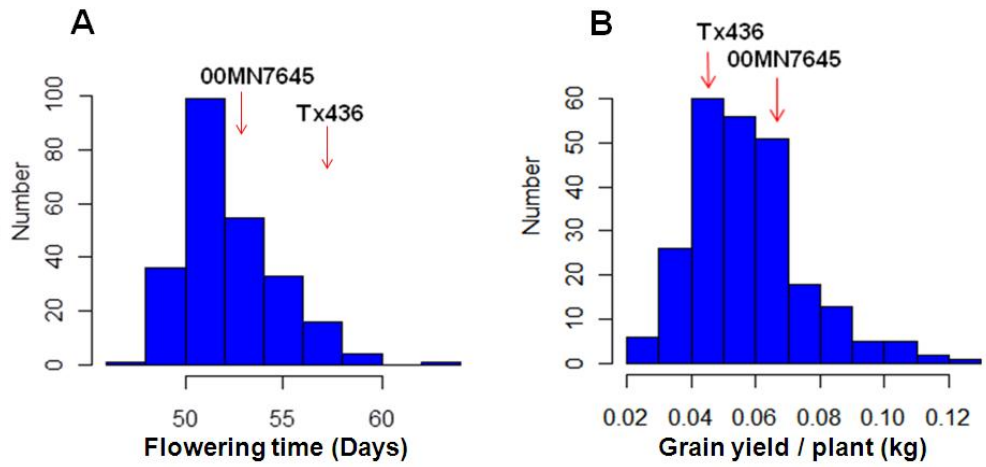


Figure 4.3 Histogram of flowering time and grain yield of 248 RILs and their parents in Hays 2011. (A) Flowering time and (B) grain yield per plant.

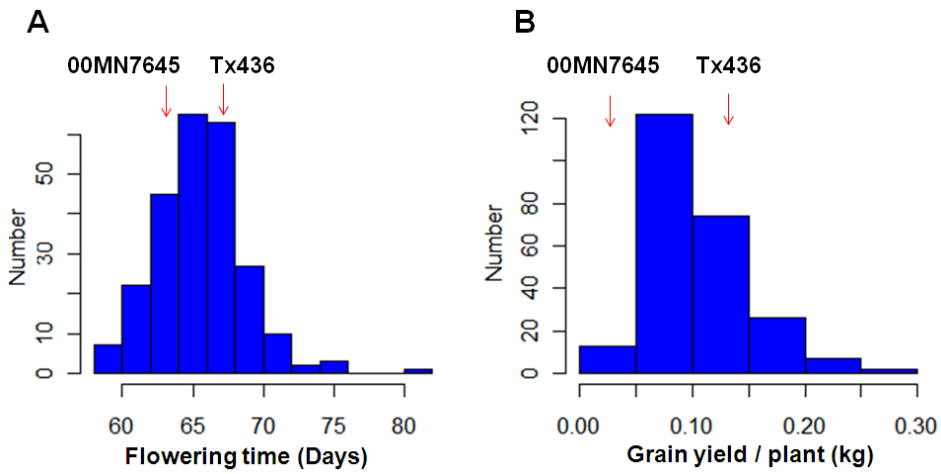


Figure 4.4 Histogram of stay green measurements (A) SPAD meter reading of 188 RILs and their parents in 2008 and 2009, (B) GLAVS of 248 RILS and their parents in Manhattan 2011, and (C) GLAVS of 248 RILS and their parents in Hays 2011.

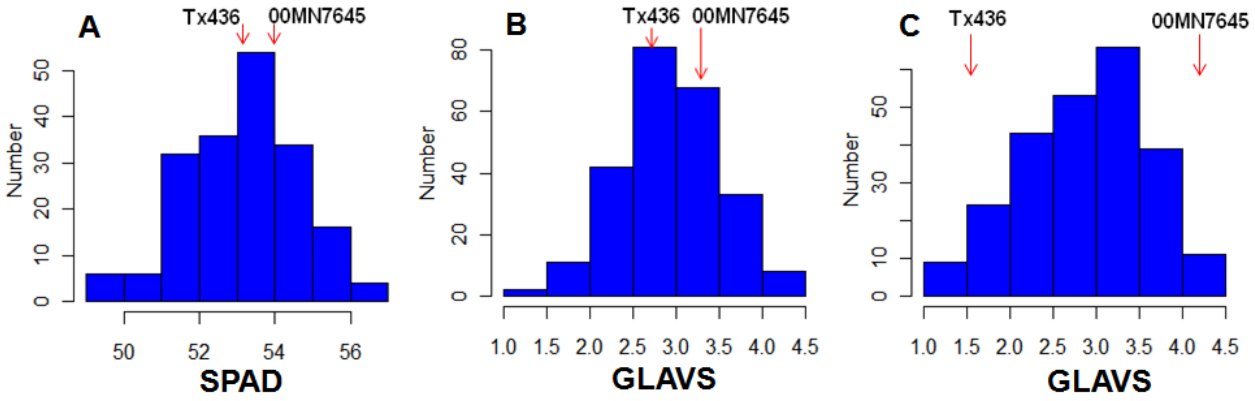


Figure 4.5 Correlations between green leaf area visual scoring (GLAVS) and chlorophyll content (SPAD meter readings) of 248 RILs and their parents under drought conditions in 2011(A) Manhattan (B) Hays .

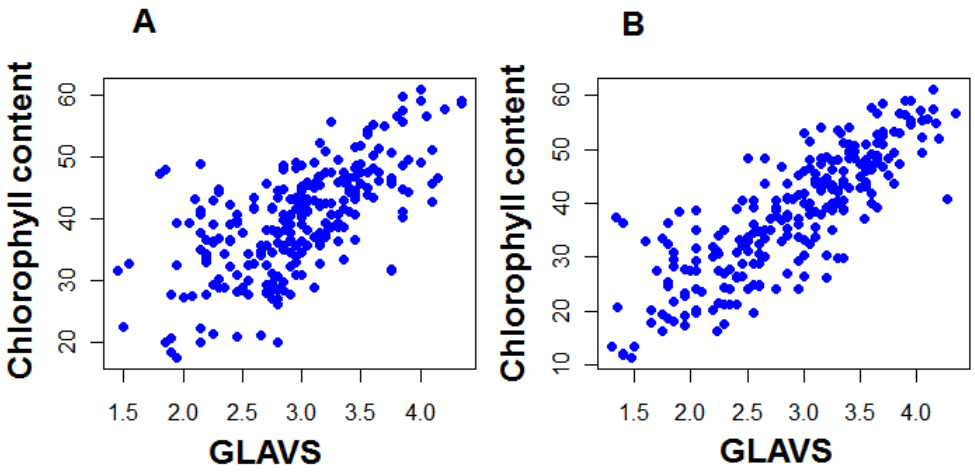


Figure 4.6 SNP distributions across 10 chromosomes in sorghum. Black dots represent the sequenced part of chromosomes, blue dots represent segregating SNPs in the 248 RILs, and red dots represent the SNPs used for linkage map construction.

