GENETIC ANALYSIS OF STEM COMPOSITION VARIATION IN *SORGHUM BICOLOR*

A Dissertation

by

JOSEPH PATRICK EVANS

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2012

Major Subject: Biochemistry

GENETIC ANALYSIS OF STEM COMPOSITION VARIATION IN *SORGHUM BICOLOR*

A Dissertation

by

JOSEPH PATRICK EVANS

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee, John Mullet
Committee Members, Timothy Devarenne
Gregory Reinhart
William Rooney
Head of Department, Gregory Reinhart

August 2012

Major Subject: Biochemistry

ABSTRACT

Genetic Analysis of Stem Composition Variation in *Sorghum bicolor*. (August 2012)

Joseph Patrick Evans, B.S., Pacific University

Chair of Advisory Committee: Dr. John Mullet

Sorghum (*Sorghum bicolor* [L.] Moench) is the world's fifth most economically important cereal crop, grown worldwide as a source of food for both humans and livestock. Sorghum is a C4 grass that is well adapted to hot and arid climes and is popular for cultivation on lands of marginal quality. Recent interest in development of biofuels from lignocellulosic biomass has drawn attention to sorghum, which can be cultivated in areas not suitable for more traditional crops, and is capable of generating plant biomass in excess of 40 tons per acre. While the quantity of biomass and low water consumption make sorghum a viable candidate for biofuels growth, the biomass composition is enriched in lignin, which is problematic for enzymatic and chemical conversion techniques.

The genetic basis for stem composition was analyzed in sorghum populations using a combination of genetic, genomic, and bioinformatics techniques. Utilizing acetyl bromide extraction, the variation in stem lignin content was quantified across several sorghum cultivars, confirming that lignin content varied considerably among sorghum cultivars. Previous work identifying sorghum reduced-lignin lines has involved

the monolignol biosynthetic pathway; all steps in the pathway were putatively identified in the sorghum genome using sequence analysis.

A bioinformatics toolkit was constructed to allow for the development of genetic markers in sorghum populations, and a database and web portal were generated to allow users to access previously developed genetic markers. Recombinant inbred lines were analyzed for stem composition using near infrared reflectance spectroscopy (NIR) and genetic maps constructed using restriction site-linked polymorphisms, revealing 34 quantitative trail loci (QTL) for stem composition variation in a BTx642 x RTx7000 population, and six QTL for stem composition variation in an SC56 x RTx7000 population.

Sequencing the genome of BTx642 and RTx7000 to a depth of ~11x using Illumina sequencing revealed approximately 1.4 million single nucleotide polymorphisms (SNPs) and 1 million SNPs, respectively. These polymorphisms can be used to identify putative amino acid changes in genes within these genotypes, and can also be used for fine mapping. Plotting the density of these SNPs revealed patterns of genetic inheritance from shared ancestral lines both between the newly sequenced genotypes and relative to the reference genotype BTx623.

ACKNOWLEDGMENTS

We are grateful to Dr. William Rooney for allowing us access to the sorghum diversity materials, and further for the use of his oven and grinding apparatus. We would also like to thank Dr. Dan Cosgrove at Meadwestvaco for his generous provision of Indulin AT, from which we prepared our lignin standard curve.

TABLE OF CONTENTS

Page

Page

Page

LIST OF FIGURES

Page

Page

Page

LIST OF TABLES

Page

INTRODUCTION AND LITERATURE REVIEW

Sorghum: Background, Taxonomy, and Origins

Sorghum (*Sorghum bicolor* (L.) Moench) is a C4 monoecious grass species that diverged from maize approximately 12 million years ago (Swigonova et al., 2004). Sorghum is the fifth most important grain crop in the world, and the third most important in the United States after wheat and corn (Doggett, 1988). Over 200,000 acres of sorghum were planted for silage in the United States in 2011, for a total yield of over 2 million tons (USDA, 2012). Sorghum grain is also valuable, with more than 5.4 million acres of grain sorghum planted in the United States in 2011. Sorghum grain is primarily utilized as a livestock feed in the US, but is an important grain crop for human consumption in Africa and Asia.

The genus Sorghum encompasses multiple species, including the rhizomatous *S. halapense* and *S. propinquum*, as well as the non-rhizomatous *S. bicolor*. Floral morphology has traditionally been used in the identification of closely related taxa of plants, and in *S. bicolor* the spikelets (as Poaceae lack pedicels) have been used to distinguish five primary *S. bicolor* landraces (Harlan and de Wet, 1972). The five landraces so described are Bicolor, Guinea, Kaffir, Caudatum, and Durra (Fig. 1), and each originates in a distinct region of Africa.

The Bicolor landrace is believed to have arisen in central Africa (Dahlberg, 1995), and as it spread throughout Africa, introgressions with various subspecies and

---

This dissertation follows the style of Plant Physiology.

adaptation to varying environmental conditions produced the five primary landraces (Dahlberg, 1995; Smith and Frederickson, 2000). Durra is thought to have arisen from an introgression between Bicolor and *aethiopicum* that allowed the offspring to thrive in the more arid conditions in central Africa (Dahlberg, 1995). Caudatum is one of the most agronomically important sorghum races, and is believed to have originated in central Africa through an introgression between early *bicolor* and a wild sorghum species. (Dahlberg, 1995). The Guinea landrace is found in more humid conditions in western Africa, and is the result of an introgression between S. *bicolor* and S. *arundiaceum* (Dahlberg, 2000). Kafir is believed to have originated in northern Africa as an introgression between S. *bicolor* and S. *verticilliflorum* and was transported south by humans, and now forms the primary agricultural sorghum of southern Africa (Dahlberg, 1995). These landraces have been extensively cultivated and hybridized, resulting in the approximately 70 groups utilized for sorghum classification today (Murty and Govil, 1967; Dahlberg et al., 2004).

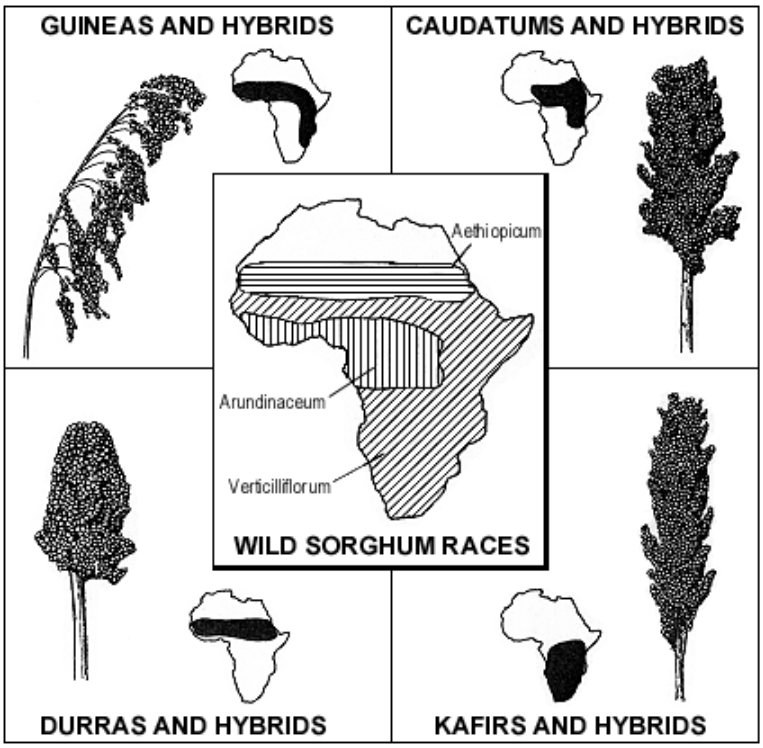


Figure 1: Sorghum landraces as determined based on spikelet morphology. (Smith and Frederickson, 2000).

Many of the sorghum cultivars used for breeding in the United States are or involve products of the Sorghum Conversion Project, which was a program to convert tall, exotic, and late flowering lines into shorter, earlier flowering forms more suitable for cultivation in the United States (Rooney and Smith, 2001). These exotic lines were selected for various traits of interest such as yield and resistance to biotic and abiotic stress and backcrossed to BTx406, a dwarf early-flowering variant suitable for US agriculture (Rosenow et al., 1997). The inbred lines produced can be roughly bulked into groups based on marker analysis: durra-kafir derivative males, feterita derivative males, zerazera derivative males, zerazera females and kafir females, with male and

female assigned to fertility-restoring "R" lines and female assigned to non-fertility restoring "A" and "B" lines (Menz et al 2004).

Sorghum: Genome and Genetics

Sorghum is a diploid organism, with a genome of approximately 700 MBp organized into 10 chromosomes (Paterson et al., 2009) making it the smallest of the sequenced C4 cereal species. The sequence of *Oryza sativa* is considerably smaller, however rice is representative of C3 cereals and lacks many desirable traits present in C4 plants, such as tolerance to high temperatures and drought.

The *Sorghum bicolor* genome sequencing project was completed in 2009, with a reference sequence generated from the inbred cultivar BTx623 (Paterson et al., 2009). While limited biochemical data is available on sorghum genes, *de novo* analysis and comparison to the related cereals *Orzya sativa* and *Zea mays*, as well as EST libraries, places the estimated gene complement of sorghum at approximately 27,600 genes (Buchanan et al., 2005; Salzman et al., 2005; Paterson et al., 2009). While the sorghum genome is approximately 2.5 times the size of the rice genome, they have similar quantities of euchromatin (252 MBp for sorghum and 309 MBp in rice), indicating that the bulk of the sorghum genome is made up of heterochromatin (Fig. 2) (Kim et al., 2005a; Paterson et al., 2009).

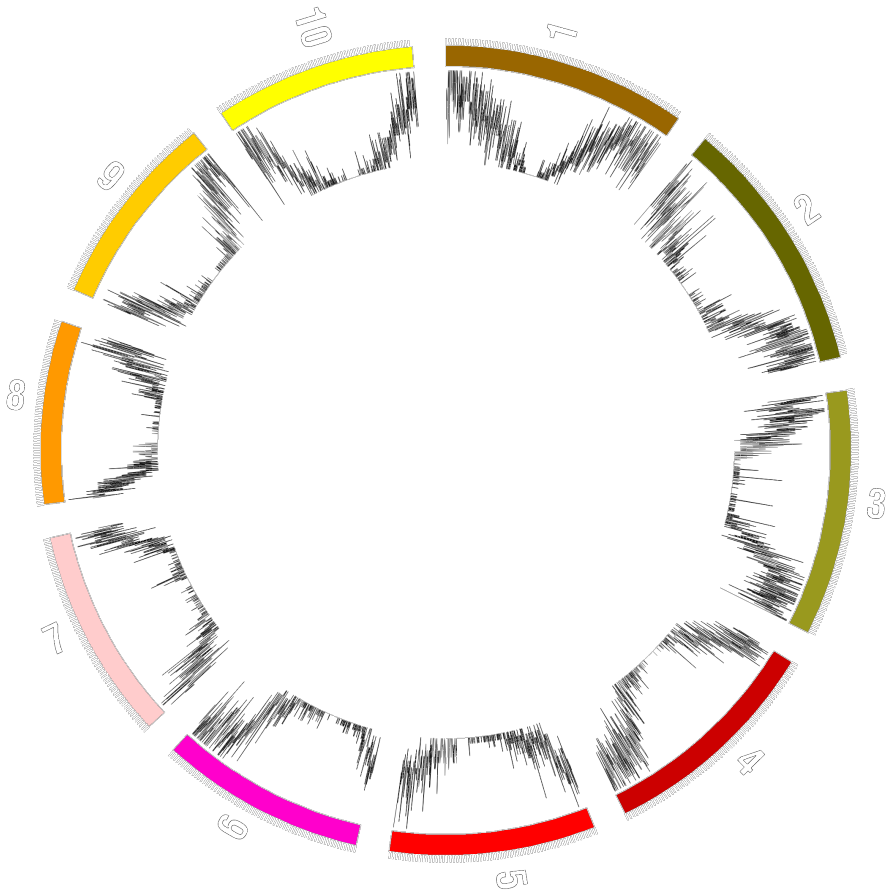


Figure 2: Plot of the sorghum genome with gene density graphed beneath each chromosome.

The relation between sorghum and other grasses is quite evident at the genomic level, with both *Sorghum bicolor* and *Sorghum propinquum* displaying significant levels of macro- and microcolinearity with both rice and maize (Bowers et al., 2005; Kim et al., 2005b). This colinearity reflects the common ancestor shared by cereals approximately 42-47 Mya, and is a valuable tool for trait mapping, as regions containing trait loci often correspond between cereal species (Paterson et al., 1995; Paterson et al., 2004). Unlike maize, however, sorghum is largely a self-pollinating species. The lack of outcrossing

involved in the propagation of sorghum leads to a much lower level of sequence variation and a correspondingly higher level of linkage disequilibrium when compared to maize, both of which are valuable traits for molecular genetics (Hamblin et al., 2004).

The sorghum genome is enriched in heterochromatic and repeat-rich regions, with such regions making up approximately two thirds of the sorghum genome (Paterson et al., 2009). Such a high quantity of heterochromatin made initial genetic studies of sorghum challenging to relate to the physical genetic location, as 97-98% of the recombination occurs within the euchromatic regions (Fig. 3) (Bowers et al., 2005; Paterson et al., 2009). This enrichment in repeat-heavy regions also provides challenges for short-read sequencing, as such regions are difficult to align with high confidence.

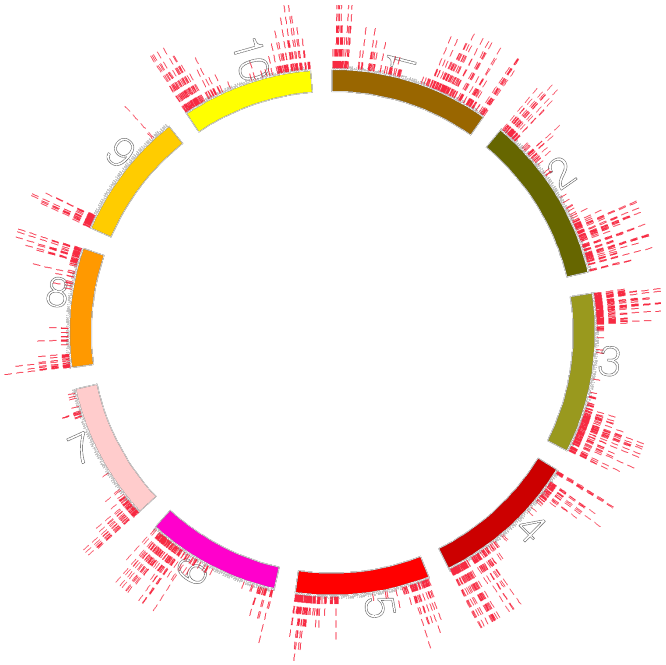


Figure 3: Physical locations of sorghum genetic markers

Fortunately, the ability of sorghum to self-pollinate allows for the rapid and relatively simple ability to develop inbred lines and take advantage of sorghum's high level of linkage disequilibrium. This characteristic has been used to develop genetic maps for traits such as favorable biofuels characteristics, disease resistance, and drought tolerance (Haussmann et al., 2002; Parh et al., 2008; Shiringani et al., 2010; Shiringani and Friedt, 2011).

Stem Physiology

Plant stems are multipurpose organs that vary considerably from plant to plant. In general, the stem is one of the primary axes of plant growth, responsible for elevating the leaves for optimal illumination and properly positioning the flowers and reproductive organs for pollination and seed dispersal. While stems vary significantly from plant to plant, the general morphology of stems is similar. An outer layer of dermal tissue protects the stem and provides a mechanism for controlling water and gas exchange with the surrounding environment. Filling the bulk of the stem is ground tissue, a mixture of parenchyma, collenchyma, and sclerenchyma cells that provide a variety of functions throughout the plant's lifespan (Moore et al., 1998). Parenchyma is the most common cell type in ground tissue, making up the cortex and pith tissues, while collenchyma and sclerenchyma cells are heavily lignified and provide structural support to the stem. Vascular tissues are found in bundles in the stem, and serve to transport water and nutrients, as well as providing structural support to the stem (Raven et al., 1986).

Monocot and dicot stems exhibit quite different morphologies. Dicot stems have an inner pith, surrounded by vascular bundles arranged in a ring structure, while monocots tend toward randomly distributed vascular bundles concentrated near the epidermis (Fig. 4). Dicot stems can also exhibit secondary growth, where cell division from lateral meristems allows the girth of the stem to increase. Monocots typically do not exhibit secondary growth in this manner, though exceptions such as palm trees can expand their girth through division and expansion of stem parenchyma cells.

The Plant Cell Wall

The plant cell well is a complex structure that lends support to plant cells, allowing them to resist osmotic pressure as well as prevent entry by pathogenic microorganisms (Raven et al., 1986). The structure of the cell wall varies between organisms and cell types, but in general it is a thick matrix of layered cellulose fibrils, crosslinked with hemicellulose, lignin, and pectin. (Fig. 5)

Plant cell walls generally develop in two stages. The primary cell wall is deposited prior to cell maturation, and retains flexibility to allow the cell to expand. The secondary cell wall is deposited once the cell reaches maturity and is generally much thicker than the primary cell wall. The secondary cell wall may also contain lignin, cutin, and/or suberin in addition to the components found in the primary cell wall (Buchanan et al., 2000).

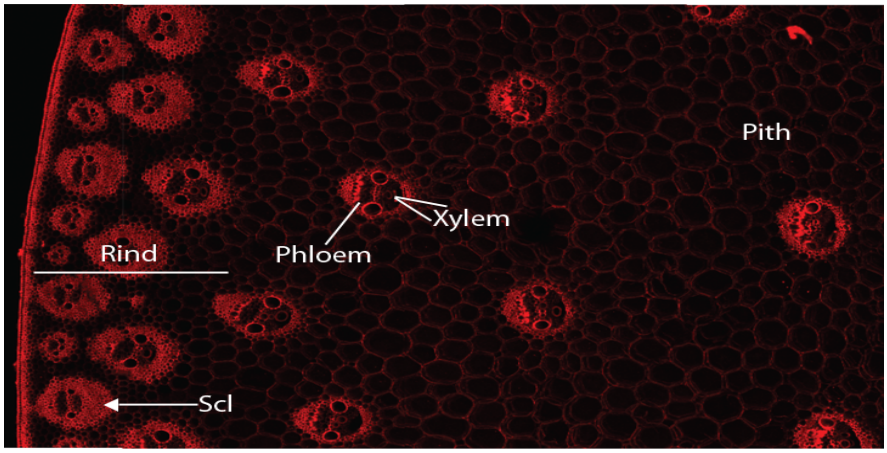


Figure 4: Confocal image of mature sorghum internode cross-section. Lignified tissues provide fluorescence, false colored red.

Primary cell walls retain flexibility, allowing cells to continue to grow and expand while still maintaining their rigidity due to turgor pressure. These cell walls are generally between 100nm and 1 μm in thickness. Primarily constructed of interlocking polysaccharide filaments and cellulose microfibrils, the primary cell wall presents a lattice appearance that can be adjusted to allow for cell growth and expansion (Carpita and Gibeaut, 1993; Cosgrove, 2005).

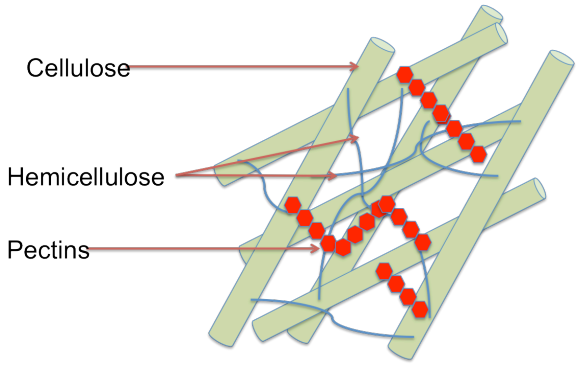


Figure 5: Schematic representation of plant primary cell wall.

Mature plant cells generate secondary cell walls once their growth is complete. These secondary cell walls share many compounds with primary cell walls, with the main difference being the addition of lignin. The secondary cell wall is deposited inwards of the primary cell wall, and provides an additional layer of protection and structural support. Grasses may feature the inclusion of silica crystals to protect the plant from herbivory (Carpita, 1996).

Cell Wall Composition

Determining and manipulating the composition of plant cell walls has been an ongoing process for several decades (Jung et al., 2012). Significant variation exists between plant species, and also between primary and secondary cell walls.

Primary cell walls are primarily composed of cellulose, arabinoxylan, uronic acids and protein (Burke et al., 1974). A significant portion of this protein content are extensins: hydroxyproline-rich structural glycoproteins embedded and intertwined throughout the primary cell wall (Lamport et al., 2011). These proteins are transported to the cell membrane from the Golgi apparatus, where they undergo crosslinking via free-radical oxidation to form structural supports for the growing cell wall (Lamport 1977). These proteins also play a role in plant pathogen response, with dense networks of extensins generated rapidly in response to elicitor molecules (Esquerre-Tugaye and Mazau, 1974; Bradley et al., 1992; Brady and Fry, 1997).

Also prevalent in primary cell walls are pectins, believed to form some of the most complex polysaccharide structures in the world (Willats et al., 2001). Pectins are heterogenous mixtures of homogalacturonan, rhamnogalacturonans, galactans,

arabinans, and a wide variety of other polysaccharides that form multiple recognizable domains, which serve differing functions related to cell growth, expansion, rigidity, and cell adhesion (Willats et al., 2001; Vincken et al., 2003). These pectins, among other functions, can form hydrated gels that serve to physically spread the mesh of cellulose microfibrils during cellular expansion, and cement the fibrils into position once the gels dehydrate. Pectins are also present in the middle lamella, a region between the primary cell walls of neighboring cells that functions to adhere the cells together (Iwai et al., 2002).

Secondary cell walls are less diverse in the variety of materials involved in their construction. While primary cell walls maintain cell integrity and shape during cell growth, secondary cell walls provide structural support and hydrophobicity to mature cells. The most well studied secondary cell walls are those found in the xylem, the element of the stem that carries water from the roots to the leaves. Secondary cell walls are primarily constructed from cellulose, hemicellulose, and lignin, the last of which provides the hydrophobic element critical to maintaining vascular integrity. In maize, development of secondary cell walls is initially concentrated on the protoxylem. As development progresses, secondary cell wall thickening is localized in the protoxylem, metaxylem, and eventually in the parenchyma and schlerenchyma, with the xylem and schlerenchymal tissues showing evidence of lignification (Jung and Casler, 2006).

Despite being present only in the secondary cell wall, lignin forms a considerable percentage of the total stem biomass (Iiyama and Wallis, 1990). While necessary for vascular hydrophobicity, the inclusion of lignin in secondary cell wells provides

challenges to efficient use of stem biomass. Lignin is a complex polyphenolic compound, made up of many lignin monomers that are covalently linked within the secondary cell wall (Fig. 6). The energy content of lignin is quite high, however lignin has so far proved exceedingly intractable to processing. The heterogenous structure of lignin has yet to be resolved, and the high number of carbon-carbon bonds, combined with the aromatic nature of the monolignols, yields a compound that is difficult to degrade enzymatically, and extremely energy intensive to degrade chemically (Boerjan et al., 2003; Chang 2007).

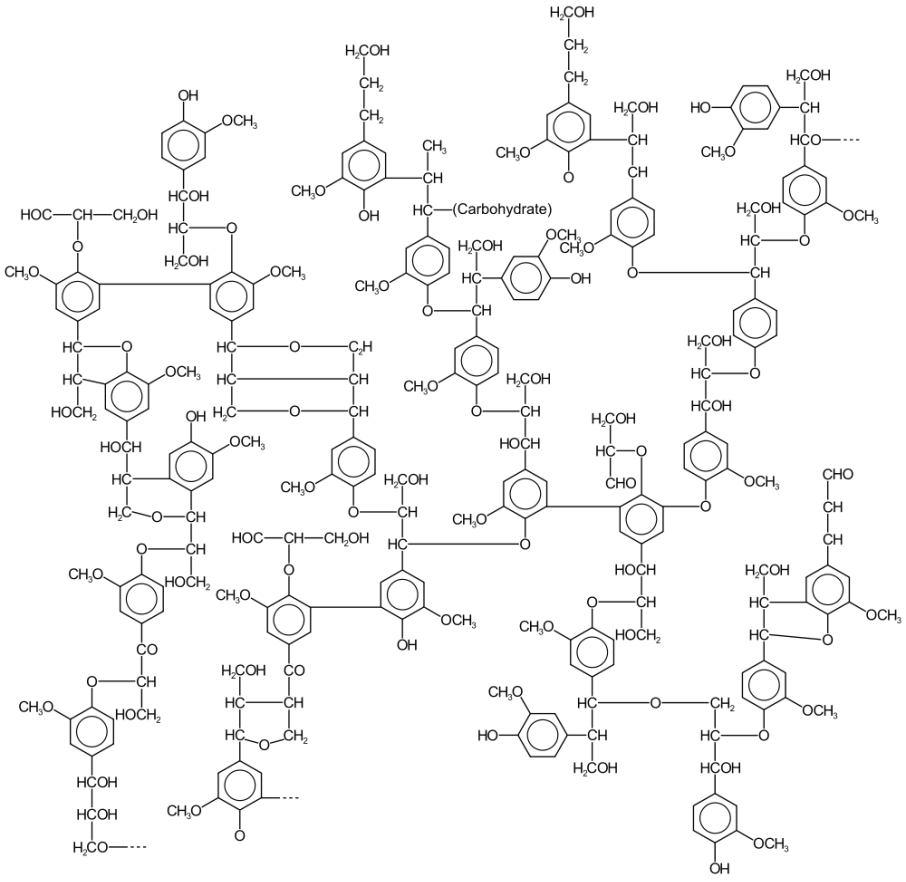


Figure 6: Example of a lignin structure. (Glazer and Nikaido, 1995).

Cellulose Biosynthesis and Deposition

Cellulose is the primary component of plant cell walls, and the most abundant biopolymer on earth. Cellulose is a polysaccharide molecule formed through the covalent β1-4 linkage of glucose molecules into long chains. In plants, cellulose is typically formed into higher-order structures called cellulose microfibrils. Each microfibril is made of roughly 36 hydrogen-bonded chains of cellulose, each containing between 500 and 14,000 covalently linked glucose molecules (Somerville, 2006).

The exact method of cellulose microfibril synthesis and deposition has not yet been determined. What is known is that in vascular plants, cellulose synthase is a transmembrane protein complex, with each complex made of six rosettes, and each rosette assembled from six or more subunits (Kimura, 1999). Each of these rosettes is believed to synthesize between six and ten glucan chains, which then crystallize into the roughly 36-chain crystals that make up microfibrils (Herth, 1983).

The most well known components of cellulose synthase are the CESA family of proteins. These proteins were first identified in higher plants through sequence analysis of mRNA in developing fibers, and share sequence identity with bacterial cellulose synthase proteins (Pear et al., 1996). These proteins contain eight transmembrane domains, and are believed to interact through a conserved RING-type zinc finger (Holland, 2000; Richmond, 2000; Somerville, 2006). UDP-glucose - channeled to the cellulose synthase complexes by sucrose synthase – is added to the nonreducing end of the growing cellulose polymers through an as-yet uncharacterized glycosyltransferase action (Koyama et al., 1997; Lai-Kee-Him et al., 2002).

Early observations of cellulose deposition in higher plants revealed that the orientation of the cellulose microfibrils in expanding cells followed that of the cortical microtubules located nearest the plasma membrane (Ledbetter and Porter, 1963). Live-cell confocal imaging reveals that cellulose synthase complexes are functionally associated with the cortical microtubules of expanding plant cells, and follow these microtubules in order to establish a regular pattern of cellulose deposition (Paredez et al., 2006). Current observations indicate cellulose synthase travels adjacent to, but not on top of, the cortical microtubules, and that it is capable of bidirectional travel, establishing parallel arrays of cellulose deposition (Giddings and Staehelin, 1988). The translocation of the complex is not believed to be linked to the association with the microtubule, but instead the motive force is provided by the cellulose polymerization itself (Robinson and Quader, 1981; Lloyd, 1984).

Hemicellulose Biosynthesis and Deposition

Hemicellulose is the name given to the complex network of polysaccharides (excluding cellulose and pectin) that can be found in plant primary and secondary cell walls and contain an equatorial β-(1-4)-linked backbone structure. Many plant cell walls contain a wide variety of hemicelluloses, including xyloglucans, xylans, mannans, glucomannans, and in the case of grasses, β-(1-3,1-4)-glucans (Scheller and Ulvskov, 2010).

Xyloglucan is the most common hemicellulose, found in every land plant species so far examined, and is the main hemicellulose present in primary cell walls in most plant species, excluding grasses (Moller et al., 2007; Popper, 2008). Xyloglucan can

exhibit significant branching and substitution, which is believed to be a determinant of the role xyloglucan will be playing in the cell wall (Scheller and Ulvskov, 2010). The xyloglucan backbone is synthesized by the CSLC family of glycosyltransferases that share similarity with the CesA family of cellulose synthase genes, and branches are added by galactosyltransferases, fucosyltransferases, and members of the GT47 and GT34 families of glycosyltransferases (Perrin et al., 1999; Cavalier and Keegstra, 2006; Cocuron et al., 2007).

Xylans are also very common in cell walls, making up a large family of polysaccharides with the common characteristic of a β-(1-4)-linked backbone of xylose residues. Variations in the composition of xylan are characteristic between the secondary cell walls of grasses and dicots: grasses feature a preponderance of glucuronic acid and arabinose residues linked to the backbone, while dicots mainly favor glucuronic acid (Thomas et al., 1987; Pauly and Keegstra, 2008; Vogel 2008). The biosynthesis of xylan has not yet been achieved *in vitro*, but it is known that the CSLC family involved in xyloglucan synthesis does not catalyze the formation of the xylan backbone (Scheller and Ulvskov, 2010). Characterization of xylem deficient mutants in Arabidposis indicates the involvement of members of the GT43 and GT47 families of glycosyltransferases in the formation of the xylan backbone, but attempts to show xylan synthase activity in these genes have not been successful (Brown et al., 2007; Lee et al., 2007; Brown et al., 2009;). Similarly, the exact players functioning in the addition of branch resides to the xylan backbone have yet to be determined. Members of the GT61 family in grasses are suspected to be involved in the formation of

glucuronoarabinoxylans, but *in vitro* experiments have so far not confirmed this hypothesis (Porchia et al., 2002; Mitchell and Dupree, 2007; Zeng et al., 2008).

β-(1-3,1-4)-linked glucans are a phenomenon so far unique to grass species (Scheller and Ulvskov, 2010). Present mainly in the primary cell walls, mixed-linkage glucans are usually formed of cellotriosyl and cellotetrasyl units linked by β-(1-3) bonds (Stone, 1992). These polysaccharides are believed to play a role in the primary cell wall during cell expansion (Gibeaut et al., 2005). Mixed-linkage glucans have been shown to be synthesized by members of the CSLF and CSLH families of cellulose synthase-like proteins (Burton et al., 2006; Doblin et al., 2009).

With the exception of mixed link glucans, hemicellulose is synthesized in the Golgi apparatus, and exported to the cell wall once assembly and modification is complete. One exception to this may be mixed-linkage glucans – the protein synthetic machinery has been localized in the Golgi apparatus, but the polysaccharide itself has not. Whether this indicates masking by acetylation or additional synthesis steps after exiting the Golgi has yet to be determined (Wilson et al., 2006).

