ANALYSIS OF BIOMASS COMPOSITION IN A SORGHUM DIVERSITY PANEL

by

Patrick K. Sweet

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Agronomy

West Lafayette, Indiana

December 2018

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Cliff Weil, Chair

Agronomy Department

Dr. Mitch Tuinstra

Agronomy Department

Dr. Clint Chapple

Biochemistry Department

Dr. Min Zhang

Statistics Department

Approved by:

Dr. Ronald Turco

Head of the Graduate Program

ACKNOWLEDGEMENTS

The education and growth I have experienced while at Purdue was made possible through the support and friendship of several people. I am thankful to my committee members: Dr. Chapple, Dr. Tuinstra, Dr. Zhang and my advisor Dr. Cliff Weil for taking me on as his student back in May 2016. The chance he took on me helped narrow my interests from general plant sciences to crop genetics. I am grateful for his patience as I adapted to new schools of thought and new social norms. I am also very grateful for granting me freedom to explore scientific avenues and to make mistake (and learn from them).

This project involved a lot of hands on work from a lot of people and could not have been accomplished without them. I thank David Schlueter and Jacquee Anderson for their contributions as technicians. David orchestrated and single handedly processed thousands of Sorghum stalks for grinding. Jacquee was very helpful around the lab and I counted on her for ordering lots of materials. Furthermore, I am grateful to a great many undergraduates who provided their assistance throughout this project; from fieldwork to lab work and everything in between.

There are at least three organizations that made this project possible. First, and foremost I am grateful to the US Department of Energy for financially supporting this research. Second, I thank the National Renewal Energy Laboratory for providing compositional analysis on thousands of Sorghum powder samples and for their technical responses to a flurry of questions. Third, I learned a lot and saved a lot of time and frustration because of the assistance of Purdue research computing (RCAC). I am grateful for many email dialogues and coffee hour consultations.

Last and certainly not least, I am indebted to Dr. Addie Thompson for her mentorship. Addie introduced me to the world of R and spent many hours at my side. For her genuine care and friendship, I am very thankful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Author: Sweet, Patrick K. MS
Institution: Purdue University
Degree Received: December 2018
Title: Analysis of Biomass Composition in a Sorghum Diversity Panel
Committee Chair: Cliff Weil

Plant biomass is an abundant source of renewable energy, but the efficiency of its conversion into liquid fuels is low. One reason for this inefficiency is the recalcitrance of biomass to extraction and saccharification of cell wall polysaccharides. This recalcitrance is due to the complex and rigid structure of the plant cell wall. A better understanding of the genes effecting cell wall composition in bioenergy crops could improve feedstock quality and increase conversion efficiency. To identify genetic loci associated with biomass quality traits, we utilized genome-wide association studies (GWAS) in an 840-line *Sorghum* diversity panel. We identified several QTL from these GWAS including some for lignin composition and saccharification. Linkage disequilibrium (LD) analysis suggested that multiple polymorphisms are driving the association of SNPs within these QTL. Sequencing and further analysis led to the identification of a SNP within the coding region of a gene encoding phenylalanine ammonia-lyase (PAL) that creates a premature stop codon and co-segregates with an increase in the ratio of syringyl (S) to guaiacyl (G) lignin. A comparison of net PAL activity between lines with and without the mutation revealed that this mutation results in decreased PAL activity.

CHAPTER 1. LITERATURE REVIEW

In recent years there have been growing social and political demands to increase production of renewable energies to reduce our dependence on fossil fuels. Although there is no federal mandate for renewable energy production, several states have passed such legislation. Some of the mandates for renewable energy production includes legislation in New York and New Jersey for 50% renewables by 2030 (New York S5549C; New Jersey, P.L. 2018, c. 17), Vermont for 75% by 2032 (30 V.S.A. § 8005), Hawaii for 100% by 2045 (Hawaii Renewable Energy Initiative), and California for 100% by 2045 (California SB 100, 2018). One class of renewable energies is fuels. Renewable fuels encapsulate a variety of combustible products that can be produced from several biomass feedstocks. These biomass feedstocks include plant material, algae, animal byproducts, and an assortment of waste materials. Plant materials used as renewable feedstocks include crop residues and dedicated bioenergy crops. The most prominent biofuel in the United States is ethanol derived from corn starch. Unfortunately, food security is also a pressing issue and corn starch is a major portion of caloric intake for humans and livestock. As an alternative to using the edible, starchy portion of a plant, the leaf and stem material, termed lignocellulosic biomass, also holds great potential as a biofuel feedstock.

Lignocellulosic biomass can be converted into a variety of liquid biofuel compounds. The efficiency and yield of a specific bioconversion method are affected by the composition of the starting material. There are three general categories of methods used to convert lignocellulosic biomass into liquid biofuels: thermal, chemical, and biological. A refinery may utilize a combination of two or all three of these strategies to improve bioconversion efficiency.

One thermochemical conversion method is syngas generation (McKendry, 2002) followed by Fischer-Tropsch synthesis (FTS) (Ail and Dasappa, 2016). Pure syngas is composed of CO and H_2. The process of converting biomass to syngas can be explained in four steps; Drying, pyrolysis, gas-solid reactions, and gas-phase reactions (Bain and Broer, 2011). After water is removed in the drying step, pyrolysis involves the thermal decomposition of biomass in the absence of oxygen, typically at 400-500°C. This step produces permanent and condensable vapors as well as char. Following pyrolysis, the gas-solid reactions and the gas-phase reactions occur simultaneously as oxygen and steam are admitted to the gasifier. Water vapor and molecular oxygen react with the char, further liberating carbon from solid to gas phase. The

primary gas-phase reactions are oxygenation and methanation of carbon monoxide through reactions with water vapor and molecular hydrogen, respectively. The final step in syngas generation is purification in which all molecules besides CO and H_2 are removed from the product.

Fischer-Tropsch Synthesis is the production of hydrocarbons and oxygenated hydrocarbons (e.g. alcohols, ketones, and aldehydes) from syngas using either a cobalt or iron catalyst (Dayton et al., 2011). Generally, low temperature (200-240°C) synthesis yield high molecular weight hydrocarbons, while high temperature (300-350°C) synthesis yield lower molecular weight hydrocarbons. FTS produces a wide range of mostly linear paraffins, olefins, and oxygenated species. Although product selectivity is impossible to completely control, variables such as temperature, pressure, H_2/CO ratio, and catalyst composition can influence product selectivity. FT derived fuels are attractive because they are free of sulfur and contain a lower content of aromatics and nitrogen, resulting in fewer harmful emissions than petroleum derived fuels (Dupain et al., 2006; van Vliet et al., 2009).

Another thermal conversion technique is fast pyrolysis. Like syngas generation, drying of biomass is required and fast pyrolysis reaction conditions are anaerobic. As the name implies fast pyrolysis is carried out more rapidly than the pyrolysis step in gasification and it is done at atmospheric pressure, resulting in a larger fraction of liquid product and a smaller fraction of permanent gasses. Temperatures for this process range from 300-600°C. The liquid products of fast pyrolysis, called bio-oil, primarily composed of mixed species oxygenated hydrocarbons including carbonyls, alcohols, heterocyclics, sugars, and aromatics (Evans and Milne, 1987; Anca-Couce, 2016).

A third method of thermal conversion, hydrothermal liquefaction (HTL, also known as direct liquefaction), is essentially pyrolysis in hot pressurized water (Behrendt et al., 2008; Elliott, 2011). The product of HTL, called bio-crude, is much different than the pyrolysis product. Bio-crude is more deoxygenated, less dense and more viscous than bio-oil (Elliott et al., 2015). This conversion technique is more optimal for aqueous biomass such as algae.

These three biomass liquefaction processes, fast-pyrolysis, hydrothermal liquefaction, and Fischer-Tropsch synthesis, produce liquid products that are not suitable as direct transportation fuels. They must undergo additional processing to improve the fuel quality (i.e. remove undesirable characteristics). Some undesirable characteristics include acidity, particulate

matter (char), high viscosity, nitrogen, low H:C ratio, odor, sulfur, phase separation, and low volatility (Elliott, 2007; Bridgwater, 2011). Depending on composition of the crude products and the desired end product, there are multiple processes that can be utilized for upgrading to suitable fuels including hydrotreating, hydrocracking, hydroisomerization, hydrodeoxygenation, and product separation.

A low temperature thermochemical method demonstrated the utility of a bimetallic catalyst to achieve high yield of product. Lignin from pulverized poplar biomass was converted with a yield of 95% to propylcyclohexane using methanol solvent, H_2 gas, and a Zn/Pd catalyst (Parsell, 2015). This process simultaneously converted lignin to a hydrocarbon fuel component, and, by removal of lignin, made the polysaccharide fraction of biomass more amenable to depolymerization.

As opposed to the lignin fraction, another study demonstrated a chemical method to convert pure glucose to liquid hydrocarbon fuels, specifically alkanes. This conversion method utilized a solid Pt/SiO2-AlO3 catalyst, methanol and water as solvents, and H_2 to convert glucose into C9 – C15 alkanes with a 90% yield (Huber et al., 2005). This study did not discuss the issue of cellulose depolymerization to glucose, but another group detailed a chemical method to produce high quality liquid hydrocarbon fuel from the polysaccharide fraction of whole biomass in high yield (Robinson et al., 1999). After polysaccharides are depolymerized the key step in this process was the use of hydroiodic acid to reduce polyhydric alcohols to hydrocarbons and subsequent removal and recycling of the halogen. As an economic advantage, no drying step was necessary for this process, therefore wet biomass could be used.

Biological conversion of lignocellulosic biomass refers to fermentation of the sugar fraction of the cell wall most commonly to ethanol. Although corn starch is widely used as the glucose source for fermentation, corn is also a food commodity, thus there exists competition between food security and energy security; an incentive to develop second-generation biofuels. Unlike corn starch, cellulose is not directly fermentable. Pretreatment of biomass is necessary to separate cellulose from lignin followed by a hydrolysis step to depolymerize cellulose into its monomeric glucose units. There are several different pretreatments that can be used including acid, alkali, and thermal treatments (Klem et al., 1998). Hydrolysis can be carried out either enzymatically or chemically. After pretreatment and hydrolysis glucose can be fermented into ethanol using yeast, or to higher energy density products such as butanol using engineered

microbes (Liu et al., 2015). Metabolically engineered microorganisms can also be used to produce non-fermentative products such as hydrocarbons (Schirmer et al., 2010), branched-chain alcohols (Atsumi et al., 2008), and acetoin (Liu et al., 2015), which can then be efficiently converted into C9 – C14 alkanes (Zhu et al., 2016). In addition to cellulose some of the sugars from hemicelluloses can also be used for biological conversion. Hemicelluloses are prone to acid hydrolysis, but due to their complex saccharide linkages are not particularly labile to enzymatic hydrolysis. In order to utilize more sugars from hemicellulose, recombinant microorganisms have been engineered to utilize arabinose and xylose in their fermentation pathways (Bettiga et al., 2009; Vilela et al., 2015).

Each of these strategies for the conversion of biomass to various biofuels has advantages and disadvantages. No single process for the production of second-generation biofuels is widely implemented due to economic and technical barriers. A key impactor of conversion efficiency is the quality of the biomass. Plants can be genetically modified, bred, and selected to produce more desirable feedstocks.

Sorghum is a crop of worldwide importance especially as a cereal and forage crop in Africa. In the United States, *Sorghum* has been proposed as a bioenergy feedstock and many of the same traits that are important for forage *Sorghum* are also important for bioenergy *Sorghum* so breeding for sorghum biomass quality can have double impact. Another attractive feature of *Sorghum* is the wide diversity within the species (*S. bicolor*) (Deu et al., 2006; Mace et al., 2013). Natural variation in *Sorghum* is evident by the wide range of morphologies represented across the species. Although *Sorghum* was not initially domesticated in the western hemisphere a large number of diverse lines have been converted to photoperiod insensitive varieties through the *Sorghum* conversion program (Stephens et al., 1967) which enables *Sorghum* breeding in temperate regions. Many varieties of sorghum exhibit several characteristics that makes it an attractive biofuel feedstock including its high drought tolerance (Bawazir and Idle, 1989; Amelework et al., 2015), its high biomass yield (Turhollow et al., 2010), its relatively low fertilizer requirement, and its similarity to maize allowing it to fit into existing agricultural systems. Sorghum is a monocot that is native to regions of Africa and Asia. Taxonomically, Sorghum is the name of the genus classified under the family Poaceae, tribe, Andropogoneae, and subtribe Sorghinae. There are three species under the *Sorghum* genus, *S. halepense*, *S. propinquum*, and *S. bicolor* (Sorghum: Origin, History, Technology, and Production, 2000). All

cultivated varieties are found in *S. bicolor*, under which are classified three subspecies, S. bicolor bicolor, S. bicolor drummondii, and S. bicolor verticilliflorum. Furthermore, S. bicolor bicolor, under which all cultivated varieties are found, is divided into 5 races, Bicolor, Guinea, Caudatum, Kafir, and Durra, based on panicle and spikelet morphology (Harlan and De Wet, 1972). The diversity of cultivated sorghum is incredible as about 40,000 accessions have been collected (Maunder, 1999). To organize this diversity, cultivated sorghum is generally categorized into three groups; grain sorghum, forage sorghum, and sweet sorghum.

Grain sorghum, as the name suggests, is bred as a food crop for human and livestock consumption. It has similar uses as corn but is grown at greater efficiency than corn in hot and dry climates. Grain sorghum is bred for traits that improve grain yield and ease of harvesting the seed head, resulting in plants that tend to have large grain heads and short stalks. Sweet sorghum is bred for biomass and simple sugar accumulation in the stalk. Large quantities of sucrose, fructose and other sugars accumulate in the stalks of sweet sorghum plants (Yuvraj et al., 2013) which is used to make syrup and is used to produce ethanol by fermentation. Forage sorghum is a drought tolerant alternative to corn primarily used for animal silage. The grain of forage sorghum is significantly reduced compared to grain sorghum and is a slightly lower quality silage crop than corn. Forage sorghum is the primary interest when breeding for the large quantities of biomass needed for lignocellulosic conversion to biofuels.

Since Sorghum is a crop of worldwide importance a variety of genetic resources have been developed for its improvement. Since the publication of the Sorghum reference genome, BTx623 (Paterson et al., 2009), whole genomic sequences of dozens more lines have become available (Paterson et al., 2009; Mace et al., 2013; McCormick et al., 2018). Diverse collections of Sorghum have also been analyzed by genotyping-by-sequencing (GBS), generating reduced representation genomic libraries for population analyses and genetic mapping (Morris et al., 2013; Thurber et al., 2013; Brenton et al., 2016). Another genetic resource in sorghum are ethyl methanesulfonate (EMS)-mutagenized populations developed for forward genetic analyses (Addo-Quaye et al., 2017). This study produced a database (https://www.purdue.edu/sorghumgenomics/) containing functionally annotated EMS induced SNPs from a large M4 population derived from the *Sorghum* line BTx623. These genetic resources are publicly available for researchers to use for the improvement of Sorghum germplasm.

The quality of *Sorghum* biomass depends on traits related to its composition. Biomass composition refers to the polymeric makeup of the cell wall. These polymers are synthesized and partially assembled within the apoplast and secreted outside of the cell membrane and into intercellular spaces. There, cell wall polymers contribute to structure and morphology, protect the cell from biotic and abiotic stresses, and facilitate water transport by providing support to vascular bundles. The biogenesis of the cell wall is a complex, highly regulated process. Knowledge of these biosynthetic and regulatory processes can serve as a baseline for identifying candidate genes that control them and forming hypotheses from association studies.

Lignin is a cell wall polymer composed of phenylalanine-derived monolignols. The primary monolignols in grasses are coniferyl alcohol and sinapyl alcohol which are referred to a guaiacyl (G) and syringyl (S) units when incorporated into the lignin polymer, but many other phenylpropanoid precursors are capable of being incorporated. The genetic and biochemical mechanisms of the phenylpropanoid pathway have been extensively described in *Arabidopsis* (Fraser and Chapple, 2011). Compared to *Arabidopsis*, orthologous lignin biosynthetic genes in *Sorghum* are duplicated (Xu et al., 2009). The ten phenylpropanoid enzymes analyzed in this study were encoded by 63 genes in *Arabidopsis* and 141 genes in *Sorghum*. Gene duplication makes the implementation of genetic modifications less translational across species, especially between distantly related species.

There have been many reports of modifications to phenylpropanoid genes that impact biomass composition in bioenergy plants. For example, in poplar, overexpression of ferulate 5-hydroxylase (F5H) increases S/G ratio (Stewart et al., 2009) while downregulation of a 4-coumarate:coenzyme A ligase (4CL) gene increased growth and cellulose content (Hu et al., 1999). In bioenergy grasses (i.e. maize, *Sorghum*, and switchgrass), several lignin modified mutants exhibit a brown midrib (bmr) phenotype. *Sorghum* bmr mutants improve sugar conversion and ethanol yield (Dien et al., 2009) and mutations have been identified at three loci, *bmr2*, *bmr6*, and *bmr12*. The bmr2 phenotype was mapped to a gene encoding 4CL (Saballos et al., 2012). Mutations in a gene encoding cinnamyl alcohol dehydrogenase are responsible for the bmr6 phenotype (Saballos et al., 2009; Sattler et al., 2009). Lastly, mutations in a gene encoding caffeic acid O-methyltransferase are responsible for the bmr12 phenotype (Bout and Vermerris, 2003; E. Sattler, 2012).

Generally, total lignin content is negatively correlated with saccharification efficiency of biomass due to its cross linking with cell wall polysaccharides. Cinnamyl alcohol dehydrogenase modifications leading to reduced lignin content improved saccharification efficiency and ethanol yield in *Brachypodium*, switchgrass, and maize (Fu et al., 2011; Fornale et al., 2012; Poovaiah et al., 2014). However, the impact of the monomer composition of that lignin on biomass saccharification is much less clear. Some studies have reported correlations between lignin composition and saccharification efficiency (Fontaine et al., 2003; Davison et al., 2006; Corredor et al., 2009; Studer et al., 2011) while others found no relationship between the two (Reddy et al., 2005; Chen and Dixon, 2007). The nature and extent of pretreatment also effects this relationship (Li et al., 2016). While increased S/G ratio had no impact on hydrolysis of untreated biomass, it increased hydrolyzability of pretreated biomass (Li et al., 2010; Studer et al., 2011; Mansfield et al., 2012).

In addition to lignin content and composition, polysaccharide composition is an important trait to consider for improvement of biomass quality. Polysaccharides of plant secondary cell walls are generally categorized as cellulose or hemicellulose. Cellulose is a crystalline array of glucan chains synthesized by a multi-subunit cellulose synthase complex (CES) at the plasma membrane (Somerville, 2006). Hemicelluloses are a broader class of heterogeneous polysaccharides that vary among species. In most dicots the primary hemicellulose is xyloglucan (Scheller and Ulvskov, 2010). Alternatively, in many commelinid monocots, such as maize and *Sorghum*, xyloglucan is only a minor component of the hemicellulose, whereas arabinoxylan is much more abundant (Wilkie, 1979). Several glycosyltransferases are involved in the synthesis of hemicelluloses within the lumen of the Golgi apparatus (Pauly et al., 2013). Still, many of the genes involved in hemicellulose biosynthesis remain unknown. Many cell wall related genes have been annotated in *Sorghum* by comparative analysis of functional domains in predicted ptotein sequences. For example, one study identified 11 cellulose synthase genes (CESA) and 104 genes involved in hemicellulose biosynthesis (Rai et al., 2016).