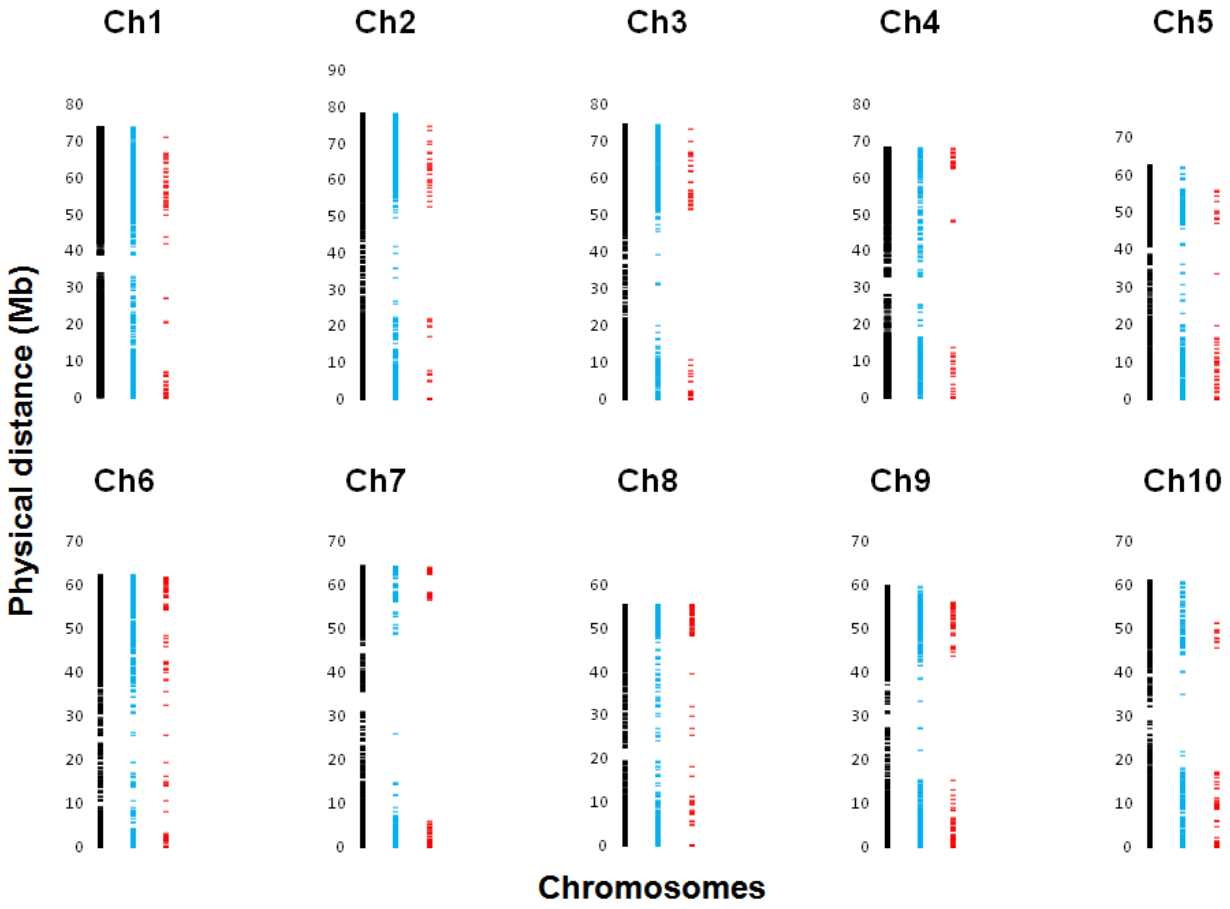


Figure 4.7 Linkage map constructed using 800 SNPs in sorghum: Linkage groups for chromosome 1 (C1) to chromosome 5 (C5).

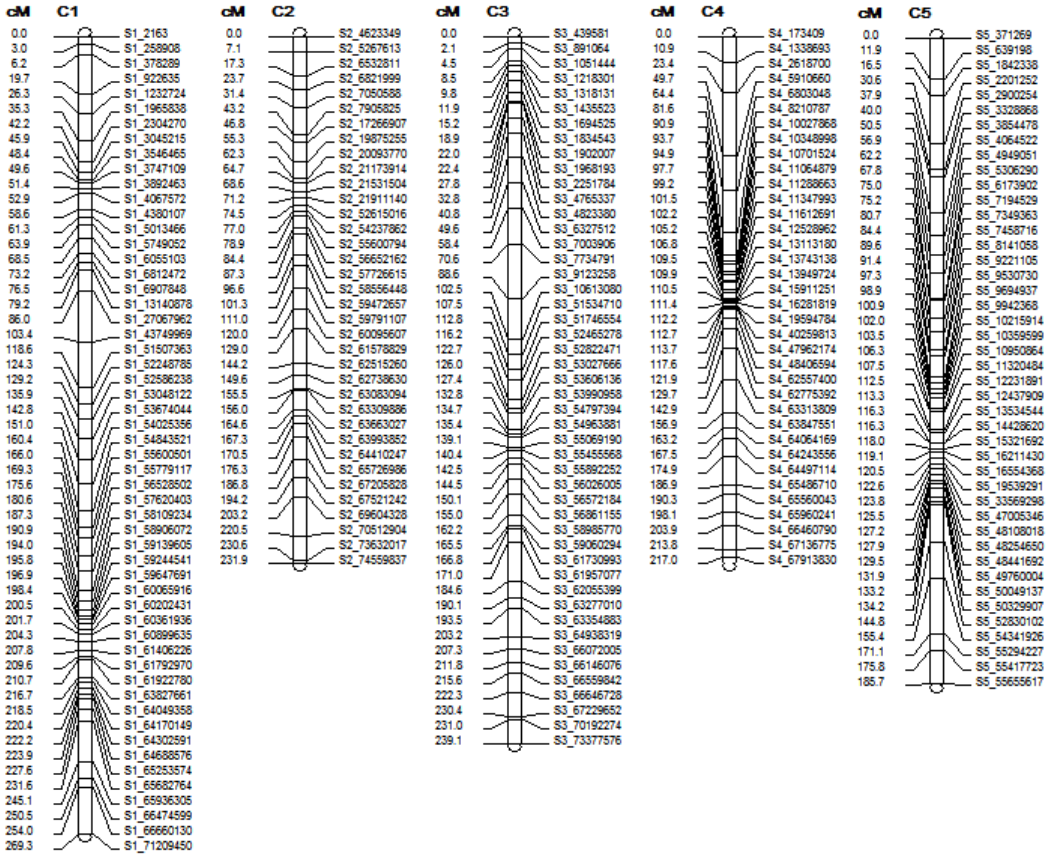


Figure 4.8. Linkage map constructed using 800 SNPs in sorghum. Linkage groups for chromosome 6 (C6) to chromosome 10 (C10).