Lignin Biosynthesis and Deposition

Lignin is a complex polyphenolic compound formed through the covalent linkage of monolignol subunits derived from the amino acid phenylalanine. Lignin is primarily deposited in the secondary cell walls of parenchyma, xylem, and schlerenchyma cells, where lignin contributes structural support and hydrophobicity to the cells. Development of the phenylpropanoid pathway was critical to the colonization of land by plants, contributing critical shielding to damaging ultraviolet radiation as well

as support and hydrophobicity for the tracheary elements. Lignin is now one of the largest carbon sinks for plants, and is estimated to represent up to 30% of the total biomass in the biosphere (Lowry et al., 1980; Bateman et al., 1998; Boerjan et al., 2003).

Monolignol biosynthesis begins with the deamination of phenylalanine by phenylalanine ammonia lyase (PAL), and then the immediate conversion of the product, cinnamate, to *p*-coumaric acid through the action of cinnamate-4-hydroxylase (C4H). These steps are believed to occur through an interaction of PAL and C4H in the endoplasmic reticulum, likely in an attempt to reduce concentrations of cinnamate, which has been shown to act as an ionophore and inhibitor of auxin-based cell growth (McLaughlin and Dilger, 1980; Rasmussen and Dixon, 1999; Achnine et al., 2004; Wong et al., 2005). Ligation of CoA by 4-Hydroxycinnamoyl-CoA ligase (4CL) produces p-coumaroyl-CoA, which either proceeds through monolignol biosynthesis or proceeds into flavonoid biosynthesis.

Generation of *p*-coumaryl alcohol, the primary component of H type lignin, involves two additional steps. *P*-coumaroyl-CoA is first reduced to *p*-coumaraldehyde through the action of hydroxycinnamoyl-CoA reductase (CCR), and then reduced again to form *p*-coumaryl alcohol (Boerjan et al., 2003).

Formation of coniferyl and sinapyl alcohol involve the creation of p-coumaroyl shikimate esters, catalyzed through the action of hydroxycinnamoyl-CoA:shikimate hydroxycinnamoyl transferase (HCT). The resulting ester is hydroxylated at the 3' position by *p*-coumaroyl shikimate-3' hydroxylase, and then de-esterified via HCT. The newly formed hydroxyl group is replaced by an O-methyl group by caffeoyl-CoA O-

Methyl transferase, and the actions of CCR and CAD generate coniferyl alcohol, the primary component of G lignins (Boerjan et al., 2003; Weng and Chapple, 2010).

Sinapyl alcohol is the primary component of S lignin, which is present in most flowering plants (Weng et al., 2008). Sinapyl alcohol is generated through the actions of ferulic acid/coniferaldehyde/coniferyl alcohol 5-hydroxylase (F5H) on coniferaldehyde and/or coniferyl alcohol, which are generated during coniferyl alcohol synthesis. The products are hydroxylate on position 5, and then the hydroxyl group is replaced by an O-methyl group by caffeic acid O-methyl transferase. Starting with coniferaldehyde yields sinapaldehyde, which is then reduced to sinapyl alcohol by CAD. If the initial substrate was coniferyl alcohol, the product is sinapyl alcohol.

The transport mechanisms for monolignols to the cell wall have yet to be fully elucidated. It has been shown that lignol glycosides accumulate in lignifying tissues, but the function of these glycosides in transport has yet to be confirmed (Lim et al., 2001). Expression studies tentatively support the function of monolignol glycosides in transport through the involvement of members of the ABC family of transport proteins (Ehlting et al., 2005). However, much work remains to be done in order to confirm the role of these proteins.

Once the monolignols arrive at the cell wall, a variety of dehydrogenation reactions convert the monolignols into the lignin polymer (Boerjan et al., 2003). Many different classes of proteins have been suggested to be responsible for the generation of monolignol radicals, including peroxidases, laccases, and polyphenol oxidases, though recent studies have focused more closely on peroxidases as the primary agent

responsible for free radical generation (Onnerud et al., 2002; Blee et al., 2003). Once free-radical generation has taken place the polymerization step takes place largely without biochemical constraints, generating a complex three dimensional, racemic lignin molecule (Ralph et al., 1999).

Ferulic acid is also produced via the phenylpropanoid biosynthetic pathway alongside monolignols, and can be found in the primary and secondary cell walls of many graminaceous species (Harris and Hartley, 1976; Ou and Kwok, 2004). Comprising nearly 3% of the dry weight of graminaceous cell walls, ferulate is most often found covalently bound via an ester linkage to the arabinose residues of arabinoxylan polysaccharides (Ralph and Helm, 1993; Saulnier et al., 1999). Ferulic acid has also been shown to form linkages with lignin, and it is believed that it may serve as a nucleation site for lignin polymerization (Iiyama et al., 1994; Wallace and Fry, 1995)

Nonstructural Carbohydrates

While the bulk of the biomass in sorghum stems originates in the plant cell wall, non-structural carbohydrates contribute significantly to the total biomass. Nonstructural carbohydrates (NSC) are composed of readily extractable carbohydrates such as starch and sucrose, and serve as energy storage and transport for the plant, especially during grain filling (Arai-Sanoh et al., 2011). Many crops grown for sugar or syrup, such as sugarcane and sweet sorghums, have been selected for high stem NSC content, though grain crops retain significant NSC despite grain yield optimization (Vietor and Miller, 1990; Murray et al., 2008).

The three primary soluble sugars found in sorghum stems are glucose, fructose, and sucrose. Glucose is produced as a product of photosynthesis, where the plant uses solar energy to fix carbon in the atmosphere into forms that can be metabolized when the sun is not present. Sucrose is synthesized from glucose and fructose through the action of sucrose synthase and is used for transporting sugars through the phloem, and makes up the primary stem NSC in mature sweet sorghums (Murray et al., 2008).

Starch is formed through the α(1-4) linkage of glucose molecules, catalyzed by starch synthase. Starches are primarily separated into amylose and amylopectin – amylose is composed of unbranched α(1-4)-linked glucose residues, while amylopectin also contains α(1-6)-linked branches of glucose residues. Amylose content in sorghum stems shows significant variation across the developmental cycle of sorghum, with high levels of starch present in the stem prior to anthesis, and levels rapidly decreasing thereafter (McBee and Miller, 1993). The presence of starch in sorghum stems has not been analyzed in detail, and much regarding the temporal and physical distribution remains unknown.

Regulation of Secondary Cell Wall Growth

While the chemical structures involved in plant cell walls are relatively well understood, and the biosynthetic complexes that create them have been studied for decades, an understanding of the regulation of these complexes is recent and still incomplete.

Secondary cell walls vary in function, and therefore vary in composition and timeframe of deposition. These variations require a complex regulatory system to

establish secondary cell walls in such varying cell types as endothecium cells, guard cells, and trichomes. Additionally, these regulatory systems are responsible for laying down the complex layers of cellulose, xylan, and lignin involved in secondary cell wall thickening. While many of the players involved in the regulation of the biosynthetic genes are known, the "master switches" that initiate the developmental programs themselves remain uncharacterized (Zhong and Ye, 2007).

Initially characterized in endothecium cells, members of the *Arabidopsis* NAC family of transcription factors play a key role in secondary cell wall development (Mitsuda et al., 2005). Two members of the NAC family, NAC SECONDARY WALL THICKENING PROMOTING FACTOR1 (NST1) and NST2 were shown to both act in the thickening of endothecium cell walls; NST1 and NST2 appear to act redundantly, requiring loss of function in both genes to cause a secondary cell wall deficient phenotype. Interestingly, overexpression of either of these genes in parenchyma tissue causes upregulation of secondary cell wall biosynthetic machinery, and ectopic deposition of secondary cell walls, supporting the concept of an upstream control mechanism (Zhong and Ye, 2007).

Other important players in the regulation of secondary cell wall biosynthesis are members of the MYB family of transcription factors. Mutation of the *Arabidopsis* MYB26 gene in endothecium has been shown to cause loss of secondary cell wall thickening, resulting in anther dehiscence (Steiner-Lange et al., 2003; Yang et al., 2007). Similar to NST1 and NST2, overexpression of MYB26 in parenchyma cells also results in an ectopic deposition of secondary cell wall. The same mutation that abolished cell

wall thickening also caused a down regulation of NST1 and NST2, and overexpression of MYB26 causes an up regulation of NST1 and NST2, indicating that they may be targets of MYB26, though whether they are targeted directly is unknown.

NST1, along with NST3 (also called SND1) have also been shown to be involved with the secondary cell wall biosynthesis of fibers in *Arabidopsis* (Zhong et al., 2006; Mitsuda et al., 2007; Zhong et al., 2007a). Much like NST1 and NST2 in the endothecium, SND1 and NST1 act in a redundant fashion to promote secondary cell wall thickening in fibers, and overexpression of SND1 leads to ectopic secondary cell wall deposition. The normal function of these transcription factors also appears to be tissue specific, as SND1 is expressed in intrafascicular and xylary fibers only, and not in the vessel elements (Zhong et al., 2006; Zhong et al., 2007a). An additional MYB factor, MYB46, has been shown to be a direct target of SND1, and plays a role in the regulation of secondary cell wall biosynthesis in fibers through regulation of the biosynthetic machinery (Zhong et al., 2007b).

Vessels (protoxylem and metaxylem) also rely on NAC transcription factors to establish the program of secondary wall deposition. The VASCULAR-RELATED NAC DOMAIN genes VND6 and VND7 are closely related to SND1, and have been shown to be upregulated in protoxylem and metaxylem, respectively (Kubo et al., 2005). Similar to NST1 and NST2, overexpression of VND6 and VND7 results in ectopic deposition of secondary cell walls in parenchyma cells, strengthening the position that VND6 and VND7 act as high level controls for secondary cell wall deposition in their respective tissues.

While many of the transcription factors involved in secondary cell wall deposition have been identified, the interactions between these genes and the biosynthetic genes for the various cell wall components remain largely unidentified. Overexpression of SND1 is known to upregulate the expression of a variety of NAC and MYB factors, indicating the presence of a transcriptional network (Fig. 7) that may be involved in the development of secondary cell wall biogenesis (Zhong et al., 2006). Further support is provided by the determination that MYB46 is a direct target of SND1, and that overexpression of MYB46 causes increases in the expression of MYB85 and KNAT7 (Mitsuda et al., 2005).

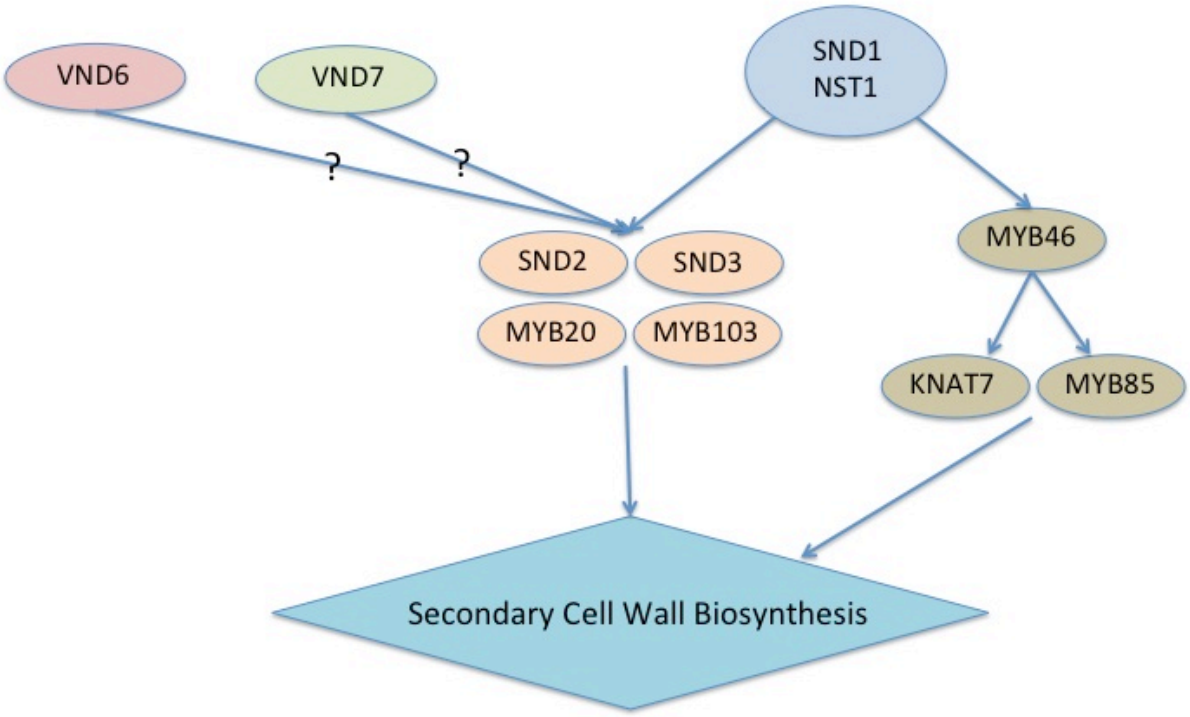


Figure 7: Diagrammatic representation of regulation of secondary cell wall biosynthesis in plant stems. Question marks represent uncertainty in downstream factors.

SORGHUM GERMPLASM SCREENING AND LIGNIN ANALYSIS

Introduction

Among the traits relevant to biofuels crops, lignin content is one of the most important. High levels of lignin in biomass destined for enzymatic or microbial degradation increases their recalcitrance to digestion, necessitating energy expenditure in the form of elaborate pretreatment before the biomass can be converted to fuel (Dien et al., 2006; Chen and Dixon, 2007; Galbi and Zacchi, 2007). Lignin additionally poses challenges for non-enzymatic biofuels methods such as pyrolysis although it can be converted into an etherized gasoline additive using thermochemical conversion (Dautzenberg et al., 2011). Lignin molecules are highly oxygenated, and the presence of oxygenated compounds in the resulting bio-oils causes these products to be unstable and require further treatment (Mohan et al., 2006). Though recent advances in fast pyrolysis and zeolite upgrading have taken steps towards providing a reliable method for stabilizing the resulting bio-oil, the ability to regulate the quantity of lignin in biomass going into a conversion process remains of critical importance (Adjaye and Bakhshi, 1995; Jae et al., 2011),

Lignin serves an important biological role in plants by providing structural support for stems and trunks, as well as waterproofing tracheary elements to allow for water transport. Previous experiments involving transgenic alteration of monolignol synthesis have caused undesirable dwarfing due in part to the loss of this waterproofing (Hoffman et al 2004), and mutations in Sorghum monolignol biosynthetic genes can result in a loss of plant structural stability (Oliver et al*.,* 2005).

In this study, diverse sorghum accessions were analyzed to determine the extent they varied in lignin content and in the distribution of lignin in stems.

Results

Flowering

In the spring of 2007, Dr. William Rooney planted 2000 diverse sorghum accessions in 20 ft plots at the Texas A&M University farm in College Station Texas (located approximately 30.550, -96.438). Plants were grown through until October under non-irrigated conditions. Sampling was performed in October, with 351 total lines selected for analysis on the basis of distinct visible phenotypes (lodging, growth phase, stem thickness and stem height). Individual plants were selected from the chosen plots and growth stage and lodging status recorded (Fig. 8). Approximately half of the lines examined had reached anthesis (Fig. 8) with many of these lines (Fig. 8) exhibiting the formation of additional grain heads arising from the stem nodes.

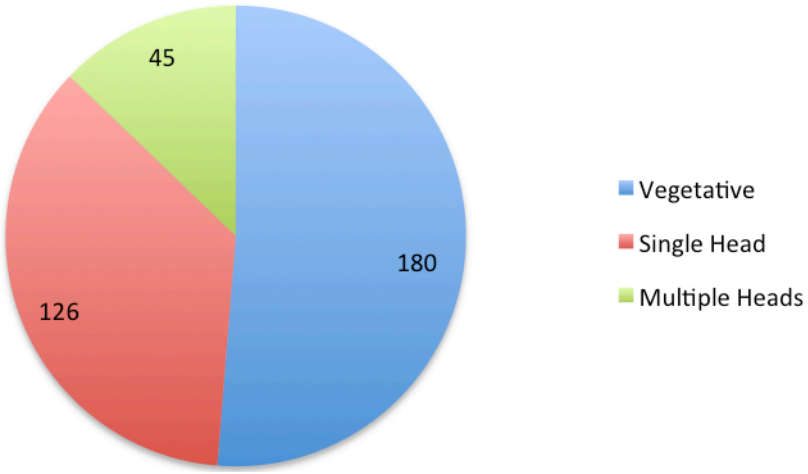


Figure 8: Flowering status.

Lodging

Lodging is an agronomically relevant trait that can be responsible for significant yield loss. It was hypothesized that lodging may be correlated to lignin content so lodging status of observed lines was recorded. A plot was considered lodged if more than half of the plants had fallen over (root lodging) or been broken below the peduncle (stalk lodging). A line was considered partially lodged if more than 25% but less than 50% of the plants had lodged. Nearly 2/3 of the examined lines exhibited some level of lodging, with the majority of those being considered fully lodged (Fig. 9).

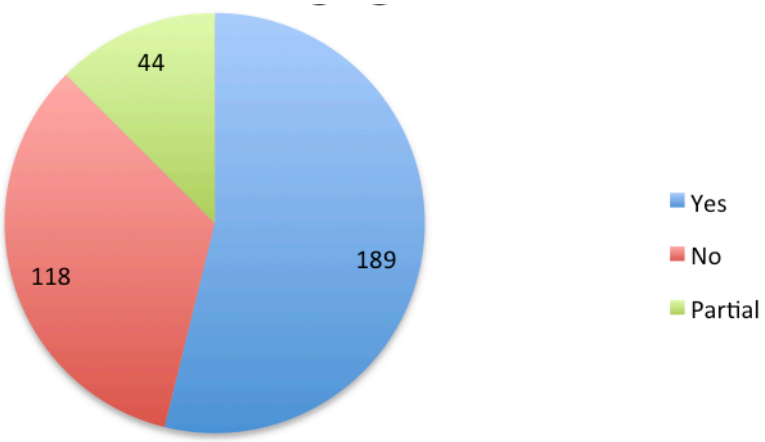


Figure 9: Lodging status.

Stem Lignin Staining

Stem segments were collected from all the lines examined for lodging and flowering time. Samples were sectioned by hand using a razor blade, and the resulting sections stained with a mixture of 20% HCl and 2% phloroglucinol in ethanol. Phloroglucinol reacts with sinapyl and cinnamyl aldehydes in the presence of acid to create a reversible dye reaction, temporarily staining the lignified tissues pink and red-brown, respectively (Pomar et al., 2004). The staining solution was applied to the samples with a Pasteur pipette and the sections photographed after 60 mins of staining. It was observed that the majority of the lignification detected by this method in sorghum stems is localized near the epidermis, with additional lignification distributed throughout the ground parenchyma (Fig. 10A). Subsequent examination of stems using confocal microscopy revealed a layer of lignin deposited in the epidermis and surrounding cells comprising vascular bundles located directly below the epidermis (Fig. 10B). Further

staining was mainly localized to interior scattered vascular bundles, though some samples evinced staining in the ground parenchyma as well (Fig. 10A,B). This staining pattern is consistent with previously observed lignification patterns in grasses such as maize (Jung and Casler, 2006).

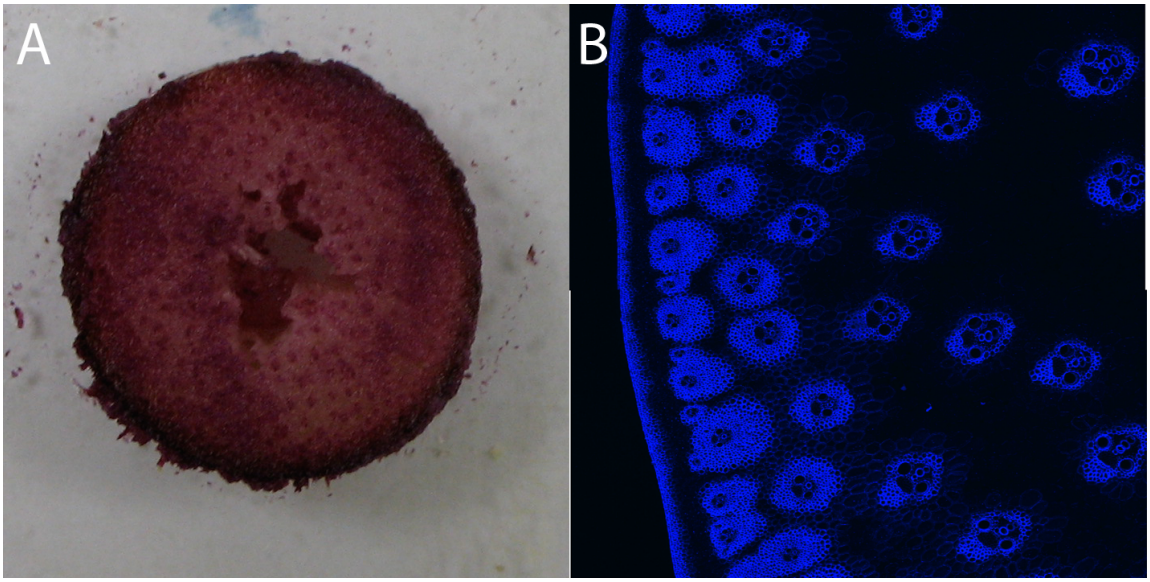


Figure 10. Methods for visualization in mature sorghum stem. (A) Mature sorghum stem internode stained with phloroglucinol-HCl. (B) Mature sorghum stem internode lignin autofluorescence observed with confocal microscopy.

From these observations, staining categorization was separated into Epidermal Staining (ES) and Interior Staining (IS). ES was defined as the intensity of staining at the epidermal surface of the section, rated on a subjective scale of 0-3, with 0 representing no staining to 3 representing the darkest, most intense staining. IS was defined as the quantity and intensity of staining of cells in the interior of the stem section, considering

both vascular bundles and ground parenchyma cell walls. The majority of the lines analyzed showed high levels of IS (Fig. 11), while much more variation was present in the level of ES (Fig. 12).

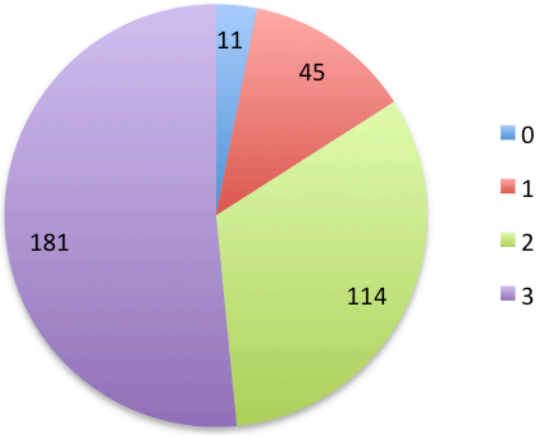


Figure 11: Interior staining, rated from 0 (no staining) to 3 (most intense staining).

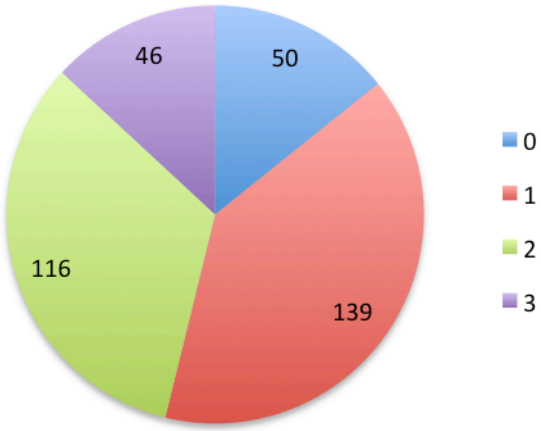


Figure 12: Epidermal staining, rated from 0 (no staining) to 3 (most intense staining).

There was no significant correlation between lodging and ES/IS or between flowering status and ES/IS (Data not shown).

Stem Lignin Quantitation

Acetyl bromide extraction followed by spectrophotometry was used to quantitate the amount of lignin present in the stems of a subset of 41 lines chosen from the set of 351 total sampled lines (Hatfield et al., 1999). Samples were chosen to represent the range of lignin staining observed and subsequently ground to ~1mm particle size. Samples were washed, treated with acetyl bromide and suspended in glacial acetic acid for spectrophotometry (Morrison, 1972). Absorbance values at 280 nm were compared to standards prepared from Indulin as per Fukushima et al. (1991) to determine total lignin percentage for each sample (Fig. 13).

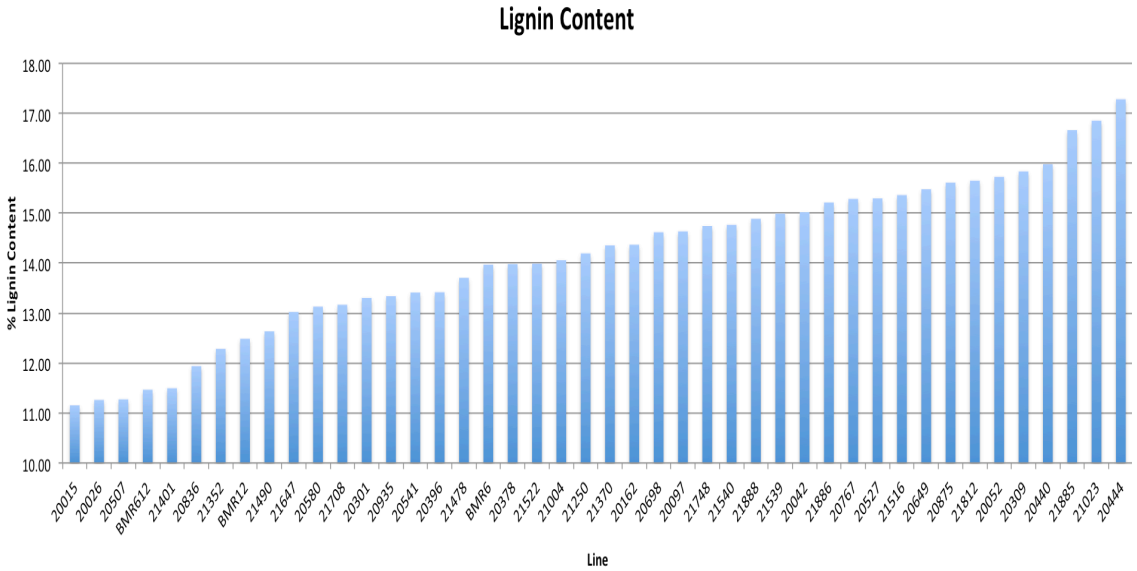


Figure 13: Lignin content measured as a percent value of total dry weight of sorghum stem. Field plot numbers are labeled on X-axis.

Lignin content varied from ~11% of total stem mass to over 17% of total stem mass in the samples examined. Also included in the analysis were three sorghum monolignol synthesis deficient lines, *bmr6, bmr12,* and *bmr612*. *bmr6* contains a premature stop codon in SbCAD4 which encodes a cinnamyl alchohol dehydrogenase responsible for the conversion of hydroxycinnamoyl aldehydes into monolignols (Sattler et al., 2009). *bmr12* contains a premature stop codon in the soghum COMT gene which encodes an O-methyl transferase responsible for the conversion of hydroxyconiferaldehyde into sinapylaldehyde (Bout and Vermerris, 2003),. *bmr612* is a cultivar with stacked *bmr6* and *bmr12* alleles. All *bmr* lines contain reduced lignin content and thus are useful for reference when examining lignin levels.

Discussion

Diverse sorghum accessions exhibit high levels of variability in lignin content and distribution in stems. Surprisingly, fifteen of the lines examined had lignin levels lower than *bmr*6, and 3 lines had lower lignin than the *bmr612* stacked line, highlighting the amount of variation inherent in the wide range of uncharacterized sorghum germplasm. Lignin levels as detected by acetyl bromide extraction also did not correlate with either internal or epidermal phloroglucinol staining. This lack of correlation may indicate a lack of sensitivity in the staining method, but most likely represents either bias in the categorization of the degree of staining or a lack of adequate homogeneity in the preparation of the stems for staining.

Lignification is a relatively tissue-specific phenomenon; secondary cell wall thickening occurs primarily in xylem and fiber cells in the stem. This provides a

mechanism for lignin variation beyond secondary cell walls containing more lignin: variation may exist in the number and lignification of vascular bundles and fibers. Jung and Casler (2006) also observed that different secondary wall tissues lignified at different rates. While all samples were harvested at the same date, the differences in flowering indicate that variation existed in the developmental state of the cultivars observed, which may also contributed to the variation as it may be challenging to identify partially lignified from fully lignified xylem elements through direct observation.

It has been previously observed that lignin content is negatively correlated with growth rate and that lignin biosynthesis is energetically expensive for the plant (Niemann et al*.,* 1992; Amthor 2003; Novaes et al., 2010). Given that plants can only assimilate a certain amount of carbon during a growing season, they must partition this carbon between the various cell components of vegetative growth and also allocate a sufficient quantity for grain development. Within the stem components, limited carbon must be split between cellulose, hemicellulose, lignin, and nonstructural carbohydrates such as sucrose and fructose. While the techniques used in this project were not able to identify the quantities of these components, newer methods such as near-infrared reflectance spectroscopy will allow for a more holistic view. This will allow for the evaluation of lignin quantity in the context of total stem carbohydrates, rather than in isolation.

Large levels of variation in the Sorghum germplasm also illuminate sorghum's value as a potential bioenergy crop. With the ability to pyramid desirable traits through

breeding (as demonstrated with *bmr612*) and such a wide variety of lignin content, it should be possible to adjust the lignin content to fit the desired industrial application. While traditional approaches to lignin modification have attempted to reduce the overall amount of lignin present in feedstocks, lignin is a very energy dense material, with an energy content 30% higher than that of cellulose (White, 1987). Exploiting both the high and low lignin cultivars in the collection will allow for flexible construction of lignin-optimized feedstocks.

Using identified differences in lignin quantity, the next step is to identify the genetic regions responsible for this variation. Once identified, the genic basis of the variation can be determined; these might be monolignol biosynthetic genes, secondary cell wall-associated transcription factors, or a combination of both.

IDENTIFICATION OF SORGHUM MONOLIGNOL BIOSYNTHETIC GENES

Introduction

Lignin is a complex macromolecular structure composed of covalently linked phenolic compounds that is intertwined with the cellulose microfibrils in the secondary cell walls of plants. The basic compounds that make up the lignin polymer are monolignols, aromatic alcohols synthesized from the amino acid phenylalanine. The synthesis of a given monolignol from phenylalanine consists of between five and eleven steps, depending on the monolignol being synthesized, beginning with the deamination of phenylalanine, proceeding through hydroxylation of the aromatic ring, O-methylation of the aromatic ring, and finally reduction of the C-terminus (Boerjan et al., 2003)

Traditional lignin engineering efforts have focused on modifying the function of the monolignol biosynthetic genes (Vanholme et al., 2008). Previously identified lignin deficient mutants in *Sorghum bicolor* were determined to be the result of gene truncations in monolignol biosynthesis genes (Bout and Vermerris, 2003; Sattler et al., 2009). Research on other organisms has shown that manipulation of the monolignol synthesis genes is a valid method for establishing changes in lignin levels and composition (Boudet et al., 2003; Vanholme et al., 2008;Grabber et al., 2010).

