pleiotropic effects of genes involved in cell wall lignification on agronomic characters

Yongsheng Chen
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd
Part of the Agronomy and Crop Sciences Commons

Recommended Citation
Chen, Yongsheng, "pleiotropic effects of genes involved in cell wall lignification on agronomic characters" (2011). Graduate Theses and Dissertations. 10415.
https://lib.dr.iastate.edu/etd/10415

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Pleiotropic effects of genes involved in cell wall lignification on agronomic characters

by

Yongsheng Chen

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Genetics

Program of Study Committee:
Thomas Lübberstedt, Major Professor
Paul Scott
William Beavis
Phil Becraft
Dan Nettleton

Iowa State University
Ames, Iowa
2011
# TABLE OF CONTENTS

ABSTRACT................................................................................................................................. v

CHAPTER 1. GENERAL INTRODUCTION....................................................................................... 1
  Introduction ............................................................................................................................... 1
  Literature review ..................................................................................................................... 2
  References ............................................................................................................................... 15

CHAPTER 2. MOLECULAR BASIS OF TRAIT CORRELATIONS......................................................... 26
  Abstract ................................................................................................................................. 26
  Trait correlations and plant breeding .................................................................................... 26
  Methods to study the molecular basis of genetic trait correlations ....................................... 28
  Genetic causes of trait correlations ....................................................................................... 32
  Examples for resolution of intragenic polymorphisms as causes of genic pleiotropy in plants 34
  Consequences for plant breeding ......................................................................................... 37
  Acknowledgement ................................................................................................................. 39
  References .............................................................................................................................. 40
  Tables ....................................................................................................................................... 50
  Figures ..................................................................................................................................... 54

CHAPTER 3. IDENTIFICATION OF NOVEL BROWN-MIDRIB (BM) GENES IN MAIZE BY TESTS OF ALLELISM ............................................................................................................. 58
  Abstract ................................................................................................................................. 58
  Introduction and results .......................................................................................................... 59
  Acknowledgements ............................................................................................................... 63
  Authors’ Contributions ......................................................................................................... 63
  References .............................................................................................................................. 64
  Tables ....................................................................................................................................... 67

CHAPTER 4. GENETIC AND PHYSICAL FINE MAPPING OF THE NOVEL BROWN MIDRIB GENE BM6 IN MAIZE TO A 180KB REGION ON CHROMOSOME 2 ........................................................................... 69
  Abstract ................................................................................................................................. 69
  Introduction ............................................................................................................................. 70
  Materials and methods......................................................................................................... 73
  Results ................................................................................................................................... 78
  Discussion ............................................................................................................................... 83
CHAPTER 7. EXTENSIVE GENETIC DIVERSITY AND LOW LINKAGE DISEQUILIBRIUM WITHIN THE COMT LOCUS IN GERMLASM ENHANCEMENT OF MAIZE POPULATIONS ................................................................. 175
Abstract .................................................................................................................. 175
Introduction ............................................................................................................. 176
Materials and methods .......................................................................................... 178
Results .................................................................................................................... 185
Discussion .............................................................................................................. 194
Acknowledgement ................................................................................................. 200
Authors’ Contributions ......................................................................................... 200
References ............................................................................................................ 202
Tables .................................................................................................................... 207
Figures .................................................................................................................. 212

CHAPTER 8. GENERAL CONCLUSIONS ................................................................. 222
ACKNOWLEDGEMENT ......................................................................................... 226
ABSTRACT