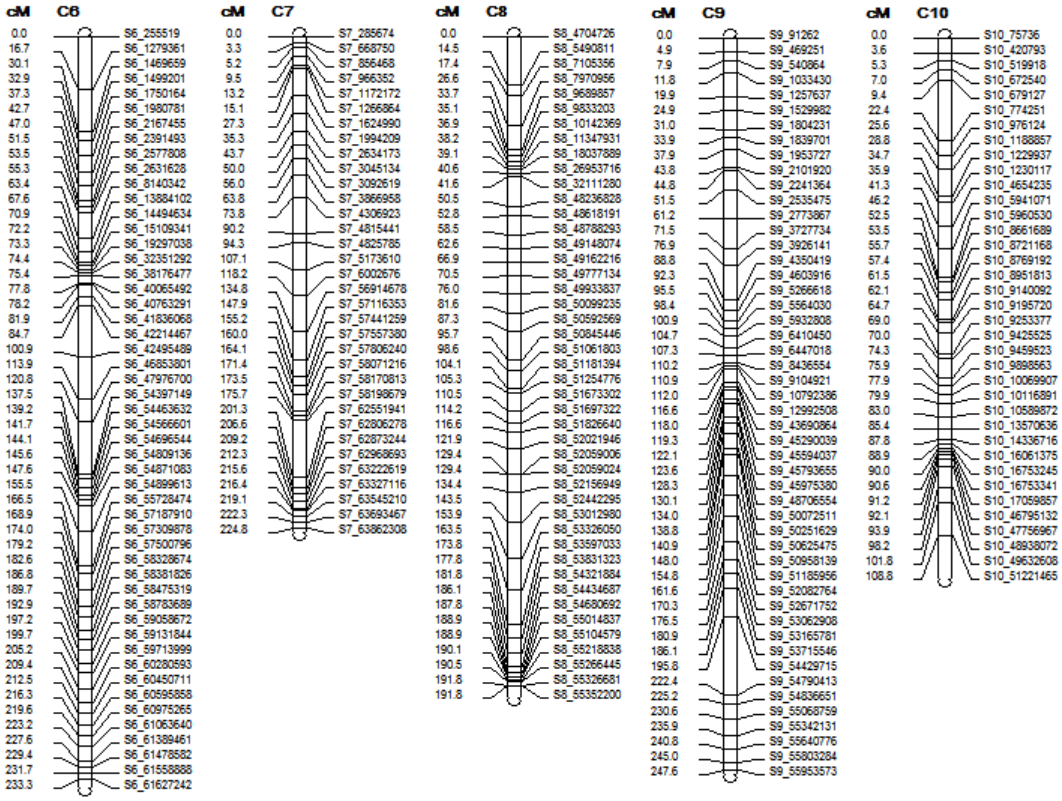


Figure 4.9. Comparison of physical positions and genetic distance of ten sorghum chromosomes. Centromere region in each chromosome has lower recombination rate and thus smaller increase in genetic distance than either end of the chromosome.

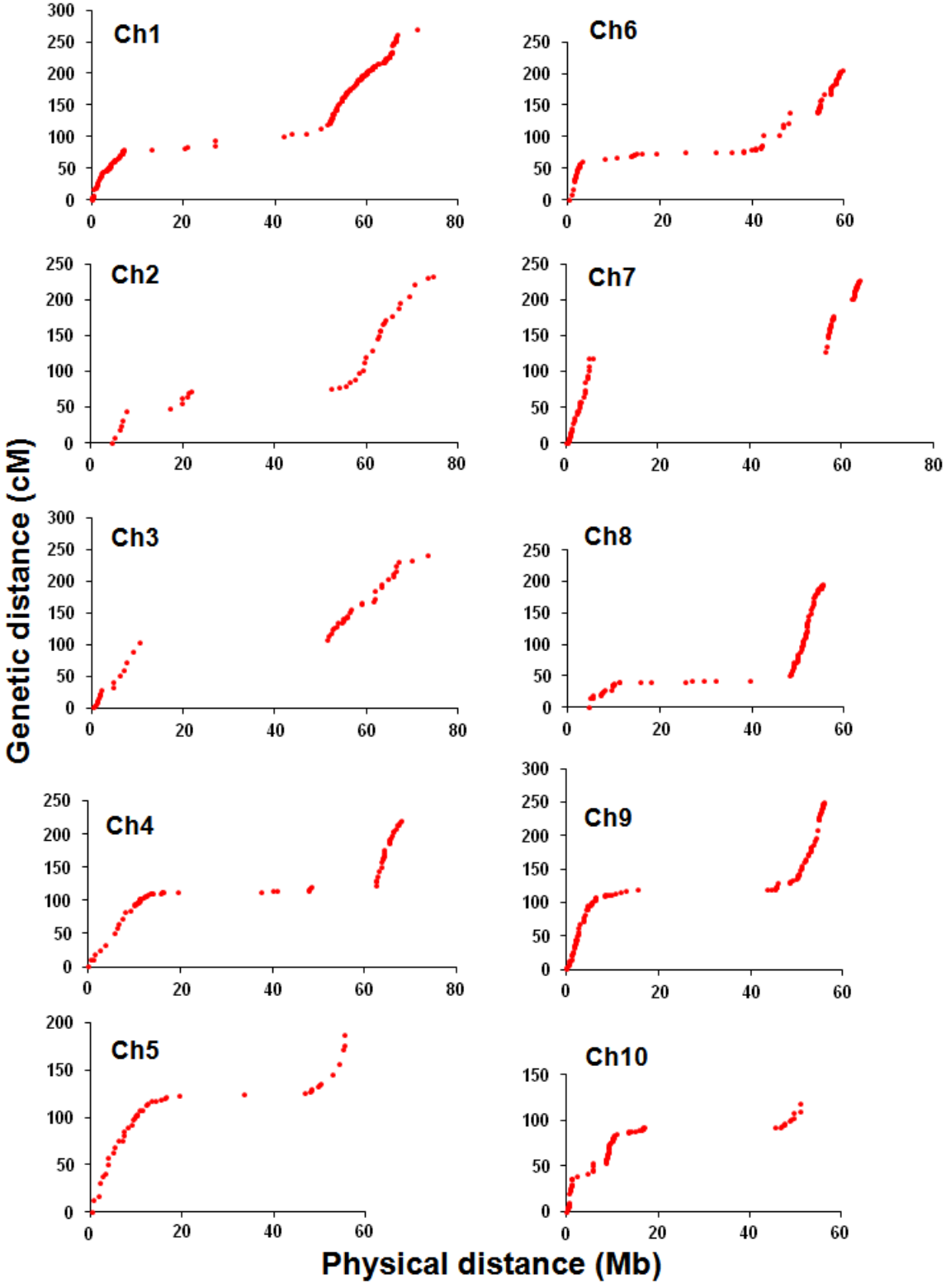


Figure 4.10 QTL identified for grain yield in chromosome 8 under favorable conditions using 188 RILs in 2008-09.

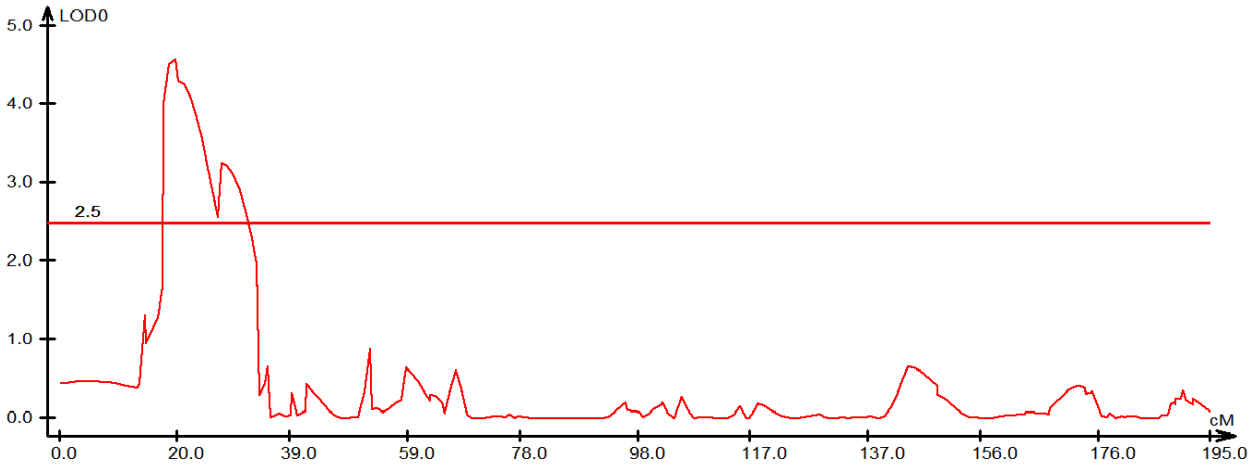


Figure 4.11. QTL identified for flowering time in chromosome 9 under favorable conditions using 188 RILs in 2008-09.