The genome of *Sorghum bicolor* was assembled in 2009, with most gene annotations established via *ab initio* predictions, with some additional data from previously identified genes in maize, rice and sugarcane (Paterson et al., 2009). As a

result, many of the predicted genes are assigned into groups based on general function (i.e. dehydrogenase action) but not specific substrate or product. While some genetic analyses have been performed on sorghum lignin variation (Shiringani and Friedt, 2011) relatively little has been done to identify the monolignol synthesis genes themselves in sorghum.

Results

Identification of Monolignol Biosynthesis Genes in Literature

The monolignol biosynthetic pathway consists of 10 enzymes (Fig. 14): phenylalanine ammonia lyase (PAL), cinnamate 4-hydroxylase (C4H), 4-coumarate-CoA ligase (4CL), hydroxycinnamoyl-CoA: quinate shikimate hydroxycinnamoyl transferase (HCT), coumarate-3 hydroxylase (C3H), caffeoyl-CoA *O*-methyl transferase (CCoAOMT), cinnamoyl CoA reductase (CCR), ferulate 5-hydroxylase (F5H), caffeic acid/5-hydroxyconiferaldehyde *O*-methyltransferase (COMT), cinnamyl alcohol dehydrogenase (CAD) (Chen et al., 2006). Exploration of the monolignol biosynthetic pathway has been extensive but has been spread across many species, necessitating each gene be identified in the organism being studied (Boerjan et al., 2003).

Genes involved in various pathways have been inferred using the existing annotations (http://www.gramene.org/pathway/sorghumcyc.html) but as sparse experimental evidence for gene function exists in sorghum, and many of the annotations only list similarity to other putative genes, it is valuable to attempt to establish a more direct link with experimentally demonstrated gene functions. Examples of each gene

were identified where protein function had been established, if such examples existed, with validation of expressed mRNA used if no protein validation was available.

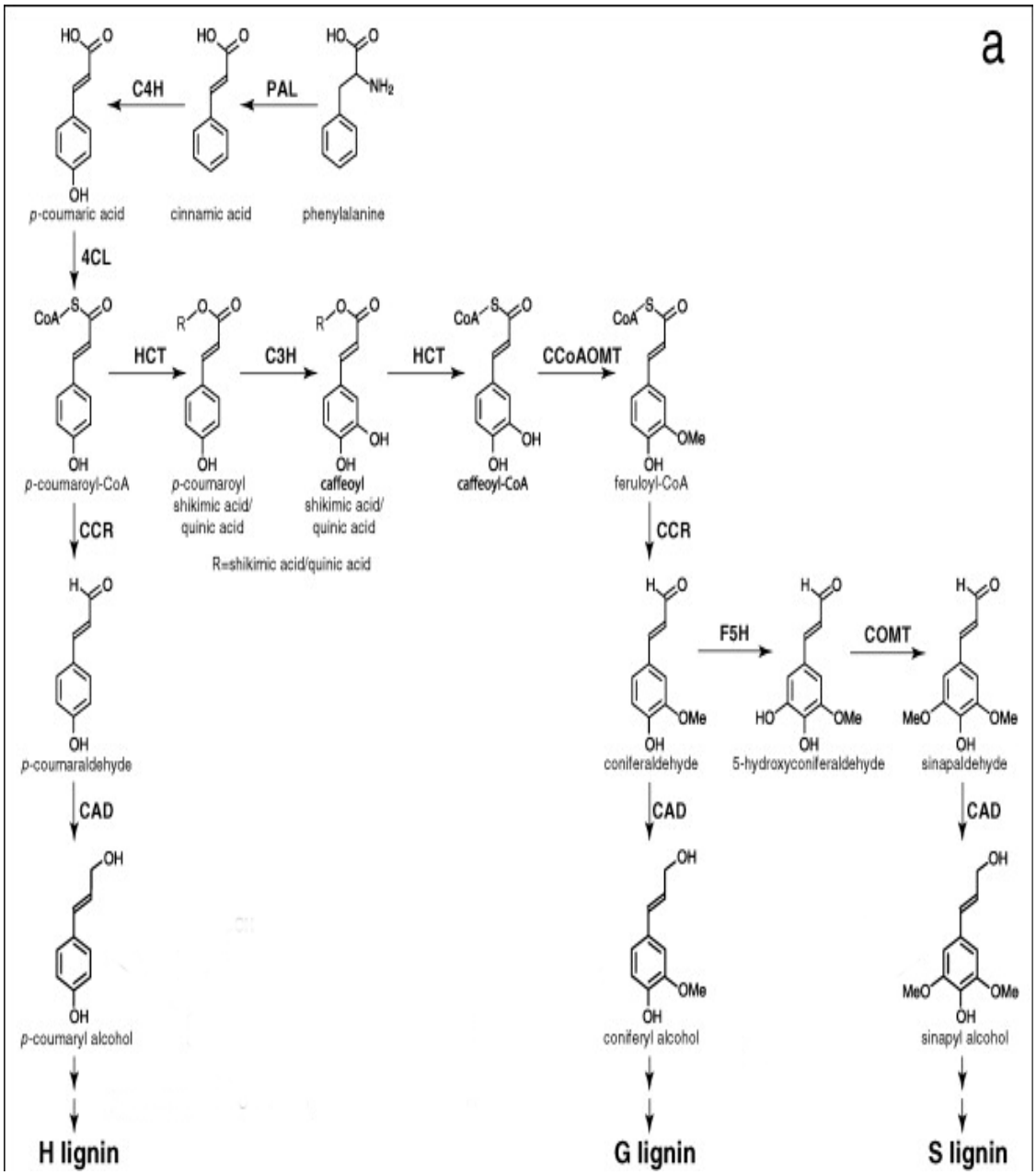


Figure 14: Monolignol biosynthetic pathway. (Vanholme et al., 2008).

Table 1: Identification of putative monolignol biosynthetic genes in Sorghum bicolor.

Gene Product Originating Organism Gene ID (Biosynthetic Product)	**Confirmation**	**Sorghum Gene(s)**	**% Coverage**	**% Identity**
PAL1_ORYSJ *Oryza sativa* Os02g0626100 (PAL)	Minami et al., 1989 Komatsu et al., 2004	Sb04g026510.1 Sb04g026560.1 Sb04g026550.1 Sb04g026540.1 Sb04g026530.1 Sb04g026520.1 Sb06g022740.1 Sb06g022750.1 Sb01g014020.1	84 82 82 82 82 82 84 79 80	90 82 82 82 82 83 89 82 79
CYP73A5 *Arabidopsis thaliana* At2g30490.1 (C4H)	Mizutani et al., 1997	Sb02g010910.1 Sb03g038160.1 Sb04g017460.1	79 87 54	70 71 75
At4CL1 *Arabidopsis thaliana* At1g51680 (4CL)	Ehlting et al., 1999	Sb10g026130.1 Sb07g022040.1 Sb04g005210.1 Sb07g007810.1 Sb04g031010.1	52 61 52 52 69	68 71 66 65 64
CAD47830 *Nicotiana tabacum* AJ507825.1 (HCT)	Hoffmann et al., 2004	Sb04g025760.1 Sb04g035780.1 Sb06g021640.1	100 100 99	64 62 43
AY149607 *Hordeum vulgare* AY149607.1 (CCR)	Larsen, K., 2004	Sb07g021680.1 Sb10g005700.1 Sb02g014910.1 Sb04g005510.1 Sb04g036780.1 Sb04g036770.1	77 77 76 73 73 73	88 81 75 72 71 71
Bmr6 *Sorghum bicolor* Sb04g005950.1 (CAD)	Sattler et al., 2009	Sb04g005950.1 Sb06g001430.1 Sb02g024210.1 Sb02g024190.1 Sb02g024220.1 Sb07g006090.1	* 93 93 94 93 93	* 65 65 65 67 65

Table 1 Continued

Gene Product Originating Organism Gene ID (Biosynthetic Product)	Confirmation	Sorghum Gene(s)	% Coverage	% Identity
AY107051 *Zea mays* AY107051.1 (C3H)	Riboulet et al., 2008	Sb03g037380.1 Sb09g024210.1	93 82	92 79
AY675076 *Sorghum bicolor* AY675076.1 (F5H)	Boddu et al., 2004	Sb04g024710.1 Sb04g024750.1 Sb04g024730.1	58 59 65	100 97 93
CAB45149 *Zea mays* AJ242980.1 (CCoAOMT)	Joan, 1999	Sb10g004540.1 Sb07g028520.1 Sb02g027930.1 Sb07g028530.1 Sb07g028490.1	97 60 59 60 59	90 74 74 73 72
AAO43609 *Sorghum bicolor* HQ668169.1 (COMT)	Bout and Vermerris, 2003	Sb07g003860.1	92	99

Identification of Sorghum Homologs of Monolignol Biosynthetic Genes

Monolignol biosynthetic gene sequences were obtained where expression and, if possible, function had been experimentally confirmed (Table 1). Sequences were compared using the discontiguous megablast function against the published *Sorghum bicolor* genome. Putative genes were taken from the closest available phylogenetic relative of Sorghum, and the putative genes were only selected if they showed at least 50% sequence coverage and an e value no higher than 1e-50. The sole exception to this is the sorghum COMT and F5H genes: full length mRNA sequences have been obtained from sorghum, but the predicted genes that underlie that genomic region do not match

perfectly. This most likely represents errors in the gene prediction algorithms used to identify putative sorghum genes. In the case of COMT, the alignment is an incorrect match to a single predicted gene, perhaps caused by a small error in the gene or mRNA sequence. For F5H, the alignment predicts partial matches to a cluster of nearby putative genes. Given the high identity and fractional nature of the alignment, it is likely that one or more of these putative genes are actually the same gene, incorrectly annotated.

Not surprisingly, many of these genes appear to be present in multiple copies. Gene duplication and redundancy are common in grasses and C3H and C4H are members of the cytochrome p450 monooxygenase-dependent family of enzymes which have undergone significant duplication in sorghum (Paterson et al., 2009).

Many of the biosynthetic genes appear together in clusters, most notably on chromosomes 3, 4, 6, 7 and 10 (Fig. 15). The cluster of CCR, 4CL, and CAD on the p arm of chromosome 4 are within 700 kb of each other, a detail that will prove relevant if future mapping indicates a QTL within that region.

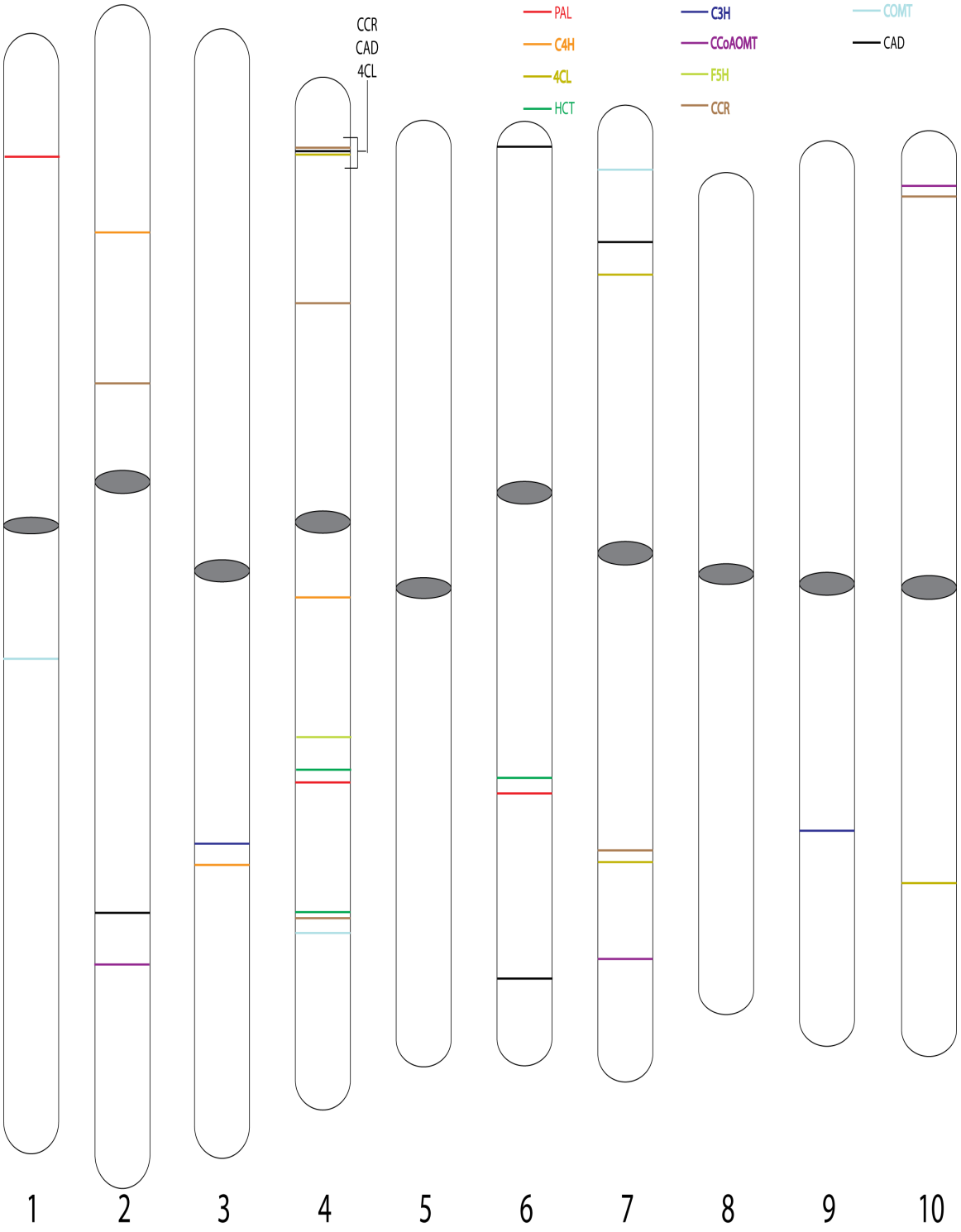


Figure 15: Physical locations of putative sorghum biosynthetic genes. Numbers on base of figure indicate chromosome number. Grey ovals represent centromeric regions.

Discussion

Identification of the genes involved in monolignol biosynthesis will prove to be crucial to the development of assays involved in identifying the genetic basis of lignin variation in sorghum. In addition to having a visible reference for determining whether a particular biosynthetic gene lies within a QTL, even trans-acting QTL will need at some point to interact with the biosynthetic genes. While these QTL are as yet undetermined, once established having a monolignol reference will make candidate gene discovery more rapid.

Previous work involved in the modification of lignin content through perturbation of the monolignol biosynthetic pathway has shown that changes in the enzymes themselves can have unpredictable and sometimes undesirable effects on the organism as a whole (Vanholme et al., 2008). It is likely that QTL analysis will reveal that naturally occurring lignin variation is the result of trans-acting factors that regulate the levels of expression of the biosynthetic enzymes, rather than the enzymes themselves (Patzlaff et al., 2003; Legay et al., 2007; Zhong and Ye, 2009). In this case, identification of the monolignol biosynthetic genes is even more critical, as it will be necessary to identify which of the enzymes are being regulated by each transcription factor.

Lignin composition is also of concern for those plants desired for industrial application, and regulation of the biosynthetic enzymes would be directly relevant toward favoring one type of monolignol over another. In such a situation, being able to identify where in the pathway regulation is taking place would be of high importance.

DEVELOPMENT OF GENOMIC TOOLS

Introduction

One of the challenges facing molecular biologists and biochemists is the difficulty of identifying the genetic locus for a trait present in a naturally occurring population. Unlike T-DNA insertion lines, there is not a sequence tag that can be easily located to identify where a sequence variant has occurred, and unlike TILLING populations, differences in sequence are not arising in a genetically uniform background. In the case of stem composition variants, the problem is further compounded by the challenge of rapidly identifying the phenotypic variation in question.

One technique that is used to identify the genetic basis for variation in traits is Quantitative Trait Loci (QTL) analysis. QTL analysis is a statistical method that allows researchers to combine genetic data (in this case molecular marker alleles) with phenotypic data to determine the genetic basis for the variation of complex traits segregating in a population (Falconer and Mackay, 1996, Kearsey 1998). This is a powerful tool for localizing the genes/alleles causing for phenotypic variation, but it requires a high fidelity genetic map and good phenotyping data.

Traditionally, markers such as restriction fragment length polymorphisms (RFLPs), short sequence repeats (SSRs) and amplified fragment length polymorphisms (AFLPs) were used in the creation of genetic maps, all of which were markers that could be physically identified during gel electrophoresis (such as SSRs), or identified using fluorescence tagging (such as AFLPs) (Gupta and Rustgi, 2004). Deriving these markers is expensive and time consuming, and analysis is generally limited to a

relatively small number of markers for each group of samples. Fortunately these methods do not require the possession of a genome sequence for them to function, allowing for the construction of genetic maps in any organism. Before the *Sorghum bicolor* genome was sequenced, much of the genetic information available originated from such sources (Boivin et al., 1999; Bhattramakki et al., 2000; Menz et al., 2002; Kim et al., 2005).

Utilizing next-generation sequencing, millions of sequences can be obtained simultaneously, and sample multiplexing allows for many different samples to be analyzed at the same time. The challenge in this process is dealing with the enormous amount of data that is produced: each lane on an Illumina GaIIx system produces over three gigabytes, and the newer HiSeq systems can produce sixteen gigabytes per lane. Such a quantity of data requires the development of a rapid analysis platform.

With the publication of the *Sorghum bicolor* genome sequence, such a platform becomes much simpler to implement. This study demonstrates the development of a system for rapid genotyping of *Sorghum bicolor* using publicly available tools and datasets, producing data that can be rapidly used for genetic mapping.

Results

Initial Data Processing

Genomic DNA from parental lines and progeny is digested with the restriction enzyme *FseI* followed by the ligation of adapters containing ID tags. Running multiple samples per lane lowers the overall read count for each sample, and by only sequencing

areas flanking known restriction sites the read depth at each of these locations can be significantly increased improving accuracy.

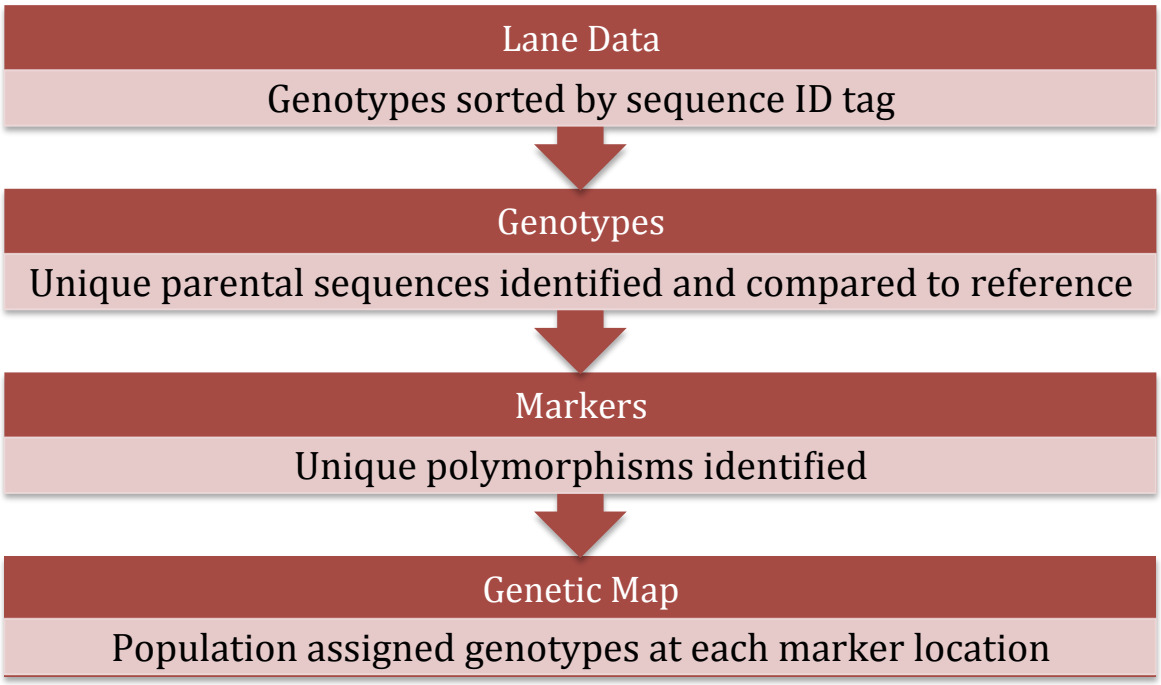


Figure 16: Generalized workflow for generation of sequence based markers from next generation sequencing data.

Multiplexing is critical to the use of next generation sequencing (NGS) technologies in the generation of genetic markers and subsequent genotyping. As such, each of the samples being analyzed can be identified by a four nucleotide ID tag that has been ligated on the 5' end of the DNA template being sequenced. Each lane is analyzed independently to allow the software to run on the memory limitations of a desktop workstation, and to allow some lanes to be used for genotyping while others are used for other projects. During the sorting process, quality control data, lane flanking

information, and ID tags are removed to prepare the data for downstream processing and to save memory. After sorting, data is saved to independent files for ease of access and downstream processing.

Marker Discovery

Parental genotype sequences are compared and all non-unique sequences are removed. Unique sequences that align to only one location in the reference genome sequence with read counts of less than five are also removed to ensure high confidence in putative markers. Sequences are then aligned to the genomic sequence using the BLAST algorithm (Altschul et al., 1990). Installing a local instance of the BLAST program is relatively easy, and provides very rapid alignment of short sequence reads. Output from BLAST is parsed, and sequences with more than one perfect alignment to the genome sequence are discarded. Sequences showing variation from the reference genome are then compared between parental lines to prevent identical polymorphisms from being declared as markers. In cases where only one parent has a detectable polymorphism, the other is assigned the reference sequence.

Digital Genotyping

Once generated, the list of alleles for each marker is loaded into memory. Each genotype's data is then read in, line by line, and compared to the stored sequences. Most of the sequences are not polymorphic between genotypes and thus cannot be used for mapping as there is no sequence variation to identify lineage (Fig. 16). Sequences that do match to one of the parental alleles are stored in memory, with a notation of how many such reads were present. If less than three reads are present, the sequence is not

scored as there is not sufficient read depth to guarantee accuracy. In some cases the genotypes being scored have sequences present from both parents at a given marker, these markers are flagged as potential residual heterozygosity if one allele is not represented at least an order of magnitude greater than the other.

Once all the genotypes have been assigned, the data is ordered by physical location on the genome and output into a readable file for further analysis (Fig. 17). After minor formatting changes, this file can be used to generate recombinational distances for QTL mapping.

Marker Database Development

One downside of the technique discussed above is the need to develop new markers each time a run is performed. Not only does this require valuable sample space and require lengthy computation, but also the low read number from each genotype ensures that some regions will not be sequenced sufficiently deeply to use them as genetic markers.

A solution to this problem is to store the confirmed markers from each run in a database such that they will be accessible for subsequent runs. This serves to allow users to genotype at markers that may not have sequenced deeply enough for assignment during their particular sequencing run, and also allows users to ensure their parental lines have not suffered any contamination, since they can be compared at each locus with stored data from previous analyses.

A deep sequencing project undertaken in collaboration with Dr. Patricia Klein provided initial data. The data consisted of a large set of genotyping markers sequenced multiple times for many genotypes, and provided data similar to that described above.

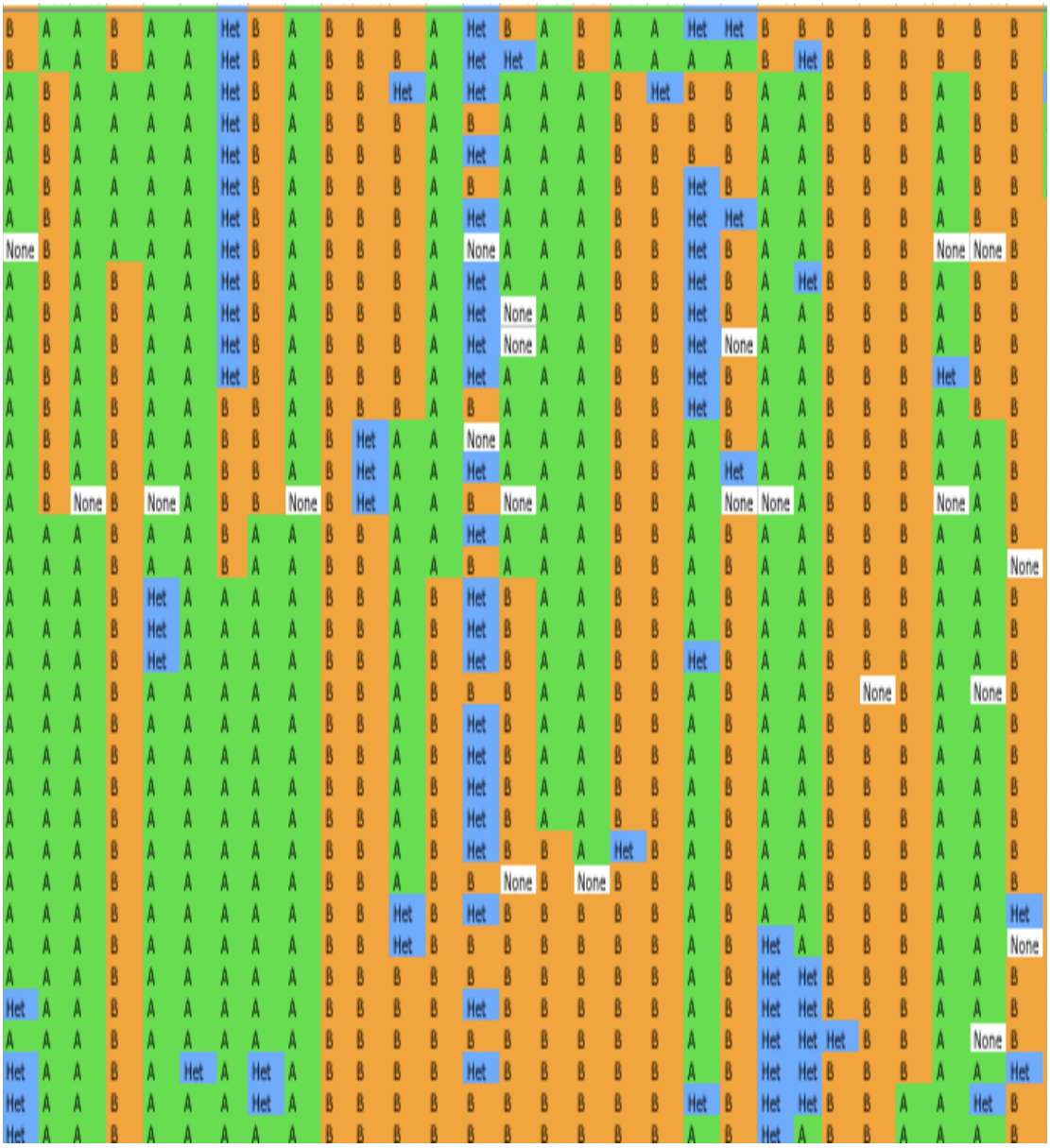


Figure 17: Display of genetic alleles within a RIL population. Color has been applied for ease of human readability.

In all, 30 genotypes of sorghum (RTx430, R07045, 80M, SC748, SC56, Rio, BTx3197, R07007, R07008, Tx7000, BHK, RTx436, P850029, SC170, 100M, R07012, BTx623, R07020, BTx642, Hegari, DYM, BTx406, M35, IS3620c, R07030, RTx2536, R07054, R07018, R07034, R07042) were sequenced at 33,126 loci, generating a set of 73,829 total unique sequences. One of the benefits of restriction site anchored sequencing is that all the sequences will be anchored to the same sites on the genome, barring those with variants within the actual restriction site. As such, the data was stored in a MySQL database in a schema similar to that illustrated (Fig. 18). The database schema itself is relatively simple, allowing for significant expansion if different genotyping techniques are to be employed.

From the establishment of a database, scripts can be used to access any relevant sections of the genome from a web interface (Fig. 19). This data can also be exported as a text file so it may be accessed by the genotyping software described above.

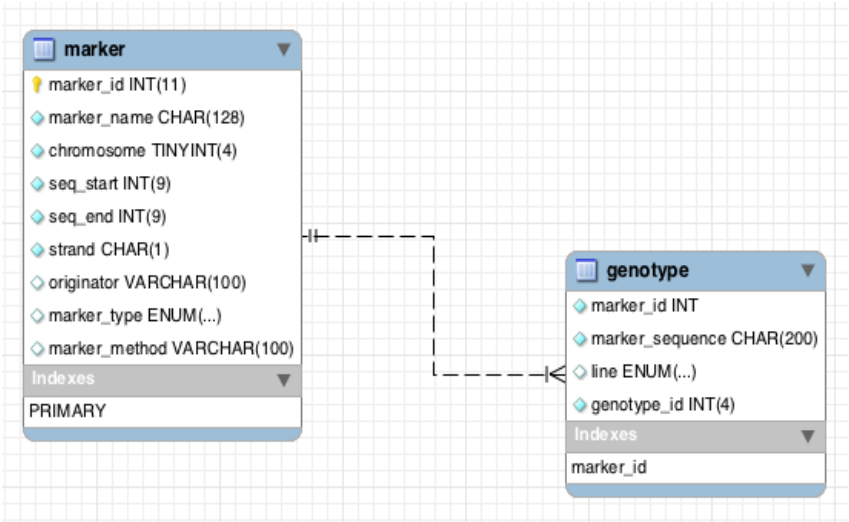
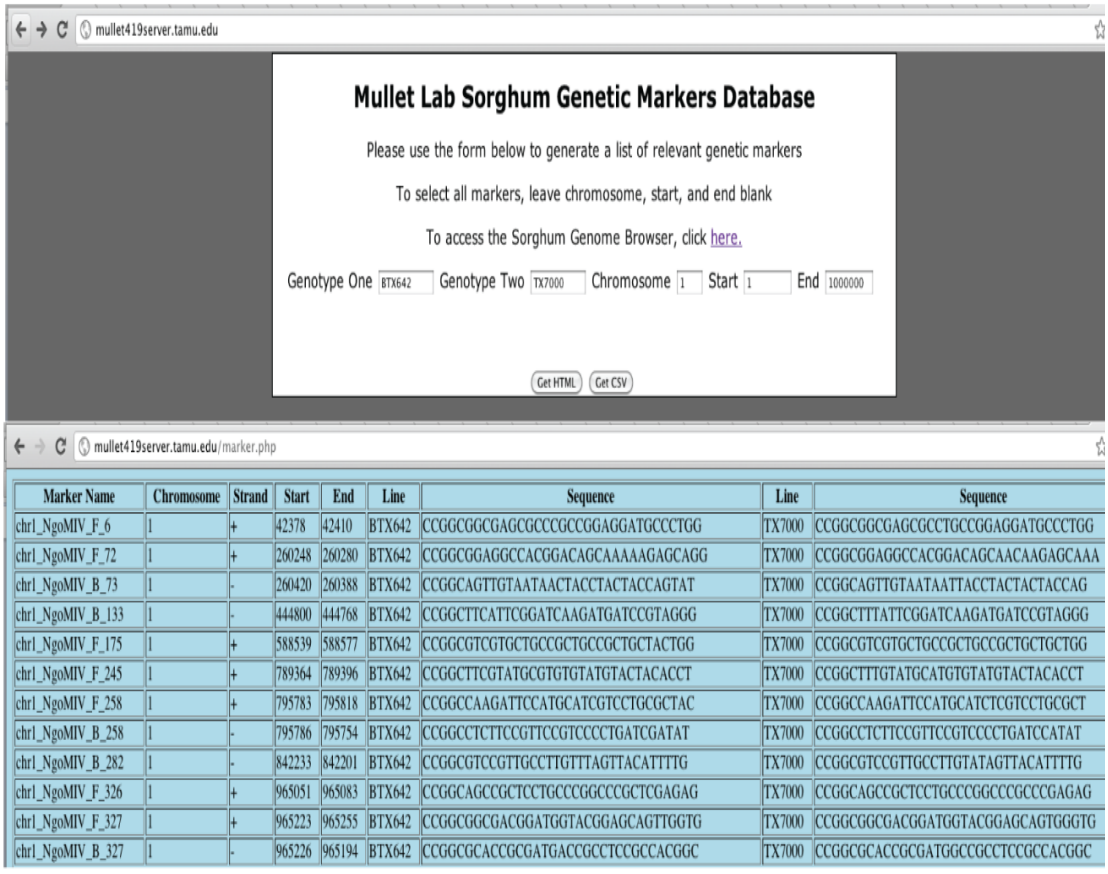


Figure 18: Graphical display of database layout.