Brown midrib (bm) mutants in maize, which are characterized with altered lignin composition, reduced lignin content and thus with enhanced cell wall digestibility (CWD), are often associated with inferior agronomic traits. To understand how the undesirable associations happen would help us better design strategies to manipulate the cell wall lignification for CWD enhancement without sacrificing biomass yield. First, I reviewed the current knowledge and approaches to study the nature of trait correlations. We proposed that dissection of the trait correlations into DNA polymorphism level is beneficial for plant breeding as intergenic, intragenic, or true pleiotropy will have different impact. Secondly, three new bm mutants were identified by allelism test and were designated as bm5, bm6, and bm7. Then I focused on characterization of bm6, which was revealed to increase CWD but suppress plant height in F2 population. With large mapping populations with about 1000 brown F2 plants, its underlying gene was delimited into a ~180kb interval referring to B73 genome, wherein 10 predicted gene models reside. Besides using natural mutants, we also employed candidate association approach, which suggested that the pleiotropic effects of monolignol biosynthetic genes are most likely due to intragenic linkage of quantitative trait polymorphisms (QTPs). Therefore, optimal haplotypes which combine QTPs beneficial for both CWD and biomass yield might exist. In order to build up suitable materials to confirm this finding, we characterized the COMT gene sampled from the germplasm enhancement of maize (GEM). Much higher genetic diversity of COMT alleles at both DNA and predicted amino acid level and extensive lower linkage disequilibrium (LD) were observed in GEM than in inbred lines. The higher genetic variation suggests GEM is a valuable genetic resource to broaden genetic variation for breeding. And the extensive lower LD indicates the higher resolution of association mapping in GEM derived materials.
CHAPTER 1. GENERAL INTRODUCTION

Introduction

Cell Wall Digestibility (CWD) is a major target for both improving silage quality and lignocellulosic ethanol production in maize. Lignification has negative effects on the cell wall digestibility in silage maize and other forage crops (Barrière and Argillier 1993; Casler and Jung 1999). Reduced levels of lignification increase accessibility of cellulosics and hemicellulosics by enzymatic procedures and, therefore, produce more fermentable sugars. Accordingly, increasing CWD by reduced lignin content or altered lignin composition will improve forage quality and lignocellulosic ethanol conversion. However, brown midrib (bm) mutants of maize and sorghum as well as mutants of Arabidopsis in genes involved in lignin monomer biosynthesis and polymerization show that these genes have pleiotropic effects on lignin structure and biomass yield related and agronomic traits. Understanding the nature of pleiotropic effects of genes involved in cell wall lignifications will benefit future forage and bioenergy breeding.

In my dissertation, I reviewed related literatures to provide the background and significance of my research (Chapter 1). Subsequently, the current knowledge on the molecular causes of trait correlations due to either pleiotropy or linkage was reviewed (Chapter 2), which will help breeders to design their strategy when dealing with a pleiotropic quantitative trait locus (QTL). The following chapters used methods mentioned in Chapter 2 to study the pleiotropic effects of genes involved in cell wall lignification. Genetic mutants are valuable resources to dissect genetic correlations. Besides bm1-4 mutants in maize, there are still other bm mutants known as brown midrib star mutants in maize, which we classified against the bm1-4 and each other to search new bm mutants (Chapter 3). Validation and characterization of new bm would advance our knowledge of cell wall lignification, genetic correlation between cell wall digestibility and
biomass yield, and might also obtain new targets for lignin structure modification for forage and biofuel crop breeding. In Chapter 4, we characterized the pleiotropic effect of \( bm6 \) and precisely mapped the underlying gene into a small physical interval, which will facilitate the isolation of the \( bm6 \) gene. Alternatively, candidate gene association analysis is promising to find causative polymorphisms for complex traits of interest and dissect the genetic correlations at the level of polymorphisms. Previous studies (Guillet-Claude et al. 2004a; Andersen et al. 2007, 2008) revealed candidate causative polymorphisms for CWD in monolignol biosynthetic genes. A similar approach was employed in my research to associate polymorphisms in monolignol biosynthesis genes and biomass yield-related and agronomic traits, to answer whether these genes are pleiotropic with regard to CWD and biomass at the level of polymorphisms (Chapter 5 and Chapter 6). These association studies indicated that the pleiotropic effect in monolignol genes might be due to intragenic linkage of causative polymorphisms for different traits. Therefore, optimal haplotypes can be identified in exotic germplasm for maximizing the forage and bioethanol yield per unit land area. \( bm3 \) gene (COMT) was shown to have the strongest effect on cell digestibility and plant height among known \( bm \) mutants. In addition, it might be a single copy gene in maize genome (Guillaumie et al. 2007a). Thus it is a good candidate gene for exploiting genetic diversity, searching for optimal haplotypes and validating its phenotypic effects in \textit{vivo}. Therefore, we randomly sampled COMT alleles from genetic enhancement of maize, and first analyzed its genetic diversity, linkage disequilibrium in exotic races (Chapter 7). Finally, I summarized my dissertation and provided general conclusions based on my research (Chapter 8).