Genes involved in plant secondary cell wall biosynthesis are controlled by a hierarchy of transcription factors (Nakano et al., 2015). This regulatory network largely includes transcription factors with NAC, MYB, and WRKY DNA-binding domains. MYB transcription factors are negative regulators of lignin biosynthesis (Tak et al., 2017) and overexpression of PvMYB4 increases ethanol yield in switchgrass 2.6-fold (Shen et al., 2013). Several common cis-binding

elements have been identified in the promoters of secondary cell wall biosynthetic genes of several species, although there is slight variation in these motifs. These cis-elements include several AC-rich elements in parsley, tobacco, and *Arabidopsis* (Lois et al., 1989; Hatton et al., 1995; Raes et al., 2003), MYB responsive elements in *Arabidopsis* (Kim et al., 2012; Zhong and Ye, 2012), the W box motif in *Medicago truncatula* and *Arabidopsis* (Wang et al., 2010), and NAC binding elements in *Arabidopsis* (Zhong et al., 2010). In both dicots and monocots NAC transcription factors are master regulators of cell wall biosynthesis. Specifically, SECONDARY WALL-ASSOCIATED NAC DOMAIN1 (SND1) is an activator of secondary cell wall biosynthesis. Ectopic expression of SND1 in *Arabidopsis* results in dwarfed plants with very thick secondary cell walls (Zhong et al., 2006) and this mutation is complemented by expression of SND1 from *Miscanthus* (Golfier et al., 2017). Conversely, the transcription factor WRKY12 is a global repressor of secondary cell wall biosynthesis (Wang et al., 2010).

Alone, knowledge of genes involved in cell wall biogenesis is not particularly useful to breeders. However, the identification (or induction) of specific polymorphisms in these genes that have a measurable effect on a bioenergy trait is of great value. Such genetic polymorphisms can be discovered via association studies and used in marker-assisted selection to expedite the breeding process.

With the advent of large quantities of genotypic data generated via Next Generation sequencing (NGS) or SNP chips, the limitation of association studies and breeding programs generally has become the generation of phenotypic data. Due to the laborious and expensive nature of phenotyping, a researcher must often make compromises between sample number and accuracy. Biomass composition traits are particularly laborious to measure. For example, there are a variety of procedures used for direct quantification of lignin. Despite theoretically quantifying the same molecule, these methods provide inherently different measures of lignin content. Three routine methods, Klason lignin, acetyl bromide lignin, and acid soluble lignin, highlight this discrepancy. Klason lignin, which is known to overestimate lignin content in grasses (Hatfield and Fukushima, 2005), is a gravimetric technique that weighs lignin after solubilizing and washing away polysaccharides from the cell wall (Browning, 1967). Acid soluble lignin is determined by quantifying the UV absorption from the portion of lignin that is solubilized by sulfuric acid during the Klason lignin procedure (Hatfield and Fukushima, 2005).

The acetyl bromide method quantifies the UV absorbance of dissolved lignin at 280 nm (Johnson et al., 1961). All three methods are so time consuming that experiments involving large numbers of samples can become impractical.

Indirect quantification can provide higher throughput but sacrifices some accuracy. This thesis relies on the prediction of lignin content and composition by pyrolysis molecular beam mass spectrometry (PyMBMS). Developed by the National Renewable Energy Lab (NREL), PyMBMS provides estimations of Klason lignin and lignin monomer composition (S/G ratio) by quantifying lignin-derived pyrolysates from thermally degraded biomass (Penning et al., 2014). Another method for indirect quantification of biomass composition is prediction using light spectroscopy. The advantage of this method is that it is nondestructive, so traits can be measured on the same plant at multiple stages during its growth cycle. There have been some advances on predicting biomass composition with light spectroscopy. For example, using a combination of Raman, near, and mid-infrared spectroscopy S/G ratio could be predicted with an R^2 of 0.62-0.83 in eucalypt trees (Lupoi et al., 2014). These high throughput methods for genotyping and phenotyping are necessary to facilitate association studies.

Genome wide association studies exploit the natural variation within a diverse germplasm to identify DNA polymorphisms that are associated (cosegregated) with a trait of interest. Although some polymorphisms have a biological impact on a trait, many SNPs become associated with a trait due to their high linkage disequilibrium (LD) with biologically relevant polymorphisms. GWAS has proven useful for identifying genetic loci associated with many quantitative traits in many different species. It has been used to discover QTL for traits including yield components (Wang et al., 2017), disease susceptibility (Bartoli and Roux, 2017), and morphological characteristics (Liu et al., 2015). In *Sorghum* implementation of GWAS in diversity panels has uncovered QTL for traits such as panicle architecture, plant height (Morris et al., 2013), grain yield components (Boyles et al., 2016), and photosynthesis-related traits (Ortiz et al., 2017).

GWAS models the relationship between genotype and phenotype. Due to the confounding effect of population structure, simple linear models cannot be used to draw these associations. Genetic relatedness should be accounted for to avoid spurious associations that can arise from ‘unlinked LD’ within subpopulations. The GAPIT software accounts for genetic relatedness by including kinship as a random effect in the GWAS model (Lipka et al., 2012).

GWAS relies on the linkage disequilibrium (LD) between markers and causative polymorphisms and it identifies statistical associations between SNPs and traits. The magnitudes of these associations are heavily dependent on sample size. Minor alleles can only be detected in very large populations. Therefore, a limitation of GWAS is that allele frequency affects the ability to detect even large effect polymorphisms.

To facilitate large genomic studies, a large amount of highly accurate genotyping data needs to be generated. One cost effective way to generate such data is genotyping by sequencing (GBS) (Elshire et al., 2011). An advantage of GBS is that by ligating barcoded adapters to fragments of genomic DNA, many samples can be sequenced in the same Illumina sequencing flow cell. These genomic DNA fragments are generated by digesting genomic DNA with a type I restriction enzyme. Thus, GBS generates reduced representation libraries because the marker density for each sample depends on the frequency of a restriction site. For this reason, a disadvantage of GBS is that such library construction results in lower resolution of QTL. Another undesirable consequence of multiplexing is that each genotype ends up with a significant amount of missing SNP calls. A partial solution to this problem is imputation which uses calls from linked markers to predict missing SNPs.

The work of this thesis, funded by the Department of Energy, takes a genomics approach to study *Sorghum* as a potential bioenergy crop. Discovering genetic loci associated with biomass composition could help improve biomass quality and the feasibility of cellulosic biofuels.

CHAPTER 2: ANALYSIS OF SORGHUM BIOMASS COMPOSITION

Introduction

There is an ever-increasing demand for renewable fuels as the issues surrounding climate change and energy dependence continue to grow. Ethanol is a commonly used supplement in gasoline to improve gasoline quality and reduce our reliance on fossil fuels. In the U.S., ethanol is primarily the product of fermenting sugars derived from corn starch, but this practice puts additional pressure on the food supply, driving up the price of food commodities. However, there are alternative feedstocks. Sugars are not only found in the starchy tissue of plants but are also ubiquitous in the cellulosic tissue. Cellulose is part of a complex polymeric network within the plant cell wall. A crystalline polymer of glucose, cellulose is the most abundant biopolymer on the planet; however, it is not readily fermentable. Prior to fermentation, cellulose must be extracted from plant tissue and hydrolyzed to glucose. The process is expensive, energy intensive, and can involve harsh chemical pretreatment (Chaturvedi and Verma, 2013). The poor efficiency of this process has hindered the ethanol industry from utilizing cellulosic material as a sugar feedstock (Cai et al., 2013). Part of the solution may lie in genetic modification and plant breeding to optimize biomass for efficient conversion to fermentable sugars.

In addition to cellulose, the plant cell wall is composed primarily of various hemicelluloses and lignin. The abundance and composition of the polymers varies among species and affects the recalcitrance of biomass to extraction and saccharification of cell wall polysaccharides. Confounding results among studies relating biomass composition to saccharification efficiency suggest the interaction among these components plays a large role in biomass recalcitrance (Sorek et al., 2014).

Genome-wide association study is a tool for identifying genetic polymorphisms associated with variation in a trait of interest. Such polymorphisms can be used by plant breeders for marker-assisted selection to select improved varieties efficiently. Mapping biomass composition traits using GWAS has been limited due to the need for large sample size combined with the laborious nature of measuring such traits. Nevertheless, polymorphisms associated with bioenergy traits in several plants, such as poplar, barley straw, and maize stover have been identified by GWAS (Penning et al., 2014; Fahrenkrog et al., 2017; Naz et al., 2017).

Sorghum is a perennial C_4 cereal crop with potential to be an important bioenergy crop. Features such as its high nitrogen use efficiency (Muchow, 1998), drought tolerance (Amelework et al., 2015), abundant biomass (Brenton et al., 2016), and extensive diversity (Deu et al., 2006; Morris et al., 2013) make it an attractive bioenergy model crop. There remains great potential in *Sorghum* to understand and improve biomass quality for utilization as a bioenergy feedstock.

Materials and Methods

Phenotyping and data analysis

A *Sorghum* diversity panel (840 lines) was grown at the Agronomy Center for Research and Education, West Lafayette, IN during the 2015 and 2016 summers. Two stalks from each genotype were harvested and dried in a forced air dryer at 50 °C until completely dry. For each genotype the stalk samples were pooled and ground to pass through a 0.5 mm screen. Biomass composition traits were measured at the National Renewable Energy Lab (NREL) using pyrolysis molecular beam mass spectrometry (PyMBMS) (Penning et al., 2014), high throughput compositional analysis (Sluiter et al., 2012), and high throughput pretreatment and enzymatic hydrolysis (Resch et al., 2015).

PyMBMS is a high throughput method for determining lignin content and composition (Penning et al., 2014). After mean normalization of each mass spectrum, mass fragments attributed to syringyl (154, 167, 168, 182, 194, 208, and 210) and guaiacyl (124, 137, 138, 150, 164, and 178) derivatives were summed to quantify each lignin monomer. Lignin content (Klason lignin) was quantified by summing syringyl units, guaiacyl units, mz_120, mz_152, and mz_181 (Table 1) and corrected to a known quantity of lignin (19.2 %) in switchgrass standards. Next, since instrument fluctuations between the years was evident (Figure 1) each trait was adjusted to normalize the differences in the switchgrass standards using the 2015 trait values as a baseline. Lastly, to account for the effects of confounding variables, best linear unbiased predictors (BLUPs) were calculated for each trait using the lme4 package (Bates et al., 2015) in R. Tray number, vial number, and growth environment (year) were included as random effects in a mixed linear model.

NREL's two-stage acid hydrolysis for compositional analysis (Sluiter et al., 2012) was used to quantify the amount of structural glucose and xylose released from *Sorghum* biomass after acid hydrolysis. Briefly, this process involves extraction of soluble sugars followed by a

two-step hydrolysis in sulfuric acid. Following hydrolysis, the free sugars were spectroscopically quantified by the absorbance of the hydrolyzed sample at 510 and 340 nm for glucose and xylose, respectively. BLUPs were also calculated for these traits by fitting environment (year), plate, well, and time as random effects in a linear model. The amount of glucan and xylan was inferred by dividing these values by the mole fraction of water removed when going from sugar to polysaccharide (1.11 and 1.14), respectively.

The amount of glucose and xylose released after hot water pretreatment (180 °C, 17.5 min) and hydrolysis with cellulases (72 h, overload enzyme) was also measured by NREL. Only a single year of data was collected for these traits since data from the 2016 season was not available due to technical difficulties. BLUPs were calculated for sugar composition and release from the 2015 season by adjusting for plate as a random effect in a linear model. Sugar yield was calculated by dividing the amount of sugar released by amount of sugar from compositional analysis.

Reflectance Spectroscopy for Prediction of Lignin Composition

Part of each ground *Sorghum* stalk that was sent to NREL from the 2015 set for compositional analysis were reserved for analysis by reflectance spectroscopy. Each sample was poured into a thin cup and compressed with the hand clamp attachment of a HR1024i spectrometer. Two spectra were captured in reflectance mode for each sample, turning the sample 180° between scans. Each spectral scan collected 976 bands between 350 and 2500 nm and spectra were interpolated to single nanometer resolution. A previously published R-script was used for implementation of partial least squares regression (Couture et al., 2016). For implementation of this method the lignin monomer composition of each sample, as determined by PyMBMS, was used as the response vector. The predictor matrix was composed of every 10^{th} band from the interpolated reflectance spectra within the range of 1000 nm and 2500 nm (i.e. 1000, 1010, 1020, etc.). The optimal number of components, from one to 15 components, was determined by a leave-one-out cross-validation procedure through reduction of the predicted residual sum of squares (PRESS).

Genotyping

Genotyping-by-sequencing (GBS) data aligned to the *Sorghum* reference genome (Phytozome version 3.0) and imputed with the Beagle software (Browning and Browning, 2016) was supplied by Dr. Patrick Brown (University of Illinois). There were a total of 107421 imputed SNPs across 839 genotypes (one genotype was missing GBS data) in the unfiltered genotype matrix.

GWAS

Prior to conducting association studies, the genotypes were filtered to exclude SNPs with a minor allele frequency below 0.025. The resulting genotype matrix contained 80090 SNPs across 838 genotypes (one genotype had no phenotypic data). For each trait, BLUPs were calculated using a linear model to adjust for the year effect. GWAS was conducted using a compressed mixed linear model with the GAPIT software (Lipka et al., 2012). FDR adjusted p-values were used to create Manhattan plots using the qqman package in R (Turner, 2014). Genome-wide significance thresholds were set at 0.01 (False positive rate).

DNA extraction

The *Sorghum* Diversity Panel (840 lines) were grown for ~1 week in the greenhouse. One to two seedlings per genotype were harvested on silica beads and dried at 50 °C overnight. Dry leaves were transferred to a 2 mL tube, two stainless steel beads were added to each tube, and the tissue was shaken in a TissueLyser II (Qiagen) at 30 s^{-1} for 4 min (or until tissue was powdered). Powdered tissue was frozen in a -80 °C freezer for ≥ 1 hour. DNA was extracted by a modified CTAB extraction protocol as follows: 750 µL extraction buffer (1.2 mM NaCl, 100mM Tris HCl pH 8, 20 mM EDTA, 2% w/v CTAB, 0.1% BME) was added to the frozen tissue and the samples were vortexed for 30 s. The homogenized samples were incubated in a 60 °C water bath for 1 h and inverted every 15 min. Samples were incubated at room temperature for 10 min then 750 µL 24:1 chloroform: isoamyl alcohol was added to each tube and samples were mixed by inversion. Samples were centrifuged at 3250 x g for 25 min and the aqueous phase was transferred to a new 2 mL tube. 1 mL of dilution buffer (100 mM Tris-HCl pH 8, 20 mM EDTA, 2% CTAB w/v) was added to each tube and samples were mixed by inversion. Samples were incubated in a 60 °C water bath for 30 min, centrifuged at 5000 x g for 10 min, and the

supernatant was discarded. The DNA-CTAB pellet was resuspended in 0.5 mL wash buffer (7:3 TE buffer: ethanol), incubated at room temperature for 15 min, and centrifuged at 5000 x g for 10 min. After discarding the supernatant, the pellet was resuspended in 0.5 mL of high salt TE buffer (1M NaCl, 10mM TRIS-HCl pH, 2 mM EDTA, 50 µg/mL) and incubated in a 60 °C water bath for 15 min. The solution was then diluted with 250 µL of water and 1.2 mL ethanol and chilled at -20 °C overnight. The next day the solution was centrifuged at 15,000 rpm for 10 min and the supernatant was discarded. The DNA pellet was washed with 500 µL of cold 70% ethanol, centrifuged at top speed for 5 min, the supernatant was discarded, and the samples were dried by leaving the tubes open in a fume hood for several hours to overnight. Lastly, the purified DNA was dissolved in 50 µL of water.

Genotyping of PAL alleles

A polymorphism inducing a stop codon in gene Sobic.006G148800 at position 51041582 on chromosome 6 also creates a *Bfa*I restriction site (CTAG). The SNP was genotyped in the diversity panel by PCR amplifying a segment of the gene, then using the restriction enzyme to distinguish between the two alleles (Figure 2). Each PCR contained 2 µL gDNA, 1 µL of each 10 µM primer (GCATCTCAATGCCGGAATCTT, GTGTACTCAGGCTTGCCGTT), 10 µL DreamTaq master mix (Thermo), and 6 µL water. Initial denaturation was 95°C for 4 min. The cycling program was 40 cycles of 95°C for 30 s, 62.2°C for 30 s, and 72°C for 1 min. Final extension was 72°C for 5 min. Unpurified PCR products were used for *Bfa*I digestion reactions. Each digestion reaction contained 1.5 µL PCR product, 2.5 µL CutSmart buffer (NEB), 1 µL *Bfa*I, and 20 µL water. Control reactions were identical, but with water replacing *Bfa*I. Each reaction was incubated at 37°C for 3 h. Digested reactions were mixed with 2 µL of 10X loading dye and 15 µL was loaded into a 1 % agarose gel containing GelRed (Biotium). Gels were electrophoresed for 1 h at 90 V (Figure 3).

PAL activity assays

Plants for enzyme activity assays were grown in the greenhouse with 2 h of supplemental light in the morning and another 2 h in the evening. Plants were harvested at first sign of boot stage (i.e. when the sheath surrounding the panicle began to unfold). For harvest, leaf sheaths were removed and internodes 2 and 3 from the top were collected. Each internode was cut in half lengthwise, flash frozen in liquid nitrogen, and stored in a -80°C freezer. Each half internode was

finely ground in liquid nitrogen with a mortar and pestle. 0.25-0.35 g of ground tissue was transferred to a 2 mL microfuge tube and again stored at -80°C until analysis.

Protein was extracted from each tube by adding 1 mL of chilled 100 mM sodium phosphate buffer (2 mM EDTA, 2% w/v polyvinylpyrolidone, 4 mM dithiothreitol, pH 6) to each tube of ground tissue. Samples were gently vortexed every 30 s for 5-6 min (placing tubes in ice between cycles) until all frozen tissue was homogenized. Samples were centrifuged at 21,130 x g for 25 min at 4°C. Protein was quantified in each sample using the Bradford method (Bradford, 1976).

Enzyme activity assays were carried out in triplicate on UV-transparent 96-well plates. Each 200 µL reaction consisted of 50 µL dH_2O, 90 µL sodium borate buffer (10 mM), 10 µL protein extract, and 50 µL substrate (10 mM phenylalanine or 5 mM tyrosine in sodium borate buffer). Control reactions were identical except contained no substrate. Reaction absorbances at 271 and 310 nm were monitored at 13 time points over an hour in a Biotech Epoch plate reader at 30°C.

Results

Trait Correlations

Stalk samples were analyzed for variation to better understand the genes that effect *Sorghum* biomass composition. PyMBMS analysis was used to identify differences in the relative amount of S to G lignin. For quality control NREL includes switchgrass standards with its PyMBMS analysis. There were significant differences in the standards for lignin content and composition between the 2015 and 2016 analysis (Figure 1). These differences were attributed to instrumental fluctuations and adjusted for prior to downstream analyses.

Phenotypic relationships suggest that lignin content and composition do not greatly affect saccharification in this population. Across all biomass samples analyzed, there were positive correlations among the theoretical lignin pyrolysates derived from S units (Figure 4) as well as among the pyrolysates derived from G units (Figure 5). This supported the notion that these pyrolysates were derived from the respective monolignol. There was little to no relationship between either lignin trait and sugar release or yield (Figure 6), nor was there a strong relationship between lignin content and composition (Figure 6). However, there were positive correlations between glucose release/yield and xylose release/yield (Figure 6).

Prediction of Lignin Composition from Light Spectroscopy

An attempt was made to build a model to predict lignin composition from NIR and SWIR reflectance spectra of ground *Sorghum* biomass. The first part of the model building procedure was variable selection by leave-one-out cross validation through reduction of the PRESS statistic (Figure 7). The PRESS was lowest in models with a single component. On average across 1000 simulations, the first component explained 73% of the variation in the predictor variables while explaining only 0.1% of the response variable, S/G ratio. Since there was such a low level of prediction, no latent variable model was constructed.

GWAS and Candidate Genes

Genome-wide association studies revealed several QTL for bioenergy traits; three for S/G ratio (Figure 8), one shared by glucose release (Figure 9), xylose release (Figure 10), and glucose yield (Figure 11), and one shared by two lignin pyrolysates (mz_167 and mz_168; Figures 12-13).