Marker Name	Chromosome	Strand	Start	End	Line	Sequence	Line	Sequence
chr1_NgoMIV_F_6	1	+	42378	42410	BTX642	CCGGCGGCGAGCGCCCGCCGGAGGATGCCCTGG	TX7000	CCGGCGGCGAGCGCCTGCCGGAGGATGCCCTGG
chr1_NgoMIV_F_72	1	+	260248	260280	BTX642	CCGGCGGAGGCCACGGACAGCAAAAAGAGCAGG	TX7000	CCGGCGGAGGCCACGGACAGCAACAAGAGCAAA
chr1_NgoMIV_B_73	1	-	260420	260388	BTX642	CCGGCAGTTGTAATAACTACCTACTACCAGTAT	TX7000	CCGGCAGTTGTAATAATTACCTACTACTACCAG
chr1_NgoMIV_B_133	1	-	444800	444768	BTX642	CCGGCTTCATTCGGATCAAGATGATCCGTAGGG	TX7000	CCGGCTTTATTCGGATCAAGATGATCCGTAGGG
chr1_NgoMIV_F_175	1	+	588539	588577	BTX642	CCGGCGTCGTGCTGCCGCTGCCGCTGCTACTGG	TX7000	CCGGCGTCGTGCTGCCGCTGCCGCTGCTGCTGG
chr1_NgoMIV_F_245	1	+	789364	789396	BTX642	CCGGCTTCGTATGCGTGTGTATGTACTACACCT	TX7000	CCGGCTTTGTATGCATGTGTATGTACTACACCT
chr1_NgoMIV_F_258	1	+	795783	795818	BTX642	CCGGCCAAGATTCCATGCATCGTCCTGCGCTAC	TX7000	CCGGCCAAGATTCCATGCATCTCGTCCTGCGCT
chr1_NgoMIV_B_258	1	-	795786	795754	BTX642	CCGGCCTCTTCCGTTCCGTCCCCTGATCGATAT	TX7000	CCGGCCTCTTCCGTTCCGTCCCCTGATCCATAT
chr1_NgoMIV_B_282	1	-	842233	842201	BTX642	CCGGCGTCCGTTGCCTTGTTTAGTTACATTTTG	TX7000	CCGGCGTCCGTTGCCTTGTATAGTTACATTTTG
chr1_NgoMIV_F_326	1	+	965051	965083	BTX642	CCGGCAGCCGCTCCTGCCCGGCCCGCTCGAGAG	TX7000	CCGGCAGCCGCTCCTGCCCGGCCCGCCCGAGAG
chr1_NgoMIV_F_327	1	+	965223	965255	BTX642	CCGGCGGCGACGGATGGTACGGAGCAGTTGGTG	TX7000	CCGGCGGCGACGGATGGTACGGAGCAGTGGGTG
chr1_NgoMIV_B_327	1	-	965226	965194	BTX642	CCGGCGCACCGCGATGACCGCCTCCGCCACGGC	TX7000	CCGGCGCACCGCGATGGCCGCCTCCGCCACGGC

Figure 19: Access page to request marker data from sorghum database and results of displayed query.

Graphical Access to Marker Data

While having access to marker data in tabular text format is sufficient for developing genetic maps and QTL analysis, it is not ideally suited for human browsing. As such, the marker database was converted into a format readable by the ubiquitous GBrowse genome visualization program (Stein et al., 2002). This allows all known markers to be plotted as a track in the GBrowse instance, and custom scripts allow the user to extract sequence variants and genotype information at each marker location (Fig. 20A). By using the MySQL database backend as demonstrated above, over 900,000 sequence variants have been added for convenient browsing. (Fig. 20B)

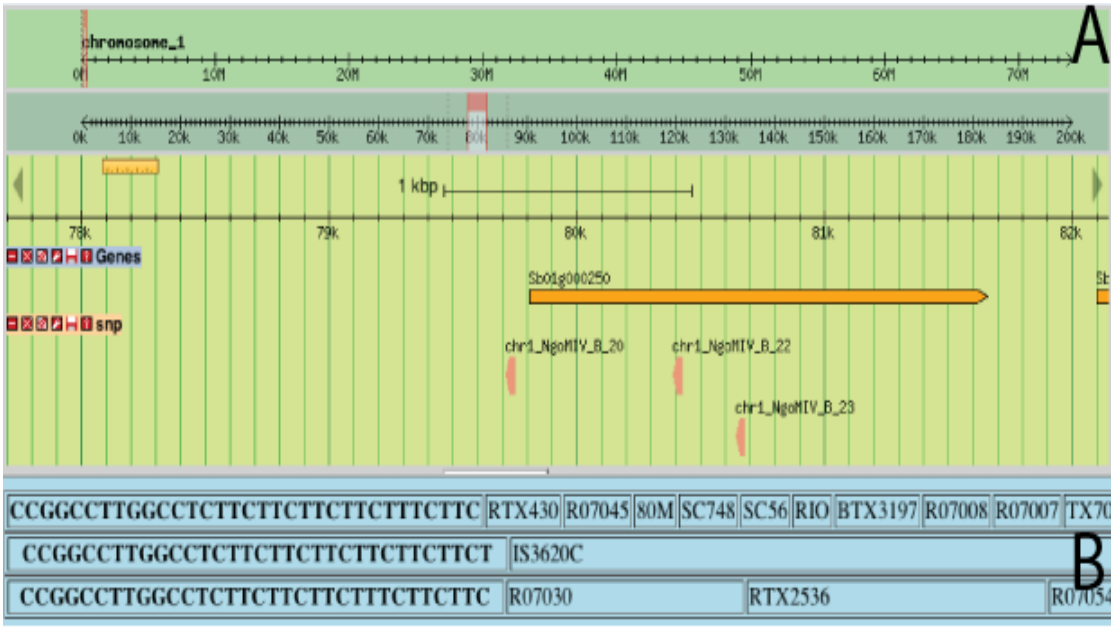


Figure 20: (a) GBrowse display of available sorghum genetic information, including structural annotations and markers. (b) Selection of a genetic marker results in display of marker sequence and all known alleles.

Discussion

Sequence-based genetic markers allow researchers to rapidly uncover the physical locations of the QTL responsible for phenotypes of interest. The physical locations of the markers are determined during the initial phases of marker discovery, and this allows researchers to quickly anchor the QTL to the chromosome and uncover the genes within QTL intervals causing phenotypic variation. As more researchers gain access to high throughput sequencing platforms, the ability to process this data will become more important, as will the ability to easily visualize and share the results.

Here has been shown a rapid technique for extracting genetic markers from NGS platforms, suitable for execution on a standard desktop workstation. All of the components involved (GBrowse, MySQL, various Perl and PHP scripts) are freely available, avoiding the high costs of other genomics solutions. Average processing time for genotyping analysis (processing raw data to generating data suitable for generating genetic distances) is less than 24 hours on currently available desktop computers.

QUANTITATIVE TRAIT LOCUS MAPPING FOR STEM COMPOSITION TRAITS

Introduction

Along with other C4 grasses such as miscanthus (*Miscanthus* spp.), sugarcane (*Saccharum officinarum* L.), and switchgrass (*Panicum virgatum* L.) sorghum is a promising candidate for future biofuels production (Farrell et al., 2006). Approximately 80% of the biomass of energy sorghum is present in stems at the end of the growing season. There are numerous technologies for the conversion of stem carbohydrates to biofuels. Optimizing the stem carbohydrate composition for each of these methods could potentially allow for greater conversion efficiency and reduced need for land usage (Hamelinck et al., 2005). Stem carbohydrates are separated into the structural and nonstructural categories, with non-structural carbohydrates consisting primarily of simple sugars such as sucrose, as well as starch and structural carbohydrates consisting primarily of cellulose and hemicellulose. Intertwined with the structural cell wall carbohydrates is a complex mesh of lignins, which usually have a negative effect on stem biofuel conversion efficiency (Hamelinck et al., 2005; Ragauskas et al., 2006; Weng et al., 2008). Identifying the genetic basis for control of these traits would allow for selective optimization of stem composition components, which are ideal for biomass destined for different conversion programs.

Sorghum is an important source of annual feed silage, as well as a promising source of biomass for biofuels production. Sorghum is drought resistant, capable of

providing reasonable yields despite reduced water supply, and is often grown as a "hedge crop" by farmers worried about water deficit (Sanderson et al., 1992; EPA, 2000). Sorghum silage is, as a whole, less digestible for cattle than maize silage, largely thought to be a result of higher amounts of lignins present in the stem. Lignin serves multiple important purposes in sorghum, such as providing waterproofing for the tracheary elements, but its presence is also detrimental to the ability for cattle and enzymatic systems to break down the cellulose and extract energy from the silage material (Humphreys and Chapple, 2002).

Despite over 200,000 acres of sorghum being planted for silage in the US in 2011, with a total yield of over 2 million tons (USDA, 2012), very little is understood about the genetic control of stem composition in sorghum. Progress has been made previously in identifying the biosynthetic components of lignin biosynthesis in, but the genetic basis of larger scale control of secondary cell wall biosynthesis remains largely unknown sorghum (Bucholtz et al., 1980; Bout and Vermerris, 2003). An additional challenge to developing a framework for the control of stem composition is the difficulty of identifying phenotypes. Lignin is particularly intransigent to measurement (Hatfield and Fukushima, 2005), though the other main stem components (cellulose, hemicellulose, starch, sucrose, and protein) are somewhat more easily quantified (Hatfield et al., 1999).

Near-Infrared Reflectance Spectroscopy (NIR) is a technique that has been developed to allow for the rapid, non-destructive analysis of sample composition (Poke and Raymond, 2006; Labbé et al., 2008). Once a calibration curve has been generated

through standard chemical techniques this calibration curve can then be applied to NIR scans of new material to create an accurate estimate of the components in that material (Labbé et al., 2007; Sluiter et al., 2008). This allows for quick analysis, since the material needs only be ground and scanned rather than subjected to extensive chemical analysis. Using NIR, several hundred samples can be analyzed per day, allowing for analysis of whole populations in a reasonable timeframe.

Once composition data is determined, it can be combined with genetic marker data to identify regions of the genome that modulate phenotypic variation. Once sequence markers have been established for a population, they can be combined with phenotypic trait data such as stem composition to establish quantitative trait loci (QTL), which are the genes/genic regions responsible for phenotypic variation in a population (Geldermann, 1975). QTL have the benefit that they are able to identify multiple sources of variation that contribute to a continuously varying trait, allowing for the potential determination of many genes involved in the variation of stem composition (Kearsey and Pooni, 1996).

Results

Construction of Genetic Maps

Two populations were analyzed in this study, a population of recombinant inbred lines (RILs) originally generated from a cross between the sorghum cultivars SC56 and RTx7000, and an additional RIL population generated from a cross between the sorghum cultivars BTx642 and RTx7000. DNA from RILs of both populations were sequenced on an Illumina GAIIx system and unique polymorphic sequences identified and their

physical locations found in the genome (Weers, 2011). This analysis yielded a total of 566 markers for the BTx642 x RTx7000 population (Fig. 21, Weers 2011).

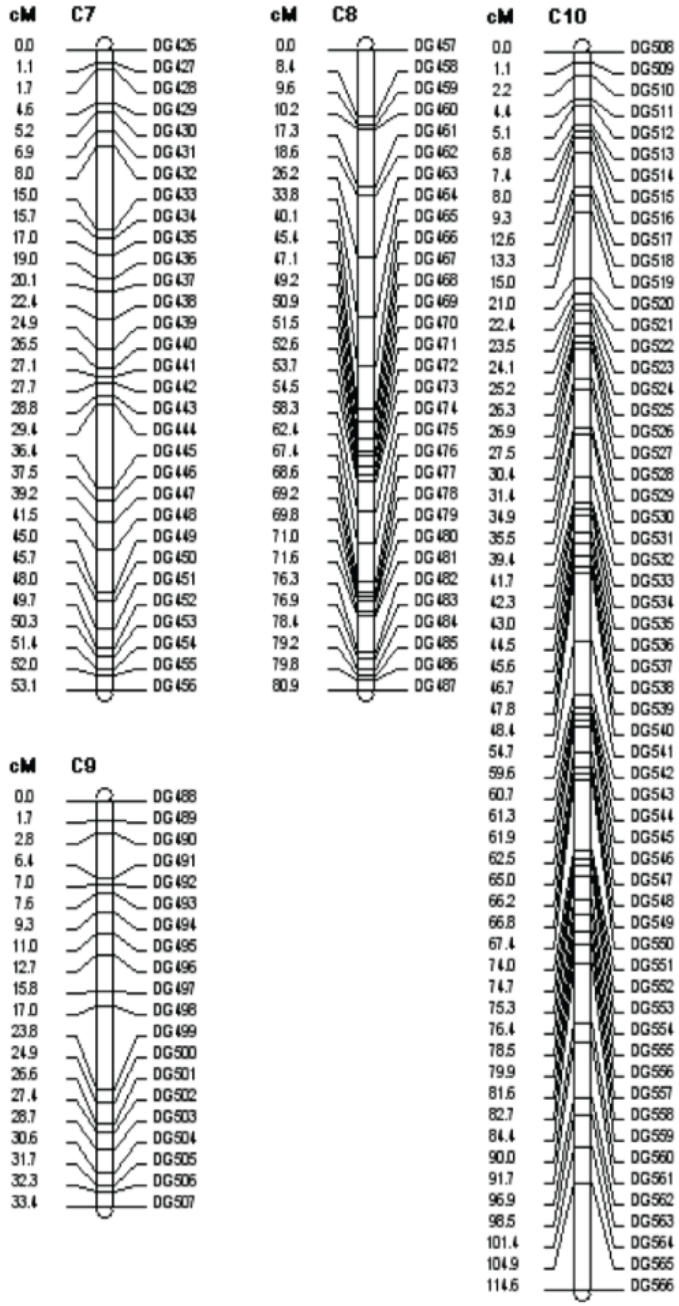


Figure 21: Genetic map of chromosomes 7-10 for BTx642xRTx7000 RIL population. Genetic distance is listed to the left of each chromosome.

Similar analysis was performed on the SC56 x Tx7000 RIL population, generating 392 markers (Fig. 22, Appendix Fig. A-1 and A-2). Genetic distances were then determined using the Kosambi mapping function provided by the MapMaker software.

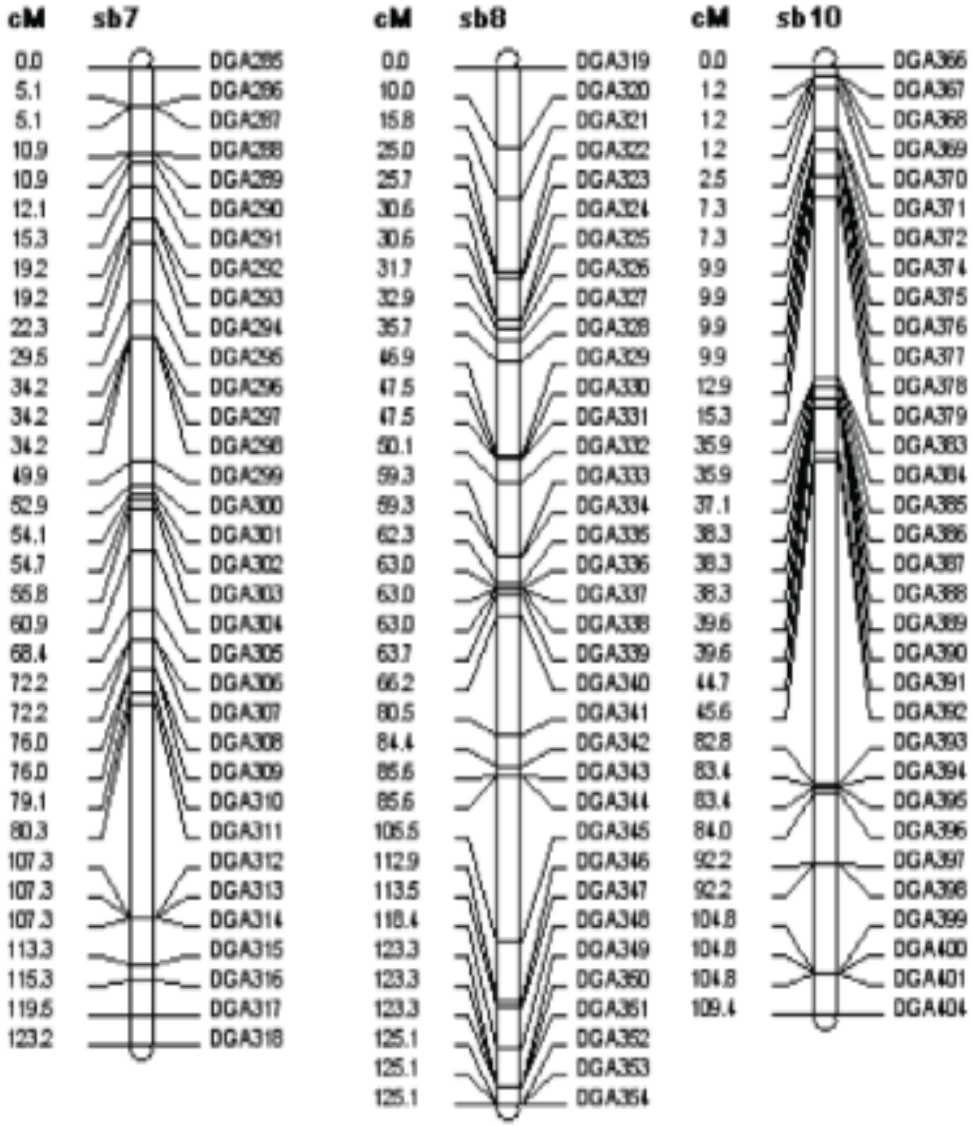


Figure 22: Genetic map of chromosomes 7, 8,and 10 for SC56xRTx7000 RIL population. Genetic distance is listed to the left of each chromosome.

Trait Measurement

For each population, 90 lines were grown in College Station in summer of 2009 (SC56 x RTx7000) and 2010 (BTx642 x RTx7000) and harvested in the field, with five individuals selected from each line. Samples were measured, and the leaves and leaf sheaths removed from the stems. Stems were then dried in a forced air oven for 72 hours, and ground in a UDY cyclone sample mill (UDY Corporation, Fort Collins, CO) until passing through a 1mm screen. Samples were then scanned on a FOSS XDS Rapid Solids Analyzer and the resultant spectra interpreted via the ISIScan software package and a calibration curve provided by the National Renewable Energy Laboratory.

NIR analysis reveals the relative amount of various stem components, expressed as a percent value of the total amount of material scanned (Appendix Tables 2 and 3). Each of the nine traits examined (ash, protein, sucrose, lignin, xylan, cellulose, water extractives, alcohol extractives, and starch) was then moved forward for QTL analysis.

Composite Interval Mapping

Once phenotypes were identified and the genetic map constructed, the composite interval mapping function of QTL Cartographer was implemented to generate a map of the QTL responsible for the variation in each of the stem composition traits. Composite interval mapping (CIM) is useful in this case since there are likely multiple, possibly interacting QTL for each phenotype, and CIM allows for these QTL to be more readily distinguished (Jansen, 1996).

Analysis of the BTx642xTx7000 population revealed a total of 34 QTL for stem composition traits, mainly centered on chromosomes 1,3,6 and 8 (Fig. 23). Several other

peaks are visible but did not meet the significance threshold established by permutation tests (1000 permutations, significance value 0.05). Previous work has demonstrated the presence of a flowering time locus on chromosome six, which would be located at approximately 33.2 cM, near the marker DG376 (Murphy et al., 2011).

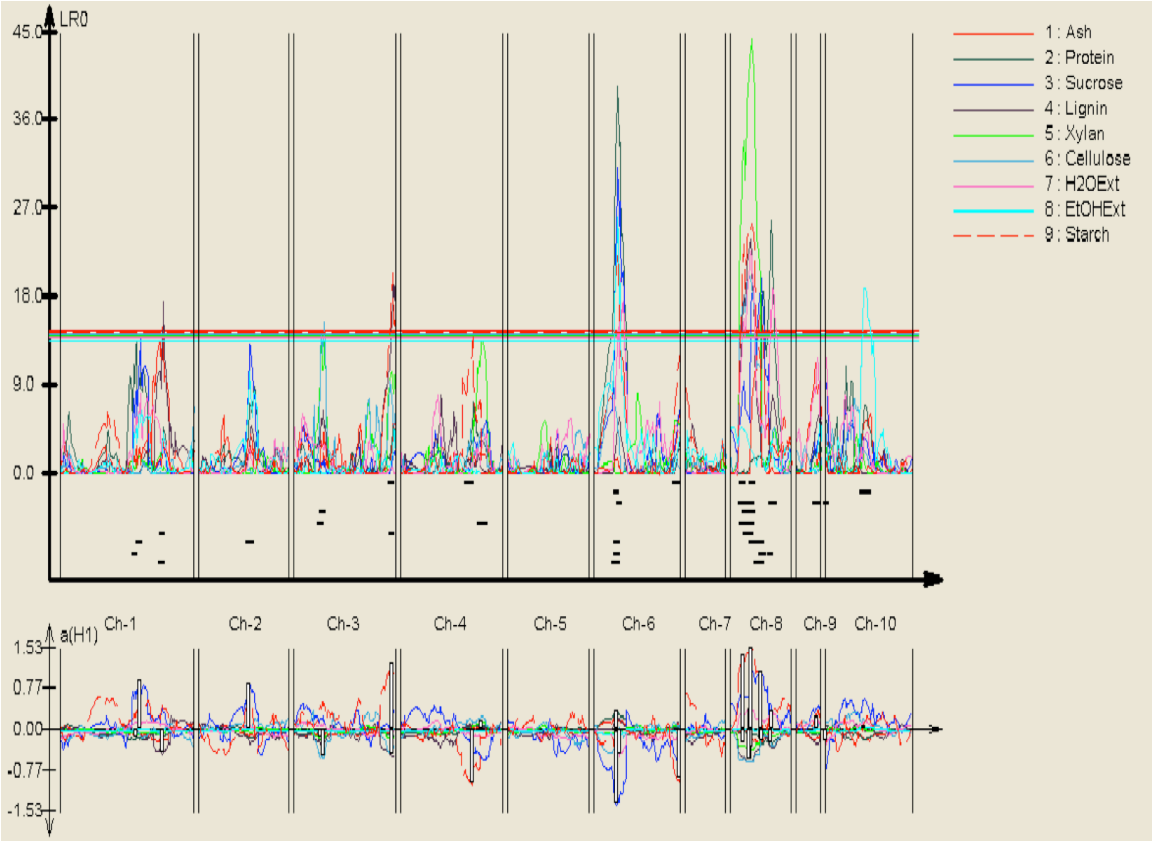


Figure 23: QTL map of BTx642 x Tx7000 RIL population determined via composite interval mapping. Height of peak indicates confidence of QTL identification. Horizontal colored bars indicate significance threshold.

Plotting QTL reveals the peak of the primary QTL cluster on chromosome six to be located between 30.1 and 33.2 cM (Table 2). While both BTx642 and Tx7000 come from *ma1* backgrounds,recent work has shown that BTx642 has a weak allele of *ma1,*

causing crosses to segregate at the Ma1 locus (Smith and Frederikson, 2000; Harris, 2007; Weers, 2011). This segregation is believed to be responsible for the QTL cluster centered on DG376 on chromosome six.

Table 2: Location of QTL for stem composition traits in BTx642 x Tx7000 RIL population as determined by CIM. Additive effect denotes the amount of variation that locus is responsible for, in units of percent total stem volume. R2 indicates the amount of variation that locus is responsible for in terms of variation within the population. LOD1 left and right columns indicate the physical locations of the left and right boundaries of the QTL, in this case meaning the area wherein the LOD score of the QTL is within one point of that of the peak.

Trait	QTL#	Chromosome	Marker	Position (cM)	LOD	Additive	R^2	LOD1 Left	LOD1 Right
Ash	1	1	68	135.8	3.56	-0.2027	0.1029	60210749	60802865
Ash	2	6	7	30.1	4.8286	0.2421	0.1572	4711341	40120199
Ash	3	8	9	40.1	3.6478	-0.2133	0.1151	6254008	47392453
Protein	1	1	51	105.5	2.8984	-0.1826	0.075	51826648	52825635
Protein	2	6	7	30.1	8.5785	0.3424	0.26	4711341	40120199
Protein	3	8	9	42.1	4.0421	0.2411	0.1358	41623973	47392382
Protein	4	8	16	53.6	5.6167	0.2542	0.1533	48946234	49740291
Sucrose	1	1	54	105.8	2.9943	0.9201	0.0783	52825635	54216748
Sucrose	2	2	28	66.5	2.8805	0.8565	0.0744	55580245	58017106
Sucrose	3	6	7	30.1	6.7977	-1.3888	0.1968	4711341	40120199
Sucrose	4	8	7	28.2	3.9732	1.1147	0.1271	5330131	6254008
Sucrose	5	8	9	40.1	4.3531	1.0811	0.118	6254008	47392453
Lignin	1	1	69	136.4	3.8047	-0.447	0.1024	60210749	60802865
Lignin	2	3	88	131.1	4.4324	-0.4826	0.1212	72173211	72711436
Lignin	3	8	7	26.2	5.191	-0.569	0.148	3339660	6254008
Xylan	1	3	26	35.9	3.0251	-0.154	0.0767	7940280	8485419
Xylan	2	4	64	107	2.9397	0.1583	0.0787	62339651	63335625
Xylan	3	8	5	17.3	7.4016	-0.273	0.2389	2476297	3339660
Xylan	4	8	7	27.2	9.6496	-0.3229	0.3126	3339660	6254008
Cellulose	1	3	31	39.4	3.3641	-0.5152	0.1152	8348932	9985015

Table 2 continued

Trait	QTL#	Chromosome	Marker	Position (cM)	LOD	Additive	R2	LOD1 Left	LOD1 Right
Cellulose	2	8	6	24.6	4.3728	-0.5736	0.1421	3339660	6254008
H2O Extractives	1	6	10	33.2	4.3267	-0.4695	0.1091	40120270	41856240
H2O Extractives	2	8	5	17.3	3.4179	0.3196	0.0849	2476297	3339660
H2O Extractives	3	8	7	26.2	4.938	0.3685	0.12	3339660	6254008
H2O Extractives	4	8	17	55.5	4.1066	0.3338	0.1022	49587492	50817616
H2O Extractives	5	9	15	27.4	2.5674	0.223	0.0504	2970840	4281712
H2O Extractives	6	10	1	0	2.6012	-0.2435	0.058	91868	597201
EtOH Extractives	1	6	7	30.1	5.716	-0.0835	0.1931	4711341	32354861
EtOH Extractives	2	10	33	50.4	4.1237	0.0762	0.1385	8592781	12145020
Starch	1	3	87	130.5	4.4541	1.2446	0.1231	71832292	72711436
Starch	2	4	52	95	3.0402	-1.0095	0.0814	58864301	61030074
Starch	3	6	58	112.6	2.6868	-0.9233	0.0714	58810683	61702168
Starch	4	8	5	17.3	5.0969	1.3695	0.1482	2476297	3339660
Starch	5	8	7	28.2	5.5492	1.5061	0.1782	5330131	6254008

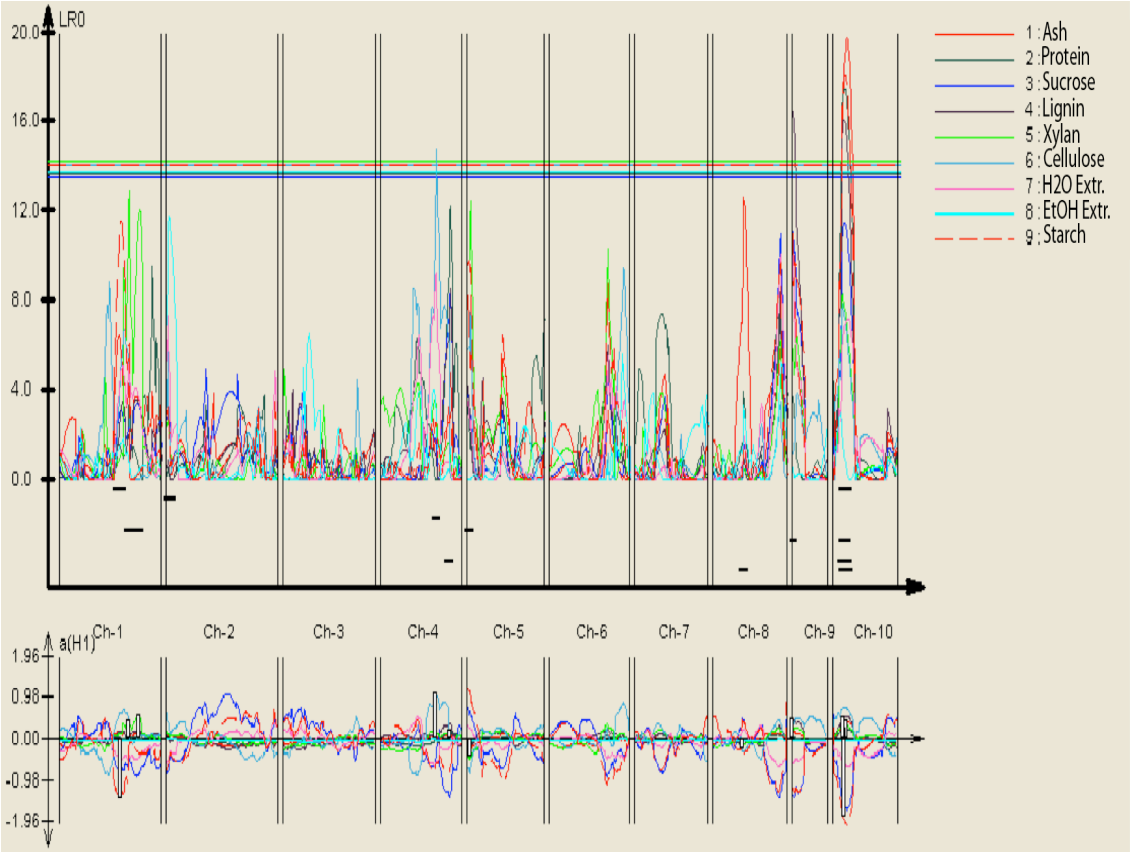


Figure 24: QTL map of SC56 x Tx7000 RIL population determined via composite interval mapping. Height of peak indicates confidence of QTL identification. Horizontal colored bars indicate significance threshold.