\textbf{Literature review}

In this section, I will review current status and future perspectives of biofuel and silage corn
production, how cell wall digestibility affects both lignocellulosic traits, and how to improve cell wall digestibility by manipulation of lignin structure. Subsequently, I will describe the genes involved in lignin biosynthesis as well as their pleiotropic effect in relation to cell wall digestibility and biomass related traits.

**Current and future biofuels industry**

The development of clean and renewable biofuels from abundant, domestic biomass is currently receiving considerable attention due to growing concerns over energy security and climate change. Currently, 90% of biofuels is biodiesel and bioethanol (Röttig et al. 2010). Current biofuels are called the first generation biofuels, as they are mainly produced from food (Sims et al. 2010). Specifically, bioethanol is mainly produced from sugar and corn starch (Demain 2009), and biodiesel is produced from vegetable oils or animal fats (Marchetti et al. 2007). Increasing the yield of biofuels requires more land use for biofuel crops, which will compete with food production.

To produce profitable biofuel, abundant biomass should be produced with low agricultural input on land with low agricultural value, and the biomass should be converted into biofuels with low input energy (Hill et al. 2006). Current corn based bioethanol as well as soybean based biodiesel does not satisfy these criteria. Communities have realized that producing bioethanol from starch is not attractive on a long term perspective (Somerville and Milne 2005). Alternatively, the second generation biofuels produced from lignocelluloses, e.g., agricultural and forest residues, and dedicated bioenergy crops (Sims et al. 2010), is a very attractive substitute for petroleum fuel as well as the first generation biofuels in future. Its promising future is reflected in the Energy Independence and Security Act of 2007 (Public Law 110-140-Energy Independence and Security) which defined a 36 billion gallon target for renewable fuels by 2022
in USA, with 21 billion gallons coming from non-corn-starch feedstocks. The European Union also set a goal to substitute 10% of traditional fuel with bioethanol by 2020 (The European plan on climate change).

Among diverse lignocellulosic biomass as feedstock for bioethanol production, maize stover is attractive, as it does not conflict with food production, and abundantly available. For example, 80-100 million dry tons of stover could be collected per year in the U.S., of which most is available for ethanol plants (Kadam and McMillan 2003). Maize accumulated extensive genetic resources, including exotic germplasm and various mutants, and rich sequence information, e.g., abundant EST, gene expression profile, and whole genome sequences (Schnable et al. 2009), which makes it a good model for developing biofuel crops. Undoubtedly, corn stover will play an important role in lignocellulosic ethanol production in future. However, biomass is complex and recalcitrant to conversion (Hayes et al. 2009). As the profit of bioethanol production depends on the price of ethanol relative to other fuels and conversion efficiency (Piston et al. 2010), enhancing corn stover biomass yield and quantity maize stover for bioethanol conversion will also be a major concern for future maize breeding.

**Current and future forage corn silage production**

Besides for grain production, maize is an important forage crop due to consistent quality and higher yield and energy content than other forages, as well as less labor and machinery time needed for harvest and less cost per ton of dry matter than other harvested forages (Lauer 1995). Corn silage is produced by harvesting the whole corn plant a few weeks before maturity and ensiling.