The most prominent association peak for S/G ratio was located on chromosome 6 with the peak GBS SNP located at position 51072215 (Figure 8). This SNP was also associated with the abundance of several lignin pyrolysates (mz_167, Figure 12; mz_168, Figure 13; mz_154, Figure 14; mz_181, Figure 15). Analysis of this locus revealed tight linkage with a tandem repeat of genes encoding phenylalanine ammonia-lyase (PAL; Sobic.006G148800 and Sobic.006G148900; Figure 16). It was hypothesized that variation in one of these genes is driving the association of the linked SNPs with S/G ratio because PAL is the first enzyme in the phenylpropanoid pathway and previous studies in *Arabidopsis* and tobacco have reported altered lignin composition in *pal* mutants (Sewalt et al., 1997; Rohde et al., 2004). Sanger sequencing of these PAL genes in four lines high for S/G ratio and four lines low for S/G ratio revealed cosegretation of a SNP within the coding sequence of Sobic.006G148800 with the trait (Figure 2). The allele of this SNP in the lines high for S/G ratio induces a premature stop codon. Genotyping of this SNP in the entire diversity panel (840 lines) revealed a significant association (p-value = 4.8×10^{-26}) with S/G ratio (Figure 17). The mutant stop codon allele increases the mean S/G ratio by 0.05 in the *Sorghum* diversity panel (Figure 17), though not all lines carrying the allele have high S/G lignin ratios. When this SNP was included in the genotypic dataset and GWAS was rerun for each trait, the SNP was even more strongly associated with S/G ratio

(Figure 8) and abundance of several lignin pyrolysates (mz_167, Figure 12; mz_168, Figure 13; mz_154, Figure 14; mz_181, Figure 15). Furthermore, this PAL mutation was significantly associated with two additional S-pyrolysates (mz_182, Figure 18; mz_194, Figure 19).

To study the biological impact of this PAL mutation, PAL activity was measured in the same eight lines used in cosegregation analysis. As there are seven copies of the PAL gene is sorghum the hypothesis was that this nonsense mutation reduces net PAL activity. It has been reported that PALs in grasses are bifunctional enzymes that also exhibit tyrosine ammonia lyase (TAL) activity (Rosler et al., 1997; Barros et al., 2016). Therefore, both PAL and TAL activity was measured in each plant. This bifunctionality (PTAL activity) in *Sorghum* was supported by the positive linear relationship between PAL and TAL activity in the samples tested (Figure 20). These enzyme activity assays showed on average a decreased PTAL activity in the lines carrying the nonsense allele of the PAL SNP, but the significance was marginal (p-value = 0.08) (Figure 21).

A second association peak for S/G ratio was found on chromosome 9 (Figure 8). Analysis of the linkage disequilibrium among the five significant SNPs in this peak suggested that it contains two independent QTL (Figure 22). One of these chromosome 9 QTL peaks at position 56642557. Bioinformatic analysis (NCBI-BLAST) of the genes in this region suggested that a possible candidate gene driving the association of the linked SNPs encodes a cinnamate glucosyltransferase (Sobic.009G224100) which uses trans-cinnamate, an intermediate in the phenylpropanoid pathway, as a substrate and conjugates it to glucose. This gene resides 10.4 kb upstream from the peak associated SNP (Figure 23).

The second QTL on chromosome 9 for S/G ratio, which peaks at position 57212539, is also associated with glucose release, xylose release and glucose yield (Figures 9-11). Analysis of the genes at this locus show tight linkage of a tandem repeat of genes encoding fasciclin-like arabinogalactan proteins (FLAs; Figure 24). These genes were hypothesized as candidates for driving the association of the linked SNPs since previous research has shown altered sugar content in the stems of Arabidopsis FLA mutants (MacMillan et al., 2010). Similar to the cosegregation analysis for the PAL candidate genes, Sanger sequencing of the coding sequences of these FLA genes was used to look for polymorphisms that cosegregated with the trait in four lines high for glucose release and four lines low for glucose release; however, no coding

sequence variation was found in either FLA gene that segregated with glucose release or xylose release.

Although no significant association was observed for total lignin content in these lines (Figure 25), there was a significant association on chromosome 7 for the abundance of two S-derived lignin pyrolysates (mz_167 and mz_168) that did not appear for any other traits (Figures 12-13). This QTL was centered on SNP S07_59456807. Despite bioinformatic analysis of the 15 genes around this locus (Table 2), a strong candidate gene was not identified.

Lastly, no associations were found for glucan content (Figure 26), xylan content (Figure 27), or xylose yield (Figure 28).

In summary, phenotypic relationships suggest that lignin content and composition do not greatly affect saccharification in this population. Difficulties with prediction of lignin composition from reflectance spectra are summarized. Lastly, GWAS led to the identification of QTL and candidate genes for several compositional traits.

Discussion

As discussed in chapter 1 there are consistent reports on the negative correlation between lignin content and saccharification (Dien et al., 2009; Fu et al., 2011; Fornale et al., 2012; Poovaiah et al., 2014), but the reports are mixed on lignin composition (Li et al., 2016). However, the ratio of the primary monolignols (S/G ratio) seems to have a more profound effect on saccharification when there is more than 50 % syringyl units (Fontaine et al., 2003). As the highest S/G ratio observed in our Sorghum diversity panel was relatively low (0.78), the lack of relationship with saccharification is consistent with literature (Reddy et al., 2005; Chen and Dixon, 2007). Hot water pretreatment, as used by NREL, can affect the relationship between lignin content and saccharification by breaking β-O-4’ bonds of lignin and disrupting the cell wall matrix (Samuel et al., 2013). Hot water pretreated *Arabidopsis* stems had a positive relationship between S/G ratio and saccharification (Li et al., 2010). Although the relationship between lignin composition and saccharification is weakened by increasing severities of biomass pretreatment (Li et al., 2016), we found no strong relationship between the two.

There was a moderate correlation among the sugar release and sugar yield traits (Figure 6). This partially explains the shared QTL on chromosome 9 for glucose release (Figure 9), xylose release (Figure 10), and glucose yield (Figure 11), but, although xylose yield is also

correlated with the other sugar traits, it does not share the chromosome 9 association (Figure 27). This may be a consequence of the inaccuracies of the high throughput acid hydrolysis method, but it could also suggest that the causative polymorphism is more related to cellulose biosynthesis than xylan biosynthesis. Tightly linked to this QTL is a tandem repeat of FLA which, in *Arabidopsis*, is affiliated with sugar traits (MacMillan et al., 2010) and highly coexpressed with cellulose synthase (Persson et al., 2005). The fasciclin domain is involved in cell adhesion and is hypothesized to be involved in cell identity (Johnson et al., 2003; Johnson et al., 2011). Although no FLA coding sequence polymorphisms cosegregated with any sugar release traits, the experiments conducted here cannot exclude the possibility of coding sequence variation driving these association peaks. Such causative coding sequence variation could have a statistically significant effect in the whole population yet may not cosegregate perfectly with glucose and xylose release in the 4 highs and 4 lows sequenced. Alternatively, non-coding variation or variation in another linked gene could be driving these associations.

The key finding in this thesis is the coding sequence mutation in PAL that effects S/G ratio. As PAL is not at a fork in the phenylpropanoid pathway, the relationship between a PAL mutation and S/G ratio is not immediately clear. PAL mutants in *Arabidopsis* and tobacco also displayed an increased S/G ratio although these changes were accompanied by decreased overall lignin content (Sewalt et al., 1997; Rohde et al., 2004). In contrast, the data here do not suggest a significant relationship between total lignin content and composition in *Sorghum* (Figure 6). One explanation for this difference could be inherent inaccuracies of PyMBMS for lignin quantification. In support of this idea, validation of PyMBMS as a high throughput screen for lignin reported an R^2 of 0.72 for estimating Klason lignin in maize stover (Penning et al., 2014). Penning et al. modeled Klason lignin abundance in maize stover rather than *Sorghum*. Since the same pyrolysate profile (Table 1) was used in this thesis for Klason lignin quantification, the results may have been skewed by differences that exist between maize and *Sorghum* in non-lignin contributions to these pyrolysates. This inaccuracy may also be the reason for the lack of QTL discovered for Klason lignin (Figure 25). An attempt was made to discover more lignin-related QTL by conducting GWAS for individual lignin pyrolysates (listed in Table 1). A QTL on chromosome 7 was discovered for the abundance of the 167 and 168 mass fragments, but a clear candidate for driving this association peak among the genes in the region (Table 2) was not identified. There are several possible explanations for the lack of the identification of a single

candidate gene including that this was a false positive association peak, there is poor annotation of the linked genes, or one of the linked genes has an interaction with lignin biosynthesis beyond my knowledge.

To follow up the discovery of the PAL mutation enzyme activity assays were conducted with crude extracts from the top internodes of eight *Sorghum* lines (shown in Figure 2). There are seven PAL homologs in *Sorghum*. The hypothesis for this experiment was that a nonsense mutation in one copy of PAL would reduce net PTAL activity, potentially altering the ratio of S lignin to G lignin. A challenge of this experiment was the non-uniform background of the lines tested. Differences in plant development and genetic regulation of PTAL activity among these lines were confounded with the effect of the mutation on net PTAL activity. Harvesting the plants at a recognizable developmental stage and using the maximum activity among four stem segments (for each plant) for comparisons helped normalize developmental differences. The means of the maximum activities for the four lines in each allelic group were calculated and the lines with the nonsense mutation (T allele) had a lower PTAL activity than the lines with the WT allele (C allele; p-value = 0.08; Figure 21). Although this p-value is not statistically significant below the generally accepted level (0.05) the biological impact of this nonsense mutation on PAL activity seems to have a relatively large impact on lignin composition (Figure 17).

Association studies, such as the ones presented here, are a tool for academia and industry alike. In the academic realm association mapping can be a valuable tool to discover genes involved in biochemical pathways. The variation in these genes can then be used for marker-assisted selection to improve crops in breeding programs. A major criticism of GWAS is that it is limited to discovery of major alleles and large effect alleles. Nevertheless, alleles such as the PAL allele observed here, can provide a foundation for selection of improved varieties.

The attempted modelling of lignin composition from hyperspectral bands failed to produce a reasonable model due to lack of covariation between the spectral bands and S/G ratio. The first component explained 73% of the spectral variation and only 0.1% of the phenotypic variation. Since the latent predictor variables of PLSR are principal components (linear combinations of the original predictors), each consecutive component account for a decreasing variation. Typically, the PRESS decreases with increasing number of components as more of the model variance is explained by the predictors. The reason we don’t see a similar trend (Figure 7)

is because the first component explained a trivial amount of phenotypic variation, so any additional components added will also only account for a trivial amount of phenotypic variation.

There are at least two possible explanations for the lack of covariation between the spectral data and the compositional data. First, PyMBMS is an indirect method for lignin quantification. Characteristic mass fragments from pyrolyzed biomass that are correlated (some better than others) with the abundance of lignin units are summed to predict lignin monomer abundance. It was shown that the PyMBMS method is able to predict the abundance of CuO oxidized S and G lignin units from maize stover with an R^2 of .89 and 0.85, respectively (Penning et al., 2014). To the extent these maize values translate to sorghum, using PyMBMS data as our reference value for spectral modelling sets the upper limit of predictability and this may not have been sufficiently accurate. Second, the light absorbed in the NIR and SWIR regions are only vague features of molecular vibrations. The intensity of the absorbance at each band depend on elemental composition and the bonds among those elements. Although theoretically straightforward patterns can be recognized for solutions of simple molecules, lignin, and plant biomass even more so, are complex and heterogeneous.

In addition to using a more accurate phenotyping method, a better suited experiment for modelling lignin composition would involve reducing the amount of confounding variation across the population so that more of the spectral variation is due to lignin variation. Rather than a *Sorghum* diversity panel, a population of near-isogenic lines, with a wide range and uniform distribution of lignin composition, may be more ideal for this experiment.

Despite the lack of success in hyperspectral modelling of lignin composition, the results presented in this thesis provide insights about genes and QTL related to biomass composition that can inform future efforts in research and breeding.

FIGURES AND TABLES

Table 1. PyMBMS fragments

Lignin Pyrolysates	Source	Abundant ion
Phenol	lignin	94
Vinylphenol	h,g,s	120
Guaiacol	g	124
ethylguaiacol, homovanillin, coniferyl alcohol	g	137
methylguaiacol	g	138
vinylguaiacol, coumaryl alcohol	g	150
4-ethylguaiacol, vanillin	g	152
Syringol	s	154
Eugenol	g	164
ethylsyringol, syringylacetone, propiosyringone	s	167
4-methylsyringol	s	168
coniferyl aldehyde	g	178
S-based, other unknown	s,g	181
syringaldehyde	s	182
4-propenylsyringol	s	194
sinapaldehyde	s	208
sinapyl alcohol	s	210

Table 2. Chromosome 7 candidate gene list

***Sorghum* Gene**	**NCBI description**	**Nearest *Arabidopsis* ortholog**
Sobic.007G158800	Patatin-like phospholipase of plants	AT2G26560.1 (PLA, IIA, PLA2A, PLP2) phospholipase A 2A
Sobic.007G158900	Sodium channel modifier	AT2G17530.2 Protein kinase superfamily protein
Sobic.007G159000	Nascent polypeptide-associated complex subunit alpha, muscle-specific form	AT2G01710.1 Chaperone DnaJ-domain superfamily protein
Sobic.007G159100	Tetraspanin-8	AT3G45600.1 (TET3) tetraspanin3
Sobic.007G159200	Histone-lysine N-methyltransferase NSD2 isoform X1 [Zea mays]	AT5G27650.1 Tudor/PWWP/MBT superfamily protein
Sobic.007G159300	Transcription factor bHLH30	AT3G25710.1 (ATAIG1, BHLH32, TMO5) basic helix-loop helix 32
Sobic.007G159450	SH3 domain-containing protein	
Sobic.007G159500	Similar to Thymidylate synthase-like	AT4G38090.2 Ribosomal protein S5 domain 2-like superfamily protein
Sobic.007G159600	SH3 domain-containing protein	
Sobic.007G159650	SH3 domain-containing protein	AT4G38080.1 hydroxyproline-rich glycoprotein family protein

Sobic.007G159700	ATP synthase subunit d, mitrochondrial	AT3G52300.1 (ATPQ) ATP synthase D chain, mitochondrial
Sobic.007G159800	AT-hook motif nuclear-localized protein 17	AT5G49700.1 Predicted AT-hook DNA-binding family protein
Sobic.007G159900	Ubiquitin carboxyl-terminal hydrolase 17 isoform X2	AT5G65450.1 (UBP17) ubiquitin-specific protease 17
Sobic.007G160050	Hypothetical protein	
Sobic.007G160200	Mitochondrial uncoupling protein 5	AT4G24570.1 (DIC2) dicarboxylate carrier 2

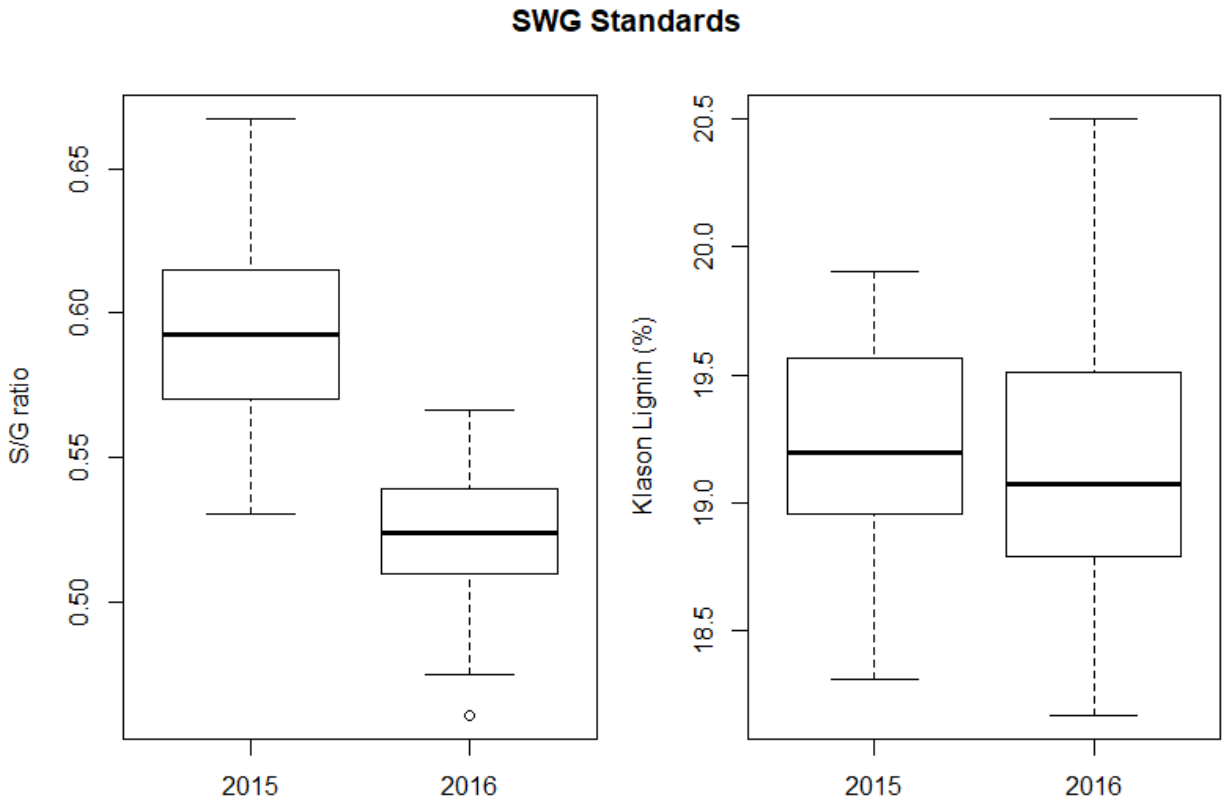

Figure 1. Instrumental variation of PyMBMS between years
Lignin content and composition for switchgrass standards between years after mean normalization.

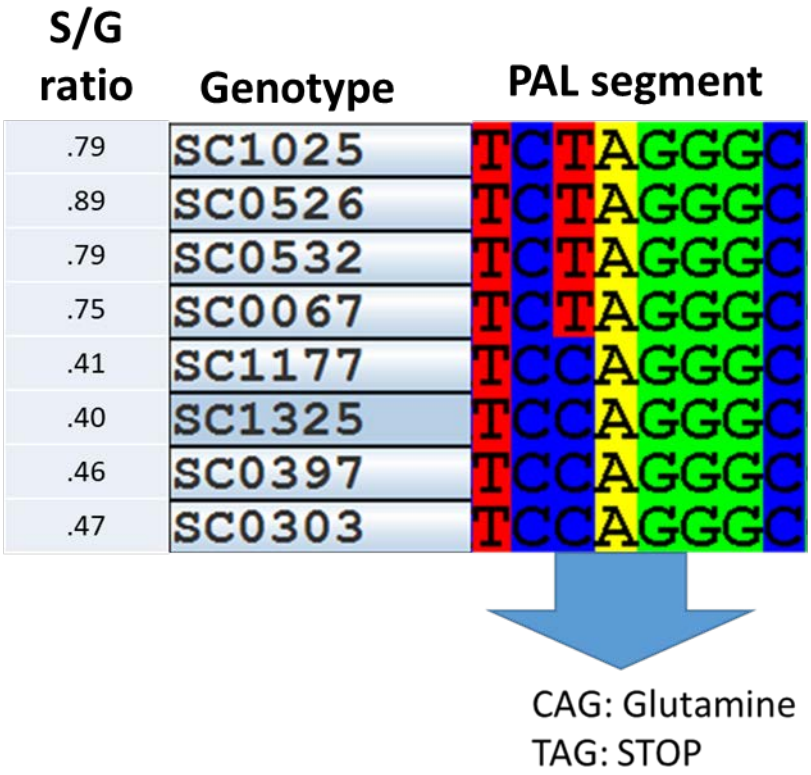

Figure 2. Cosegregation of SNP in Sobic8800 with S/G ratio
Sanger sequencing of the coding sequence of Sobic.006G148800 in four genotypes with a high S/G ratio and four lines with a low S/G ratio reveals cosegregation of a SNP. This SNP induces a premature stop codon and a *Bfa*I recognition sequence in the highs.