The same technique was used to construct a QTL map of the SC56 x RTx7000 RIL population, grown in College Station in 2009. (Fig. 24) The number of detected QTL was much lower than in the BTx642 x RTx7000 population, yielding 6 QTL that met the thresholds established by permutation tests (1000 permutations, 0.05 significance level). Several other peaks reached near-significance, suggesting that additional markers or a larger number of samples might reveal additional loci for control of stem composition.

Table 3: Location of QTL for stem composition traits in SC56 x Tx7000 RIL population as determined by CIM. Additive effect denotes the amount of variation that locus is responsible for, in units of percent total stem volume. R2 indicates the amount of variation that locus is responsible for in terms of variation within the population. LOD1 left and right columns indicate the physical locations of the left and right boundaries of the QTL, in this case meaning the area wherein the LOD score of the QTL is within one point of that of the peak.

Trait	QTL#	Chromosome	Marker	Position(cM)	LOD	Additive	R^2	LOD1 Left	LOD1 Right
Ash	1	10	13	23.3	4.3087	0.4243	0.24279	6988261	7617908
Protein	1	10	13	19.3	3.8109	0.2571	0.170686	6988261	7617908
Lignin	1	9	1	0	3.5978	0.4723	0.121241	0	1206013
Lignin	2	10	13	17.3	3.5087	0.5048	0.138044	6988261	7617908
Cellulose	1	4	32	93.5	3.2087	1.0943	0.123692	61188292	62302015
Starch	1	10	13	19.3	3.9522	-1.8521	0.182126	6988261	7617908

QTL Inspection

One of the benefits conferred by sequence based marker determination is the ability to accurately and rapidly plot regions delineated by genetic markers to their physical locations on the genome.

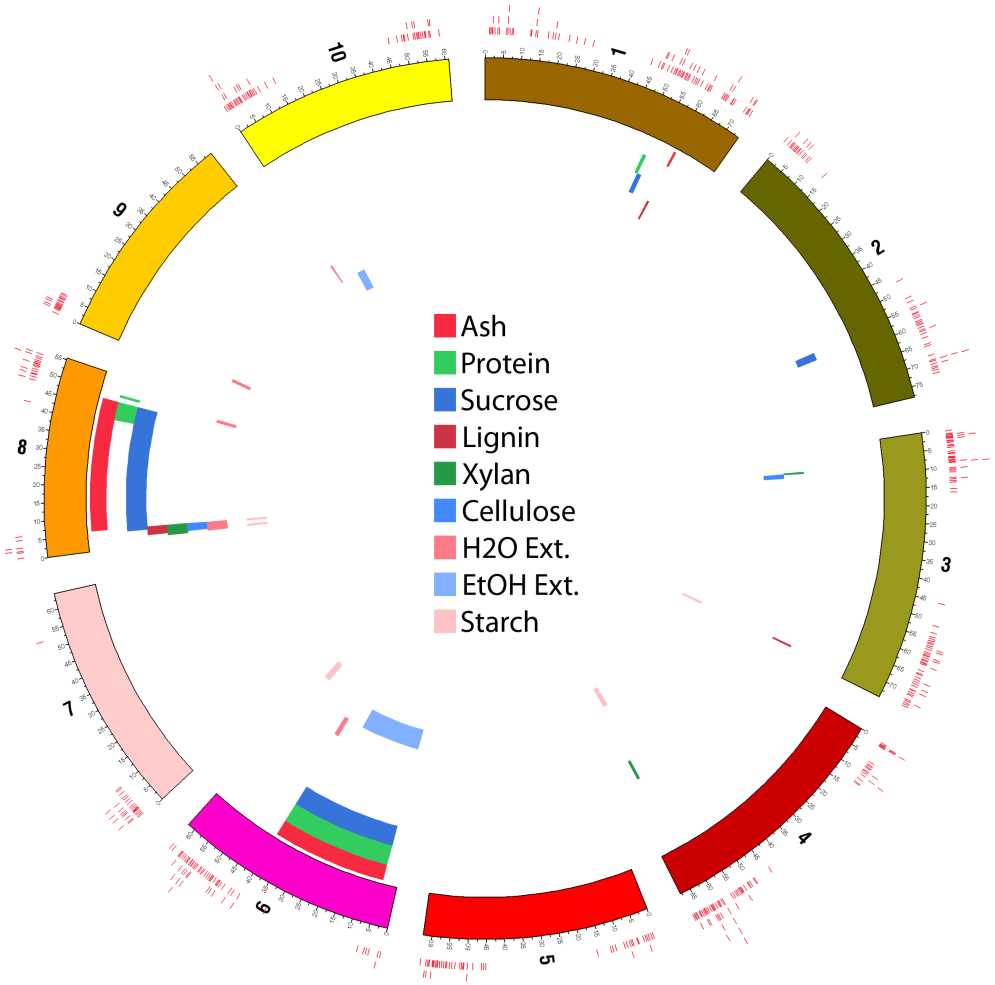


Figure 25: Physical location of QTL discovered in BTx642 x Tx7000 RIL population. Marker locations are plotted as red dashes on the outer perimeter of the illustration.

By using the known locations of the genetic markers used in QTL analysis, the QTL identified in the BTx642 x Tx7000 RIL population were plotted on a schematic of the genome (Fig. 25). Each QTL was plotted to 1 LOD distance from the peak of the QTL. Regions of low genetic recombination are clearly visible as portions of the chromosomes with small genetic distance but large physical distance, such as the region spanning from ~4.7 Mb to ~40.1 Mb on chromosome six, as well as the region spanning from ~6.5 Mb to ~47.3 Mb on chromosome eight. These regions also show low gene density, with a density of 0.66 genes/100kbp for the region on chromosome six, and 1.11 genes/100kbp for the region on chromosome eight compared to an average gene density of 3.59 genes /100kbp for sorghum in general.

Particularly noteworthy is the region between ~5.3 Mbp and ~6.3 Mbp on chromosome eight. QTL for lignin, xylan, cellulose, water extractives and starch all lie within this region, and if the LOD boundary is expanded to 1.5, QTL for ash and sucrose content also overlap. Analysis of the additive effects contributed by each allele reveals that the RTx7000 allele contributes to an increase in the amount of stem nonstructural components (starch, water extractives, and sucrose) while the BTx642 allele contributes to an increase in the amount of structural components (xylan, lignin, cellulose).

Using the current annotated S. *bicolor* genome available from http://www.phytozome.com, genes underlying the primary QTL on chromosome eight were identified and analyzed. The region spanning 5,200,000 – 6,300,000 bp on chromosome 8 contains 61 annotated genes (Appendix Table 4). Of those 61 genes, 21 genes lack a functional annotation, and the rest have annotations that have been applied

computationally as described by Paterson et al., (2009). Focusing on genes that could act in a regulatory fashion further reduced the remaining pool of 40 potential gene candidates. The large number of compositional traits affected by this single QTL indicates that the gene in question likely acts as a developmental or biosynthetic regulator, rather than as a step in the biosynthetic process itself, allowing further reduction of the candidate pool to ten (Remington and Purugganan, 2003).

Table 4: Potential gene candidates located within first QTL cluster on chromosome 8, as identified in BTx642 x Tx7000 population.

Gene Name	PFAM ID	Panther Description
Sb08g004510	PTHR22982	CALCIUM/CALMODULIN-DEPENDENT PROTEIN KINASE-RELATED
Sb08g004460	PTHR23258	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE
Sb08g004450	PTHR23258	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE
Sb08g004720	PTHR10641	MYB-RELATED
Sb08g004900	PTHR11085	CHROMATIN REGULATORY PROTEIN SIR2
Sb08g004550	PTHR23258	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE
Sb08g004940	PTHR23258	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE
Sb08g004830	PTHR23258	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE
Sb08g004700	PTHR23258	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE
Sb08g004790	PTHR12802	SWI/SNF COMPLEX-RELATED

Examination was also undertaken on the SC56 x Tx7000 population. While QTL were far less numerous in this population, the cluster of QTL located on chromosome 10 is amenable to a similar analysis. This cluster encompasses the region from approximately 7.0-7.6 Mb, and contains 64 annotated genes (Appendix Table 5), for an average gene density of 10.17 genes/100 kbp. 16 of these genes lack a functional

annotation and of the remaining 48, only a single gene, Sb10g007420.1, is annotated as a putative transcription factor.

Discussion

The shift to renewable biomass feedstock for production of biofuels will require a greater level of control over the composition of that feedstock. Identifying the genes responsible for the variation in populations will allow breeders to create hybrids with optimized stem compositions for biofuels, silage, or other products. While stem compositional estimates determined by NIR spectroscopy have not yet been correlated with biofuels production, the technique has shown its value in the ability to measure the composition of entire RIL populations in a rapid and repeatable fashion.

This study also establishes the presence of stem composition variability QTL in elite breeding lines of sorghum. Previous work on lignin variation in sorghum has focused on cultivars containing the *brown midrib* family of mutations that reduce the ability of sorghum to manufacture monolignols (Oliver et al., 2005). This work shows that QTL controlling stem lignin content can be identified in non-mutant lines, and even in similar grain-type sorghum cultivars, suggesting the existence of a stem composition control system already present in divergent sorghum lines.

Perhaps most interestingly, this work also indicates the potential for identifying the individual genes responsible for the observed phenotype. Using the sequence-derived markers described above, the physical location of the QTL can be identified. The number of markers allows the region to be reduced compared to traditional marker techniques, shrinking the pool of genes in the region. In the case of the data presented

above, the region only contains 61 genes, and only 10 of them are predicted to be able to act in a regulatory fashion. This small pool of genes allows for future experimentation in a more targeted fashion, presenting locations to look for sequence polymorphisms and/or expression variation.

SORGHUM GENOME RESEQUENCING

Introduction

Quantitative Trait Loci (QTL) mapping has been used for nearly two decades to associate complex phenotypic variation with the region of the genome responsible for that variation (Miles and Wayne, 2008). One of the limitations of this technique, however, is that the QTL often span several centimorgans (cM) and several megabase pairs (Mbp) of DNA (Miles and Wayne, 2008). Given that sorghum has a gene density of approximately 8 genes/100 kb in the euchromatic regions, even a QTL that span only a few cM could still correspond to a region containing a hundred or more genes (Kim et al., 2005). While annotations to the genome can help select potential gene candidates, dozens of genes will likely need to be screened for functional mutations that cause phenotypic variation associated with each QTL.

Once a region of the genome has been identified through QTL mapping, the process of fine mapping and gene identification becomes much more complicated, requiring the identification of additional markers and the use of larger populations (Lee and van der Werf, 2006; Nagy et al., 2007). In some organisms it may be possible to look for polymorphisms by sequencing all the genes in the region of interest, but the intron-rich nature of many genes makes this largely impractical for significant numbers of gene candidates (Mourier, 2003; Paterson et al., 2009;). Some of these issues can be avoided by sequencing cDNA representations of mRNA transcripts of these genes, but this introduces new problems in terms of splice variants and challenges in sample collection.

A promising emerging technique that may solve this dilemma is whole genome re-sequencing. Using short read next generation sequencing platforms such as Illumina and SOLiD, millions of short sequence reads can be aligned to pre-existing genome sequences rapidly, and these alignments used to identify sequence variants (Lunter and Goodson, 2011). Re-sequencing projects in rice have demonstrated their ability to identify over a million sequence variants between even closely related cultivars, allowing for rapid identification of gene coding variants that may affect function (Xu et al., 2012). While this technique may not currently be cost effective for all crop plants due to large genome sizes and repeat density, it has great potential for use in sorghum (~700Mb genome size) and rice (~400Mb), two of the most important grain crops in the world (FAO 1995).

Each re-sequencing of a cultivar also provides valuable information regarding the lineage of that cultivar and the origin of different portions of the cultivar's genome (Xu et al., 2012). Moreover, the genomes can be compared to identify regions that have undergone selection (Tajima, 1989). This is especially relevant when dealing with cultivated sorghums in the United States, as most such sorghums are products of generations of breeding under selection for height, early maturity/photoperiod insensitivity and grain yield (Quinby, 1974; Smith and Frederiksen, 2000).

The three genotypes used in this study (Tx7000, BTx623, and BTx642) share significant genetic heritage. Tx7000 was derived from Blackhull Kafir, a Kafir race sorghum from Southern Africa, and Milo, a type of Durra sorghum. BTx623 was derived from a cross of BTx3197 and SC170; BTx3197 originates from Blackhull Kafir,

Double Dwarf Kafir and Milo sorghums, while SC170 is from the Caudatum race. BTx642 was derived by conversion of IS12555 (Durra) using BTx406, a line with both Kafir and Durra background (Milo is a type of Durra), respectively (Smith and Frederiksen, 2000; Klein et al., 2008). While the origins of these genotypes is known, the portions of each progenitor present in derived material is unknown. By sequencing parental genomes and identifying variation in genetic diversity among the three genotypes, it was possible to identify with great accuracy the physical portions of the genome that are identical or nearly identical by descent.

Results

Sequencing and Mapping

The two sorghum genotypes BTx642 and RTx7000 have previously been used to identify QTL that contribute to the stay-green drought avoidance trait (Xu et al., 2000; Harris et al., 2007). Significant resources exist for these populations, including an established RIL population and abundant marker and genetic data (Harris et al., 2007; Weers 2011)

Libraries of sheared genomic DNA were constructed from BTx642 and RTx7000 and subjected to paired-end sequencing on an Illumina Hi-Seq sequencing device. After quality trimming and mapping, approximately 60 million mapped reads from each genotype (Table 5) could be aligned to the published *Sorghum bicolor* BTx623 genome. Each line averaged 11 million reads that did not align, largely due to the quality criteria established for alignment: 80% of the sequence was required to match the reference sequence, meaning that for 150 bp reads 120bp must match perfectly. When mapping

with reduced stringency of 70% sequence identity, the number of mapped reads only increased by approximately 1 million (Table 5). This indicates that the remainder of the non-aligned reads are likely not limited by poor quality but instead originate in regions of the genome where the reference genome contains no sequence data, from mitochondria or chloroplasts, or from the approximately 50 Mbp of genetic material that was not anchored to the reference genome during construction (Paterson et al., 2009)

Significant variation in read depth was observed across the genome, with approximately 10 percent of the genome having no reads (Table 5). Visualizing the read depth across each of the chromosomes shows that the bulk of zero read depth regions are localized near the regions identified as containing high levels of the Cen38 pericentromeric repeat, or regions that were not assembled due to their high repeat content during the sorghum genome sequence construction. (Figure 26, Paterson et al., 2009).

Table 5: Summary of sequencing for sorghum cultivars.

Cultivar	Similarity	# Useable Reads	Mapped Reads (bp)	Average Coverage (excluding zero coverage regions)	Fraction of reference covered	Total zero coverage length (bp)	# zero coverage regions	Max Length of zero coverage region
BTx642	80%	66,442,052	55,558,636	10.63	0.89	80,229,330	350648	4800003
Tx7000	80%	70,497,898	62,218,713	11.85	0.9	74,149,696	277809	4800044
BTx642	70%	66,442,052	56,908,091	10.74	0.89	78,700,128	360746	4800020
Tx7000	70%	70,497,898	63,239,083	11.93	0.9	73,105,855	284141	4800044

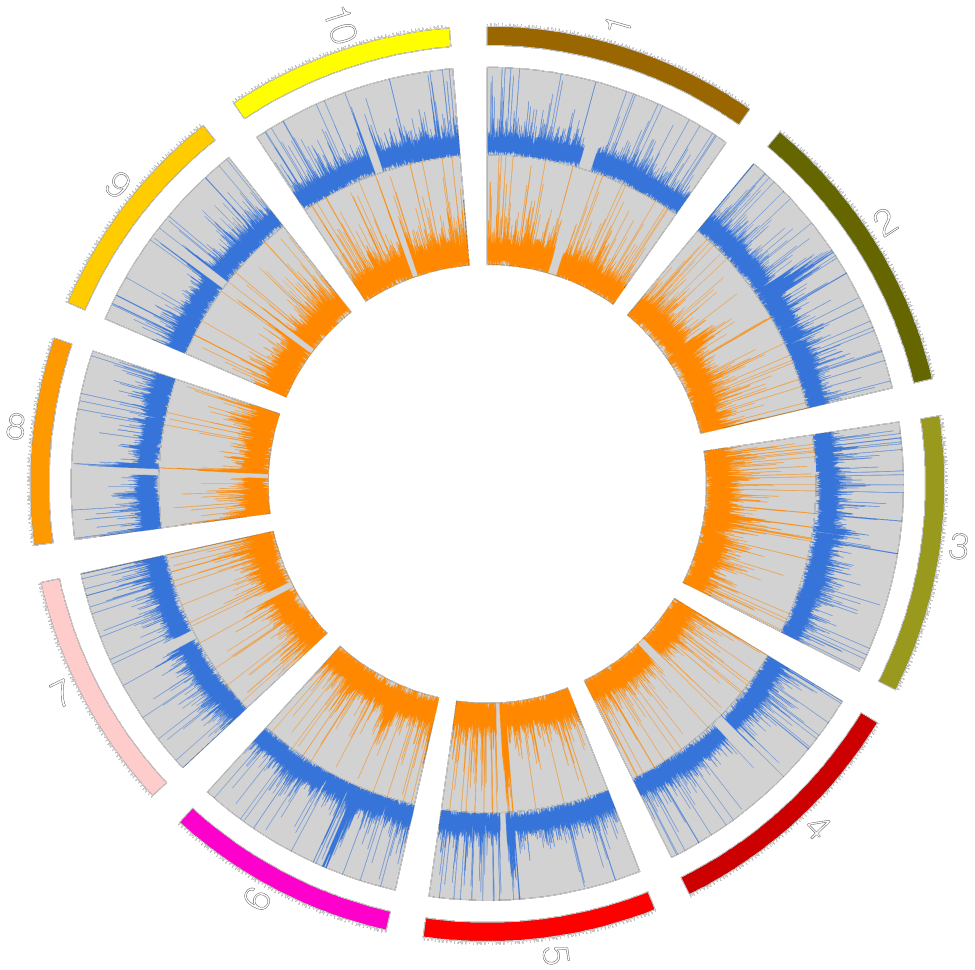


Figure 26: Read densities of Tx7000 (blue line) and BTx642 (orange line) from paired-end read assembly. Peak maximum represents depth >100 reads.

Confirmation of Sequence Variants and Estimated SNP Coverage

The BTx642 and Tx7000 re-sequenced genomes were compared to the BTx623 reference genome sequence to identify putative SNPs and indels (Table 6). Over one million SNPs and more than 100,000 indels were detected in each comparison. A higher number of SNPs and indels were found when BTx642 was compared to BTx623 than when Tx7000 was compared to BTx623. While an average of 10x coverage for both

genomes allows for the identification of a large number of SNPs, an unanswered question was how much of the sequence variation was not visible due to insufficient read depth. To determine this, the sequences of previously identified and genetically mapped Digital Genotyping sequence markers that distinguish BTx642 and Tx7000 were compared to information obtain from genome resequencing (Weers, 2011). Genome resequencing identified SNPs corresponding to DG-markers with 94% accuracy (Table 6). Surprisingly, despite the higher number of SNPs discovered in BTx642, a lower percentage of the confirmed SNPs was discovered, with only 566 of the 1572 unique variants identified (Table 6). If taken as an estimate of total SNP discovery rates across the genome, it can be seen that approximately 40% of the total SNP number was discovered in the Tx7000 genotype, versus 27% in BTx642.

Table 6: Number of SNPs and indels identified via resequencing that distinguish BTx642 or Tx7000 and BTx623.

Name	Avg. Read Depth	# SNPs	# Indel	# Variants compared	# Variants discovered	% Matching SNPs	Estimated % SNPs discovered
BTx642	9.48	1,378,502	179,628	1572	566	94.0	27.09
RTx7000	10.66	1,040,535	147,229	1009	553	94.2	40.03

Variation Across the Sorghum Genome

Initial SNP discovery was based on a comparison of the aligned reads from BTx642 or Tx7000 and the reference genome sequence of the BTx623 sorghum cultivar, so the variants discovered distinguish the newly sequenced lines and the reference sequence. By comparing the variants from BTx642 to those of Tx7000, it is possible to

identify polymorphisms that vary between these cultivars. Regions of the genome where BTx642, Tx7000 or BTx623 share variant identity could help identify genomic regions that derive from lines common to their respective pedigrees.

Tx7000 and BTx3197, the immediate progenitor of the reference sorghum genotype BTx623 are relatively closely related because both have Kafir-Milo derived genetic material their pedigrees (Figure 27; Smith and Frederiksen, 2000; Klein et al., 2008). BTx3197 was subsequently crossed to SC170, a Caudatum line, to generate BTx623. Tx7000 and BTx623 are both grain type sorghums that were selected for early flowering, high grain yield, and short stature. BTx642 originates from IS12555, a Durra (Milo) genotype that is a source of the stay-green drought resistance trait (Harris et al., 2007). BTx642 was created by crossing IS12555 to BTx406 (Kafir-Milo). BTx642 was a BC1 derived line from this cross that was early flowering and of short stature, traits inherited from BTx406. Therefore, regions of the Tx7000, BTx623 and BTx642 genomes may be quite similar if they were derived from the small group of early Kafir-Milo sorghum introductions used in the U.S. Other portions of these genomes derived from SC170 (Caudatum) and IS12555 (Durra) are likely to be quite different from regions of these genomes derived from Kafir-Milo genotypes.

To determine whether this was the case, the assembled genomes of BTx642 and Tx7000 were compared to the BTx623 reference sequence (Paterson et al., 2009) to identify sequence variants. When compared to the BTx623 reference, the Tx7000 genotype shows less variation than BTx642, with approximately 75% as many SNPs and 82% as many indels identified (Table 7).

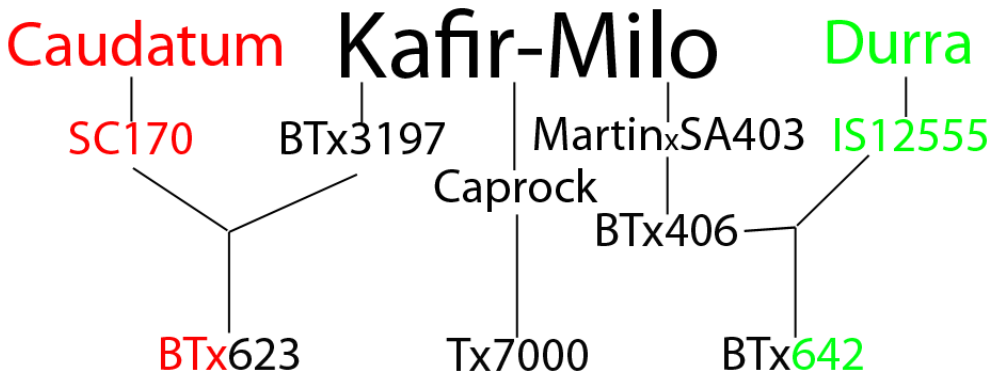


Figure 27. Simplified representation of the lineage of sorghum genotypes BTx623, Tx7000, and BTx642. The pedigrees indicate that genomic regions present in the three genotypes derived from Kafir-Milo progenitors will have greater genetic similarity than genomic regions derived from SC170 (Caudatum), present in BTx623, or IS12555 (a Durra) present in BTx642.

Table 7: Variant discovery and confirmation for sorghum cultivars.

Name	Similarity	Avg. Read Depth	# SNP	# Indel
BTx642	80%	9.48	1,378,502	179,628
RTx7000	80%	10.66	1,040,535	147,229
BTx642	70%	9.59	1,421,178	192,153
RTx7000	70%	10.75	1,071,690	157,225

BTx642 evinces significant divergence from the reference genotype, as shown by the high levels of sequence variation across much of the euchromatic regions of the genome (Fig. 28). In general, regions of high genic density (Fig. 28, black line) have correspondingly high levels of genetic variation (Fig. 28, blue and red line). Of note are the regions at the ends of chromosomes 7 and 9 that show high genic density but very low genetic variation, indicating that these regions may have originated from the same source early in the development of these genotypes.

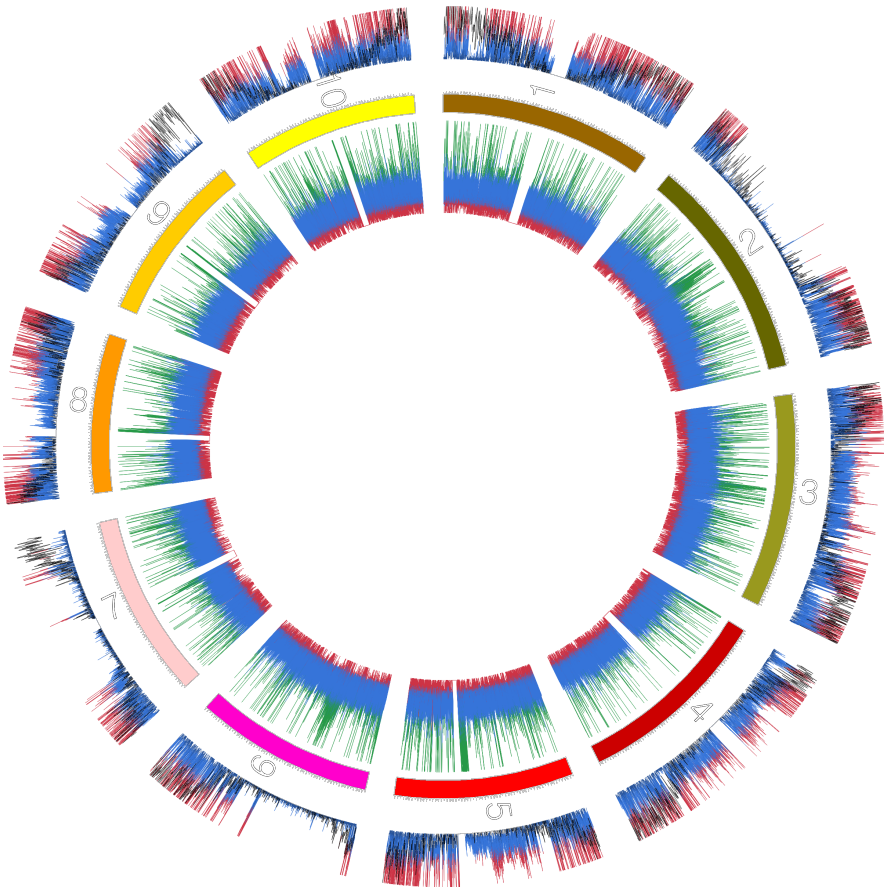


Figure 28: Read density, gene density, and the density of SNPs that distinguish the BTx642 cultivar from BTx623. Inner line graph denotes read coverage, with a graph maximum of 50, averaged over 1kb stationary windows. Green color denotes >30 average read depth, red indicates <5. Outer line graph indicates SNP density averaged over a 10kb stationary window, with a graph maximum of 60, and the red color indicating >30 SNP/10kb average. Black line indicates genic density calculated over 100kb stationary window, with a graph maximum of 15 genes/100kb.

Tx7000 is more similar genetically to BTx623 than is BTx642, as shown by the lower total number of SNPs and indels that distinguish these lines (Table 7). Notably, Tx7000 shares regions of low sequence variation and high genic density with BTx623, particularly the regions from ~7-12 Mb on chromosome 1 and the ends of chromosomes 7 and 9 (Figure 29).

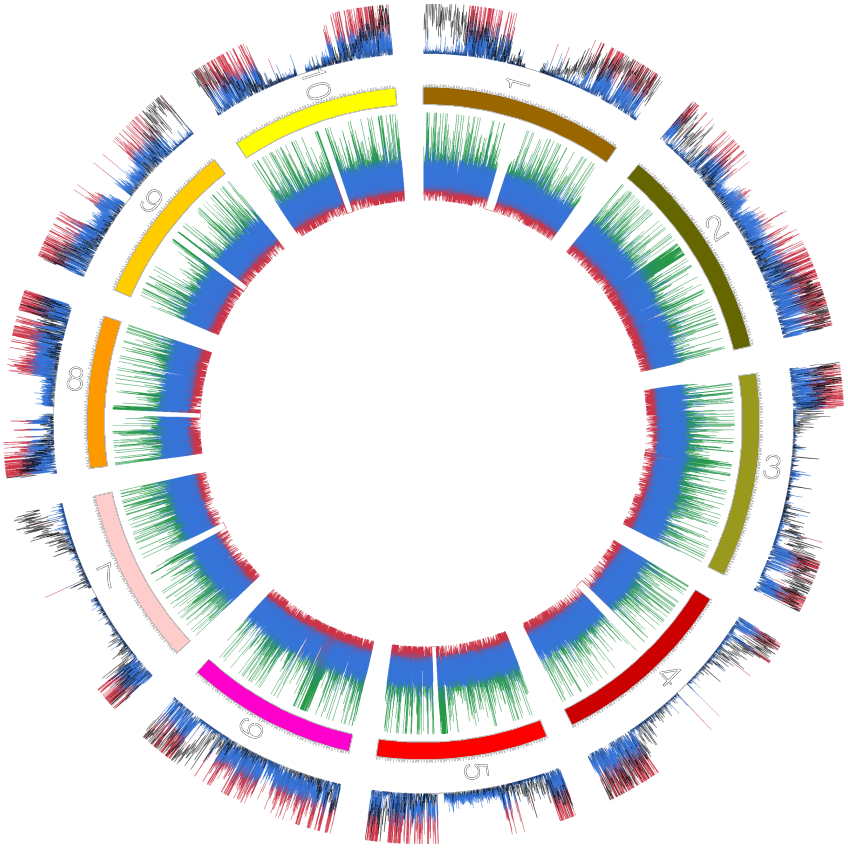


Figure 29: Read density, gene density, and the density of SNPs that distinguish the Tx7000 cultivar from the BTx623 reference. Inner line graph denotes read coverage, with a graph maximum of 50, averaged over 1kb stationary windows. Green color denotes >30 average read depth, red indicates <5. Outer line graph indicates SNP density averaged over a 10kb stationary window, with a graph maximum of 60, and the red color indicating >30 SNP/10kb average. Black line indicates genic density calculated over 100kb stationary window, with a graph maximum of 15 genes/100kb.

In order to assist in identifying genomic regions of low diversity shared by BTx642 and Tx7000, the SNP densities were plotted against each other with a genic density overlay (Fig. 30). This plot reveals several regions of low variation in euchromatic portions of the chromosomes that have high gene density. As noted

previously, the region from ~7-12 Mb on chromosome 1 is such a location, as well as the terminal ~10 Mb on chromosomes 7 and 9, and numerous additional regions (Table 8).