In Europe, corn is widely grown as forage with ~4.6 Mha surface area coverage (Pichon et al. 2006). In the past decades, breeding efforts in Europe achieved a great improvement in whole
plant yield with ~4.4 t/ha overall. This was accompanied with a substantial decrease in cell wall digestibility, which resulted in a reduced feeding value of elite maize hybrids (Barrière et al. 2004a). Selection of alleles for good stalk standability and breakage resistance helped to increase whole plant yield, but also likely eliminated alleles favorable for cell wall digestibility. Alternatively, genetic drift caused elimination of alleles favorable for good cell wall digestibility. Therefore, Barrière et al. (2004a, 2005) advocated exploiting genetic resources for high forage quality maize.

The USA has the largest corn forage production in the world (Lauer et al. 2001). Although there was improvement with 0.13 to 0.16 t/ha per year of forage yield since 1930, no improvement has been achieved for in vitro digestibility of whole plants and cell wall digestibility (Lauer et al. 2001). Based on 7399 samples collected from 1987 to 1993, Ertl and Orman (1994) found limited variability for feed quality in the hybrids and elite breeding materials used in US. Introducing exotic germplasm into inbred lines is promising to increase variability. However, the exotic composition of US corn germplasm is less than 1% in 1983 (Goodman 1984), and only increased slightly to about 3% (Goodman 1999). Increasing the exotic composition of US corn germplasm can avoid the genetic vulnerability of corn. At the north central region plant introduction station (http://www.ars.usda.gov/research/publications/Publications.htm?seq_no_115=226581), an allelic diversity project was initiated to sample the majority of ~250 known races to develop adapted resources for breeding. This project will provide very useful resources to introduce exotic alleles of potential use for forage and other breeding.

To maximize energy gain per unit land, improving forage yield as well as forage quality is necessary. University of Wisconsin (UW) has started a silage breeding program in the US public
sector and has released five inbred lines for forage quality improvement (http://cornbreeding.wisc.edu/). These five cultivars were bred by conventional breeding approaches. Vermerris et al. (2007) created three inbred lines with equal or higher fermentable sugar yield than brown midrib mutants but with good late-season standability and considerably better tolerance to damage caused by Fall Armyworm than standard research inbred lines. Their results proved that it is possible to create maize lines with both great forage quality and good agronomic traits. In future, conventional breeding in combination with genetic manipulation and/or transgenic approaches might facilitate forage quality improvement. Cell wall digestibility is a good target of forage quality improvement. Among many factors affecting cell wall digestibility, lignin content and structure are the most critical ones (Barrière et al. 2003). Several studies have shown that brown midrib mutants of maize have higher intake and digestibility than normal maize (Barrière and Argillier, 1993), but bm hybrids usually have inferior agronomic performance (Pedersen 2005).

**Impact of cell wall lignification on forage and bioethanol production**

Cell wall digestibility has been proven to be a major determinant of forage maize quality (Andrieu et al. 1993; Barrière et al. 2003, 2004b). Lorenz and Coors (2008) found that there are many similarities between improving forage quality and feedstock quality for lignocellulosic ethanol production. Therefore, cell wall digestibility is a major target for both improvement of forage quality (Barrière et al. 2005) and lignocellulosic bioethanol production.

The cell wall is a complex structure and contributes the most to plant biomass, and its major components are celluloses, hemicelluloses, and lignin (Foster et al. 2010). Lignins are likely the only cell wall component resistant to degradation by bacteria and fungi in animal guts (Monties et al. 1989; Monties et al. 1991). High cell wall digestibility can be achieved by reducing lignin
content, and be supported by the brown-midrib3 (bm3) mutant. According to Barrière (2004a), cell wall digestibility was equal to 49.4 % in a set of 31 normal hybrids, while it reached 58.1 % in their isogenic bm3 counterparts whose lignin content was only 2/3 of normal hybrids. Across many publications, the correlations between lignin content and cell wall digestibility varied from –0.3 to –0.9 (Barrière et al. 2003). However, it should be noted that the correlations depend on germplasm background and methods used for estimation of lignin content and cell wall digestibility (Barrière et al. 2003).