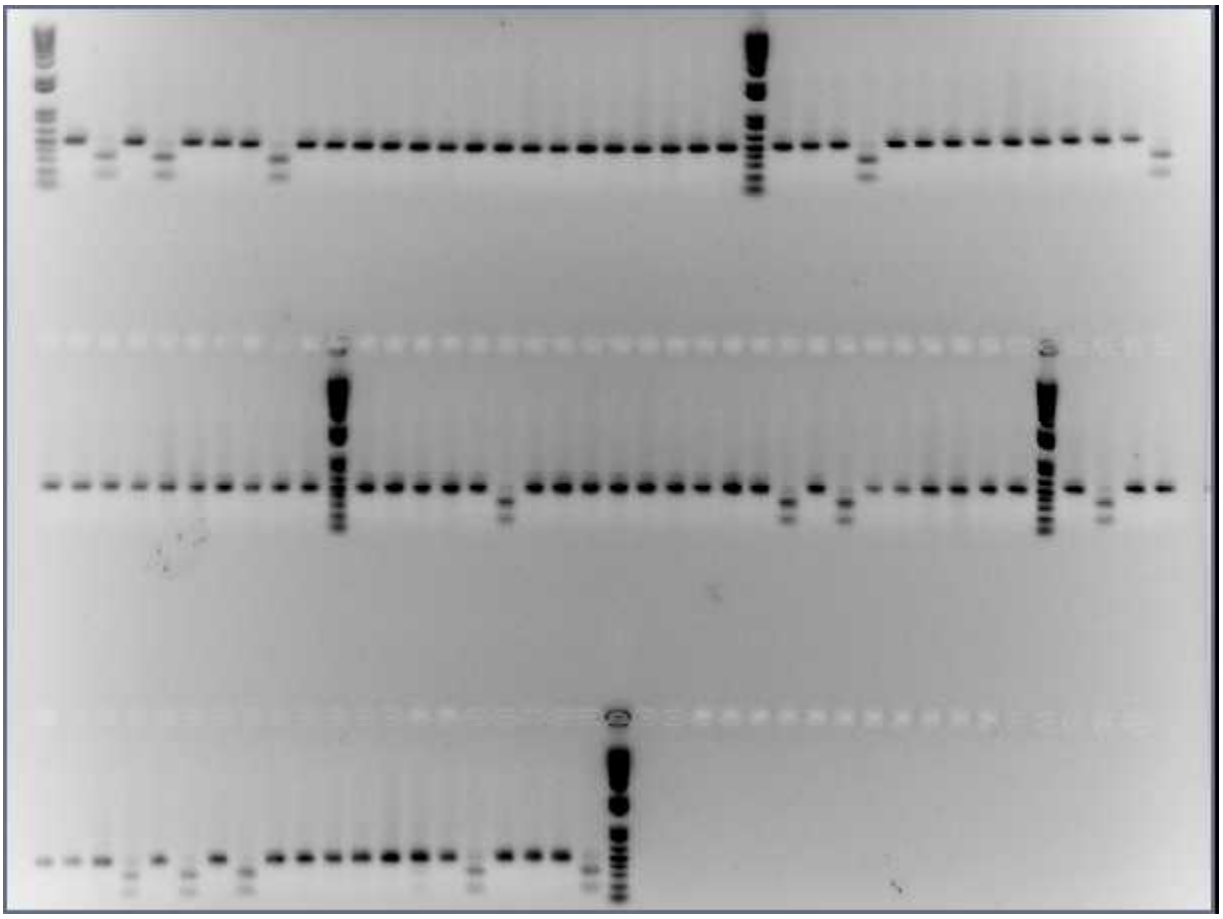

Figure 3. Gel electrophoresis of *Bfa*I digestions for genotyping PAL mutation
Example of a gel that was the result of genotyping a mutation in Sobic.006G148800 (PAL). PCR product size: 498 bp; Restriction digest fragments: 326 & 172 bp

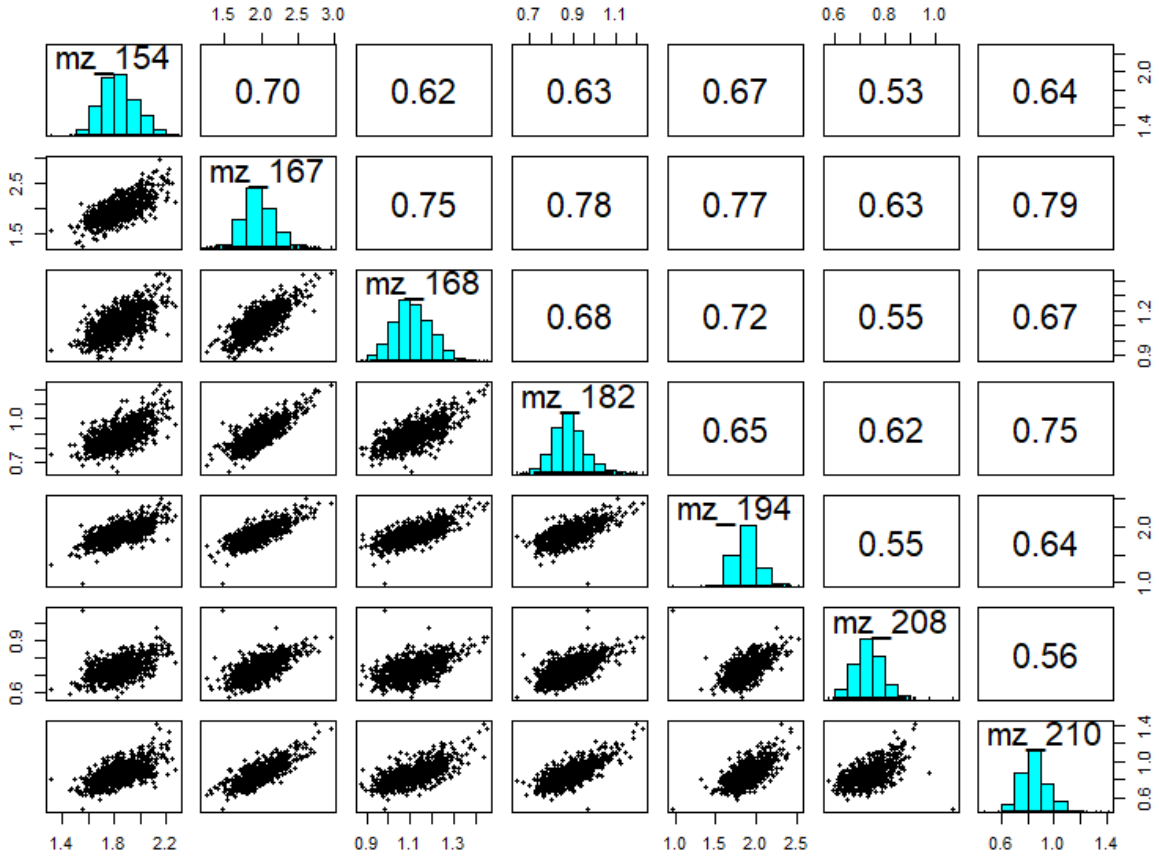

Figure 4. Correlation of syringyl-lignin derived pyrolysates
The upper panel contains the Pearson correlation coefficients.

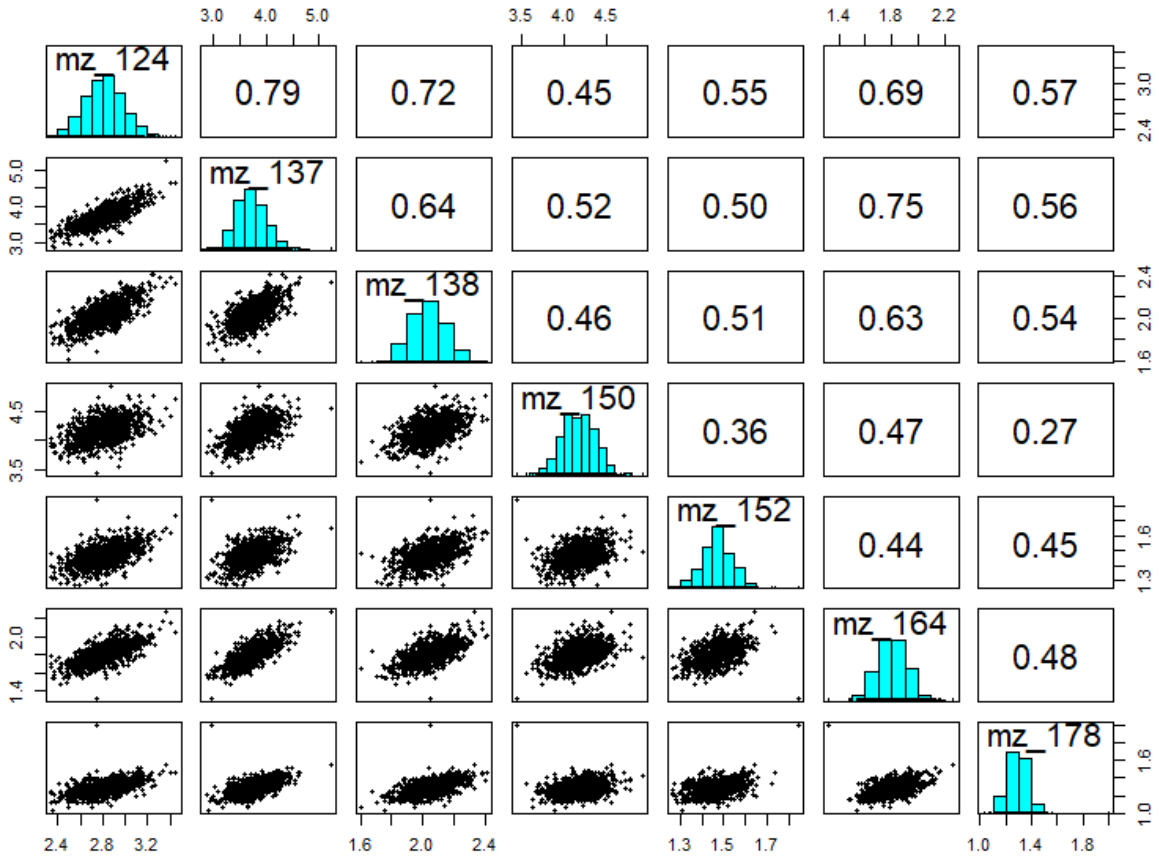

Figure 5. Correlation of guaiacyl-lignin derived pyrolysates
The upper panel contains the Pearson correlation coefficients.

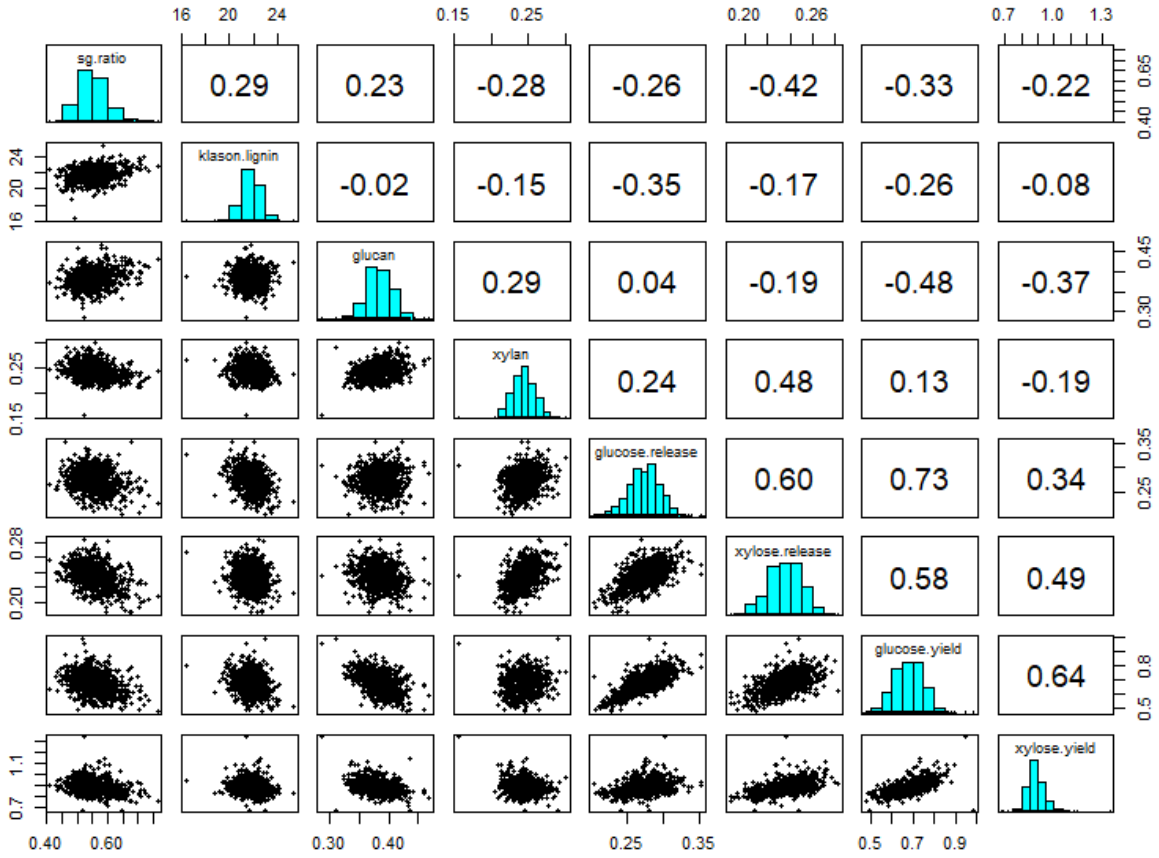

Figure 6. Correlation of biomass composition traits
The upper panel contains the Pearson correlation coefficients.

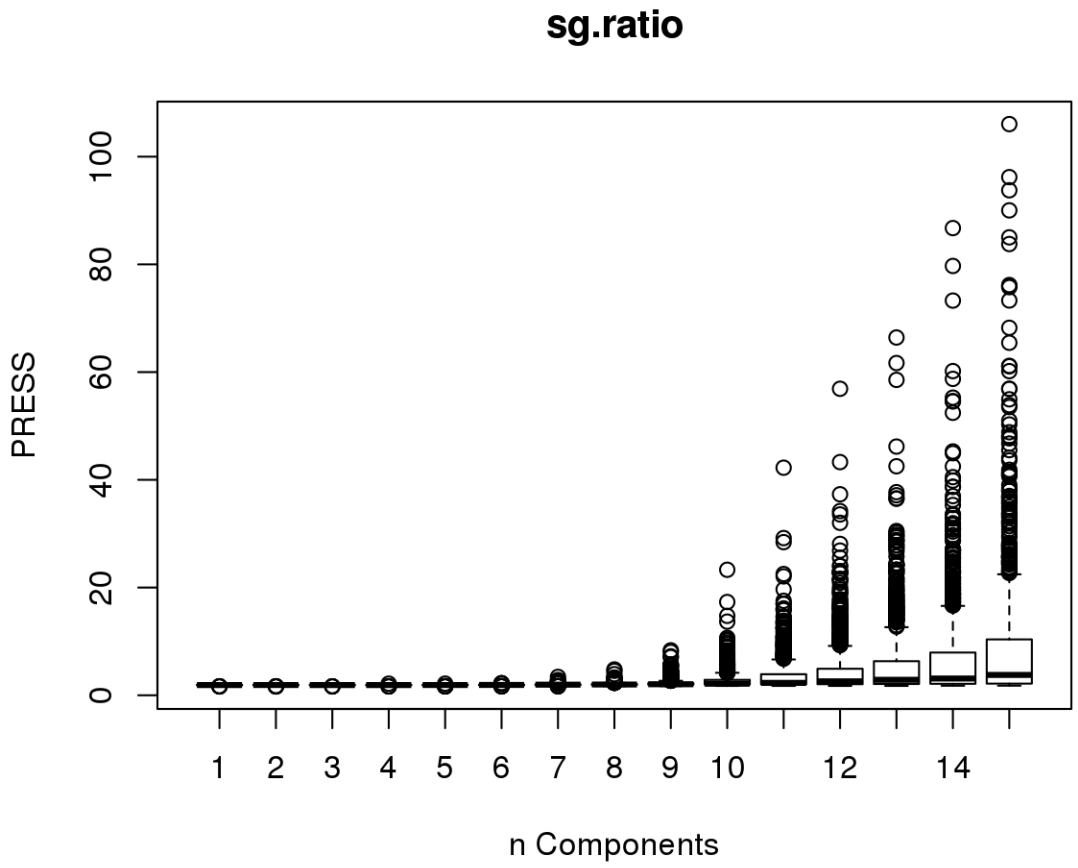

Figure 7. Cross-validation by PRESS reduction
Leave-one-out cross validation through reduction of the predicted residual sum of squares (PRESS) shows no increase in phenotypic variation explained (reduction of PRESS) as more components are added to the model. For each model with different number of components (x-axis) the variation of PRESS over 1000 simulations is shown as a boxplot.

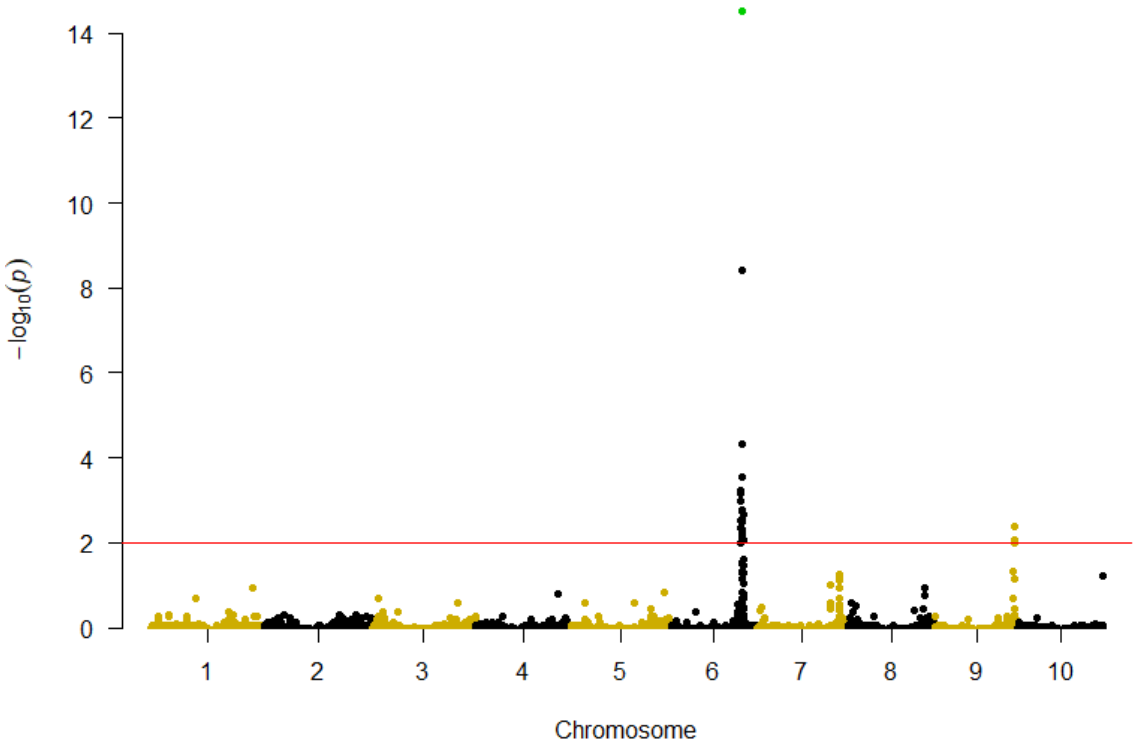

Figure 8. Manhattan plot for S/G ratio
FDR-adjusted p-values are plotted. The SNP highlighted in green was independently genotyped (not in the original GBS dataset).