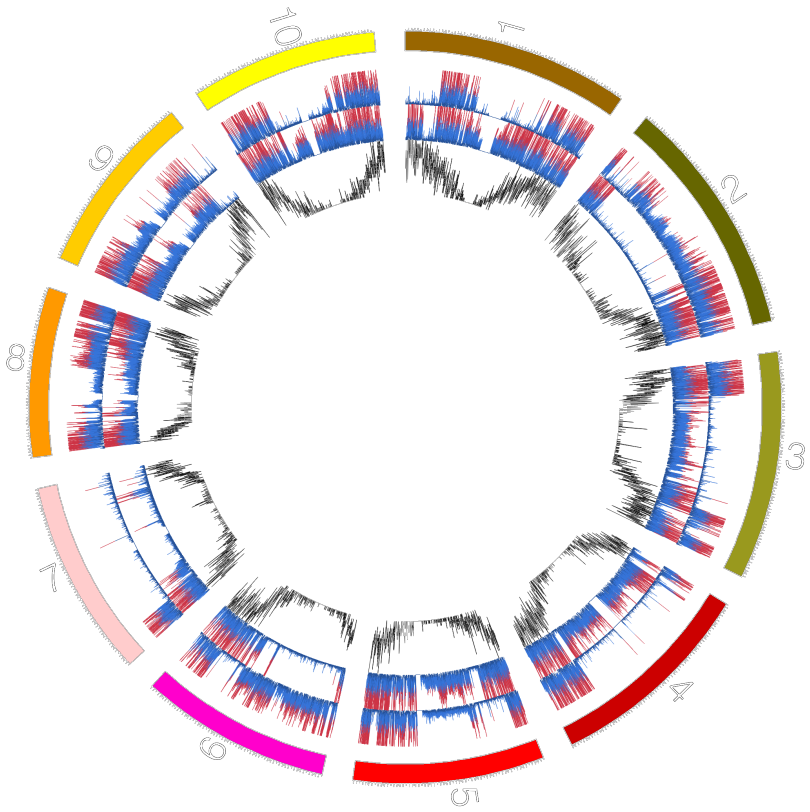


Figure 30: Plot of SNPs in BTx642 and Tx7000 vs. BTx623, with gene density as comparison. Outermost layer represents all 10 sorghum chromosomes, with base pair distances labeled on outer edge. Second layer represents Tx7000 vs BTx623 SNP density in 10Kb stationary windows. Red coloration represents >30 SNPs / 10KB, with the plot maxima set to 60 SNPs/10Kb. Third later represents BTx642 vs BTx623 SNP density, with the same settings the outer layer. Innermost black plot represents gene density per 100kb stationary window, with a plot maxima of 15 genes per 100kb.

Table 8: Regions of low genetic diversity between sequence cultivars and the reference genome

Chromosome	Approx. Location	# genes in region
1	7-11 Mb	411
2	7-8 Mb	51
4	2.5-5 Mb	257
5	12-15 Mb	68
7	53-66 Mb	585
9	50-58 Mb	890

BTx642 and Tx7000 Variant Analysis

DNA polymorphisms that distinguish BTx642 and Tx7000 were identified and their distribution analyzed. Surprisingly, the total number of identical SNPs that distinguish these genomes from BTx623 approached 400,000, representing a substantial fraction of the observed variation (Table 9). This is especially noticeable on chromosome 9, where over 75% of the SNPs identified distinguished both genotypes from BTx623.

Table 9. Distribution of common and unique SNPs across the BTx642 and Tx7000 sorghum genomes.

Line	Chrom. 1	Chrom. 2	Chrom. 3	Chrom. 4	Chrom. 5	Chrom. 6	Chrom. 7	Chrom. 8	Chrom. 9	Chrom. 10	Total
Tx7000	26803	132199	40025	38266	54422	132771	15088	73913	29720	57916	601123
BTx642	111092	43602	139005	149075	116173	45767	47751	76641	28598	121413	879117
Common	45068	34668	29214	20972	46533	22674	8843	60492	92936	37572	398972

Plotting these shared SNPs on the genome reveals that, in general, the regions of shared SNPs are interspersed with regions of SNP density unique to either BTx642 vs.

BTx623 or Tx7000 vs. BTx623 (Fig. 31). Regions of shared SNPs extend for nearly the entire length of chromosome 9, made up of nearly 70,000 common variants.

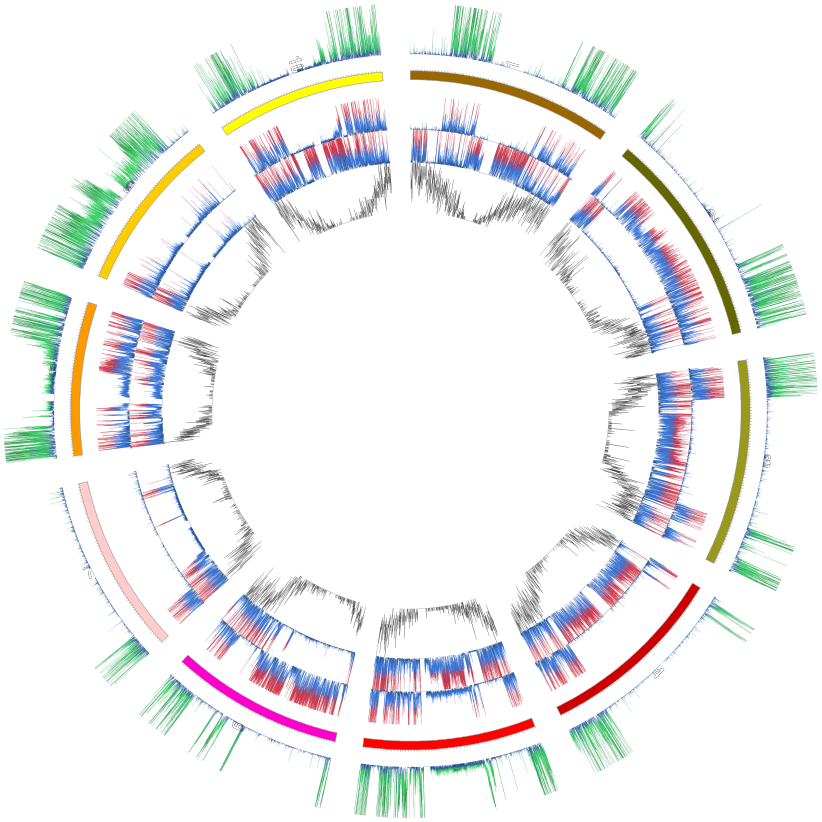


Figure 31: Plot of SNPs that distinguish Tx7000 and BTx642 (green line) from BTx623 on the sorghum genome (colored solid bars). Inner lines represent, in descending order from solid bars: SNPs uniquely discovered in Tx7000 vs BTx623, SNPs uniquely discovered in BTx642 vs BTx623, and gene density.

Looking more closely at chromosome 9, it can be seen that from 15 Mbp to 40 Mbp, a vast majority of the SNP variation relative to BTx623 is in common to both BTx642 and Tx7000 (Fig. 32). This contrasts with the region extending from ~51 Mbp to the end of the chromosome, which displays a high degree of similarity among all three

genotypes. This indicates that the region of 15-40 Mbp was inherited from a source shared by BTx642 and Tx7000, but different from BTx623. This contrasts with the region from 51 Mbp to the end of the chromosome, which is shared by all three genotypes. The presence of such large regions of similarity could indicate the presence

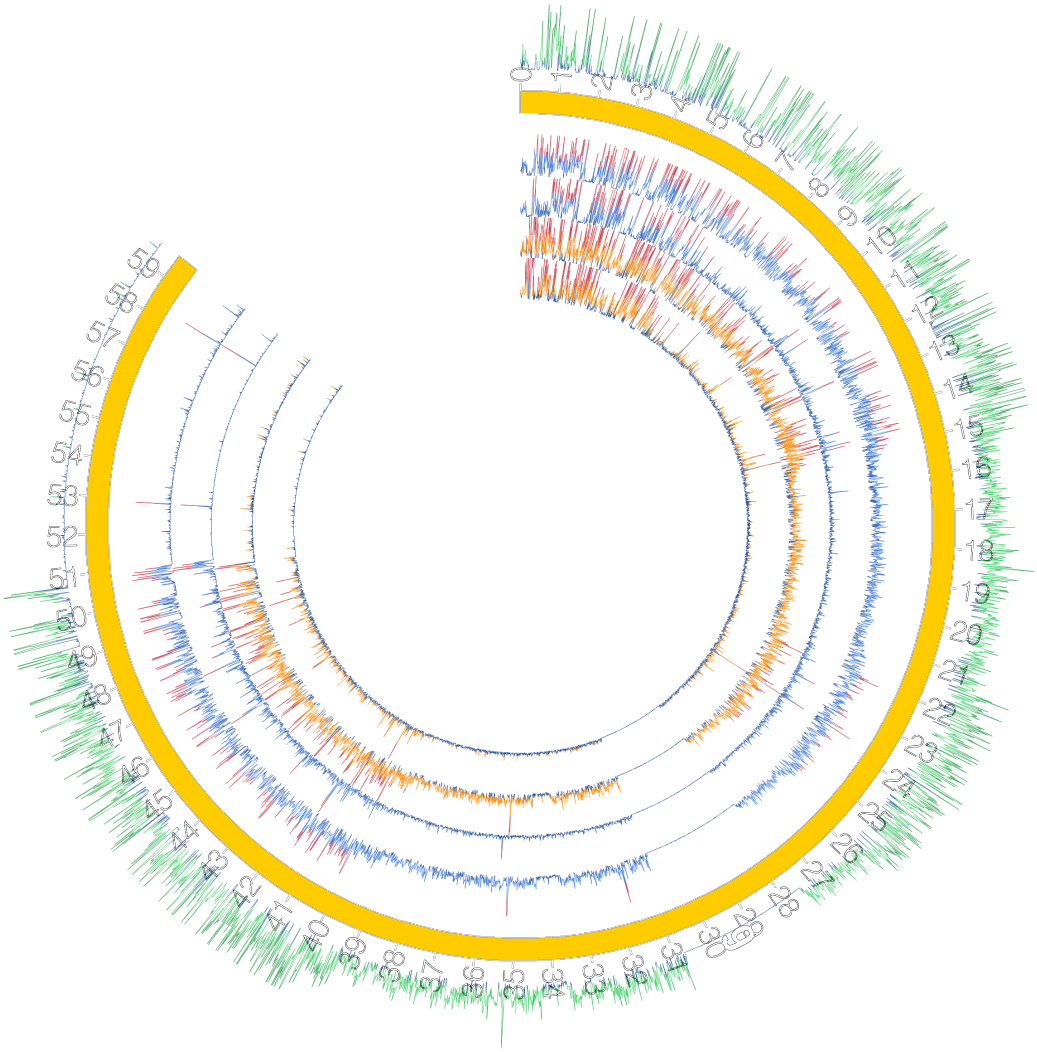


Figure 32: Plot of SNPs between BTx642 and BTx623 (orange lines) and between Tx7000 and BTx623 (blue lines) on sorghum chromosome 9. Outer green line represents SNPs in common between BTx642 and Tx7000. Outer blue and orange lines represent all SNPs present in each line versus BTx623. Inner blue and orange lines represent SNPs versus BTx623 that are unique to each genotype.

of agronomically relevant traits that have been selected from various backgrounds and retained intact through the breeding process.

Chromosome 6 is a region of the genome that encodes maturity locus 1 (*Ma1/ma1)* as well as a dwarfing locus, *Dw2/dw2* (Lin et al., 1995; Klein et al., 2008). BTx406 is a source of *dw2* and *ma1*. The region spanning these QTL in BTx642 was previously shown to be derived from BTx406 (Klein et al., 2008). Closer examination shows that there are multiple blocks of genetic material spread throughout the *ma1* locus that are probably derived from BTx406 rather than one large homogeneous region (Figure 33). Regions spanning the Ma2 locus that are probably identical by decent (IBD) with BTx623 are clearly visible in both BTx642 (~32-37 Mbp, ~40.5-42 Mbp) and in Tx7000 (~40-42 Mbp), separated by regions of non-IBD material. The identity of *ma1* has recently been determined to be SbPRR37, which is indicated in Figure 33 (Murphy et al., 2011). The region surrounding *SbPRR37* can be seen to be of different origin in BTx642 than from Tx7000, which is consistent with the presence of the *Sbprr37-1* allele in BTx406, while Tx7000 contains the *Sbprr37-2* allele originating from Blackhull Kafir (Murphy et al., 2011). Interestingly, the BTx623 allele is *Sbprr37-3*, which varies from Tx7000, but can be seen in Figure 33 to be located in a region of low variation between Tx7000 and Btx623. This is consistent with the proposal that *Sbprr37-3* arose from *Sbprr37-2* through an additional mutation (Murphy et al., 2011).

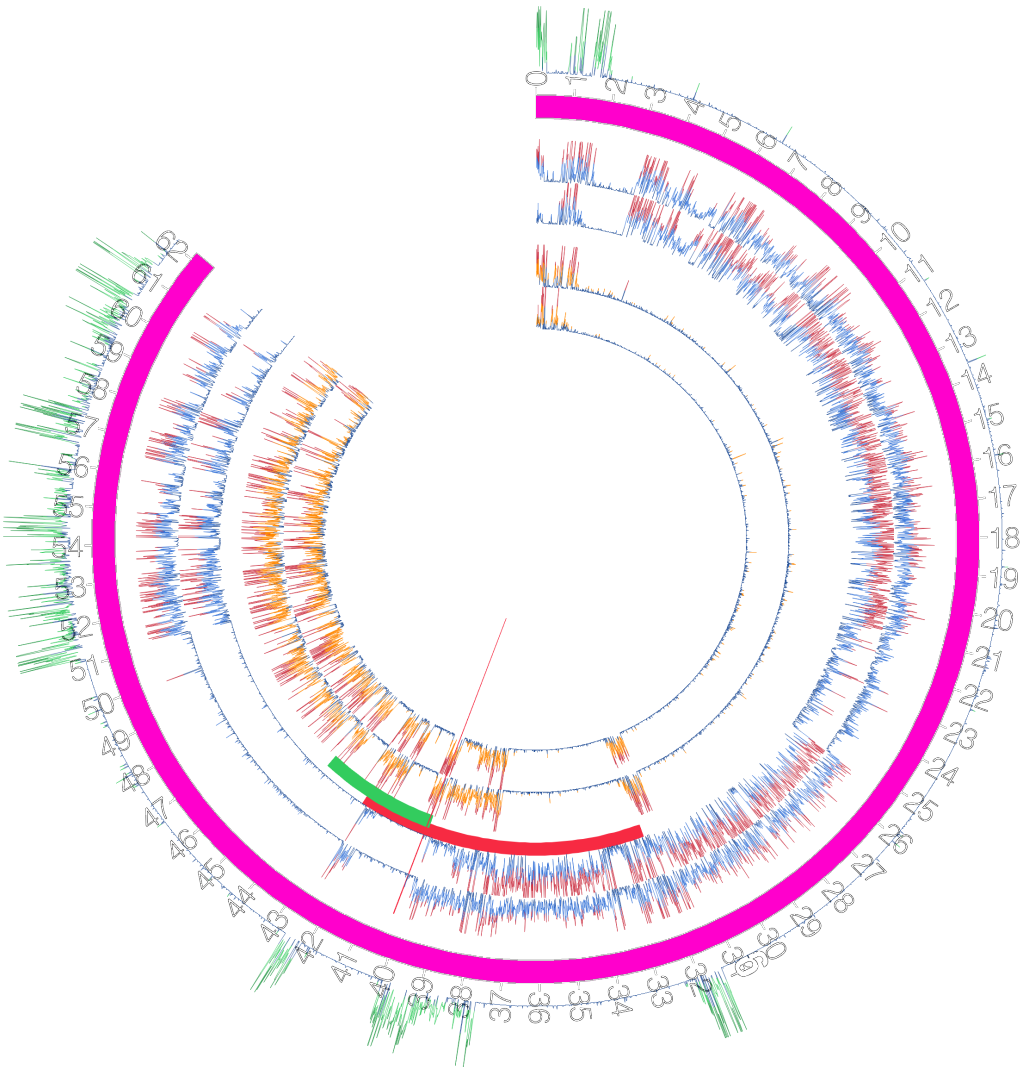


Figure 33. Plot of SNPs between BTx642 and BTx623 (orange lines) and between Tx7000 and BTx623 (blue lines) on sorghum chromosome 6. Outer green line represents SNPs in common between BTx642 and Tx7000. Outer blue and orange lines represent all SNPs present in each line versus BTx623. Inner blue and orange lines represent SNPs versus BTx623 that are unique to each genotype. Solid red bar represents QTL for ma1, solid green bar represents QTL for dw2, red line indicates position of SbPRR37 (Klein et al., 2008, Murphy et al., 2011).

Discussion

Genome-wide identification and analysis of sequence variants that distinguish three widely studied cultivars of sorghum reveals some surprising insights into the

heritage of these lines. The development of the Tx7000 and BTx642 genotypes has been well documented by the breeders, but to this point the determination of the genomic makeup has been limited to pedigree and DNA marker analysis. While flexible, genetic analysis is less precise than sequencing when characterizing the physical areas of the genome inherited from distant relatives, and genetic distances do not necessarily correlate with physical distance, especially in recombinationally poor regions.

The identification of more than 600,000 unique SNPs per sequenced line will provide an invaluable resource for future analysis, both from the standpoint of identifying variants located near or within genes, but also for identifying polymorphisms that can be used for fine mapping within an existing population.

As BTx642, Tx7000, and BTx623 are products of the sorghum conversion project, it is unsurprising to detect shared genetic material between these cultivars. What is striking, however, is the size of the retained shared material. The most visible example is chromosome 9, which appears to contain large stretches of extremely similar material from the lineage of BTx642 and Tx7000 that is not shared with BTx623. This can be contrasted with the regions detailed in Table 8, indicating material shared between all three genotypes. It is well established that these cultivars have been under significant selective pressure (Menz et al., 2004), and these regions likely contain loci responsible for traits that were desirable during the generation of these cultivars. Chromosome 6 shows both a case and a caution to this hypothesis. The area containing SbPRR37 is clearly divergent between BTx642 and Tx7000, which is consistent with the observed variation at the ma1 locus. However, observation based on genetic similarity

alone could lead to the assumption of Tx7000 and BTx623 having the same alleles at that locus, which is not the case. In such a situation, the actual sequencing data provided will allow for clarification once the gene candidate is identified, but phenotyping will be of critical importance until that stage.

MATERIALS AND METHODS

Phloroglucinol Staining and Quantitation

2000 domestic and exotic sorghum accessions were planted in College Station in spring of 2007 by Dr. William Rooney, and grown under non-irrigated conditions. In October, 351 of those lines were selected for sampling, with fresh stem segments frozen for staining. Segments were sectioned by razor blade to approximately 1 cm thickness. Sections were then stained with a mixture of 2% w/v phloroglucinol in 95% EtOH – concentrated HCl (5:1) for one hour under cover. Samples were then photographed with an Olympus E-300 camera.

Acetyl Bromide Lignin Extraction and Quantification

41 lines were selected from the 351 lines sampled in October 2007. Whole stem tissues were dried in a forced air oven at 140F for a minimum of 48 hours, and the dried tissues ground in a Wiley mill until material could pass through a 1 mm screen. Lignin extraction follows a modified protocol detailed in Iiyama and Wallis (1990). 12 mg (± 0.1mg) of dried tissue was weighed on an analytical balance and transferred to a borosilicate tube. 10 mL DI H2O added and tube placed on a heat block set to 65 C for one hour, agitating every ten minutes. Sample was then filtered through a GF/A glass fiber filter (Whatman Inc., Florham Park, NJ, USA) and rinsed three times with successive three minute rinses of DI H2O, ethanol, acetone, and diethyl ether. Sample then dried in a Teflon capped scintillation vial overnight at 70 C. 3 mL 25% (v/v) acetyl bromide in glacial acetic acid added to each vial, tightly capped, and incubated at 50 C

for two hours, agitating every 30 minutes. Transferred to 50 mL volumetric flasks containing 25 mL glacial acetic acid-NaOH (1.5:1) and filled to volume with glacial acetic acid. Allowed to settle overnight and absorbance at 280 nm taken on Beckman DU-64 UV/VIS spectrophotometer.

Generation of Lignin Standard Curve

Lignin standard curve was generated as described by Fukushima et al. (1991). In brief, Indulin AT (Meadwestvaco, Richmond, VA, USA) was washed with boiling DI H2O until effluent was colorless and dried overnight at 50 C. Material was then dissolved in acetyl bromide-glacial acetic acid as described previously, and absorbance measured at 280 nm. Intervals were plotted and a linear regression was generated (Appendix Figure 34, Appendix Table 6).

Identification of Sorghum Monolignol Biosynthetic Genes

Experimentally confirmed monolignol biosynthetic genes were identified through literature search and their nucleotide sequence compared to the published sorghum genome sequence using the discontiguous megablast function. Results were selected where matches had at least 50% sequence coverage and e-values no greater than 1e-50.

Plant Growth and DNA Extraction

Sorghum seeds from SC56 x Tx7000 F9 RIL lines were obtained from Dr. William Rooney, and germinated in Metro Mix 200 growth media. Leaf tissue was harvested 14 days after planting and genomic DNA extracted using a FastPrep extraction kit and FastPrep-24 device (MP Biomedicals LLC, Solon, OH, USA) according to manufacturer instructions.

Generation of Sequence Based Genetic Markers

Multiplex sequencing for marker generation was performed as in Morishige et al. (2012). In brief, genomic DNA is digested by the *Fse1* digestion enzyme, and each sample has an adapter unique to its destination pool ligated to the 3' end. Samples are sheared with a Bioruptor (Diagenode, Denville, NJ), purified, and size selected on an agarose gel. Illumina adapters are ligated to samples and samples are pooled and sequenced according to Illumina protocols (Illumina Inc., San Diego, CA, USA).

Once generated, scripts separated sequences into groups and assigned to their originating RIL lines based on the previously assigned adapter sequences. Perl scripts pooled identical sequences within each sample and discarded sequences present in less than 4 copies, as well as sequences not bearing the CCGGCC sequence at the 3' end to ensure high quality marker assignment. Sequences in the parental lines were compared, and identical sequences discarded as non-unique. Remaining parental sequences were then compared to the sorghum reference sequence (downloaded from ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v6.0/Sbicolor/assembly/) using a local instance of the BLAST comparison algorithm on an Apple workstation. Scripts identified uniquely mapping sequences containing polymorphisms, and compared them between parental lines, assigning each parental line either their polymorphism or the equivalent reference sequence if no equivalent polymorphism is present.

Scripts then compare each sequence from the RIL lines to the established markers, generating an output file consisting of assignments to the 'A' parent, 'B' parent, neither if no matches were found, or 'Het' if both markers were found, as well as

outputting the number of each sequence identified. This file is sorted by RIL and also by physical location on the genome, allowing for rapid visual inspection of marker assignment. Once inspection is complete, the file can easily be converted into the format used by MapMaker/EXP for genetic map generation. Sequences used in mapping populations can be viewed in Appendix Table 1 and Weers (2011).

Construction of SC56 x Tx7000 Genetic Map

An F9 population of 200 individuals formed from the cross of the SC56 and Tx7000 inbred lines from the population created by Dr. Darrell Rosenow, and for those plants was obtained from Dr. William Rooney. Genetic markers were determined as described previously, and assigned to chromosomes in the MapMaker program using the Map and Assign commands, using the Kosambi mapping function. The genetic map generated can be found as Figure 22 and Appendix Figures A-1 and A-2. Markers that had recombinational distances of 0 were removed as they can cause detrimental effects, yielding a total of 392 usable markers.

Stem Composition Determination

SC56 x Tx7000 RIL population whole stems were harvested in the field in June 2009 and 2010. After stripping leaves and leaf sheath, visible internodes 2-6 of the stem were dried in a forced air oven at 160 F for a minimum of 48 hours. Stems were then ground in a UDY mill to 1 mm size and scanned on a FOSS XDS near infrared reflectance spectroscopy system. Spectra were interpreted using a calibration curve provided by the National Renewable Energy Laboratory to yield total stem composition (Rooney and Wolfrum, 2012).

BTx642 x Tx7000 population whole stems were harvested in the field in June 2010 and 2011. After stripping leaves and leaf sheath, the whole stem was dried in a forced air oven at 160 F for a minimum of 48 hours. Stems were ground in a UDY mill to a particle size of 1mm and scanned on a FOSS XDS near infrared reflectance spectroscopy system as described for the previous population.

QTL Analysis

For both populations, genetic marker data and composition data as generated previously was analyzed using the WinQTL Cartographer software package. QTL were detected using composite interval mapping, and significance thresholds were determined through permutation tests.

Whole Genome Resequencing

BTx642 and Tx7000 genotypes were sequenced by the National Center for Genome Research on an Illumina HiSeq according to manufacturer instructions for paired end sequencing. Resulting sequences were analyzed using the CLC Genomics Workbench (CLC Bio, Cambridge, MA, USA). Reads were trimmed for quality using a quality threshold of 0.05 and an upper boundary of 2 ambiguous nucleotides per read, with a minimum read length of 20 bases. Reads were then mapped to the BTx623 reference genome downloaded from ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v6.0/Sbicolor/assembly/ with annotation track added from ftp://ftp.ensemblgenomes.org/pub/plants/release-13/gtf/sorghum_bicolor. Reads were required to have similarity of at least 0.8 to the reference with overlap of at least 0.75 in order to be mapped, with standard insertion, deletion, and mismatch costs.

Minimum paired-distance was 180 bp, and maximum distance was 1200 bp, with conflict resolution set to vote and global alignment and masking disabled.

SNP Detection

SNP detection was performed using the CLC Genomics Workbench on Illumina HiSeq reads mapped as described previously. Window length and maximum gap length and mismatch count were set to default. Minimum central and average quality were increased to 20 and 15, respectively. Minimum coverage set to 4, with minimum variant frequency set to 0.75 and maximum expected variation set to 2. SNP codon merging was disabled.

SNP and Coverage Plotting

To determine coverage data, reads were re-mapped using the Stampy alignment tool (Lunter and Goodson, 2011). Resultant coverage statistics were determined using BEDtools (Quinlan and Hunter, 2010). Custom scripts were used to determine the density of SNPs and coverage for various stationary window sizes. Resulting information was plotted in graphical form using the Circos visualization package (Krzywinski et al., 2009).

CONCLUSIONS

Sorghum Germplasm Screening and Lignin Analysis

Lignin content variation in sorghum has been observed before, but most studies have compared single genotypes with lignin deficient brown midrib lines, or with other grass or plant species (Bucholtz et al., 1980; Iiyama and Wallis, 1990; Hatfield et al., 1999; Oliver et al., 2004; Oliver et al., 2005). These studies have identified significant lignin variation from one grass species to another, or between mutant sorghum lines and commercial cultivars, but little work has been done comparing sorghum genotypes to one another.

Given the extreme variation present in other aspects of sorghum physiology, it was hypothesized that variation in the amount and distribution of lignin throughout the sorghum stem would also vary significantly between genotypes. Phloroglucinol staining of stems from 351 sorghum accessions revealed substantial variation both in the amount of staining, but also in the localization of staining. Subsequent acetyl bromide extraction and quantification of lignin from 41 of the surveyed lines supported this hypothesis. The variation in lignin content was even more pronounced than expected, with the spectrum of total stem lignin content stretching from 11% to 17%. The lignin minimum is particularly surprising, as this represents a genotype that contains less lignin than a mutant line with compromised function in two of the steps of the monolignol biosynthetic pathway. Taken together, these results indicate a potential source of both high and low lignin phenotypes, allowing for sorghum growers to customize the levels of lignin in their cultivar for various downstream applications.

Identification of Sorghum Monolignol Biosynthetic Genes

Lignin is formed through the complex covalent linkage of three monolignols, alcohols that are largely synthesized by elements of the phenylpropanoid biosynthetic pathway (Boerjan et al., 2003). The genes encoding three of the enzymes in this pathway have previously been identified in sorghum, but the remaining seven currently are not annotated (Bout and Vermerris, 2003; Boddu et al., 2004; Sattler et al., 2009).

Utilizing sequences from monolignol biosynthesis genes that had been identified in other species, it was possible to identify homologs of those genes in the sorghum genome. Discontinuous mega-BLAST was used to identify genes that have undergone substantial divergence, since several of the genes had been identified in species not closely related to sorghum, such as *Arabidopsis thaliana*.

Using a minimum coverage of 50% and an e-value below 1e-50, we were able to identify putative homologs of all the remaining genes encoding monolignol biosynthesis enzymes in the sorghum genome. The physical locations of these genes were also identified and plotted, making them easier to include or exclude from lignin QTL studies.

Development of Genomic Tools

Next generation sequencing allows for the rapid generation of millions of sequences from genomic DNA, accelerating by several orders of magnitude the rate at which genetic markers can be discovered. Once generated, these markers must then be analyzed across a population in order to determine QTL within that population, and then

the output formatted in such a way that downstream processing for QTL detection can be performed.

Custom Perl scripts were written to allow generation of markers from any two homozygous parental populations. Sequences from parental genotypes were compared and identical sequences removed, and remaining sequences compared to the reference BTx623 genotype using a local installation of the BLAST program (Altschul et al., 1990). Variants from the BTx623 reference were identified in each line and the other line assigned the value of the reference, and then these markers were compared with reads from the RIL population to be mapped. Each RIL line was assigned an allele based on marker sequence at each marker location, allowing identification regions of the genome that originated from each parent. This allele assignment was then output in a format compatible for human inspection and downstream genetic map generation.

Currently, NGS markers from parental lines are generated each time the population is analyzed. In order to reduce delays in processing and save reagents, a database was constructed using marker sequences determined for a collection of 30 sorghum accessions. By placing these markers in a MySQL database and adding a web-page frontend, users can now identify divergent markers between any two lines for any region of the genome. This allows users to bypass the marker generation process, which is the most computationally intensive step in genetic map development.

Quantitative Trait Locus Mapping for Stem Composition Traits

Sorghum has been proposed as a candidate crop for biofuels generation due to its rapid growth, drought tolerance, and high biomass yield (Farrel et al., 2006). In order to

further improve sorghum as a biofuels crop, breeders need to be able to generate sorghum varieties with varying stem composition to optimize the feedstock for various biofuels conversion methods. To allow such control, the genetic basis of stem composition must first be identified.

Two RIL populations of sorghum, SC56 x Tx7000 and BTx642 x Tx7000, were field grown in College Station. Plants were harvested and their stems measured, dried, and ground, and the composition of the subsequent material determined though near infrared reflectance spectroscopy.

Genetic maps were created for these lines. In the case of SC56 x Tx7000, genetic maps were constructed using 392 markers determined using the techniques described in Section 4. For BTx642 x Tx7000, genetic maps were constructed by Brock Weers (Weers, 2011) and consisted of 566 markers. These maps were combined with composition data to generate QTL, identifying 34 QTL in the BTx642 x Tx7000 population and 6 QTL in the SC56 x Tx7000 population.