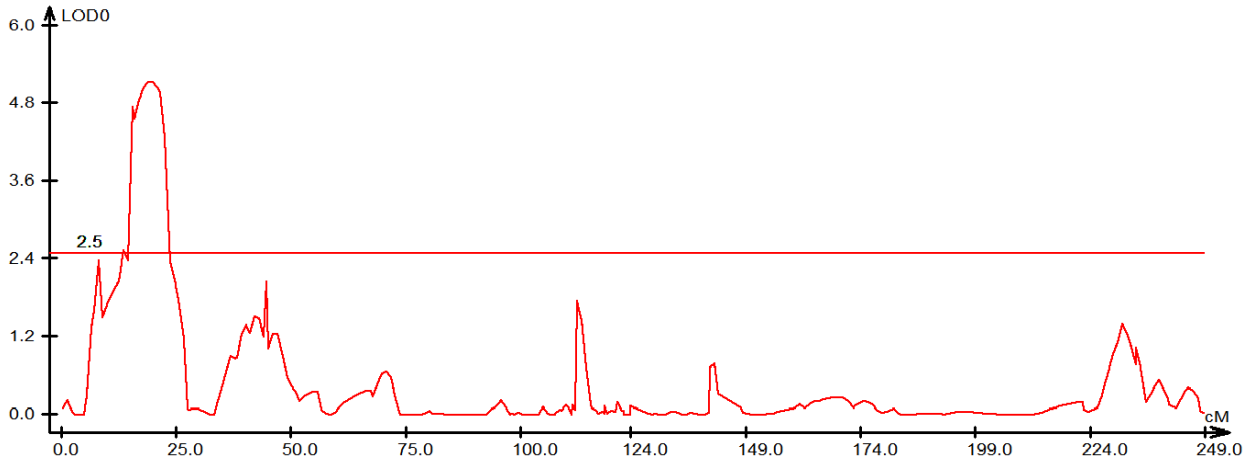


Figure 4.12 QTL for grain yield in Chromosome 8 under drought conditions using 248 RILs in Hays, 2011

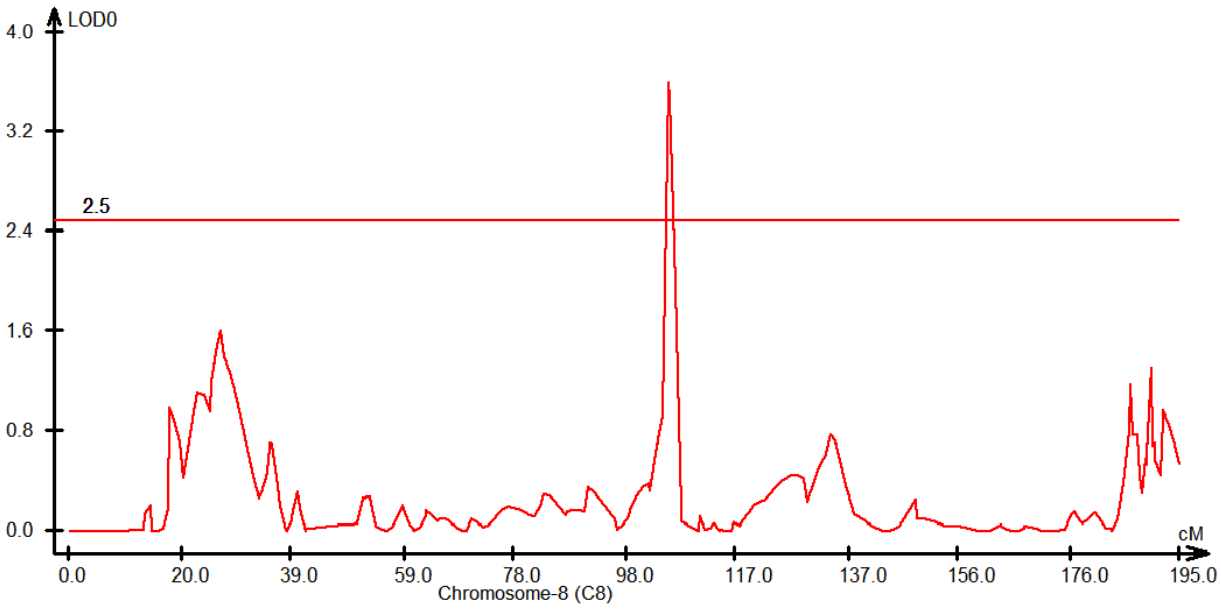


Figure 4.13 QTL identified for flowering time in chromosome 1 under drought conditions using 248RILs in Hays, 2011

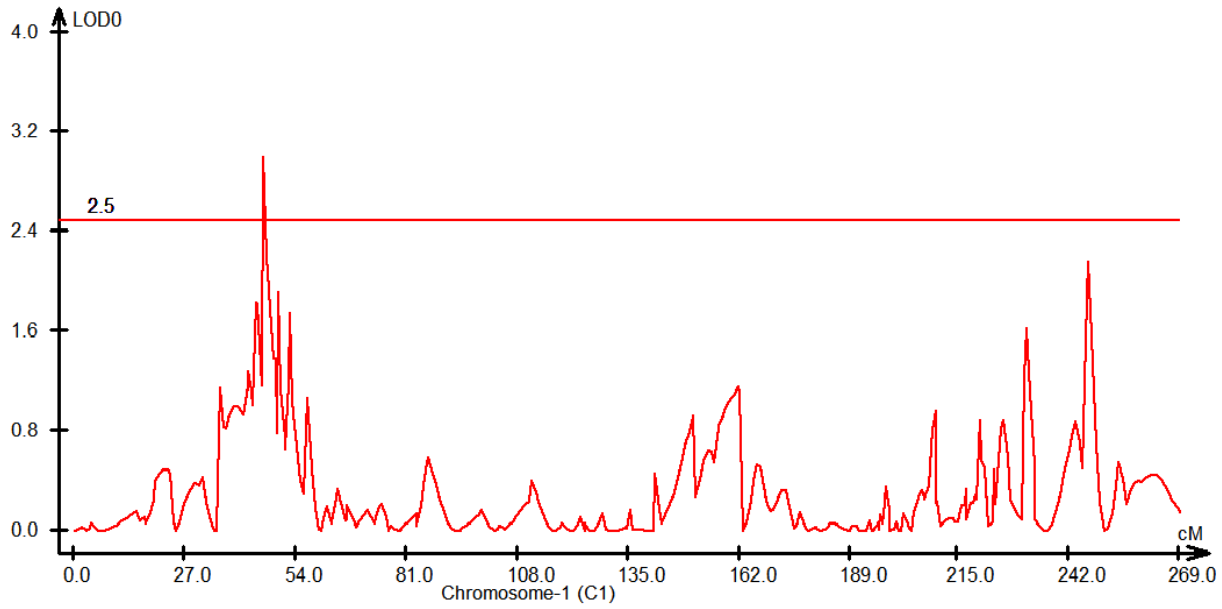


Figure 4.14 QTL for GLAVS in Chromosome 4 under drought conditions using 248 RILs in Hays, 2011.