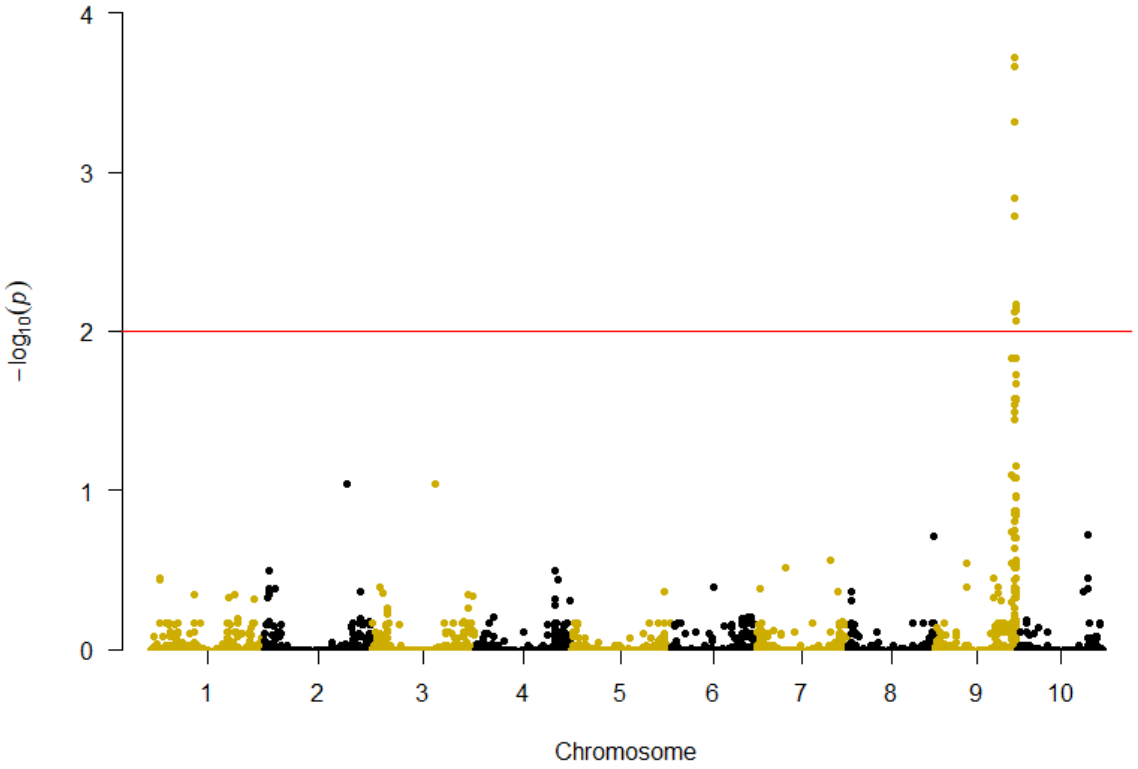

Figure 9. Manhattan plot for glucose release
FDR-adjusted p-values are plotted.

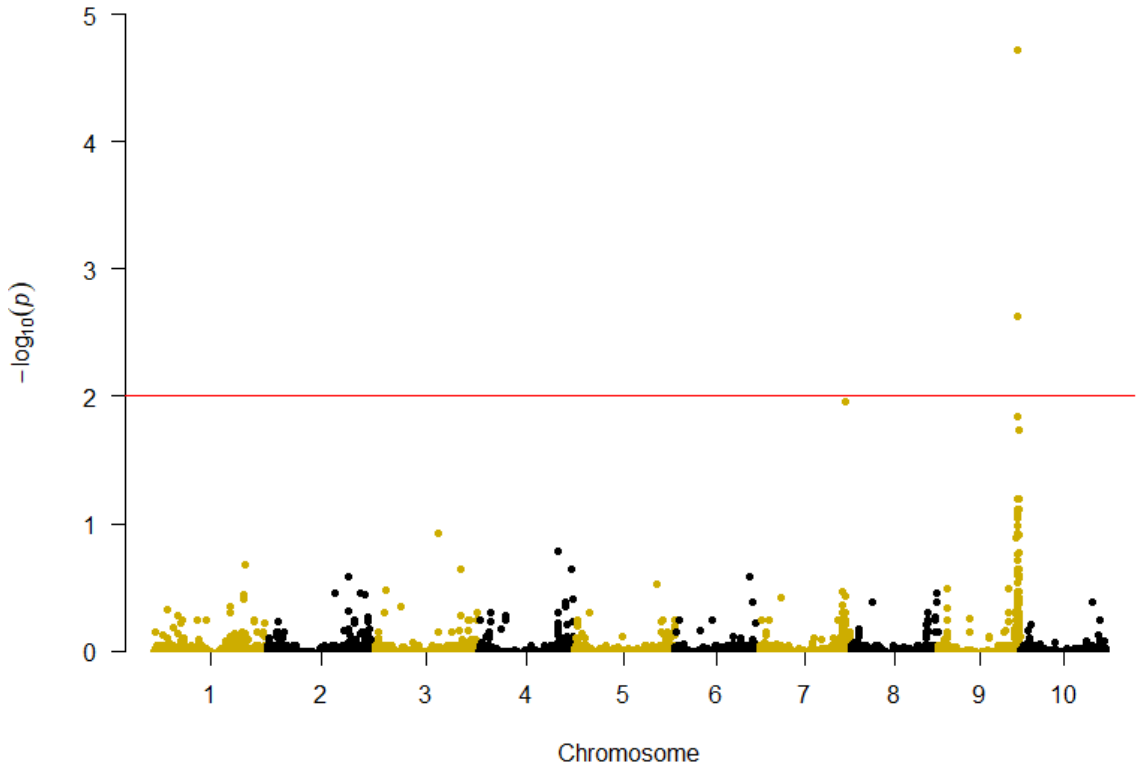

Figure 10. Manhattan plot for xylose release
FDR-adjusted p-values are plotted.

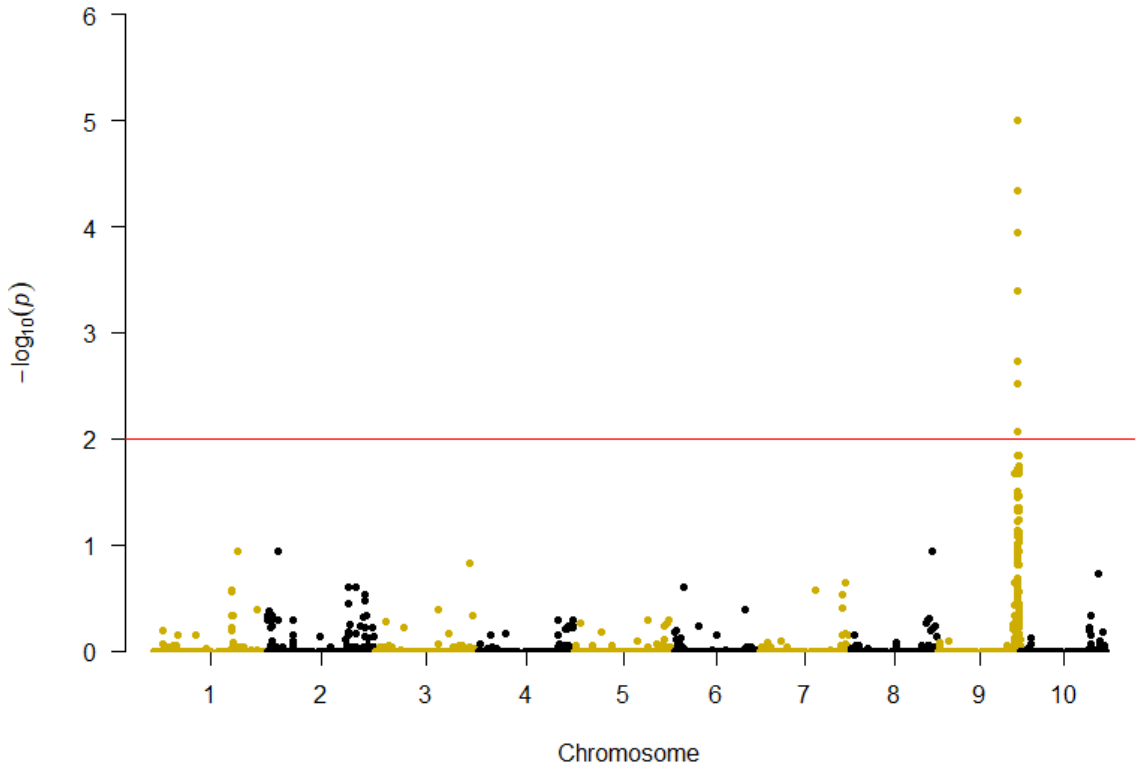

Figure 11. Manhattan plot for glucose yield
FDR-adjusted p-values are plotted.

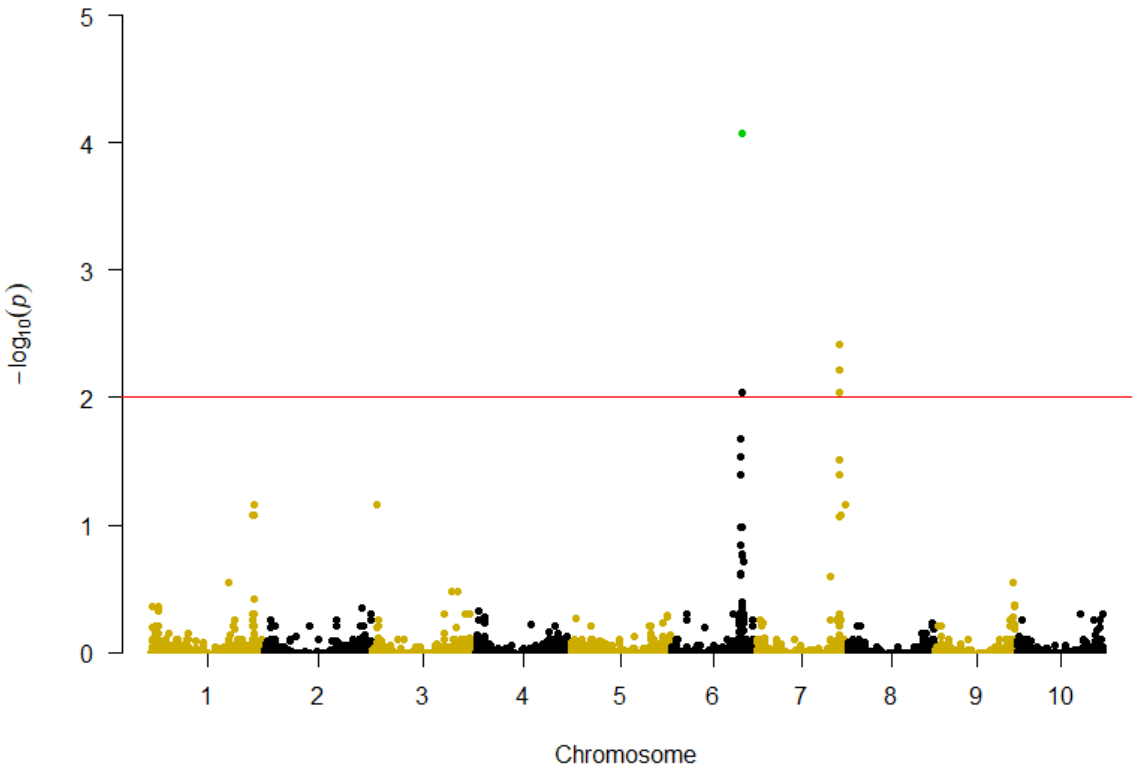

Figure 12. Manhattan plot for mz_167
Manhattan plot for the abundant ion shared by ethylsyringol, syringylacetone, and propiosyringone (mz_167). FDR-adjusted p-values are plotted. The SNP highlighted in green was independently genotyped (not in the original GBS dataset).

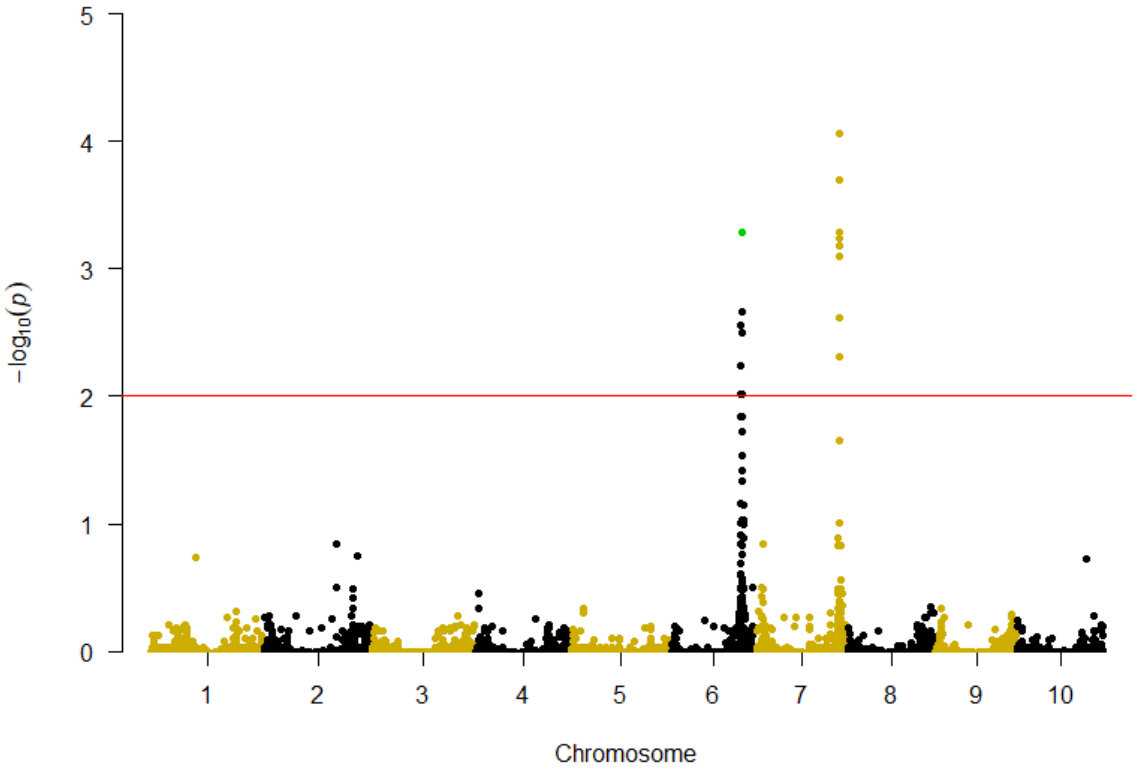

Figure 13. Manhattan plot for mz_168
Manhattan plot for 4-methylsyringol (mz_168). FDR-adjusted p-values are plotted. The SNP highlighted in green was independently genotyped (not in the original GBS dataset).

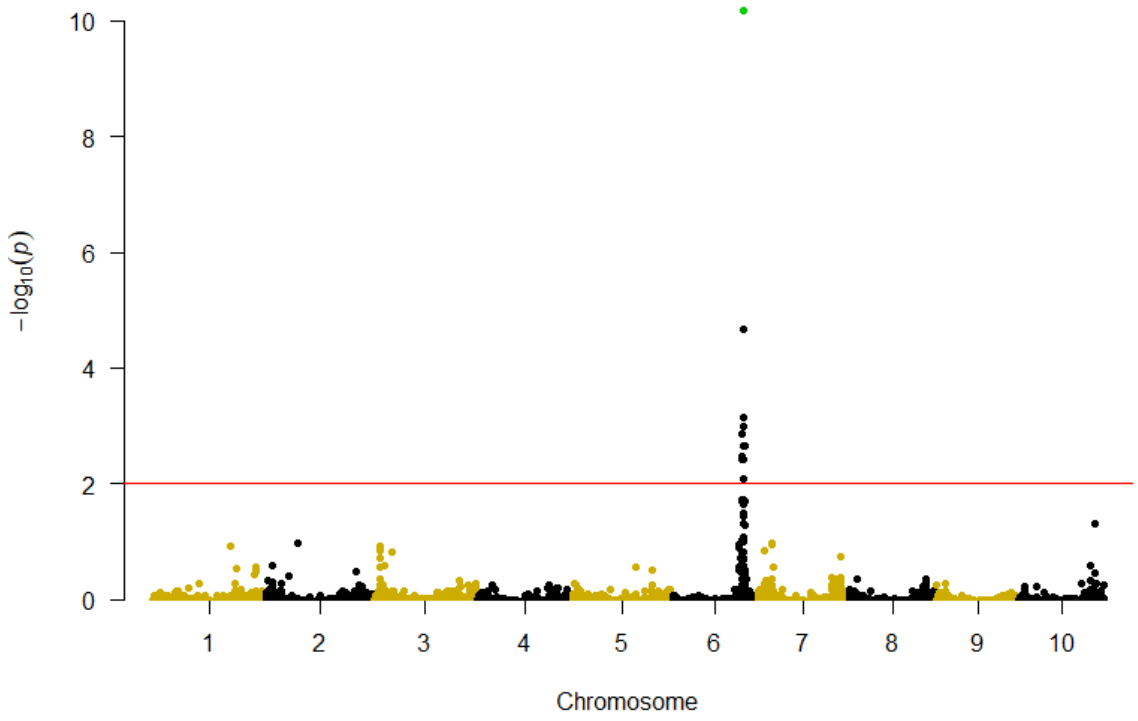

Figure 14. Manhattan plot for mz_154
Manhattan plot for syringol (mz_154). FDR-adjusted p-values are plotted. The SNP highlighted in green was independently genotyped (not in the original GBS dataset).

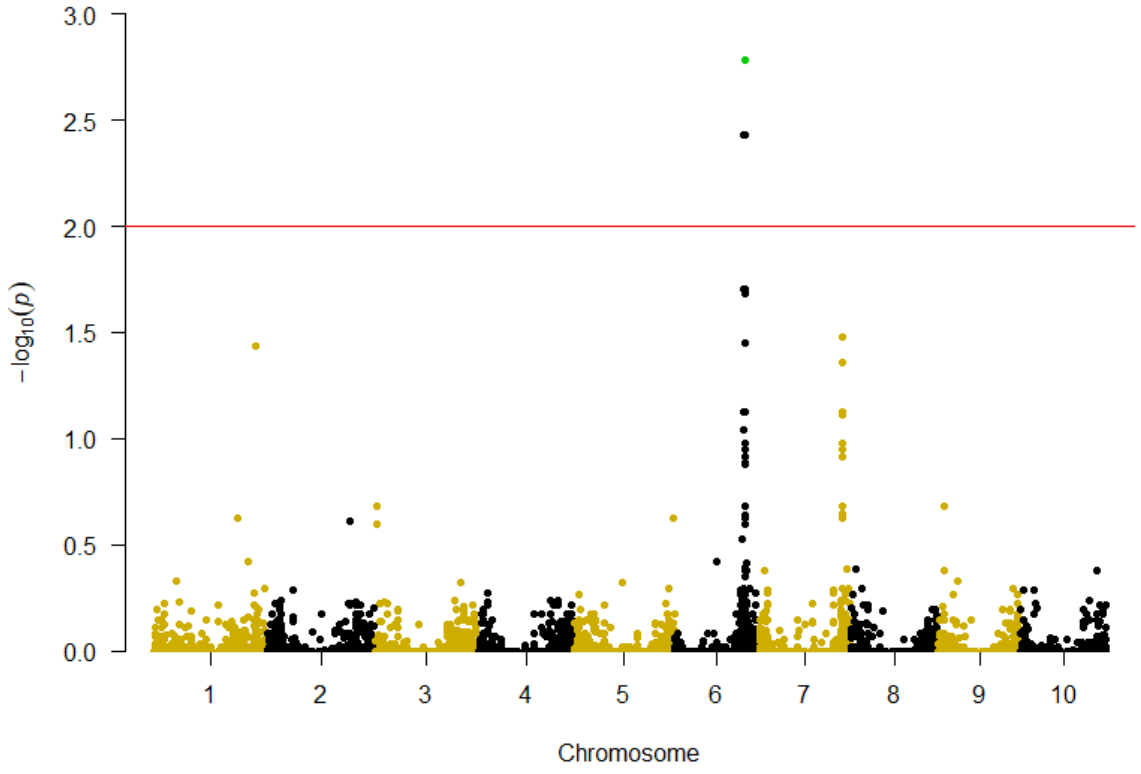

Figure 15. Manhattan plot for mz_181
Manhattan plot for unknown S-based pyrolysates (mz_181). FDR-adjusted p-values are plotted. The SNP highlighted in green was independently genotyped (not in the original GBS dataset).