A QTL cluster identified on chromosome 8 of the BTx642 x Tx7000 population represents a region of the genome that controls six of the nine components analyzed. The region under the QTL cluster contains 61 annotated genes, of which 40 are functionally annotated, and only a single gene, Sb08g004720, is annotated as a transcription factor, reducing the list of candidate genes significantly. A similar cluster on chromosome 10 in the SC56 x Tx7000 population also reveals a single transcription factor, indicating the function of transcription factors in the regulation of stem cell composition.

Sorghum Genome Resequencing

The BTx642 x Tx7000 population has been used extensively for mapping the stay-green family of drought resistance traits (Xu et al., 2000; Harris et al., 2007; Weers 2011). While a detailed genetic map exists for this population, even in a marker-rich region of the genome the requirement of QTL mapping that there be recombinational distance between these markers means neighboring markers may be separated by hundreds of kilobases, which may contain dozens or hundreds of genes. Additionally these markers generally only provide information about heredity at a locus, not information about the genes themselves within this region, resulting in the need to laboriously clone and sequence any genes of interest within such a region. By using NGS technologies to re-sequence the entire genome, many of these limitations have been overcome.

Libraries from the sorghum Tx7000 and BTx642 genotypes were sequenced and aligned to the reference sorghum BTx623 genome, yielding approximately 11x coverage. These alignments were then analyzed for sequence variants versus the BTx623 reference sequence, yielding approximately 1.4 million variants for BTx642 and 1 million variants for Tx7000. These levels of variation agree with the known lineages of these cultivars: While Tx7000 descends from Kafir-Milo sorghums (Blackhull Kafir) and BTx623 descends from a Kafir-Milo x Caudatum cross (BTx3197 x SC170), BTx642 results from a Kafir-Milo x Durra cross (BTx398 x IS12555) that was then back-crossed to the Durra background, resulting in an apparently smaller amount of genetic material in common with BTx623. Comparing these discovered variants with

known variants utilized in previous mapping studies, it was discovered that approximately 27% of the known variation had been uncovered in BTx642, and 40% in Tx7000. Future sequencing projects will determine if these numbers are indicative of the discovery rate of the total number of variants, or an artifact of low coverage in the regions surrounding these markers.

Plotting the levels of variation between each line and the reference genome allowed for visualization of regions of identical by descent genetic material between BTx642, Tx7000, and BTx623, such as the region from ~51 Mbp – 59 Mbp on chromosome 9, which is nearly identical across all three genotypes. This visualization also reveals introgressions within regions that previous marker-based genetic maps identified as being regions of single descent, such as the *Ma1/ma1* and *Dw2/dw2* locus on chromosome 6 (Klein et al., 2008). This region of the genome is known to be under selective pressure, indicating that these introgressions may be the result of attempts to introduce and retain desired traits within the region (Menz et al., 2004). The identification of such small introgressions will prove valuable for fine mapping, as the variants used to detect the introgressions in this visualization project can also be used to generate genetic markers for mapping populations.

LITERATURE CITED

Achnine L, Blancaflor EB, Rasmussen S, Dixon RA (2004) Colocalization of L-phenylalanine ammonia-lyase and cinnarnate 4-hydroxylase for metabolic channeling in phenylpropanoid biosynthesis. The Plant Cell **16:** 3098-3109

Adjaye JD, Bakhshi NN (1995) Production of hydrocarbons by catalytic upgrading of a fast pyrolysis bio-oil .2. Comparative catalyst performance and reaction pathways. Fuel Processing Technology **45:** 185-202

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology **215:** 403-410

Amthor JS (2003) Efficiency of lignin biosynthesis: a quantitative analysis. Annals of Botany **91:** 673-695

Arai-Sanoh Y, Ida M, Zhao R, Yoshinaga S, Takai T, Ishimaru T, Maeda H, Nishitani K, Terashima Y, Gau M, Kato N, Matsuoka M, Kondo M (2011) Genotypic variations in non-structural carbohydrate and cell-wall components of the stem in rice, sorghum, and sugar cane. Bioscience Biotechnology and Biochemistry **75:** 1104-1112

Bateman RM, Crane PR, DiMichele WA, Kenrick PR, Rowe NP, Speck T, Stein WE (1998) Early evolution of land plants: Phylogeny, physiology, and ecology of the primary terrestrial radiation. Annual Review of Ecology and Systematics **29:** 263-292

Bhattramakki D, Dong JM, Chhabra AK, Hart GE (2000) An integrated SSR and RFLP linkage map of *Sorghum bicolor* (L.) Moench. Genome **43:** 988-1002

Blee KA, Choi JW, O'Connell AP, Schuch W, Lewis NG, Bolwell GP (2003) A lignin-specific peroxidase in tobacco whose antisense suppression leads to vascular tissue modification. Phytochemistry **64:** 163-176

Boddu J, Svabek C, Sekhon R, Gevens A, Nicholson RL, Jones AD, Pedersen JF, Gustine DL, Chopra S (2004) Expression of a putative flavonoid 3 '-hydroxylase in sorghum mesocotyls synthesizing 3-deoxyanthocyanidin phytoalexins. Physiological and Molecular Plant Pathology **65:** 101-113

Boerjan W, Ralph J, Baucher M (2003) Lignin biosynthesis. Annual Review of Plant Biology **54:** 519-546

Boivin K, Deu M, Rami JF, Trouche G, Hamon P (1999) Towards a saturated sorghum map using RFLP and AFLP markers. Theoretical and Applied Genetics **98:** 320-328

Boudet AM, Kajita S, Grima-Pettenati J, Goffner D (2003) Lignins and lignocellulosics: a better control of synthesis for new and improved uses. Trends in Plant Science **8:** 576-581

Bout S, Vermerris W (2003) A candidate-gene approach to clone the sorghum brown midrib gene encoding caffeic acid O-methyltransferase. Molecular Genetics and Genomics **269:** 205-214

Bowers JE, Abbey C, Anderson S, Chang C, Draye X, Hoppe AH, Jessup R, Lemke C, Lennington J, Li ZK, Lin YR, Liu SC, Luo LJ, Marler BS, Ming RG, Mitchell SE, Qiang D, Reischmann K, Schulze SR, Skinner DN, Wang YW, Kresovich S, Schertz KF, Paterson AH (2003) A high-density genetic recombination map of sequence-tagged sites for sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. Genetics **165:** 367-386

Bradley DJ, Kjellbom P, Lamb CJ (1992) Elicitor-induced and wound-induced oxidative cross-linking of a proline-rich plant-cell wall protein - a novel, rapid defense response. Cell **70:** 21-30

Brady JD, Fry SC (1997) Formation of Di-isodityrosine and loss of isodityrosine in the cell walls of tomato cell-suspension cultures treated with fungal elicitors or H2O2. Plant Physiology **115:** 87-92

Brown DM, Goubet F, Wong VW, Goodacre R, Stephens E, Dupree P, Turner SR (2007) Comparison of five xylan synthesis mutants reveals new insight into the mechanisms of xylan synthesis. The Plant Journal **52:** 1154-1168

Brown DM, Zhang Z, Stephens E, Dupree P, Turner SR (2009) Characterization of IRX10 and IRX10-like reveals an essential role in glucuronoxylan biosynthesis in arabidopsis. The Plant Journal **57:** 732-746

Buanafina MMD (2009) Feruloylation in grasses: current and future perspectives. Molecular Plant **2:** 861-872

Buchanan BB, Gruissem W, Jones RL (2000) Biochemistry & molecular biology of plants. American Society of Plant Physiologists, Rockville

Buchanan CD, Lim SY, Salzman RA, Kagiampakis L, Morishige DT, Weers BD, Klein RR, Pratt LH, Cordonnier-Pratt MM, Klein PE, Mullet JE (2005) *Sorghum bicolor*'s transcriptome response to dehydration, high salinity and ABA. Plant Molecular Biology **58:** 699-720

Bucholtz DL, Cantrell RP, Axtell JD, Lechtenberg VL (1980) Lignin biochemistry of normal and brown midrib mutant sorghum. Journal of Agricultural and Food

Chemistry **28:** 1239-1241

Burke D, Kaufman P, Mcneil M, Albershe.P (1974) Structure of plant-cell walls .6. Survey of walls of suspension-cultured monocots. Plant Physiology **54:** 109-115

Burton RA, Wilson SM, Hrmova M, Harvey AJ, Shirley NJ, Medhurst A, Stone BA, Newbigin EJ, Bacic A, Fincher GB (2006) Cellulose synthase-like CslF genes mediate the synthesis of cell wall (1,3;1,4)-beta-D-glucans. Science **311:** 1940-1942

Carpita NC (1996) Structure and biogenesis of the cell walls of grasses. Annual Review of Plant Physiology and Plant Molecular Biology **47:** 445-476

Carpita NC, Gibeaut DM (1993) Structural models of primary-cell walls in flowering plants - consistency of molecular-structure with the physical-properties of the walls during growth. Plant Journal **3:** 1-30

Cavalier DM, Keegstra K (2006) Two xyloglucan xylosyltransferases catalyze the addition of multiple xylosyl residues to cellohexaose. The Journal of Biological Chemistry **281:** 34197-34207

Chang MCY (2007) Harnessing energy from plant biomass. Current Opinion in Chemical Biology **11:** 677-684

Chen F, Dixon RA (2007) Lignin modification improves fermentable sugar yields for biofuel production. Nature Biotechnology **25:** 759-761

Chen F, Reddy MSS, Temple S, Jackson L, Shadle G, Dixon RA (2006) Multi-site genetic modulation of monolignol biosynthesis suggests new routes for formation of syringyl lignin and wall-bound ferulic acid in alfalfa (*Medicago sativa* L.). Plant Journal **48:** 113-124

Cocuron JC, Lerouxel O, Drakakaki G, Alonso AP, Liepman AH, Keegstra K, Raikhel N, Wilkerson CG (2007) A gene from the cellulose synthase-like C family encodes a beta-1,4 glucan synthase. Proceedings of the National Academy of Sciences of the United States of America **104:** 8550-8555

Cosgrove DJ (2005) Growth of the plant cell wall. Nature Reviews Molecular Cell Biology **6:** 850-861

Dahlberg JA (1995) Dispersal of sorghum and the role of genetic drift. African Crop Sci Journal **3:** 143-151

Dahlberg JA (2000) Classification and characterization of sorghum. In CW Smith, RA Frederiksen, eds, Sorghum: origin, history, technology, and production. John Wiley &

Sons, New York, pp 99-130

Dahlberg JA, Burke JJ, Rosenow DT (2004) Development of a sorghum core collection: Refinement and evaluation of a subset from Sudan. Economic Botany **58:** 556-567

Dien BS, Jung HJG, Vogel KP, Casler MD, Lamb JFS, Iten L, Mitchell RB, Sarath G (2006) Chemical composition and response to dilute-acid pretreatment and enzymatic saccharification of alfalfa, reed canarygrass, and switchgrass. Biomass & Bioenergy **30:** 880-891

Doblin MS, Pettolino FA, Wilson SM, Campbell R, Burton RA, Fincher GB, Newbigin E, Bacic A (2009) A barley cellulose synthase-like CSLH gene mediates (1,3;1,4)-beta-D-glucan synthesis in transgenic Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America **106:** 5996-6001

Doggett H (1988) Sorghum, 2nd ed. John Wiley & Sons, New York

Ehlting J, Buttner D, Wang Q, Douglas CJ, Somssich IE, Kombrink E (1999) Three 4-coumarate : coenzyme A ligases in *Arabidopsis thaliana* represent two evolutionarily divergent classes in angiosperms. Plant Journal **19:** 9-20

Ehlting J, Mattheus N, Aeschliman DS, Li EY, Hamberger B, Cullis IF, Zhuang J, Kaneda M, Mansfield SD, Samuels L, Ritland K, Ellis BE, Bohlmann J, Douglas CJ (2005) Global transcript profiling of primary stems from *Arabidopsis thaliana* identifies candidate genes for missing links in lignin biosynthesis and transcriptional regulators of fiber differentiation. Plant Journal **42:** 618-640

EPA (2000) Major crops grown in the United States. http://www.epa.gov/agriculture/ag101/cropmajor.html

Esquerre MT, Mazau D (1974) Effect of a fungal disease on extensin, plant-cell wall glycoprotein. Journal of Experimental Botany **25:** 509-513

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th ed. Longman, Essex

FAO (1996) The world sorghum and millet economies facts, trends, and outlook - a joint publication by FAO and ICRISAT. http://www.fao.org/es/esc/common/ecg/63/en/Sor_Mil.pdf

Farrell AE (2006) Ethanol can contribute to energy and environmental goals. Science **312:** 1748-1748

Fukushima RS, Dehority BA, Loerch SC (1991) Modification of a colorimetric analysis for lignin and its use in studying the inhibitory effects of lignin on forage digestion by ruminal microorganisms. Journal of Animal Science **69:** 295-304

Galbe M, Zacchi G (2007) Pretreatment of lignocellulosic materials for efficient bioethanol production. Biofuels **108:** 41-65

Geldermann H (1975) Investigations on inheritance of quantitative characters in animals by gene markers .1. Methods. Theoretical and Applied Genetics **46:** 319-330

Gibeaut DM, Pauly M, Bacic A, Fincher GB (2005) Changes in cell wall polysaccharides in developing barley (*Hordeum vulgare*) coleoptiles. Planta **221:** 729-738

Giddings TH, Staehelin LA (1988) Spatial relationship between microtubules and plasma-membrane rosettes during the deposition of primary wall microfibrils in *Closterium* Sp. Planta **173:** 22-30

Glazer AN, Nikaido H (2007) Microbial biotechnology: Fundamentals of applied microbiology, 2nd ed. Cambridge University Press, Cambridge

Grabber JH, Schatz PF, Kim H, Lu FC, Ralph J (2010) Identifying new lignin bioengineering targets: 1. Monolignol-substitute impacts on lignin formation and cell wall fermentability. Bmc Plant Biology **10:** 114-126

Gupta PK, Rustgi S (2004) Molecular markers from the transcribed/expressed region of the genome in higher plants. Funct Integr Genomics **4:** 139-162

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

Hamelinck CN, van Hooijdonk G, Faaij APC (2005) Ethanol from lignocellulosic biomass: techno-economic performance in short-, middle- and long-term. Biomass & Bioenergy **28:** 384-410

Harlan JR, Dewet JMJ (1972) Simplified classification of cultivated sorghum. Crop Science **12:** 172-176

Harris K, Subudhi PK, Borrell A, Jordan D, Rosenow D, Nguyen H, Klein P, Klein R, Mullet J (2007) Sorghum stay-green QTL individually reduce post-flowering drought-induced leaf senescence. Journal of Experimental Botany **58:** 327-338

Harris KR (2007) Genetic analysis of the *Sorghum bicolor* stay-green drought tolerance trait. PhD dissertation. Texas A&M University, College Station

Harris PJ, Hartley RD (1976) Detection of bound ferulic acid in cell-walls of gramineae by ultraviolet fluorescence microscopy. Nature **259:** 508-510

Hatfield R, Fukushima RS (2005) Can lignin be accurately measured? Crop Science **45:** 832-839

Hatfield RD, Grabber J, Ralph J, Brei K (1999) Using the acetyl bromide assay to determine lignin concentrations in herbaceous plants: some cautionary notes. Journal of Agricultural and Food Chemistry **47:** 628-632

Hatfield RD, Hatfield RD, Wilson JR, Mertens DR (1999) Composition of cell walls isolated from cell types of grain sorghum stems. Journal of the Science of Food and Agriculture **79:** 891-899

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

Herth W (1983) Arrays of plasma-membrane rosettes involved in cellulose microfibril formation of *Spirogyra*. Planta **159:** 347-356

Hoffmann L, Besseau S, Geoffroy P, Ritzenthaler C, Meyer D, Lapierre C, Pollet B, Legrand M (2004) Silencing of hydroxycinnamoy-coenzyme A shikimate/quinate hydroxycinnamoyltransferase affects phenylpropanoid biosynthesis. The Plant Cell **16:** 1446-1465

Holland N, Holland D, Helentjaris T, Dhugga KS, Xoconostle-Cazares B, Delmer DP (2000) A comparative analysis of the plant cellulose synthase (CesA) gene family. Plant Physiology **123:** 1313-1323

Humphreys JM, Chapple C (2002) Rewriting the lignin roadmap. Current Opinion in Plant Biology **5:** 224-229

Iiyama K, Lam TBT, Stone BA (1994) Covalent cross-links in the cell-wall. Plant Physiology **104:** 315-320

Iiyama K, Wallis AFA (1990) Determination of lignin in herbaceous plants by an improved acetyl bromide procedure. Journal of the Science of Food and Agriculture **51:** 145-161

Iwai H, Masaoka N, Ishii T, Satoh S (2002) A pectin glucuronyltransferase gene is essential for intercellular attachment in the plant meristem. Proceedings of the

National Academy of Sciences of the United States of America **99:** 16319-16324

Jae J, Tompsett GA, Foster AJ, Hammond KD, Auerbach SM, Lobo RF, Huber GW (2011) Investigation into the shape selectivity of zeolite catalysts for biomass conversion. Journal of Catalysis **279:** 257-268

Jansen RC (1996) A general Monte Carlo method for mapping multiple quantitative trait loci. Genetics **142:** 305-311

Jung HG, Casler MD (2006) Maize stem tissues: Cell wall concentration and composition during development. Crop Science **46:** 1793-1800

Jung HJG, Samac DA, Sarath G (2012) Modifying crops to increase cell wall digestibility. Plant Science **185:** 65-77

Kearsey MJ (1998) The principles of QTL analysis (a minimal mathematics approach). Journal of Experimental Botany **49:** 1619-1623

Kearsey MJ, Pooni HS (1996) The genetical analysis of quantitative traits. Chapman & Hall, London

Kim JS, Islam-Faridi MN, Klein PE, Stelly DM, Price HJ, Klein RR, Mullet JE (2005) Comprehensive molecular cytogenetic analysis of sorghum genome architecture: Distribution of euchromatin, heterochromatin, genes and recombination in comparison to rice. Genetics **171:** 1963-1976

Kim JS, Klein PE, Klein RR, Price HJ, Mullet JE, Stelly DM (2005) Molecular cytogenetic maps of sorghum linkage groups 2 and 8. Genetics **169:** 955-965

Kimura S, Laosinchai W, Itoh T, Cui XJ, Linder CR, Brown RM (1999) Immunogold labeling of rosette terminal cellulose-synthesizing complexes in the vascular plant *Vigna angularis*. The Plant Cell **11:** 2075-2085

Klein RR, Mullet JE, Jordan DR, Miller FR, Rooney WL, Menz MA, Franks CD, Klein PE (2008) The effect of tropical sorghum conversion and inbred development on genome diversity as revealed by high-resolution genotyping. Crop Science **48:** S12-S26

Komatsu S, Kojima K, Suzuki K, Ozaki K, Higo K (2004) Rice proteome database based on two-dimensional polyacrylamide gel electrophoresis: Its status in 2003. Nucleic Acids Research **32:** D388-D392

Koyama M, Helbert W, Imai T, Sugiyama J, Henrissat B (1997) Parallel-up structure evidences the molecular directionality during biosynthesis of bacterial cellulose. Proceedings of the National Academy of Sciences of the United States of

America **94:** 9091-9095

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: An information aesthetic for comparative genomics. Genome Research **19:** 1639-1645

Kubo M, Udagawa M, Nishikubo N, Horiguchi G, Yamaguchi M, Ito J, Mimura T, Fukuda H, Demura T (2005) Transcription switches for protoxylem and metaxylem vessel formation. Genes & Development **19:** 1855-1860

Labbe N, Rials TG, Kelley SS, Martin MZ (2007) High throughput material characterization techniques: near infrared and laser induced breakdown spectroscopy in materials. In D Argyropoulos, ed, Chemicals and energy from forest biomass. Oxford University Press, Oxford, pp 495-512

Labbe N, Ye XP, Franklin JA, Womac AR, Tyler DD, Rials TG (2008) Analysis of switchgrass characteristics using near infrared spectroscopy. Bioresources **3:** 1329-1348

Lai-Kee-Him J, Chanzy H, Muller M, Putaux JL, Imai T, Bulone V (2002) In vitro versus in vivo cellulose microfibrils from plant primary wall synthases: Structural differences. Journal of Biological Chemistry **277:** 36931-36939

Lamport DTA (1977) Structure, biosynthesis, and significance of cell wall glycoproteins. In FA Loewus and VC Runeckles, eds, The structure, biosynthesis, and degradation of wood. Plenum Press, New York, pp. 79-115

Lamport DTA, Kieliszewski MJ, Chen YN, Cannon MC (2011) Role of the extensin superfamily in primary cell wall architecture. Plant Physiology **156:** 11-19

Larsen K (2004) Molecular cloning and characterization of cDNAs encoding cinnamoyl CoA reductase (CCR) from barley (*Hordeum vulgare*) and potato (*Solanum tuberosum*). Journal of Plant Physiology **161:** 105-112

Ledbetter MC, Porter KR (1963) A microtubule in plant cell fine structure. Journal of Cell Biology **19:** 239-250

Lee C, O'Neill MA, Tsumuraya Y, Darvill AG, Ye ZH (2007) The irregular xylem9 mutant is deficient in xylan xylosyltransferase activity. Plant & Cell Physiology **48:** 1624-1634

Lee SH, van der Werf JHJ (2006) An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. Genetics Selection Evolution **38:** 25-43

Lin YR, Schertz KF, Paterson AH (1995) Comparative-analysis of QTLs affecting plant height and maturity across the Poaceae, in reference to an interspecific sorghum population. Genetics **141:** 391-411

Lloyd CW (1984) Toward a dynamic helical model for the influence of microtubules on wall patterns in plants. International Review of Cytology **86:** 1-51

Lowry B, Hebant C, Lee D (1980) The origin of land plants - a new look at an old problem. Taxon **29:** 183-197

Lunter G, Goodson M (2011) Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Research **21:** 936-939

Mcbee GG, Miller FR (1993) Stem carbohydrate and lignin concentrations in sorghum hybrids at 7 growth-stages. Crop Science **33:** 530-534

Mclaughlin SGA, Dilger JP (1980) Transport of protons across membranes by weak acids. Physiological Reviews **60:** 825-863

Menz MA, Klein RR, Mullet JE, Obert JA, Unruh NC, Klein PE (2002) A high-density genetic map of *Sorghum bicolor* (L.) Moench based on 2926 AFLP, RFLP and SSR markers. Plant Molecular Biology **48:** 483-499

Menz MA, Klein RR, Unruh NC, Rooney WL, Klein PE, Mullet JE (2004) Genetic diversity of public inbreds of sorghum determined by mapped AFLP and SSR markers. Crop Science **44:** 1236-1244

Miles CM, Wayne ML (2008) Quantitative trait locus (QTL) analysis. Nature Education **1:**1

Minami E, Ozeki Y, Matsuoka M, Koizuka N, Tanaka Y (1989) Structure and some characterization of the gene for phenylalanine ammonia-lyase from rice plants. European Journal of Biochemistry **185:** 19-25

Mitchell RA, Dupree P, Shewry PR (2007) A novel bioinformatics approach identifies candidate genes for the synthesis and feruloylation of arabinoxylan. Plant Physiology **144:** 43-53

Mitsuda N, Iwase A, Yamamoto H, Yoshida M, Seki M, Shinozaki K, Ohme-Takagi M (2007) NAC transcription factors, NST1 and NST3, are key regulators of the formation of secondary walls in woody tissues of Arabidopsis. The Plant Cell **19:** 270-280

Mitsuda N, Seki M, Shinozaki K, Ohme-Takagi M (2005) The NAC transcription factors NST1 and NST2 of Arabidopsis regulate secondary wall thickenings and

are required for anther dehiscence. The Plant Cell **17:** 2993-3006

Mizutani M, Ohta D, Sato R (1997) Isolation of a cDNA and a genomic clone encoding cinnamate 4-hydroxylase from Arabidopsis and its expression manner in planta. Plant Physiology **113:** 755-763

Mohan D, Pittman CU, Steele PH (2006) Pyrolysis of wood/biomass for bio-oil: A critical review. Energy & Fuels **20:** 848-889

Moller I, Sorensen I, Bernal AJ, Blaukopf C, Lee K, Obro J, Pettolino F, Roberts A, Mikkelsen JD, Knox JP, Bacic A, Willats WGT (2007) High-throughput mapping of cell-wall polymers within and between plants using novel microarrays. Plant Journal **50:** 1118-1128

Moore R (1998) Botany, 2nd ed. WCB/McGraw-Hill, New York

Morrison IM (1972) Semi-micro method for determination of lignin and its use in predicting digestibility of forage crops. Journal of the Science of Food and Agriculture **23:** 455-463

Mourier T, Jeffares DC (2003) Eukaryotic intron loss. Science **300:** 1393-1393

Murphy RL, Klein RR, Morishige DT, Brady JA, Rooney WL, Miller FR, Dugas DV, Klein PE, Mullet JE (2011) Coincident light and clock regulation of pseudoresponse regulator protein 37 (PRR37) controls photoperiodic flowering in sorghum. Proceedings of the National Academy of Sciences of the United States of America **108:** 16469-16474

Murray SC, Sharma A, Rooney WL, Klein PE, Mullet JE, Mitchell SE, Kresovich S (2008) Genetic improvement of sorghum as a biofuel feedstock: I. QTL for stem sugar and grain nonstructural carbohydrates. Crop Science **48:** 2165-2179

Murty BR, Govil JN (1967) Description of 70 groups in genus *Sorghum* based on a modified Snowdens classification. Indian Journal of Genetics and Plant Breeding **27:** 75-91

Nagy I, Stagel A, Sasvari Z, Roder M, Ganal M (2007) Development, characterization, and transferability to other Solanaceae of microsatellite markers in pepper (*Capsicum annuum* L.). Genome **50:** 668-688

Niemann GJ, Pureveen JBM, Eijkel GB, Poorter H, Boon JJ (1992) Differences in relative growth-rate in 11 grasses correlate with differences in chemical-composition as determined by pyrolysis mass-spectrometry. Oecologia **89:** 567-573

Novaes E, Kirst M, Chiang V, Winter-Sederoff H, Sederoff R (2010) Lignin and biomass: a negative correlation for wood formation and lignin content in trees. Plant Physiology **154:** 555-561

Oliver AL, Pederson, J.F., Grant, R.J., Klopfenstein, T.J. (2005) Comparative effects of the sorghum bmr-6 and bmr-12 genes: I. Forage sorghum yield and quality. Crop Science **45:** 2234-2239

Oliver AL, Pedersen JF, Grant RJ, Klopfenstein TJ, Jose HD (2005) Comparative effects of the sorghum bmr-6 and bmr-12 genes: II. Grain yield, stover yield, and stover quality in grain sorghum. Crop Science **45:** 2240-2245

Oliver AL, Grant RJ, Pedersen JF, O'Rear J (2004) Comparison of brown midrib-6 and -18 forage sorghum with conventional sorghum and corn silage in diets of lactating dairy cows. Journal of Dairy Science **87:** 637-644

Onnerud H, Zhang LM, Gellerstedt G, Henriksson G (2002) Polymerization of monolignols by redox shuttle-mediated enzymatic oxidation: a new model in lignin biosynthesis I. The Plant Cell **14:** 1953-1962

Ou SY, Kwok KC (2004) Ferulic acid: pharmaceutical functions, preparation and applications in foods. Journal of the Science of Food and Agriculture **84:** 1261-1269

Paredez AR, Somerville CR, Ehrhardt DW (2006) Visualization of cellulose synthase demonstrates functional association with microtubules. Science **312:** 1491-1495

Parh DK, Jordan DR, Aitken EAB, Mace ES, Jun-ai P, McIntyre CL, Godwin ID (2008) QTL analysis of ergot resistance in sorghum. Theoretical and Applied Genetics **117:** 369-382

Paterson AH (1995) Molecular dissection of quantitative traits - progress and prospects. Genome Research **5:** 321-333

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang HB, Wang XY, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang LF, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob-ur-Rahman, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature **457:** 551-556

Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating

divergence of the cereals, and its consequences for comparative genomics. Proceedings of the National Academy of Sciences of the United States of America **101:** 9903-9908

Pauly M, Keegstra K (2008) Cell-wall carbohydrates and their modification as a resource for biofuels. Plant Journal **54:** 559-568

Pear JR, Kawagoe Y, Schreckengost WE, Delmer DP, Stalker DM (1996) Higher plants contain homologs of the bacterial celA genes encoding the catalytic subunit of cellulose synthase. Proceedings of the National Academy of Sciences of the United States of America **93:** 12637-12642

Perrin RM, DeRocher AE, Bar-Peled M, Zeng W, Norambuena L, Orellana A, Raikhel NV, Keegstra K (1999) Xyloglucan fucosyltransferase, an enzyme involved in plant cell wall biosynthesis. Science **284:** 1976-1979

Poke FS, Raymond CA (2006) Predicting extractives, lignin, and cellulose contents using near infrared spectroscopy on solid wood in *Eucalyptus globulus*. Journal of Wood Chemistry and Technology **26:** 187-199

Popper ZA (2008) Evolution and diversity of green plant cell walls. Current Opinion In Plant Biology **11:** 286-292

Porchia AC, Sorensen SO, Scheller HV (2002) Arabinoxylan biosynthesis in wheat. Characterization of arabinosyltransferase activity in Golgi membranes. Plant Physiology **130:** 432-441

Quinby JR (1974) Sorghum improvement and the genetics of growth. Texas Agricultural Experiment Station, College Station

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26:** 841-842

Ragauskas AJ, Williams CK, Davison BH, Britovsek G, Cairney J, Eckert CA, Frederick WJ, Hallett JP, Leak DJ, Liotta CL, Mielenz JR, Murphy R, Templer R, Tschaplinski T (2006) The path forward for biofuels and biomaterials. Science **311:** 484-489

Ralph J, Helm RF (1993) Lignin hydroxycinnamic acid polysaccharide complexes - synthetic models for regiochemical characterization. In HG Jun, DR Buxton, RD Hatfield, J Ralph, eds. Forage cell wall structure and digestibility**.** American Society for Agronomy, Madison, pp 201-246

Ralph J, Peng JP, Lu FC, Hatfield RD, Helm RF (1999) Are lignins optically active? Journal of Agricultural and Food Chemistry **47:** 2991-2996

Rasmussen S, Dixon RA (1999) Transgene-mediated and elicitor-induced perturbation of metabolic channeling at the entry point into the phenylpropanoid pathway. The Plant Cell **11:** 1537-1551

Raven PH, Evert RF, Eichhorn SE (1986) Biology of plants, 4th ed. Worth Publishers, New York

Riboulet C, Fabre F, Denoue D, Martinant JP, Lefevre B, Barriere Y (2008) QTL mapping and candidate gene research from lignin content and cell wall digestibility in a top-cross of a flint maize recombinant inbred line progeny harvested at silage stage. Maydica **53:** 1-9

Richmond T (2000) Higher plant cellulose synthases. Genome Biology **1:** REVIEWS:3001.1-3001.6

Robinson DG, Quader H (1981) Structure, synthesis, and orientation of microfibrils .9. A freeze-fracture investigation of the *Oocystis* plasma-membrane after inhibitor treatments. European Journal of Cell Biology **25:** 278-288