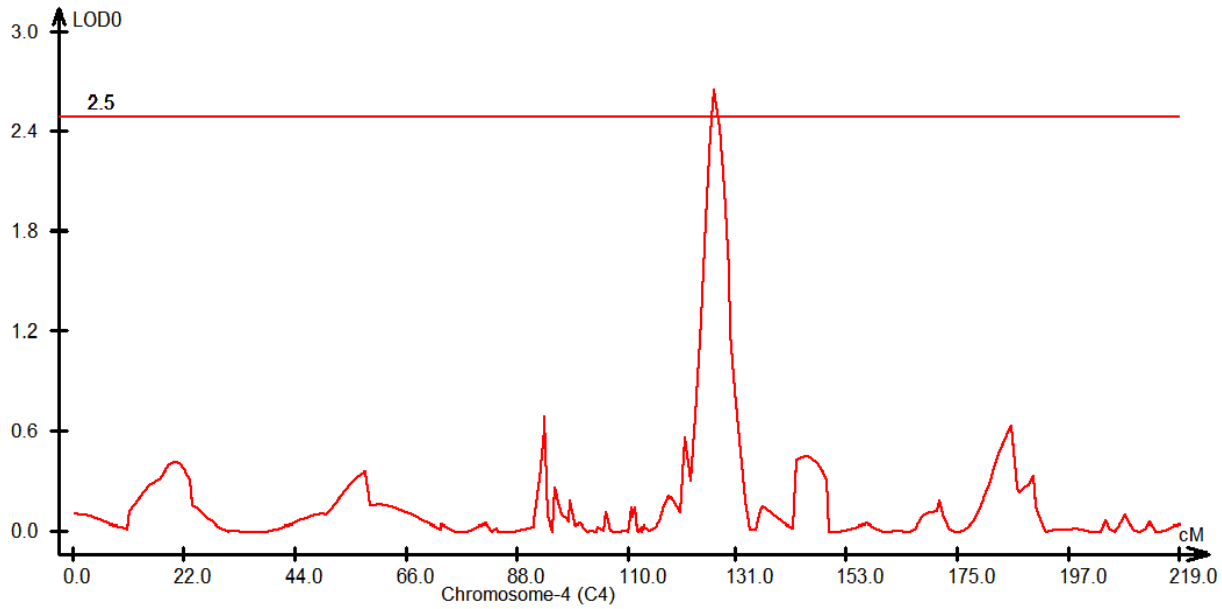


Figure 4.15 QTL for chlorophyll fluorescence in Chromosome 4 under drought conditions using 248 RILs in Hays, 2011.

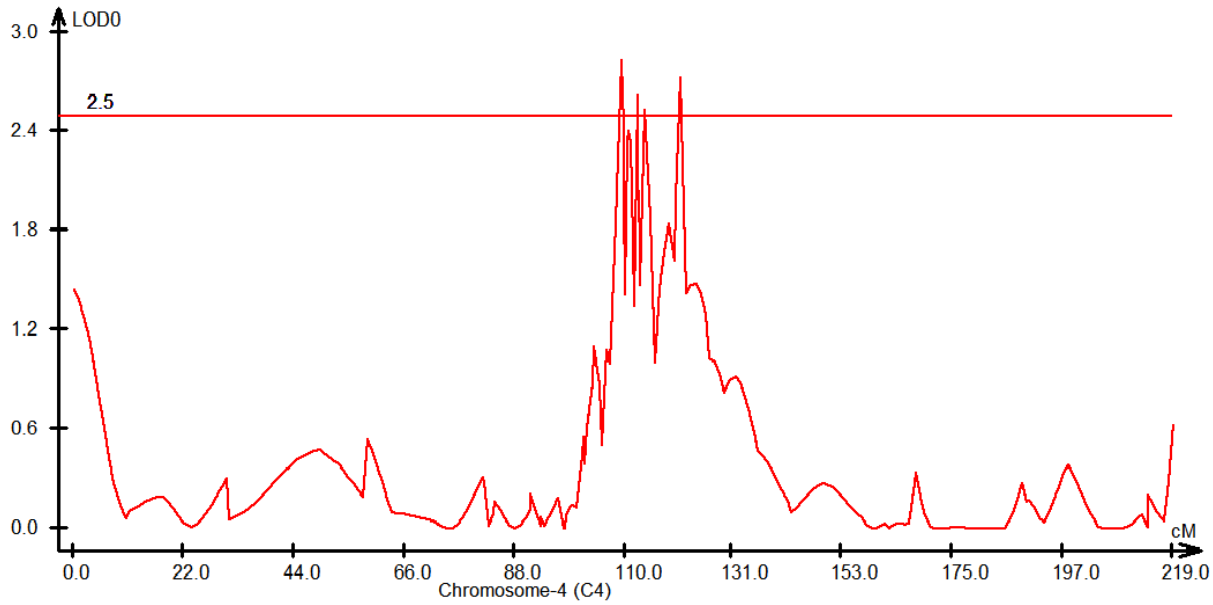


Figure 4.16 QTL for chlorophyll fluorescence ratings in Chromosome 10 under drought condition using 248 RILs in Hays, 2011.

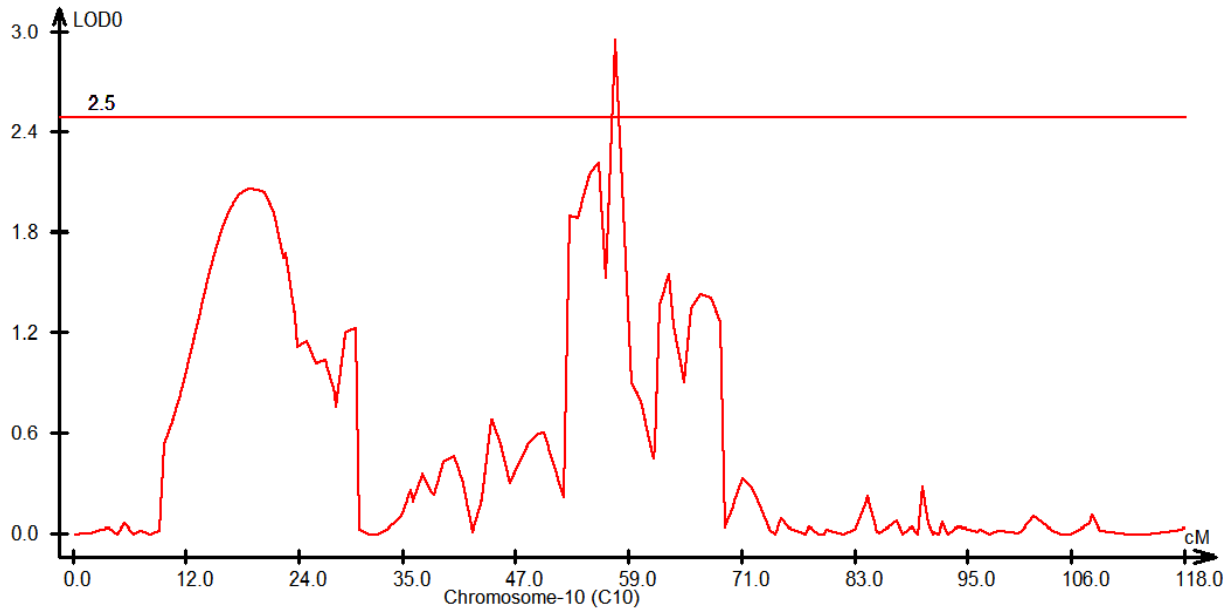


Figure 4.17 QTL for grain yield in chromosome 1 under drought conditions using 248 RILs in Manhattan, 2011.