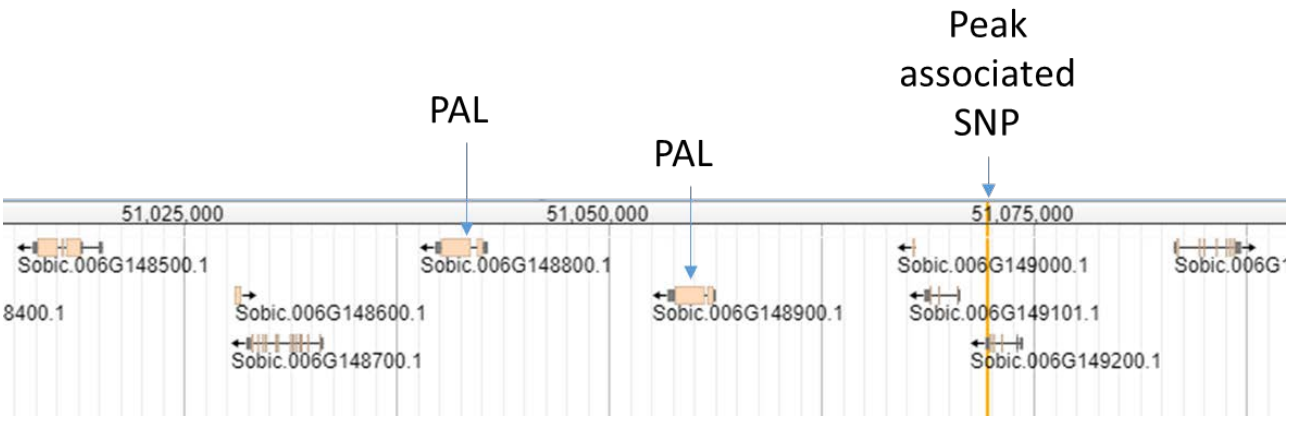

Figure 16. Chromosome 6 QTL
The locus surrounding the peak associated SNP (S06_51072215) for S/G ratio contains a tandem repeat of genes encoding phenylalanine ammonia-lyase (PAL). This image was derived from the JBrowse tool on Phytozome (phytozome.jgi.doe.gov/jbrowse).

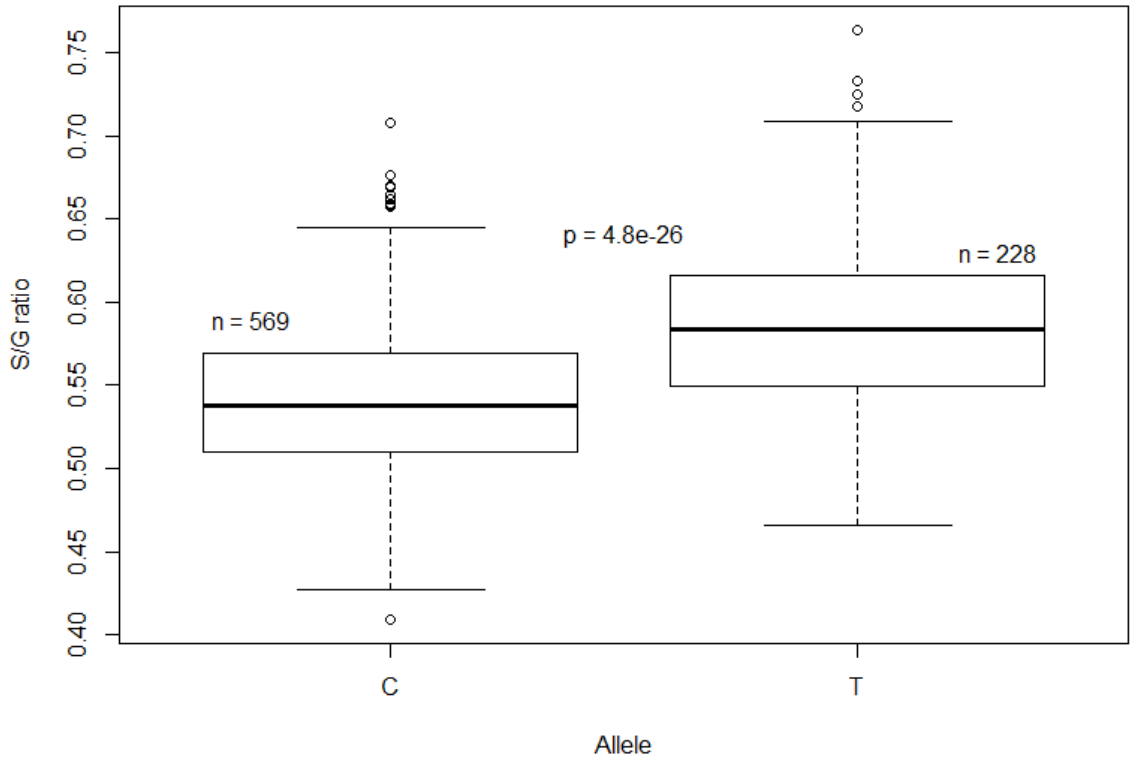

Figure 17. SNP effect of PAL mutation
P-value is the result of a Welch's t-test.

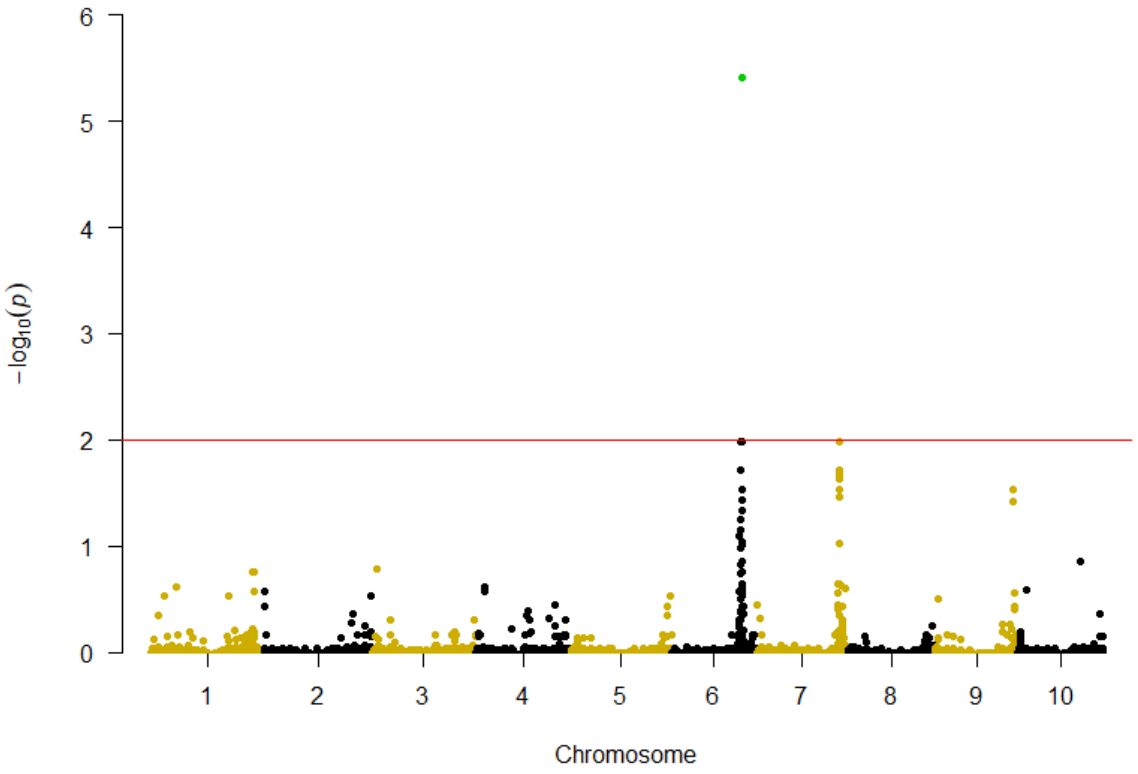

Figure 18. Manhattan plot for mz_182
Manhattan plot for syringaldehyde (mz_182). FDR-adjusted p-values are plotted. The SNP highlighted in green was independently genotyped (not in the original GBS dataset).

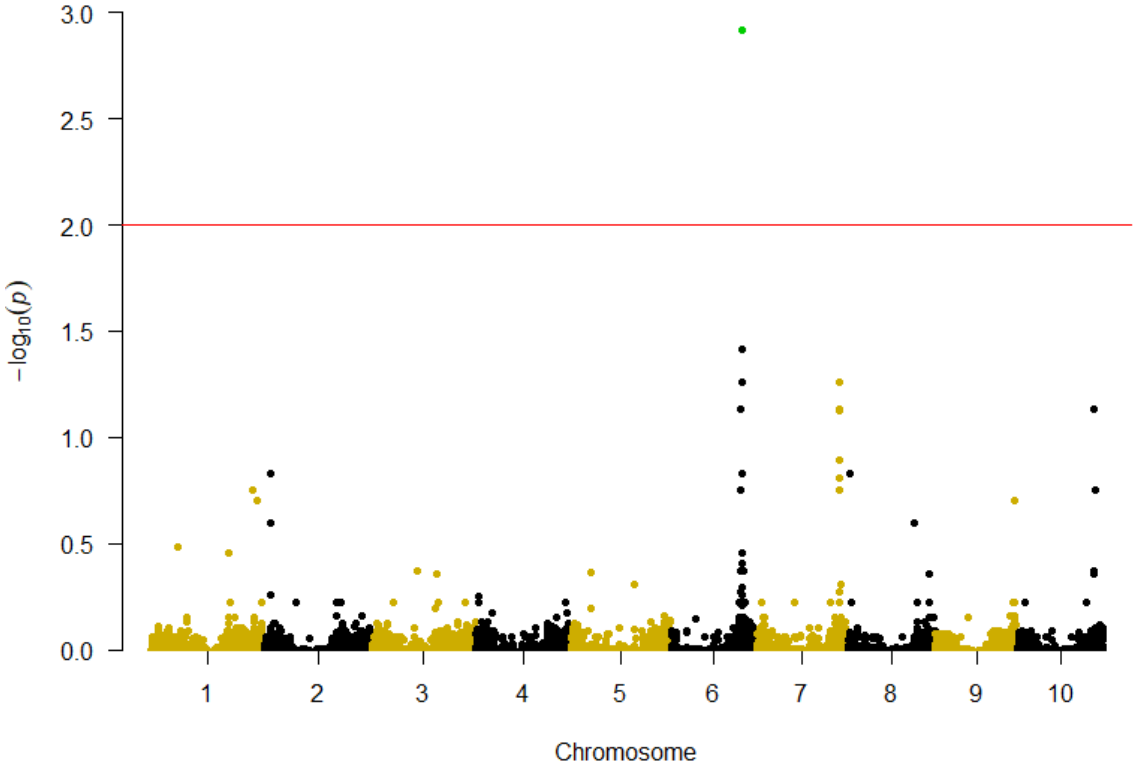

Figure 19. Manhattan plot for mz_194
Manhattan plot for 4-propenylsyringol (mz_194). FDR-adjusted p-values are plotted. The SNP highlighted in green was independently genotyped (not in the original GBS dataset).

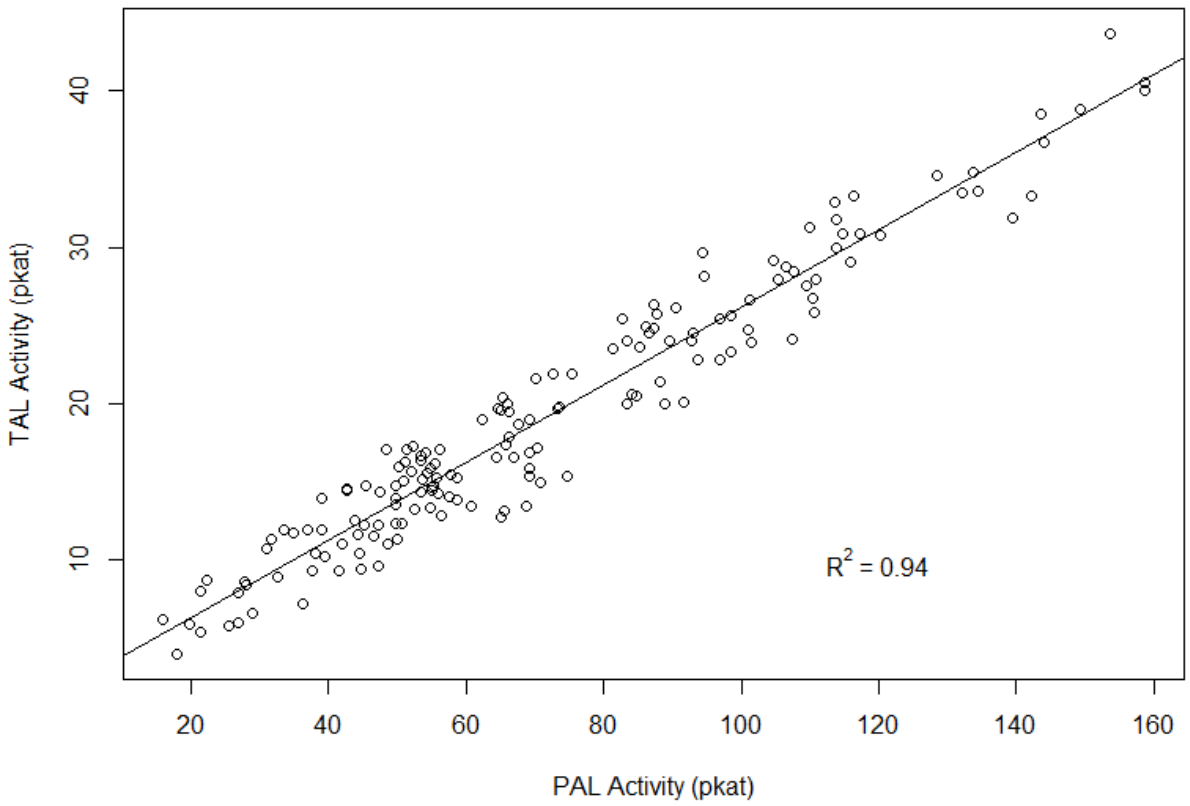

Figure 20. Correlation of PAL and TAL activity
There was a linear relationship between PAL and TAL activity in *Sorghum* stalks.

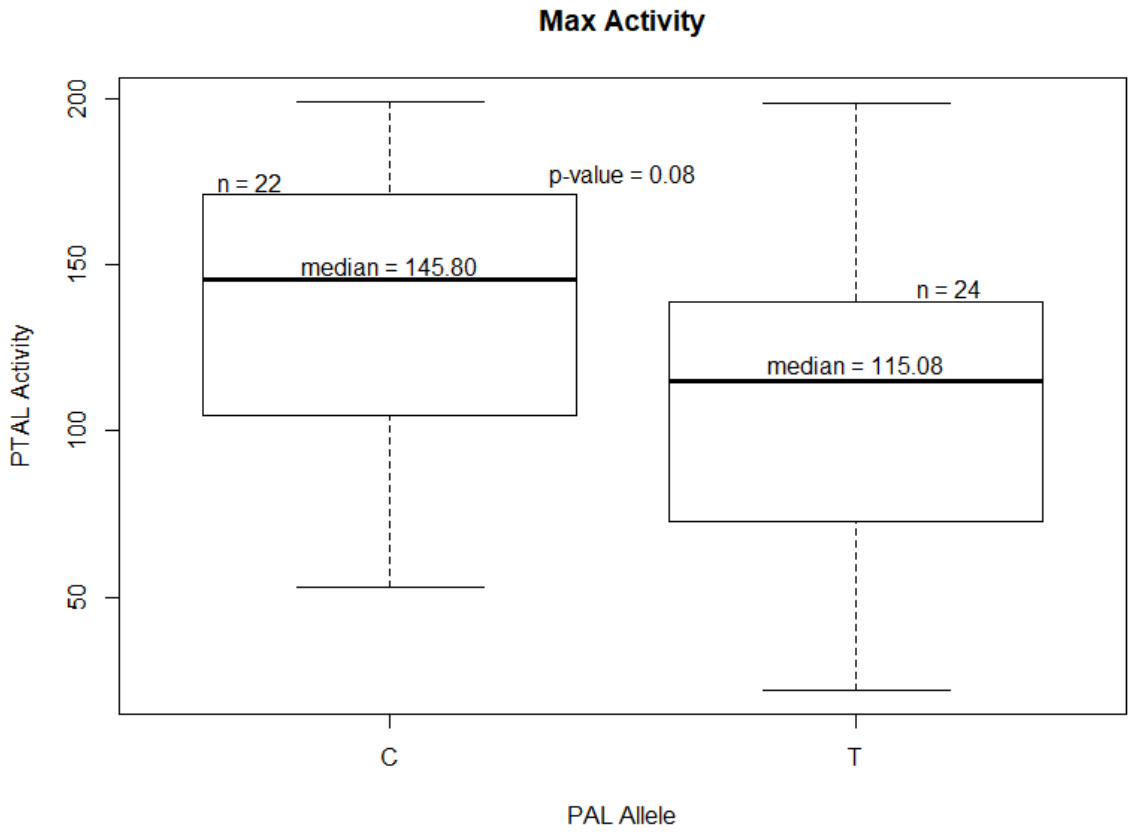

Figure 21. Allelic effect of PAL SNP on PTAL activity
A comparison of PTAL activity between samples polymorphic for the PAL mutation. Max activity of 4 internode segments within a biological rep.

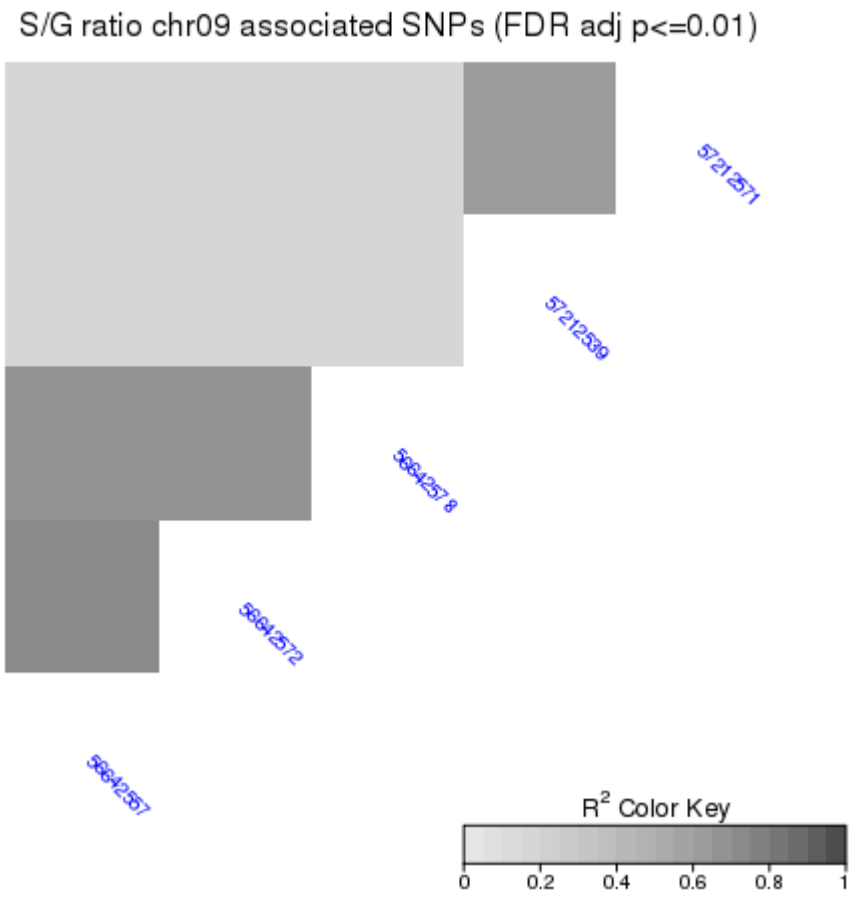

Figure 22. LD heat map for chromosome 9 association peak for S/G ratio
The LD heat map for the association peak for S/G ratio on chromosome 9 shows two distinct linkage groups. One locus (SNPs S09_57212539 and S09_57212571) is also associated with glucose release, xylose release, and xylose yield.