Rooney WL, Smith CW (2001) Techniques for developing new cultivars. In CW Smith, RA Fredericksen, eds, Sorghum: origin, history, technology, and production. John Wiley & Sons, New York, pp 329-347

Rosenow DT, Dahlberg JA, Peterson GC, Clark LE, Miller FR, SotomayorRios A, Hamburger AJ, MaderaTorres P, QuilesBelen A, Woodfin CA (1997) Registration of fifty converted sorghums from the sorghum conversion program. Crop Science **37:** 1397-1398

Salzman RA, Brady JA, Finlayson SA, Buchanan CD, Summer EJ, Sun F, Klein PE, Klein RR, Pratt LH, Cordonnier-Pratt MM, Mullet JE (2005) Transcriptional profiling of sorghum induced by methyl jasmonate, salicylic acid, and aminocyclopropane carboxylic acid reveals cooperative regulation and novel gene responses. Plant Physiology **138:** 352-368

Sanderson MA, Jones RM, Ward J, Wolfe R (1992) Silage sorghum performance trial at Stephenville. Forage Research in Texas. Report **PR-5018**:16-18

Sattler SE, Saathoff AJ, Haas EJ, Palmer NA, Funnell-Harris DL, Sarath G, Pedersen JF (2009) A nonsense mutation in a cinnamyl alcohol dehydrogenase gene is responsible for the sorghum brown midrib6 phenotype. Plant Physiology **150:** 584-595

Saulnier L, Crepeau MJ, Lahaye M, Thibault JF, Garcia-Conesa MT, Kroon PA, Williamson G (1999) Isolation and structural determination of two 5,5 '-diferuloyl oligosaccharides indicate that maize heteroxylans are covalently cross-

linked by oxidatively coupled ferulates. Carbohydrate Research **320:** 82-92

Scheller HV, Ulvskov P (2010) Hemicelluloses. Annual Review of Plant Biology **61:** 263-289

Shiringani AL, Friedt W (2011) QTL for fibre-related traits in grain x sweet sorghum as a tool for the enhancement of sorghum as a biomass crop. Theoretical and Applied Genetics **123:** 999-1011

Shiringani AL, Frisch M, Friedt W (2010) Genetic mapping of QTLs for sugar-related traits in a RIL population of *Sorghum bicolor* L. Moench. Theoretical and Applied Genetics **121:** 323-336

Sluiter A, Hames B, Ruiz R, Scarlata C, Sluiter J, Templeton D, Crocker D (2008) Determination of structural carbohydrates and lignin in biomass. National Renewable Energy Laboratory, Golden, **TP-510-42618:**1-15

Smith CW, Frederiksen RA (2000) Sorghum : origin, history, technology, and production. John Wiley & Sons, New York

Somerville C (2006) Cellulose synthesis in higher plants. Annual Review of Cell and Developmental Biology **22:** 53-78

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S (2002) The generic genome browser: a building block for a model organism system database. Genome Res **12:** 1599-1610

Steiner-Lange S, Unte US, Eckstein L, Yang CY, Wilson ZA, Schmelzer E, Dekker K, Saedler H (2003) Disruption of *Arabidopsis thaliana* MYB26 results in male sterility due to non-dehiscent anthers. Plant Journal **34:** 519-528

Stone BA, Clarke AE (1992) Chemistry and biology of 1,3-[beta]-glucans. La Trobe University Press, Victoria

Swigonova Z, Lai JS, Ma JX, Ramakrishna W, Llaca V, Bennetzen JL, Messing J (2004) Close split of sorghum and maize genome progenitors. Genome Research **14:** 1916-1923

Tajima F (1989) Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595

Thomas JR, Mcneil M, Darvill AG, Albersheim P (1987) Structure of plant-cell walls .29. Isolation and characterization of wall polysaccharides from suspension-cultured douglas-fir cells. Plant Physiology **83:** 659-671

USDA (2011) 2011 Crop production summary. http://www.usda.gov/nass/PUBS/TODAYRPT/cropan12.txt

Vanholme R, Morreel K, Ralph J, Boerjan W (2008) Lignin engineering. Current Opinion in Plant Biology **11:** 278-285

Vietor DM, Miller FR (1990) Assimilation, partitioning, and nonstructural carbohydrates in sweet compared with grain-sorghum. Crop Science **30:** 1109-1115

Vincken JP, Schols HA, Oomen RJFJ, McCann MC, Ulvskov P, Voragen AGJ, Visser RGF (2003) If homogalacturonan were a side chain of rhamnogalacturonan I. Implications for cell wall architecture. Plant Physiology **132:** 1781-1789

Vogel J (2008) Unique aspects of the grass cell wall. Current Opinion in Plant Biology **11:** 301-307

Wallace G, Fry SC (1995) In-vitro peroxidase-catalyzed oxidation of ferulic acid-esters. Phytochemistry **39:** 1293-1299

Weers BD (2011) Integrated analysis of phenology, traits, and QTL in the drought resistant sorghum genotypes BTx642 and TX7000. PhD dissertation. Texas A&M University, College Station.

Weng JK, Chapple C (2010) The origin and evolution of lignin biosynthesis. New Phytologist **187:** 273-285

Weng JK, Li X, Bonawitz ND, Chapple C (2008) Emerging strategies of lignin engineering and degradation for cellulosic biofuel production. Current Opinion in Biotechnology **19:** 166-172

Weng JK, Li X, Stout J, Chapple C (2008) Independent origins of syringyl lignin in vascular plants. Proceedings of the National Academy of Sciences of the United States of America **105:** 7887-7892

White RH (1987) Effect of lignin content and extractives on the higher heating value of wood. Wood and Fiber Science **19:** 446-452

Willats WGT, McCartney L, Mackie W, Knox JP (2001) Pectin: cell biology and prospects for functional analysis. Plant Molecular Biology **47:** 9-27

Wilson SM, Burton RA, Doblin MS, Stone BA, Newbigin EJ, Fincher GB, Bacic A (2006) Temporal and spatial appearance of wall polysaccharides during cellularization of barley (*Hordeum vulgare*) endosperm. Planta **224:** 655-667

Wong WS, Guo D, Wang XL, Yin ZQ, Xia B, Li N (2005) Study of cis-cinnamic acid in *Arabidopsis thaliana*. Plant Physiology and Biochemistry **43:** 929-937

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

Xu X, Liu X, Ge S, Jensen JD, Hu FY, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li JX, He WM, Zhang GJ, Zheng XM, Zhang FM, Li YR, Yu C, Kristiansen K, Zhang XQ, Wang J, Wright M, McCouch S, Nielsen R, Wang J, Wang W (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nature Biotechnology **30:** 105-U157

Yang CY, Xu ZY, Song J, Conner K, Barrena GV, Wilson ZA (2007) Arabidopsis MYB26/MALE STERILE35 regulates secondary thickening in the endothecium and is essential for anther dehiscence. The Plant Cell **19:** 534-548

Zeng W, Chatterjee M, Faik A (2008) UDP-xylose-stimulated glucuronyltransferase activity in wheat microsomal membranes: characterization and role in glucurono(arabino)xylan biosynthesis. Plant Physiology **147:** 78-91

Zhong R, Demura T, Ye ZH (2006) SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of Arabidopsis. The Plant Cell **18:** 3158-3170

Zhong RQ, Richardson EA, Ye ZH (2007a) Two NAC domain transcription factors, SND1 and NST1, function redundantly in regulation of secondary wall synthesis in fibers of Arabidopsis. Planta **225:** 1603-1611

Zhong R, Richardson EA, Ye ZH (2007b) The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in Arabidopsis. The Plant Cell **19:** 2776-2792

Zhong RQ, Ye ZH (2007c) Regulation of cell wall biosynthesis. Current Opinion In Plant Biology **10:** 564-572

APPENDIX

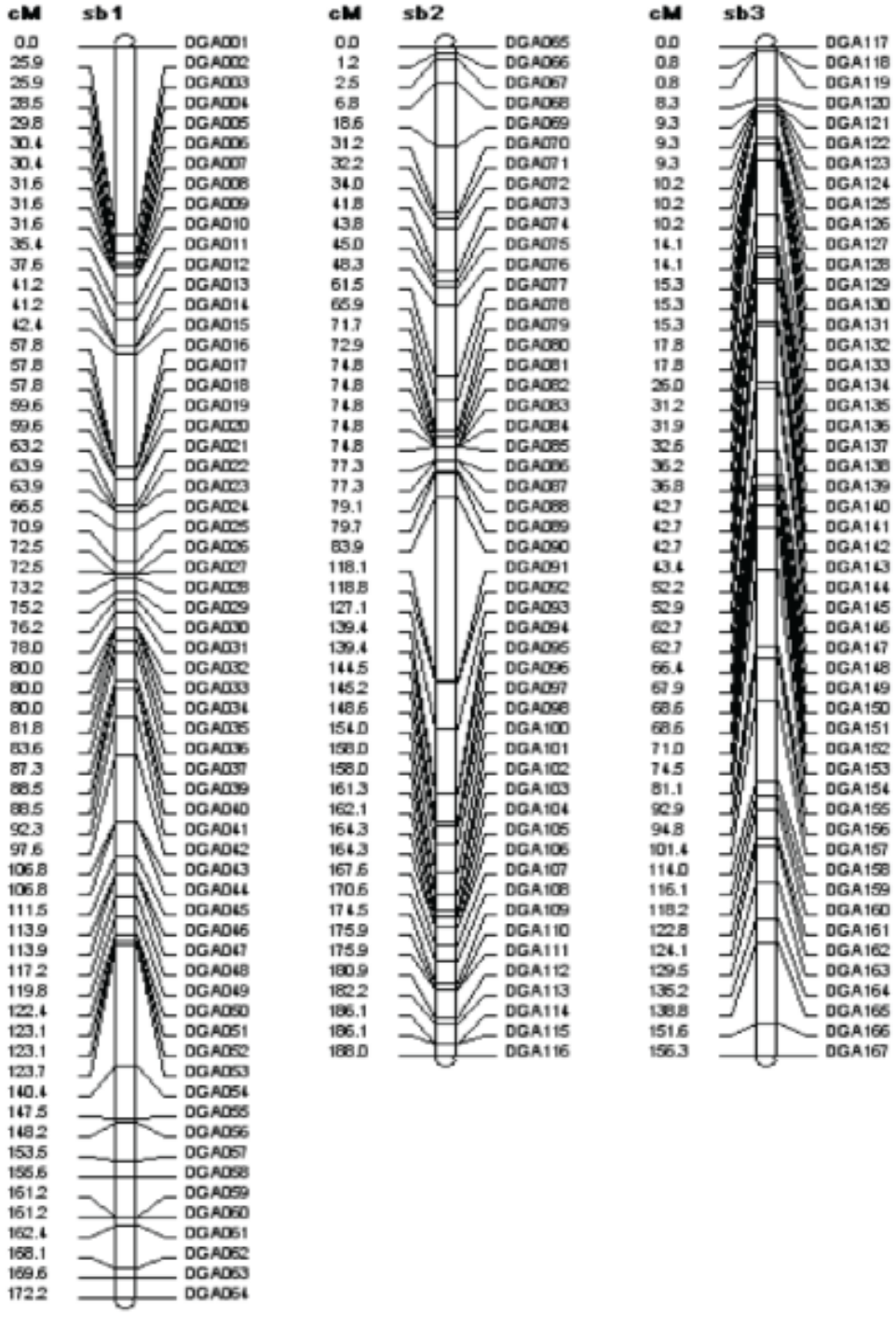


Figure A-1: Genetic map of chromosomes 1-3 for the SC56 x Tx7000 RIL population

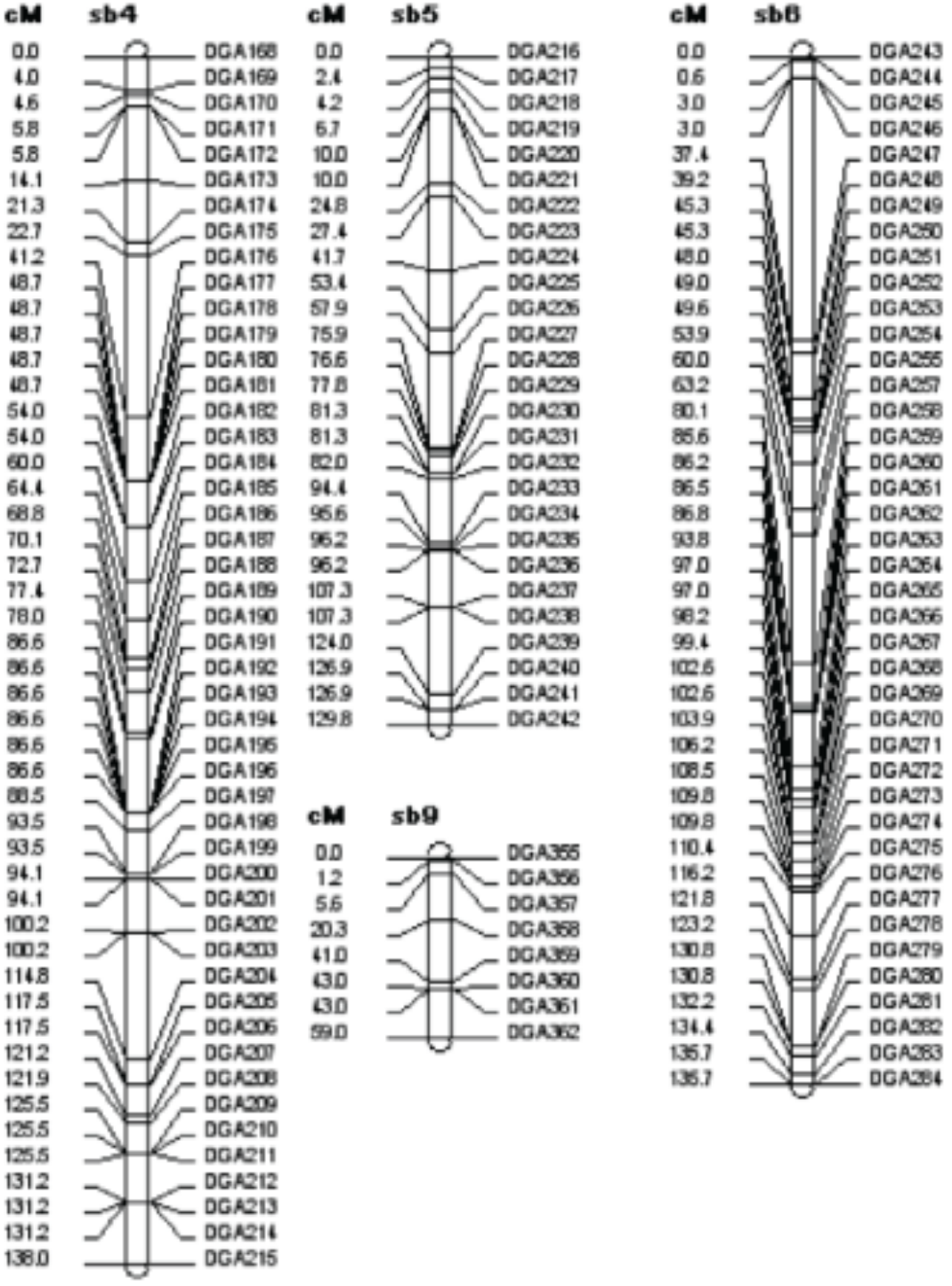


Figure A-2: Genetic map of chromosomes 4, 5, 6, and 9 for the SC56 x Tx7000 RIL population

Table A-1: List of genes located within BTx642 x Tx7000 QTL boundaries described in Section 5

Gene Name	Panther ID	Panther Description	Gene Start (bp)	Gene End (bp)
Sb08g004510	PTHR22982	CALCIUM/CALMODULIN-DEPENDENT PROTEIN KINASE-RELATED	5413386	5417550
Sb08g004460	PTHR23258	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE	5339414	5345358
Sb08g004890	PTHR11566	DYNAMIN	6067531	6074823
Sb08g004450	PTHR23258	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE	5332351	5333913
Sb08g004780	PTHR11669	REPLICATION FACTOR C / DNA POLYMERASE III GAMMA-TAU SUBUNIT	5935588	5939608
Sb08g004410	PTHR23421	BETA-GALACTOSIDASE RELATED	5275015	5291668
Sb08g004690	PTHR12154	GLYCOSYL TRANSFERASE-RELATED	5640252	5643719
Sb08g004955			6222068	6223171
Sb08g004930	PTHR19383	CYTOCHROME P450	6185629	6187809
Sb08g004910	PTHR23413	60S RIBOSOMAL PROTEIN L32 AND DNA-DIRECTED RNA POLYMERASE II, SUBUNIT N	6111731	6115200
Sb08g004720	PTHR10641	MYB-RELATED	5750343	5751112
Sb08g004745			5895704	5896828
Sb08g004560			5476332	5477128
Sb08g004600			5518924	5520240
Sb08g004990	PTHR22950	AMINO ACID TRANSPORTER	6281489	6283693
Sb08g004470	PTHR11911	INOSINE-5-MONOPHOSPHATE DEHYDROGENASE RELATED	5349758	5379686
Sb08g004825			6022410	6022649
Sb08g004900	PTHR11085	CHROMATIN REGULATORY PROTEIN SIR2	6081334	6086296
Sb08g004540	PTHR11929	ALPHA-(1,3)-FUCOSYLTRANSFERASE	5439756	5442535
Sb08g004960	PTHR22950	AMINO ACID TRANSPORTER	6226628	6228271
Sb08g004500	PTHR11627	FRUCTOSE-BISPHOSPHATE ALDOLASE	5410066	5411989
Sb08g004400	PTHR11895	AMIDASE	5263138	5267799
Sb08g004590			5497346	5499285
Sb08g004640	PTHR19446	REVERSE TRANSCRIPTASES	5570493	5572508
Sb08g004870			6052903	6056677
Sb08g004630			5561308	5567432
Sb08g004550	PTHR23258	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE	5467097	5469610
Sb08g004430			5314574	5315860
Sb08g004750			5898817	5901176
Sb08g004840	PTHR11699	ALDEHYDE DEHYDROGENASE-RELATED	6034374	6040509
Sb08g004580			5489787	5491840
Sb08g004710	PTHR19134	PROTEIN-TYROSINE PHOSPHATASE	5681324	5687411
Sb08g004760			5921558	5922712
Sb08g004860			6046227	6048110
Sb08g004915	PTHR22950	AMINO ACID TRANSPORTER	6154129	6155061
Sb08g004730	PTHR10766	TRANSMEMBRANE 9 SUPERFAMILY PROTEIN MEMBER	5816061	5821466
Sb08g004570	PTHR13734	TRNA-NUCLEOTIDYLTRANSFERASE 1	5479797	5486707
Sb08g004940	PTHR23258	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE	6195504	6198219
Sb08g004810	PTHR22950	AMINO ACID TRANSPORTER	5976226	5977674

Table A-1 continued

Gene Name	Panther ID	Panther Description	Gene Start (bp)	Gene End (bp)
Sb08g004390	PTHR10263	VACUOLAR ATP SYNTHASE PROTEOLIPID SUBUNIT	5257369	5260353
Sb08g004680	PTHR21495	NUCLEOPORIN-RELATED	5631814	5632383
Sb08g004620	PTHR14107	WD REPEAT PROTEIN	5551601	5557990
Sb08g004830	PTHR23258	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE	6026192	6028734
Sb08g004905			6110157	6111426
Sb08g004700	PTHR23258	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE	5679283	5679745
Sb08g004850	PTHR14221	WD REPEAT DOMAIN 44	6042857	6045827
Sb08g004520	PTHR10361	SODIUM-BILE ACID COTRANSPORTER RELATED	5431149	5433505
Sb08g004610			5550733	5551262
Sb08g004880			6062693	6066807
Sb08g004920			6157839	6158613
Sb08g004790	PTHR12802	SWI/SNF COMPLEX-RELATED	5942393	5949689
Sb08g004800	PTHR13683	ASPARTYL PROTEASES	5972462	5975686
Sb08g004950	PTHR11065	GUFA PROTEIN - RELATED	6214761	6218215
Sb08g004670	PTHR21495	NUCLEOPORIN-RELATED	5624451	5624987
Sb08g004650			5577657	5582630
Sb08g004570	PTHR13734	TRNA-NUCLEOTIDYLTRANSFERASE 1	5479797	5486306
Sb08g004770			5928250	5930715
Sb08g004740	PTHR11178	IRON-SULFUR CLUSTER SCAFFOLD PROTEIN NFU-RELATED	5864125	5866178
Sb08g004380	PTHR10263	VACUOLAR ATP SYNTHASE PROTEOLIPID SUBUNIT	5243063	5245675
Sb08g004926			6180403	6181200
Sb08g004420			5294545	5299196
Sb08g004980	PTHR22950	AMINO ACID TRANSPORTER	6253081	6255105

Table A-2: Listed of genes located within SC56 x Tx7000 QTL boundaries as described in Section 5

Gene Name	Panther ID	Panther Description	Gene Start (bp)	Gene End (bp)
Sb10g007290	PTHR22893	NADH OXIDOREDUCTASE-RELATED	7084056	7085665
Sb10g007500			7352447	7356315
Sb10g007710	PTHR11926	GLUCOSYL/GLUCURONOSYL TRANSFERASES	7519172	7519801
Sb10g007310	PTHR22893	NADH OXIDOREDUCTASE-RELATED	7125919	7127233
Sb10g007660			7471744	7472700
Sb10g007590	PTHR11527	SMALL HEAT-SHOCK PROTEIN (HSP20) FAMILY	7429113	7430214
Sb10g007420	PTHR12374	TRANSCRIPTIONAL ADAPTOR 2 (ADA2)-RELATED	7280665	7281800
Sb10g007750	PTHR23354	NUCLEOLAR PROTEIN 7/ESTROGEN RECEPTOR COACTIVATOR-RELATED	7605950	7610287
Sb10g007228	PTHR23024	MEMBER OF 'GDXG' FAMILY OF LIPOLYTIC ENZYMES	7015912	7017107
Sb10g007370	PTHR13104	MED-6-RELATED	7201622	7205336
Sb10g007296	PTHR22893:SF13	12-OXOPHYTODIENOATE REDUCTASE OPR	7114879	7115084
Sb10g007460	PTHR12096	NUCLEAR PROTEIN SKIP-RELATED	7318537	7320021
Sb10g007700	PTHR11926	GLUCOSYL/GLUCURONOSYL TRANSFERASES	7515796	7517602

Table A-2 continued

Gene Name	Panther ID	Panther Description	Gene Start (bp)	Gene End (bp)
Sb10g007565			7416493	7420058
Sb10g007340	PTHR23258	SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE	7152271	7153897
Sb10g007760	PTHR10502	ANNEXIN	7611042	7614980
Sb10g007410	PTHR11071	CYCLOPHILIN	7261193	7267859
Sb10g007520			7367770	7368772
Sb10g007226	PTHR23024	MEMBER OF 'GDXG' FAMILY OF LIPOLYTIC ENZYMES	7013336	7015787
Sb10g007360	PTHR16012	KINESIN HEAVY CHAIN	7162730	7178914
Sb10g007270	PTHR11731	PROTEASE FAMILY S9B,C DIPEPTIDYL-PEPTIDASE IV-RELATED	7062751	7072010
Sb10g007260	PTHR13173	WW DOMAIN BINDING PROTEIN 4	7052023	7060540
Sb10g007300	PTHR22893	NADH OXIDOREDUCTASE-RELATED	7118677	7119912
Sb10g007240			7031242	7032584
Sb10g007450	PTHR10177	CYCLINS	7306391	7308094
Sb10g007550			7393308	7394337
Sb10g007190	PTHR21568:SF1	gb def: Hypothetical protein	6990492	6990830
Sb10g007600	PTHR11527	SMALL HEAT-SHOCK PROTEIN (HSP20) FAMILY	7433097	7433914
Sb10g007210			6999381	7003729
Sb10g007620	PTHR23365	POLY-A BINDING PROTEIN 2	7437934	7441972
Sb10g007640	PTHR10483	PENTATRICOPEPTIDE REPEAT-CONTAINING PROTEIN	7447676	7449764
Sb10g007570	PTHR11527	SMALL HEAT-SHOCK PROTEIN (HSP20) FAMILY	7421552	7422367
Sb10g007440	PTHR10483	PENTATRICOPEPTIDE REPEAT-CONTAINING PROTEIN	7288988	7290966
Sb10g007280	PTHR11731	PROTEASE FAMILY S9B,C DIPEPTIDYL-PEPTIDASE IV-RELATED	7073164	7080081
Sb10g007350	PTHR10483	PENTATRICOPEPTIDE REPEAT-CONTAINING PROTEIN	7155421	7158740
Sb10g007222	PTHR23024:SF10	CARBOXYLESTERASE-RELATED	7005759	7006985
Sb10g007224	PTHR23024	MEMBER OF 'GDXG' FAMILY OF LIPOLYTIC ENZYMES	7010113	7011756
Sb10g007680	PTHR12052	MITOSIS PROTEIN DIM1	7494163	7500919
Sb10g007540	PTHR11630	DNA REPLICATION LICENSING FACTOR	7379034	7387662
Sb10g007330	PTHR22893	NADH OXIDOREDUCTASE-RELATED	7136789	7138351
Sb10g007305	PTHR22893:SF13	12-OXOPHYTODIENOATE REDUCTASE OPR	7122073	7122309
Sb10g007390	PTHR23002	ZINC FINGER CCHC DOMAIN CONTAINING PROTEIN	7220616	7223493
Sb10g007480	PTHR11821	DNAJ/HSP40	7330581	7340900
Sb10g007610	PTHR19139	AQUAPORIN TRANSPORTER	7434249	7435606
Sb10g007670			7475847	7476916
Sb10g007650	PTHR10012	PHOSPHOTYROSYL PHOSPHATASE ACTIVATOR	7466819	7470425
Sb10g007630	PTHR11711	ARF-RELATED	7442221	7443165
Sb10g007510	PTHR10026	CYCLIN	7358042	7361309
Sb10g007430			7288025	7288598
Sb10g007220			7004707	7005738
Sb10g007280	PTHR11731	PROTEASE FAMILY S9B,C DIPEPTIDYL-PEPTIDASE IV-RELATED	7073164	7079946
Sb10g007730			7559672	7559932
Sb10g007690			7506195	7507541

Table A-2 continued

Gene Name	Panther ID	Panther Description	Gene Start (bp)	Gene End (bp)
Sb10g007250			7049746	7051082
Sb10g007740			7599254	7601351
Sb10g007750	PTHR23354	NUCLEOLAR PROTEIN 7/ESTROGEN RECEPTOR COACTIVATOR-RELATED	7605950	7610287
Sb10g007530	PTHR23125	F-BOX/LEUCINE RICH REPEAT PROTEIN	7369459	7372127
Sb10g007580	PTHR11527	SMALL HEAT-SHOCK PROTEIN (HSP20) FAMILY	7425748	7426500
Sb10g007230	PTHR22915	NADH DEHYDROGENASE-RELATED	7019500	7026365
Sb10g007470			7325563	7328559
Sb10g007490	PTHR22765	RING FINGER AND PROTEASE ASSOCIATED DOMAIN-CONTAINING	7346628	7348031
Sb10g007725	PTHR11702:SF4	GTP-BINDING PROTEIN YLF2-RELATED	7531976	7540394
Sb10g007320	PTHR22893	NADH OXIDOREDUCTASE-RELATED	7130453	7131702
Sb10g007380	PTHR11945	MADS BOX PROTEIN	7213022	7220337
Sb10g007293			7111013	7111890

Table A-3: Lignin standard curve used for calibration of acetyl bromide lignin quantification as described in Section 2

Lignin Concentration (g/L)	0.060	0.040	0.020	0.010	0.005
Absorbance @ 280nm	1.334	0.682	0.497	0.202	0.134
Absorbance @ 280nm	1.400	0.827	0.554	0.209	0.083
Absorbance @ 280nm	1.302	0.871	0.483	0.357	0.144
Absorbance @ 280nm	1.396	0.864	0.457	0.242	0.223
Absorbance @ 280nm	1.405	0.793	0.467	0.265	0.105
Absorbance @ 280nm	1.068	0.829	0.595	0.214	0.133
Absorbance @ 280nm	1.024	0.835	0.490	0.144	0.072
Absorbance @ 280nm	1.327	0.798	0.400	0.214	0.320
Average Absorbance	1.282	0.812	0.493	0.231	0.152

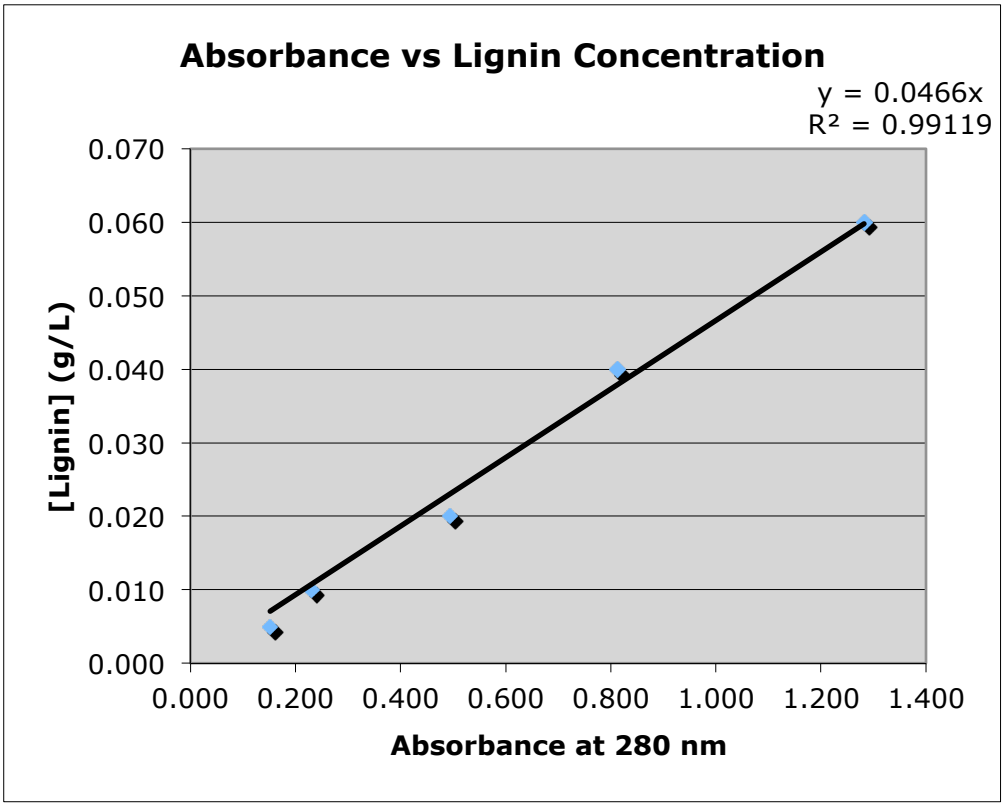


Figure A-3: Plot and regression of lignin UV absorbance as described in Section 2 and Appendix Table 6

Table A-4: 604 genotyping markers described in Section 5 are included in a separate file.

Table A-5: NIR Compositional data for SC56xTx7000 RILs harvested in 2009 is included in a separate file.

Table A-6: NIR Compositional data for BTx642xTx7000 RILs harvested in 2009 is included as a separate file.