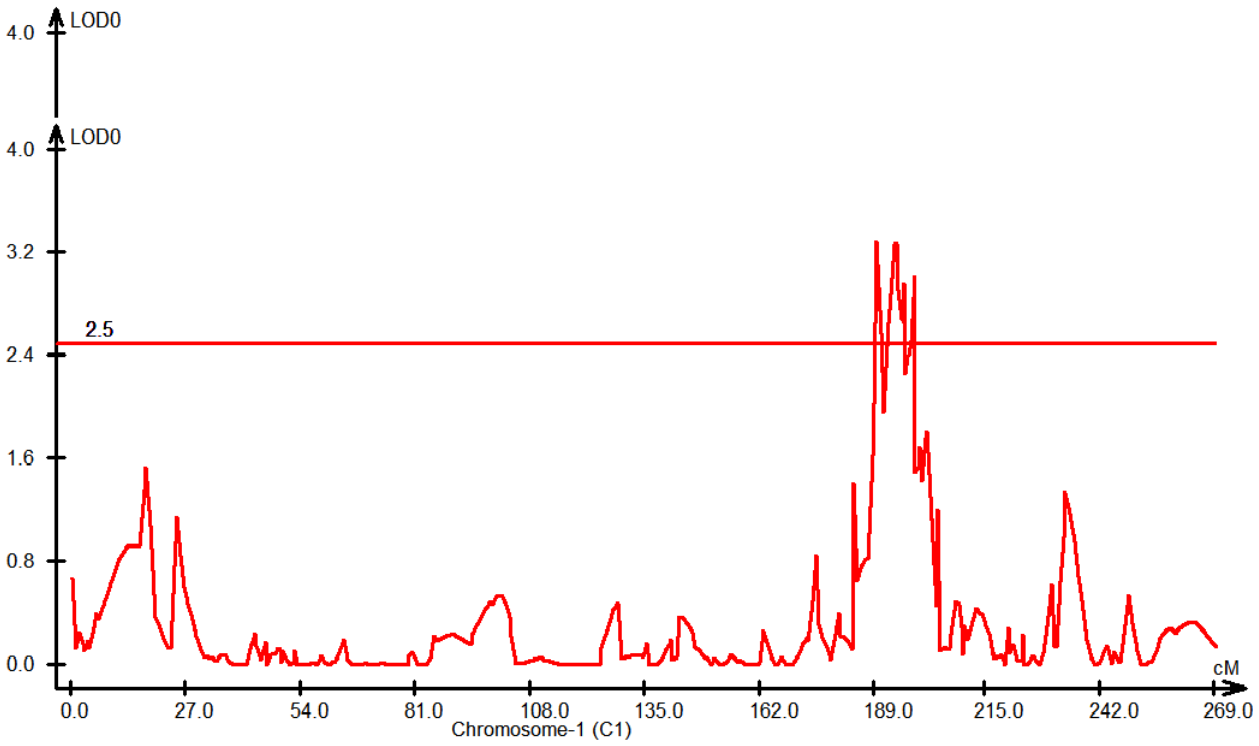


Figure 4.18 QTL for grain yield in chromosome 6 under drought conditions using 248 RILs in Manhattan, 2011.

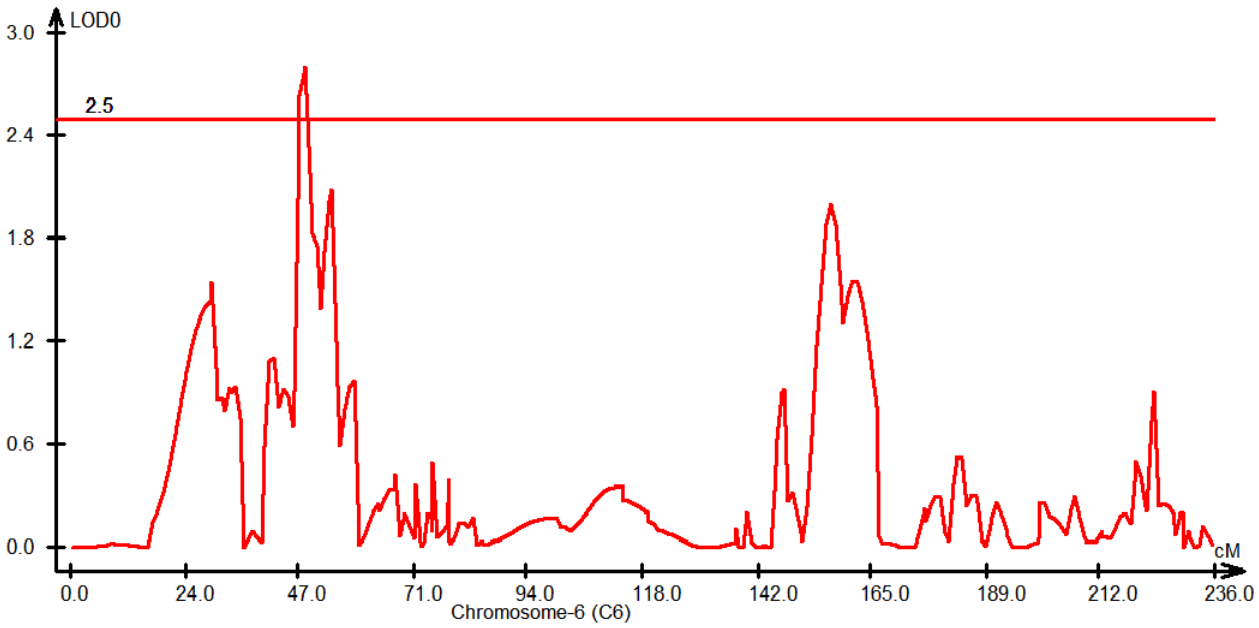


Figure 4.19 QTL for flowering time in chromosome 1 under drought conditions using 248 RILs in Manhattan, 2011.