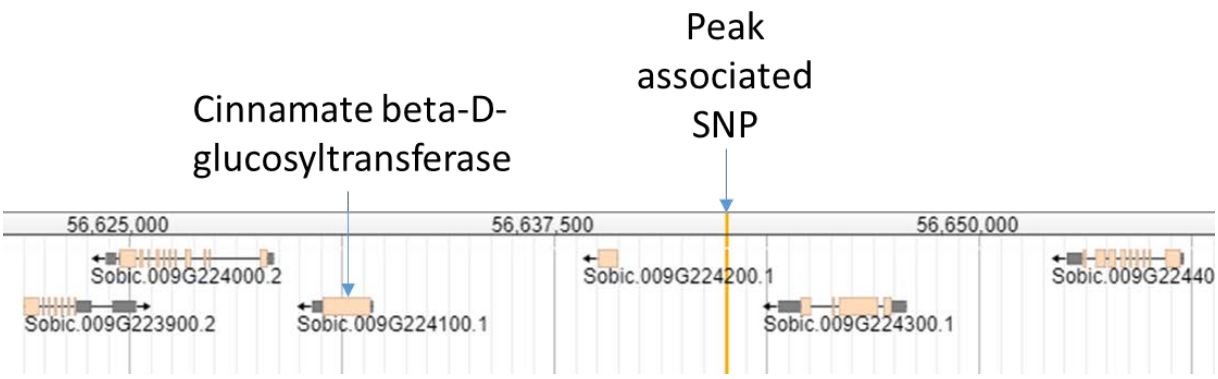

Figure 23. Chromosome 9 QTL 1
The locus of the peak associated SNP (S09_56642557) for S/G ratio. The hypothesized candidate gene in this region encodes a cinnamate beta-D-glucosyltransferase which utilizes the phenylpropanoid intermediate cinnamate as a substrate.

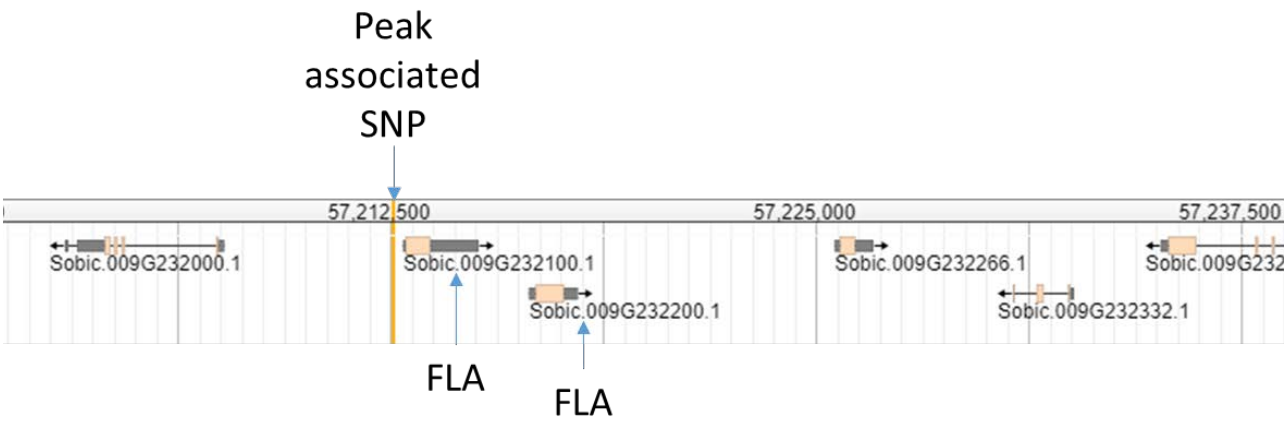

Figure 24. Chromosome 9 QTL 2
The locus of the peak associated SNP (S09_57212539) for the chromosome 9 peak for glucose release, glucose yield, xylose release, and S/G ratio contains a tandem repeat of genes encoding fasciclin-like arabinogalactan proteins (FLA). This image was derived from the JBrowse tool on Phytozome (phytozome.jgi.doe.gov/jbrowse).

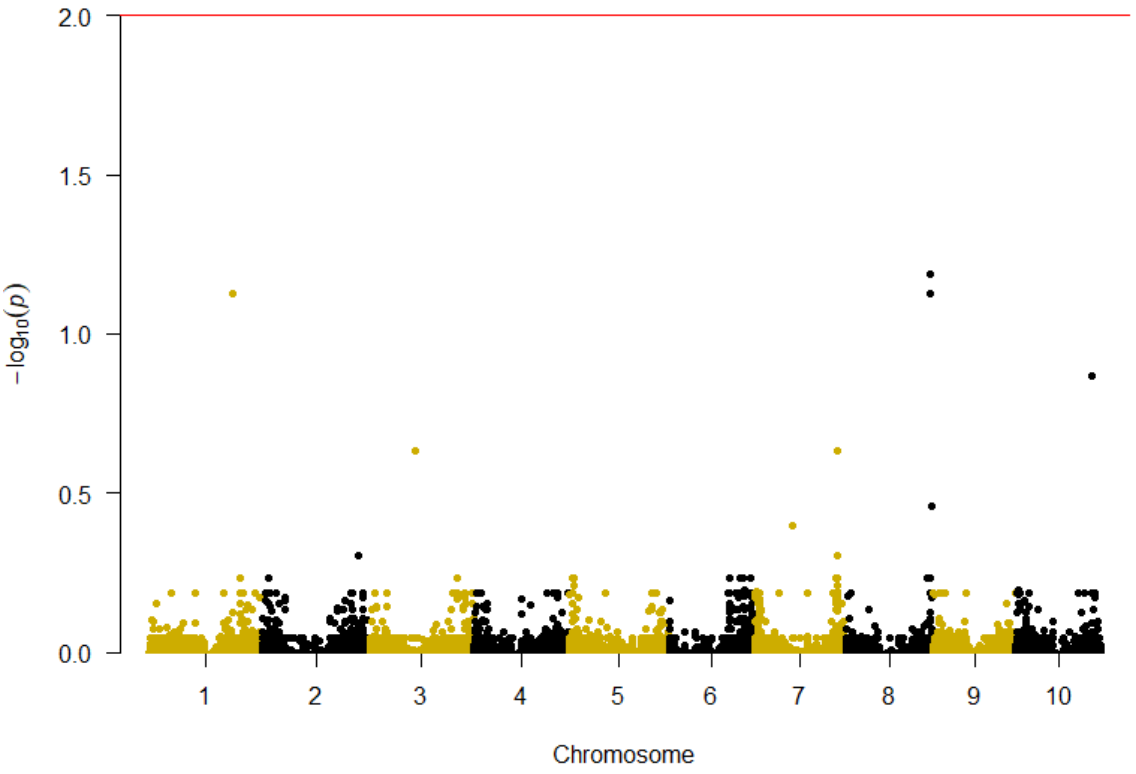

Figure 25. Manhattan plot for Klason lignin
FDR-adjusted p-values are plotted.

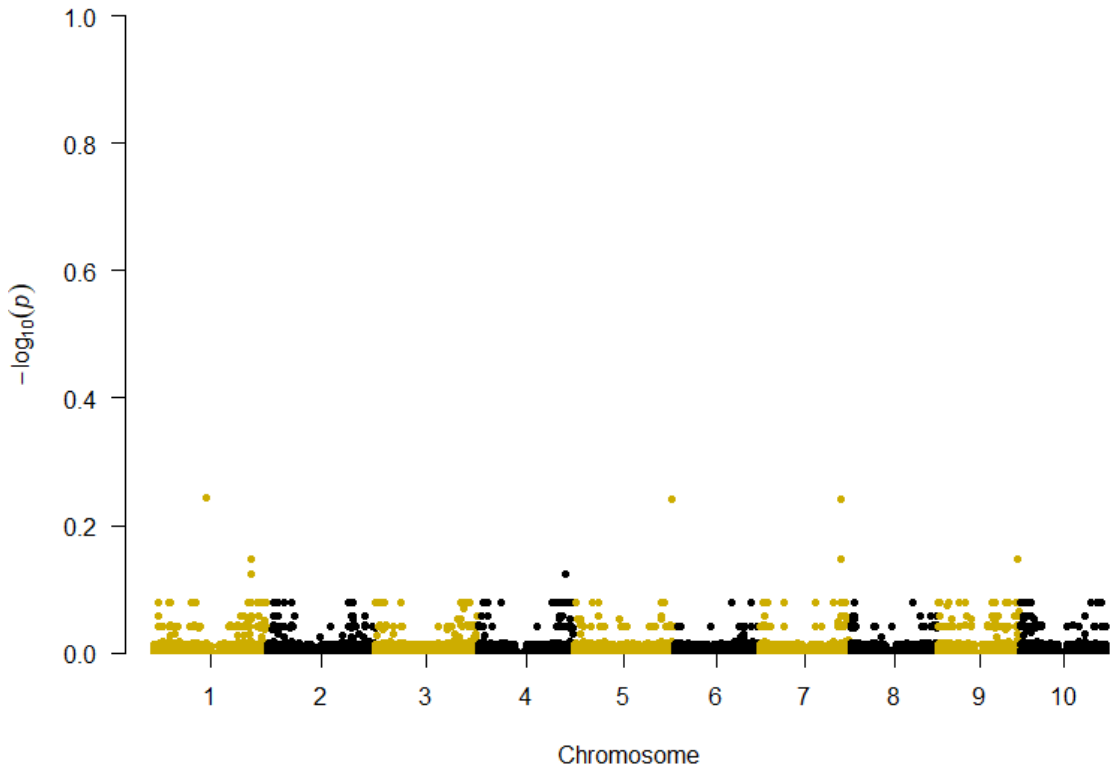

Figure 26. Manhattan plot for glucan
FDR-adjusted p-values are plotted.

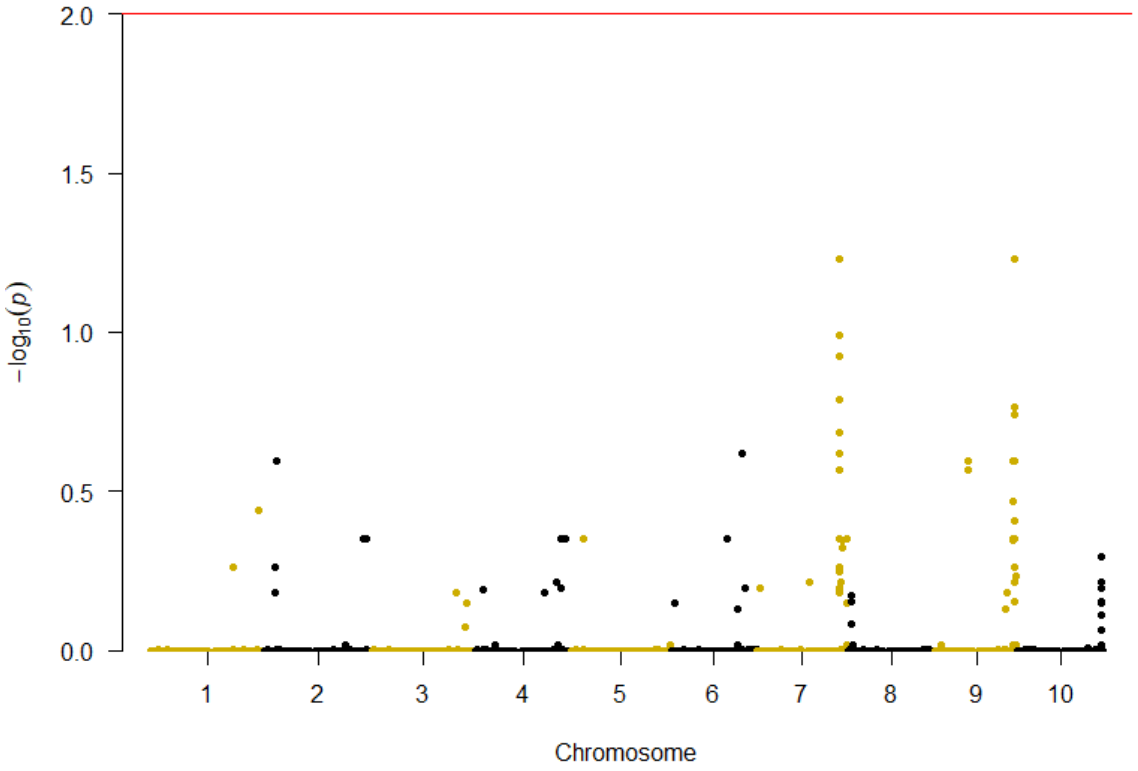

Figure 27. Manhattan plot for xylan
FDR-adjusted p-values are plotted.

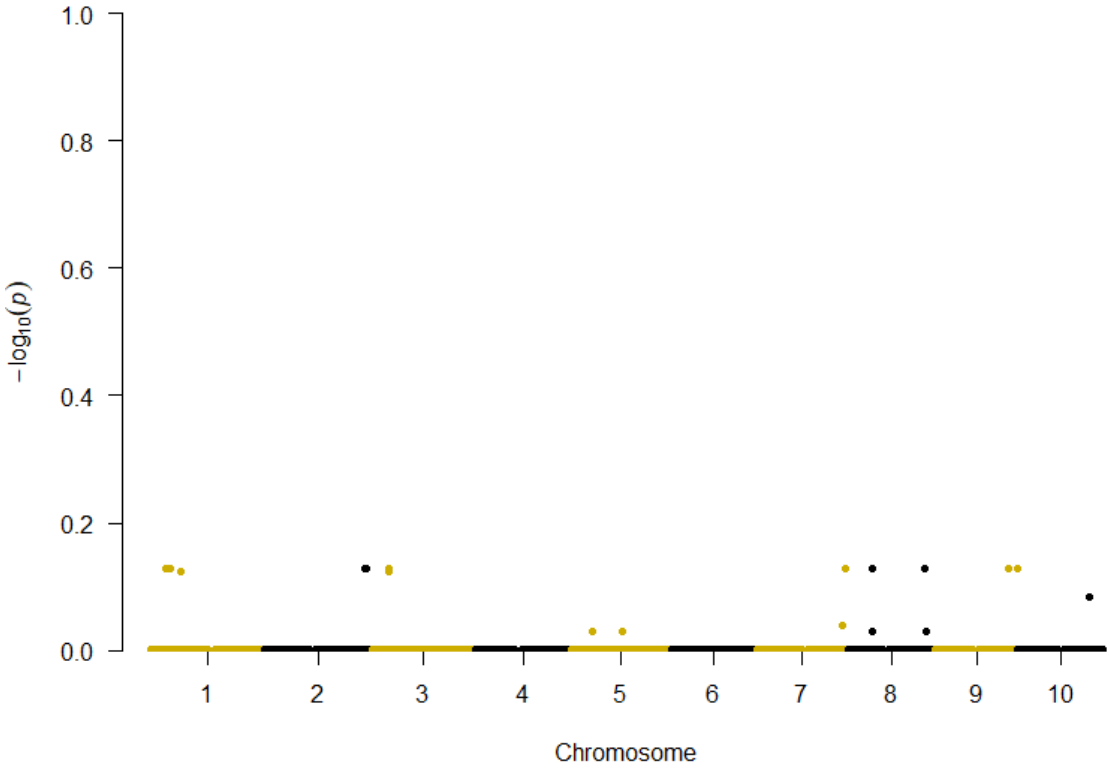

Figure 28. Manhattan plot for xylose yield
FDR-adjusted p-values are plotted.

BIBLIOGRAPHY

Addo-Quaye C, Buescher E, Best N, Chaikam V, Baxter I, Dilkes BP (2017) Forward Genetics by Sequencing EMS Variation-Induced Inbred Lines. G3 (Bethesda) **7:** 413-425

Ail SS, Dasappa S (2016) Biomass to liquid transportation fuel via Fischer Tropsch synthesis – Technology review and current scenario. Renewable and Sustainable Energy Reviews **58:** 267-286

Amelework B, Shimelis H, Tongoona P, Laing M (2015) Physiological mechanisms of drought tolerance in sorghum, genetic basis and breeding methods: A review. African Journal of Agricultural Research **10:** 3029-3040

Anca-Couce A (2016) Reaction mechanisms and multi-scale modelling of lignocellulosic biomass pyrolysis. Progress in Energy and Combustion Science **53:** 41-79

Atsumi S, Hanai T, Liao JC (2008) Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. Nature **451:** 86-89

Bain RL, Broer K (2011) Gasification. *In* RC Brown, ed, Thermochemical Processing of Biomass: Conversion into Fuels, Chemicals, and Power, Ed 1. John Wiley and Sons, Ltd., pp 47-77

Barros J, Serrani-Yarce JC, Chen F, Baxter D, Venables BJ, Dixon RA (2016) Role of bifunctional ammonia-lyase in grass cell wall biosynthesis. Nat Plants **2:** 16050

Bartoli C, Roux F (2017) Genome-Wide Association Studies In Plant Pathosystems: Toward an Ecological Genomics Approach. Front Plant Sci **8**

Bates D, Maechler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software **67:** 1-48

Bawazir AAA, Idle DB (1989) Drought resistance and root morphology in sorghum. Plant and Soil **119:** 217-221

Behrendt F, Neubauer Y, Oevermann M, Wilmes B, Zobel N (2008) Direct Liquefaction of Biomass. Chemical Engineering Technology **31:** 667-677

Bettiga M, Bengtsson O, Hahn-Hägerdal B, Gorwa-Grauslund MF (2009) Arabinose and xylose fermentation by recombinant *Saccharomyces cerevisiae* expressing a fungal pentose utilization pathway. Microbial Cell Factories **8**

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

Boyles RE, Cooper EA, Myers MT, Brenton Z, Rauh BL, Morris GP, Kresovich S (2016) Genome-Wide Association Studies of Grain Yield Components in Diverse Sorghum Germplasm. Plant Genome **9**

Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem **72:** 248-254

Brenton ZW, Cooper EA, Myers MT, Boyles RE, Shakoor N, Zielinski KJ, Rauh BL, Bridges WC, Morris GP, Kresovich S (2016) A Genomic Resource for the Development, Improvement, and Exploitation of Sorghum for Bioenergy. Genetics **204:** 21-33

Bridgwater AV (2011) Upgrading Fast Pyrolysis Liquids. *In* RC Brown, ed, Thermochemical Processing of Biomass: Conversion into Fuels, Chemicals, and Power, Ed 1. John Wiley and Sons, Ltd, pp 157-199

Browning BL (1967) Methods of Wood chemistry. Wiley-Interscience, New York

Browning BL, Browning SR (2016) Genotype Imputation with Millions of Reference Samples. Am J Hum Genet **98:** 116-126

Cai H, Dunn JB, Wang Z, Han J, Wang MQ (2013) Life-cycle energy use and greenhouse gas emissions of production of bioethanol from sorghum in the United States. Biotechnology for Biofuels **6:** 141

Chaturvedi V, Verma P (2013) An overview of key pretreatment processes employed for bioconversion of lignocellulosic biomass into biofuels and value added products. 3 Biotech **3:** 415-431

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

Corredor DY, Salazar JM, Hohn KL, Bean S, Bean B, Wang D (2009) Evaluation and characterization of forage Sorghum as feedstock for fermentable sugar production. Appl Biochem Biotechnol **158:** 164-179

Couture JJ, Singh A, Rubert-Nason KF, Serbin SP, Lindroth RL, Townsend PA (2016) Spectroscopic determination of ecologically relevant plant secondary metabolites. Methods in Ecology and Evolution