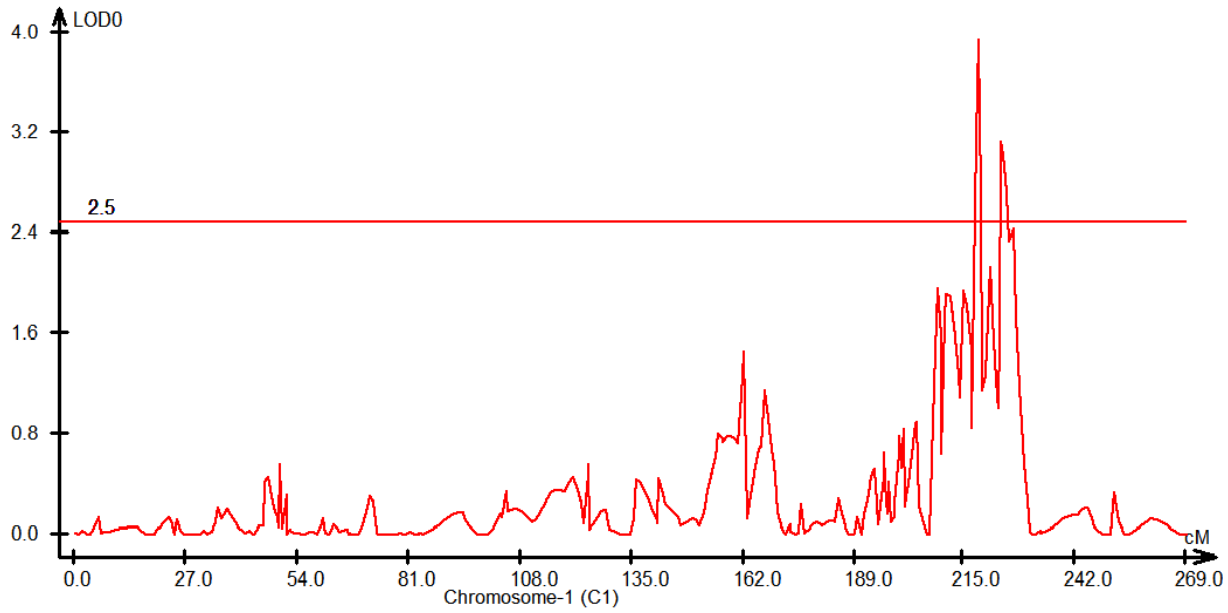


Figure 4.20 QTL for chlorophyll content in chromosome 5 under drought conditions using 248 RILs in Manhattan, 2011

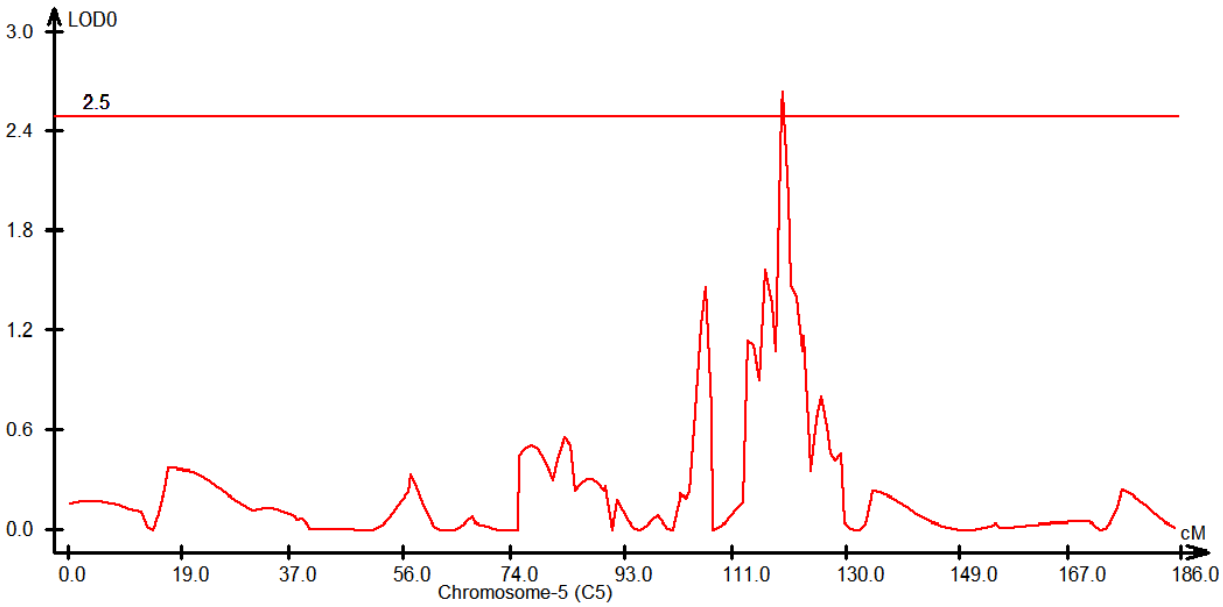


Figure 4.21 QTL for chlorophyll fluorescence in chromosome 6 under drought conditions using 248 RILs in Manhattan, 2011.

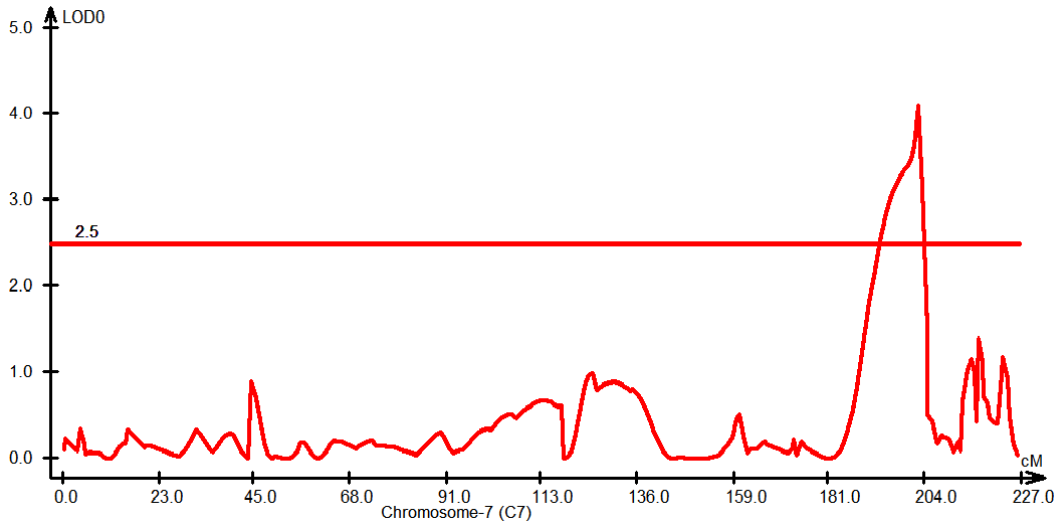
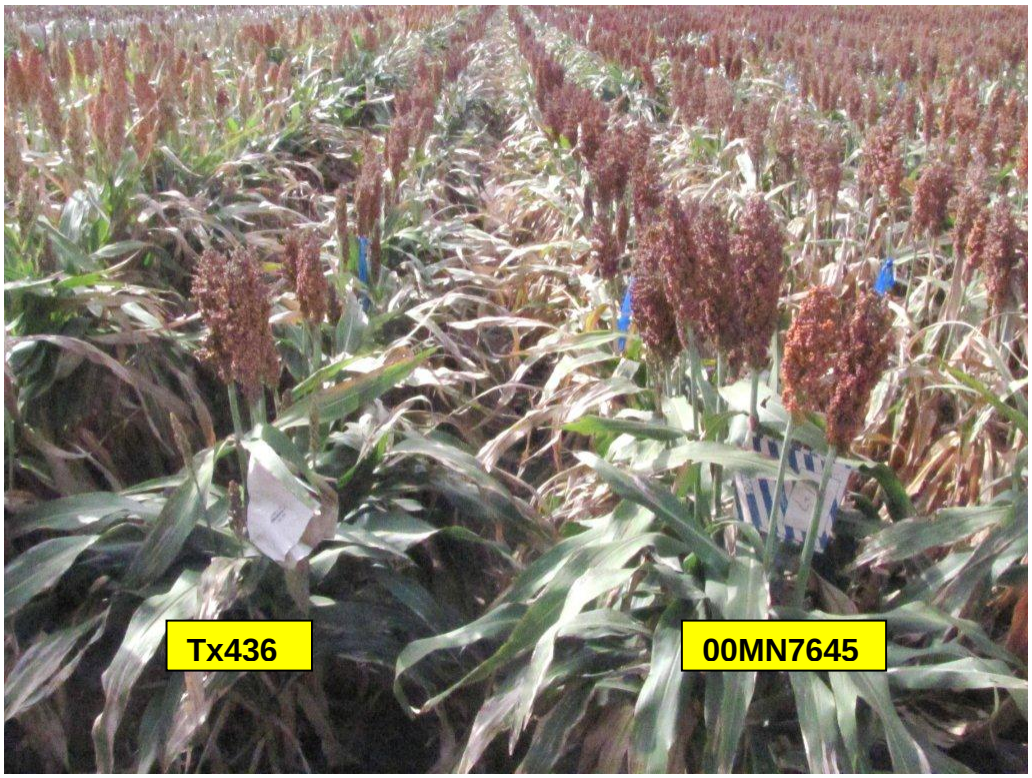


Figure 4.22 Parents of recombinant inbred lines used in the study: Tx436 and 00MN7645

Figure 4.23 Sorghum recombinant inbreds and their parents in the field in Hays, 2011

Figure 4.24 Tx436 and 00MN7645 in the hybrid background (RILTC) at maturity in Manhattan, 2011.



References

Bartels, D., and R. Sunkar. 2005. Drought and Salt Tolerance in Plants. Crit. Rev. Plant Sci. 24:23-58.

Blum, A. 2011. Drought resistance-is it really a complex trait? Funct. Pl. Biol. 38:753-757.

Childs, K.L., F.R. Miller, M.M. Cordonnier-Pratt, L.H. Pratt, P.W. Morgan, and J.E. Mullet. 1997. The Sorghum Photoperiod Sensitivity Gene, Ma3, Encodes a Phytochrome B. Plant physiol. 113:611-619.

Collins, N.C., F.o. Tardieu, and R. Tuberosa. 2008. Quantitative Trait Loci and Crop Performance under Abiotic Stress: Where Do We Stand? Plant physiol. 147:469-486.

Crasta, O.R., W.W. Xu, D.T. Rosenow, J. Mullet, and H.T. Nguyen. 1999. Mapping of post-flowering drought resistance traits in grain sorghum: association between QTLs influencing premature senescence and maturity. Mol Gen Genet. 262:579-588.

El Mannai, Y., T. Shehzad, and K. Okuno. 2012. Mapping of QTLs underlying flowering time in sorghum [Sorghum bicolor (L.) Moench]. Breed Sci. 62:151-159.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6:e19379.

Farooq, M., A. Wahid, N. Kobayashi, D. Fujita, S.M.A. Basra, E. Lichtfouse, M. Navarrete, P. Debaeke, S. Véronique, and C. Alberola. 2009. Plant Drought Stress: Effects, Mechanisms and Management Sustainable Agriculture:153-188.

Garud, T.B., N. Seetharama, S.P. Deshpande, I. Syed, M.S. Dadke, and S. Ismail. 2002. Usefulness of non-senescent parents for charcoal rot resistance breeding in sorghum. International Sorghum and Millets Newsletter 43:63-65.

Jordan, D.R., C.H. Hunt, A.W. Cruickshank, A.K. Borrell, and R.G. Henzell. 2012. The Relationship Between the Stay-Green Trait and Grain Yield in Elite Sorghum Hybrids Grown in a Range of Environments. Crop Sci. 52:1153-1161.

Kebede, H., P.K. Subudhi, D.T. Rosenow, and H.T. Nguyen. 2001. Quantitative trait loci influencing drought tolerance in grain sorghum (Sorghum bicolor L. Moench). Theor. Appl. Genet. 103:266-276.

Li, R.-h., P.-g. Guo, B. Michael, G. Stefania, and C. Salvatore. 2006. Evaluation of Chlorophyll Content and Fluorescence Parameters as Indicators of Drought Tolerance in Barley. Agric. Sci. in China 5:751-757.

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. Theor. Appl. Genet. 123:169-191.

Mir, R., M. Zaman-Allah, N. Sreenivasulu, R. Trethowan, and R. Varshney. 2012. Integrated genomics, physiology and breeding approaches for improving drought tolerance in crops. TAG Theor.Appl. Genet. 125:625-645.

Murray, M.G., and W.F. Thompson. 1980. Rapid Isolation of High Molecular-Weight Plant DNA. Nucleic Acids Res 8:4321-4325.

Passioura, J. 2006. Increasing crop productivity when water is scarce from breeding to field management. Agric. Water Management 80:176-196.

Sabadin, P., M. Malosetti, M. Boer, F. Tardin, F. Santos, C. Guimarães, R. Gomide, C. Andrade, P. Albuquerque, F. Caniato, M. Mollinari, G. Margarido, B. Oliveira, R. Schaffert, A. Garcia, F. van Eeuwijk, and J. Magalhaes. 2012. Studying the genetic basis of drought tolerance in sorghum by managed stress trials and adjustments for phenological and plant height differences. Theor. Appl. Genet.124:1389-1402.

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

Subudhi, P.K., D.T. Rosenow, and H.T. Nguyen. 2000. Quantitative trait loci for the stay green trait in sorghum (Soughum bicolor L. Moench): consistency across genetic backgrounds and environments. Theor.Appl. Genet.101:733-741.

Subudhi, P.K., G.B. Magpantay, D.T. Rosenow, and H.T. Nguyen. 1999. Mapping and marker-assisted selection to improve the stay-green trait for drought tolerance in sorghum. In: O. Ito, J. O'Toole and B. Hardy (Eds.) Genetic Improvement of Rice for Water Limited Environments. Proceedings of the Workshop on Genetic Improvement of Rice for Water-Limited Environments (Los Baños, Manilla, Philippines, 1-3 December 1998), International Rice Research Institute, Los Banõs, 353 pp.

Thomas, H., and C.M. Smart. 1993. Crops that stay green. Ann. appl. biol.123:193-219.

Tuberosa, R. 2012. Phenotyping for drought tolerance of crops in the genomics era. Front. Physio. 3:347.

vanOosterom, E.J., R. Jayachandran, and F.R. Bidinger. 1996. Diallel analysis of the stay-green trait and its components in sorghum. Crop Sci.36:549-555.

Xu, W.W., P.K. Subudhi, O.R. Crasta, D.T. Rosenow, J.E. Mullet, and H.T. Nguyen. 2000. Molecular mapping of QTLs conferring stay-green in grain sorghum (Sorghum bicolor L. Moench). Genome 43:461-469.