Davison BH, Drescher SR, Tuskan GA, Davis MF, Nghiem NP (2006) Variation of S/G ratio and lignin content in a Populus family influences the release of xylose by dilute acid hydrolysis. Applied Biochemistry and Biotechnology **130:** 427-435

Dayton DC, Turk B, Gupta R (2011) Syngas Cleanup, Conditioning, and Utilization. *In* RC Brown, ed, Thermochemical Processing of Biomass: Conversion into Fuels, Chemicals, and Power. John Wiley and Sons, Ltd., pp 78-123

Deu M, Rattunde F, Chantereau J (2006) A global view of genetic diversity in cultivated sorghums using a core collection. Genome **49:** 168-180

Dien BS, Sarath G, Pedersen JF, Sattler SE, Chen H, Funnell-Harris DL, Nichols NN, Cotta MA (2009) Improved Sugar Conversion and Ethanol Yield for Forage Sorghum (Sorghum bicolor L. Moench) Lines with Reduced Lignin Contents. BioEnergy Research **2:** 153-164

Dupain X, Krul RA, Schaverien CJ, Makkee M, Moulijn JA (2006) Production of clean transportation fuels and lower olefins from Fischer-Tropsch Synthesis waxes under fluid catalytic cracking conditions. The potential of highly paraffinic feedstocks for FCC. Applied Catalysis **63:** 277-295

E. Sattler SaPNaSAaGAaXZaSGaVWaFPJ (2012) Identification and Characterization of Four Missense Mutations in Brown midrib 12 (Bmr12), the Caffeic O-Methyltranferase (COMT) of Sorghum. **5**

Elliott DC (2007) Historical Developments in Hydroprocessing Bio-oils. Energy and Fuels **21:** 1792-1815

Elliott DC (2011) Hydrothermal Processing. *In* RC Brown, ed, Thermochemical Processing of Biomass: Conversion into Fuels, Chemicals, and Power, Ed 1. John Wiley and Sons, Ltd., pp 200-231

Elliott DC, Biller P, Ross AB, Schmidt AJ, Jones SB (2015) Hydrothermal liquefaction of biomass: Developments from batch to continuous process. Bioresource Technology **178:** 147-156

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE **6**

Evans RJ, Milne TA (1987) Molecular Characterization of the Pyrolysis of Biomass 1. Fundamentals. Energy and Fuels **1:** 123-137

Fahrenkrog AM, Neves LG, Resende MF, Jr., Vazquez AI, de Los Campos G, Dervinis C, Sykes R, Davis M, Davenport R, Barbazuk WB, Kirst M (2017) Genome-wide association study reveals putative regulators of bioenergy traits in Populus deltoides. New Phytol **213:** 799-811

Fontaine AS, Bout S, Barriere Y, Vermerris W (2003) Variation in cell wall composition among forage maize (Zea mays L.) inbred lines and its impact on digestibility: analysis of neutral detergent fiber composition by pyrolysis-gas chromatography-mass spectrometry. J Agric Food Chem **51:** 8080-8087

Fornale S, Capellades M, Encina A, Wang K, Irar S, Lapierre C, Ruel K, Joseleau JP, Berenguer J, Puigdomenech P, Rigau J, Caparros-Ruiz D (2012) Altered lignin biosynthesis improves cellulosic bioethanol production in transgenic maize plants down-regulated for cinnamyl alcohol dehydrogenase. Mol Plant **5:** 817-830

Fraser CM, Chapple C (2011) The phenylpropanoid pathway in Arabidopsis. Arabidopsis Book **9:** e0152

Fu C, Xiao X, Xi Y, Ge Y, Chen F, Bouton J, Dixon RA, Wang Z-Y (2011) Downregulation of Cinnamyl Alcohol Dehydrogenase (CAD) Leads to Improved Saccharification Efficiency in Switchgrass. BioEnergy Research **4:** 153-164

Golfier P, Volkert C, He F, Rausch T, Wolf SCe (2017) Regulation of secondary cell wall biosynthesis by a NAC transcription factor from Miscanthus. **1:** e00024-n/a

Harlan JR, De Wet JMJ (1972) A simplified classification of cultivated sorghum. Crop Science **12:** 172-176

Hatfield R, Fukushima RS (2005) Can Lignin Be Accurately Measured? Crop Science **45:** 832-839

Hatton D, Sablowski R, Yung MH, Smith C, Schuch W, Bevan M (1995) Two classes of cis sequences contribute to tissue-specific expression of a PAL2 promoter in transgenic tobacco. Plant J **7:** 859-876

Hu W-J, Harding S, Lung J, Popko JL, Ralph J, Stokke DD, Tsai C-J, Chiang VL (1999) Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. Nature Biotechnology **17:** 808-812

Huber GW, Chheda JN, Barrett CJ, Dumesic JA (2005) Production of Liquid Alkanes by Aqueous-Phase Processing of Biomass-Derived Carbohydrates. Science **308:** 1446-1450

Johnson DB, Moore WE, Zank LC (1961) The Spectrophotometric Determination of Lignin in Small Wood Samples. TAPPI **44:** 793-798

Johnson KL, Jones BJ, Bacic A, Schultz CJ (2003) The Fasciclin-Like Arabinogalactan Proteins of Arabidopsis. A Multigene Family of Putative Cell Adhesion Molecules. *In* Plant Physiol, Vol 133, pp 1911-1925

Johnson KL, Kibble NA, Bacic A, Schultz CJ (2011) A fasciclin-like arabinogalactan-protein (FLA) mutant of Arabidopsis thaliana, fla1, shows defects in shoot regeneration. PLoS One **6:** e25154

Kim WC, Ko JH, Han KH (2012) Identification of a cis-acting regulatory motif recognized by MYB46, a master transcriptional regulator of secondary wall biosynthesis. Plant Mol Biol **78:** 489-501

Klem D, Philipp B, Heinze T, Heinze U, Wagenknecht W (1998) Comprehensive Cellulose Chemistry, Vol 1. Wiley-VCH, Germany

Li M, Pu Y, Ragauskas AJ (2016) Current Understanding of the Correlation of Lignin Structure with Biomass Recalcitrance. Front Chem **4**

Li X, Ximenes E, Kim Y, Slininger M, Meilan R, Ladisch M, Chapple C (2010) Lignin Monomer Composition Affects *Arabidopsis* Cell-Wall Degradability After Liquid Hot Water Pretreatment. Biotechnology for Biofuels **3**

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics **28**

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics **28:** 2397–2399

Liu D, Chen Y, Ding F, Guo T, Xie J, Zhuang W, Niu H, Shi C, Ying H (2015) Simultaneous production of butanol and acetoin by metabolically engineered *Clostridium acetobutylicum*. Metabolic Engineering **27:** 107-114

Liu Y, Wang L, Mao S, Liu K, Lu Y, Wang J, Wei Y, Zheng Y (2015) Genome-wide association study of 29 morphological traits in Aegilops tauschii. Sci Rep **5**

Lois R, Dietrich A, Hahlbrock K, Schulz W (1989) A phenylalanine ammonia-lyase gene from parsley: structure, regulation and identification of elicitor and light responsive cis-acting elements. Embo j **8:** 1641-1648

Lupoi JS, Singh S, Davis M, Lee DJ, Shepherd M, Simmons BA, Henry RJ (2014) High-throughput prediction of eucalypt lignin syringyl/guaiacyl content using multivariate analysis: a comparison between mid-infrared, near-infrared, and Raman spectroscopies for model development. Biotechnol Biofuels **7:** 93

Mace ES, Tai S, Gilding EK, Li Y, Prentis PJ, Bian L, Campbell BC, Hu W, Innes DJ, Han X, Cruickshank A, Dai C, Frere C, Zhang H, Hunt CH, Wang X, Shatte T, Wang M, Su Z, Li J, Lin X, Godwin ID, Jordan DR, Wang J (2013) Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nat Commun **4:** 2320

MacMillan CP, Mansfield SD, Stachurski ZH, Evans R, Southerton SG (2010) Fasciclin-like arabinogalactan proteins: specialization for stem biomechanics and cell wall architecture in Arabidopsis and Eucalyptus. Plant J **62:** 689-703

Mansfield SD, Kang KY, Chapple C (2012) Designed for deconstruction--poplar trees altered in cell wall lignification improve the efficacy of bioethanol production. New Phytol **194:** 91-101

Maunder AB (1999) History of Cultivar Development in the United States. *In* CW Smith, RA Frederiksen, eds, Sorghum: Origin, History, Technology, and Production. John Wiley & Sons Inc., United States, pp 191-223

McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, Sims D, Kennedy M, Amirebrahimi M, Weers BD, McKinley B, Mattison A, Morishige DT, Grimwood J, Schmutz J, Mullet JE (2018) The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. Plant J **93:** 338-354

McKendry P (2002) Energy production from biomass (part 3): gasification technologies. Bioresource Technology **83:** 55-63

Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, Harriman J, Glaubitz JC, Buckler ES, Kresovich S (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci USA **110**

Muchow RC (1998) Nitrogen utilization efficiency in maize and grain sorghum. Field Crops Research **56:** 209-216

Nakano Y, Yamaguchi M, Endo H, Rejab NA, Ohtani M (2015) NAC-MYB-based transcriptional regulation of secondary cell wall biosynthesis in land plants. Front Plant Sci **6:** 288

Naz AA, Reinert S, Bostanci C, Seperi B, Leon J, Böttger C, Südekum K-H, Frei M (2017) Mining the global diversity for bioenergy traits of barley straw: genomewide association study under varying plant water status. GCB Bioenergy **9:** 1356-1369

Ortiz D, Hu J, Salas Fernandez MG (2017) Genetic architecture of photosynthesis in Sorghum bicolor under non-stress and cold stress conditions. J Exp Bot **68:** 4545-4557

Parsell T (2015) A synergistic biorefinery based on catalytic conversion of lignin prior to cellulose starting from lignocellulosic biomass. Green Chemistry **17:** 1492-1499

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A (2009) The Sorghum bicolor genome and the diversification of grasses. Nature **457**

Pauly M, Gille S, Liu L, Mansoori N, de Souza A, Schultink A, Xiong G (2013) Hemicellulose biosynthesis. Planta **238:** 627-642

Penning BW, Sykes RW, Babcock NC, Dugard CK, Held MA, Klimek JF, Shreve JT, Fowler M, Ziebell A, Davis MF, Decker SR, Turner GB, Mosier NS, Springer NM, Thimmapuram J, Weil CF, McCann MC, Carpita NC (2014) Genetic Determinants for Enzymatic Digestion of Lignocellulosic Biomass Are Independent of Those for Lignin Abundance in a Maize Recombinant Inbred Population. *In* Plant Physiol, Vol 165, pp 1475-1487

Penning BW, Sykes RW, Babcock NC, Dugard CK, Klimek JF, Gamblin D, Davis M, Filley TR, Mosier N, Weil CF, McCann MC, Carpita NC (2014) Validation of PyMBMS as a High-throughput Screen for Lignin Abundance in Lignocellulosic Biomass of Grasses. Bioenergy Research **7:** 899–908

Persson S, Wei H, Milne J, Page GP, Somerville CR (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc Natl Acad Sci U S A **102:** 8633-8638

Poovaiah CR, Nageswara-Rao M, Soneji JR, Baxter HL, Stewart CN, Jr. (2014) Altered lignin biosynthesis using biotechnology to improve lignocellulosic biofuel feedstocks. Plant Biotechnol J **12:** 1163-1173

Raes J, Rohde A, Christensen JH, Van de Peer Y, Boerjan W (2003) Genome-wide characterization of the lignification toolbox in Arabidopsis. Plant Physiol **133:** 1051-1071

Rai KM, Thu SW, Balasubramanian VK, Cobos CJ, Disasa T, Mendu V (2016) Identification, Characterization, and Expression Analysis of Cell Wall Related Genes in Sorghum bicolor (L.) Moench, a Food, Fodder, and Biofuel Crop. Front Plant Sci **7:** 1287

Reddy MS, Chen F, Shadle G, Jackson L, Aljoe H, Dixon RA (2005) Targeted down-regulation of cytochrome P450 enzymes for forage quality improvement in alfalfa (Medicago sativa L.). Proc Natl Acad Sci U S A **102:** 16573-16578

Resch MG, Baker JO, Decker SR (2015) Low Solids Enzymatic Saccharification of Lignocellulosic Biomass. *In*. NREL, Golden, CO

Robinson MJ, Banuelos E, Barber WC, Burgess CE, Chau C, Chesser AA, Garrett MH, Goodwin CH, Holland PL, Horne BO, Marrufo LD, Mechalke EJ, Rashidi JR, Reynolds BD, Rogers TE, Sanchez EH, Villarreal JS (1999) Chemical Conversion of Biomass Polysaccharides to Liquid Hydrocarbon Fuels and Chemicals. Preprints of Papers - American Chemical Society Division Fuel Chemistry **44:** 224-227

Rohde A, Morreel K, Ralph J, Goeminne G, Hostyn V, De Rycke R, Kushnir S, Van Doorsselaere J, Joseleau JP, Vuylsteke M, Van Driessche G, Van Beeumen J, Messens E, Boerjan W (2004) Molecular phenotyping of the pal1 and pal2 mutants of Arabidopsis thaliana reveals far-reaching consequences on phenylpropanoid, amino acid, and carbohydrate metabolism. Plant Cell **16:** 2749-2771

Rosler J, Krekel F, Amrhein N, Schmid J (1997) Maize phenylalanine ammonia-lyase has tyrosine ammonia-lyase activity. Plant Physiol **113:** 175-179

Saballos A, Ejeta G, Sanchez E, Kang C, Vermerris W (2009) A Genomewide Analysis of the Cinnamyl Alcohol Dehydrogenase Family in Sorghum [Sorghum bicolor (L.) Moench] Identifies SbCAD2 as the Brown midrib6 Gene. Genetics **181:** 783-795

Saballos A, Sattler SE, Sanchez E, Foster TP, Xin Z, Kang C, Pedersen JF, Vermerris W (2012) Brown midrib2 (Bmr2) encodes the major 4-coumarate: coenzyme A ligase involved in lignin biosynthesis in sorghum (Sorghum bicolor (L.) Moench). The Plant Journal **70:** 818-830

Samuel R, Cao S, Das BK, Hu F, Pu Y, Ragauskas AJY (2013) Investigation of the fate of poplar lignin during autohydrolysis pretreatment to understand the biomass recalcitrance. RSC Advances**:** 5305

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 Phenotype1. Plant Physiology **150:** 584-595

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

Schirmer A, Rude MA, Li X, Popova E, del Cardayre SB (2010) Microbial Biosynthesis of Alkanes. Science **329:** 559-562

Sewalt V, Ni W, Blount JW, Jung HG, Masoud SA, Howles PA, Lamb C, Dixon RA (1997) Reduced Lignin Content and Altered Lignin Composition in Transgenic Tobacco Down-Regulated in Expression of L-Phenylalanine Ammonia-Lyase or Cinnamate 4-Hydroxylase. Plant Physiol **115:** 41-50

Shen H, Poovaiah CR, Ziebell A, Tschaplinski TJ, Pattathil S, Gjersing E, Engle NL, Katahira R, Pu Y, Sykes R, Chen F, Ragauskas AJ, Mielenz JR, Hahn MG, Davis M, Stewart CN, Dixon RA (2013) Enhanced characteristics of genetically modified switchgrass (Panicum virgatum L.) for high biofuel production. Biotechnology for Biofuels **6:** 71

Sluiter A, Hames B, Ruiz R, Scarlata C, Sluiter J, Templeton D, Crocker D (2012) Determination of Structural Carbohydrates and Lignin in Biomass. *In* Laboratory Analytical Procedure. NREL

Somerville C (2006) Cellulose synthesis in higher plants. Annu Rev Cell Dev Biol **22:** 53-78

Sorek N, Yeats TH, Szemenyei H, Youngs H, Somerville CR (2014) The Implications of Lignocellulosic Biomass Chemical Composition for the Production of Advanced Biofuels. BioScience **20:** 1-10

Sorghum: Origin, History, Technology, and Production, (2000). John Wiley & Sons, Inc., New York

Stephens JC, Miller FR, Rosenow DT (1967) Conversion of Alien Sorghums to Early Combine Genotypes. Crop Science **7:** 396

Stewart JJ, Akiyama T, Chapple C, Ralph J, Mansfield SD (2009) The effects on lignin structure of overexpression of ferulate 5-hydroxylase in hybrid poplar. Plant Physiol **150:** 621-635

Studer MH, DeMartini JD, Davis MF, Sykes RW, Davison B, Keller M, Tuskan GA, Wyman CE (2011) Lignin content in natural Populus variants affects sugar release. Proc Natl Acad Sci U S A **108:** 6300-6305

Tak H, Negi S, Ganapathi TR (2017) Overexpression of MusaMYB31, a R2R3 type MYB transcription factor gene indicate its role as a negative regulator of lignin biosynthesis in banana. PLoS One **12:** e0172695

Thurber CS, Ma JM, Higgins RH, Brown PJ (2013) Retrospective genomic analysis of sorghum adaptation to temperate-zone grain production. Genome Biology **14:** R68

Turhollow AF, Webb EG, Downing ME (2010) Review of Sorghum Production Practices: Applications for Bioenergy. *In*. Oak Ridge National Laboratory, Oak Ridge, Tennessee

Turner SD (2014) qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. bioRxiv

van Vliet O, P.R., Faaij APC, Turkenburg WC (2009) Fischer–Tropsch diesel production in a well-to-wheel perspective: A carbon, energy flow and cost analysis. Energy Conversion and Management **50:** 855-876

Vilela LdF, de Araujo VPG, Paredes RdS, Bon EPdS, Torres FAG, Veves BC, Eleutherio ECA (2015) Enhanced xylose fermentation and ethanol production by engineered Saccharomyces cerevisiae strain. AMB Express **5**

Wang H, Avci U, Nakashima J, Hahn MG, Chen F, Dixon RA (2010) Mutation of WRKY transcription factors initiates pith secondary wall formation and increases stem biomass in dicotyledonous plants. Proc Natl Acad Sci U S A **107:** 22338-22343

Wang SX, Zhu YL, Zhang DX, Shao H, Liu P, Hu JB, Zhang H, Zhang HP, Chang C, Lu J, Xia XC, Sun GL, Ma CX (2017) Genome-wide association study for grain yield and related traits in elite wheat varieties and advanced lines using SNP markers. PLoS One **12:** e0188662

Wilkie KCB (1979) The Hemicelluloses of Grassses and Cereals. Advances in Carbohydrate Chemistry and Biochemistry **36:** 215-264

Xu Z, Zhang D, Hu J, Zhou X, Ye X, Reichel KL, Stewart NR, Syrenne RD, Yang X, Gao P, Shi W, Doeppke C, Sykes RW, Burris JN, Bozell JJ, Cheng MZ, Hayes DG, Labbe N, Davis M, Stewart CN, Jr., Yuan JS (2009) Comparative genome analysis of lignin biosynthesis gene families across the plant kingdom. BMC Bioinformatics **10 Suppl 11:** S3

Yuvraj, Kaur R, Uppal SK, Sharma P, Oberoi HS (2013) Chemical Composition of Sweet Sorghum Juice and its Comparative Potential of Different Fermentation Processes for Enhanced Ethanol Production. Sugar Technology **15:** 305-310

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. Plant Cell **18:** 3158-3170

Zhong R, Lee C, Ye ZH (2010) Global analysis of direct targets of secondary wall NAC master switches in Arabidopsis. Mol Plant **3:** 1087-1103

Zhong R, Ye ZH (2012) MYB46 and MYB83 bind to the SMRE sites and directly activate a suite of transcription factors and secondary wall biosynthetic genes. Plant Cell Physiol **53:** 368-380

Zhu C, Shen T, Liu D, Wu J, Chen Y, Wang L, Guo K, Ying H, Ouyang P (2016) Production of liquid hydrocarbon fuels with acetoin and platform molecules derived from lignocellulose. Green Chemistry **18:** 2165–2174