Grass cell walls contain p-coumaric (pCA) and ferulic (FA) p-hydroxycinnamic acids, which are intermediate products, or derived from intermediates of the monolignol biosynthetic pathway and account for 3 and 5% of cell wall, respectively (Grabber et al. 2004). It was shown that pCA is negatively correlated with cell wall digestibility and explains a larger portion of cell wall digestibility than lignin content (Méchin et al. 2000; Fontaine et al. 2003). Higher pCA content indicates higher cell wall maturity and a higher proportion of secondary cell walls in tissue, which results in reduced cell wall digestibility (Barrière et al. 2008). Ferulic units tend to esterify with non-cellulosic polysaccharides, for example glucuronoarabinoxylans, via diferulate and oligoferulate bridges, which also link with lignins (Piston et al. 2010). Such cross-linkages have a strong negative effect on cell wall digestibility (Piston et al. 2010). In addition, FA can suppress yeast’s activities on fermentation, which has a negative effect on ethanol production by yeast fermentation (Larsson et al. 2000; Persson et al. 2002).

In the process of conversion of lignocellulosic biomass into ethanol, the following steps are required (Hamelinck et al. 2005): pretreatment to clean and size biomass feedstock, to damage cell wall structure; hydrolysis of polysaccharides to free sugars; fermentation of the free sugars to ethanol; and finally, distillation to produce pure ethanol. Several properties of plant materials,
e.g., amount, composition, and structure of saccharides and lignin, have been proposed or shown to affect the efficiencies of each of these steps (Chang et al. 2000). Cell lignification hampers the accessibility of celluloses and hemicelluloses by enzymatic procedures, thus the lignin content and composition substantially influences the conversion of cellulose into fermentable sugar (Saballos et al. 2008). Theoretical maximum ethanol yields from biomass are highly correlated \( (r^2=0.9) \) with acid detergent lignin concentration (Isci et al. 2008). Removing lignin by oxidative pretreatment could significantly increase the release of available sugars in subsequent enzyme hydrolysis compared to the untreated control (Duncan et al. 2009). Vermerris et al. (2007) reported that un-pretreated stover of bm1 and bm3 isogenic lines in A619 background released 40-50% more glucose than A619 most likely due to reduced lignin content. With pretreatment, which destroyed the lignin’s barrier effect, the bm1 and bm3 isogenic lines as well as controls produce almost equal glucose. In maize, a 1% increase in available cellulose is expected to increase the potential ethanol production from 101.6 to 103.3 gallons per dry ton of biomass, as calculated using the U.S. Department of Energy’s Theoretical Ethanol Yield Calculator and Feedstock Composition Database (www1.eere.energy.gov/biomass/for_researchers.html).

In addition, other aspects of cell wall lignification like the ratio of S to G units affect cell wall digestibility (Grabber et al. 2004; Fontaine et al. 2003) and, therefore, likely affect ethanol production. For instance, the S to G ratio impacts the efficiency of stover hydrolysis in forage sorghums (Corredor et al. 2009). In normal plants, H units accounts for less than 5% of cell wall, but it substantially impact cell wall properties because it increases the frequency of resistant inter-unit bonds (Cabane et al. 2004). Laccases and peroxidases are thought to be involved in oxidative polymerization of monolignols into lignin in plants (Boerjan et al. 2003; Boudet 2003). Guillet-Claude et al.( 2004b) reported that deficiency of maize peroxidase gene ZmPox3 could
lead to high cell wall digestibility, which indicates that the way of oxidization coupling of monolignols might also impact forage quality and bioethanol production.

**Genes responsible for cell wall lignification**

The phenolic polymer lignin is formed by oxidative coupling of three monolignols: \( p \)-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) by laccases and/or peroxidases (Boerjan et al. 2003). Monolignols are synthesized by the phenylpropanoid pathway, which starts from the deamination of L-phenylalanine by phenylalanine ammonia lyase (PAL) to cinnamic acid, and ends with biosynthesis of three monomers: \( p \)-coumaryl alcohol, coniferyl alcohol and sinapyl alcohol by cinnamyl alcohol dehydrogenase (CAD). Between them, other enzymes involved are: cinnamate 4-hydroxylase (C4H), 4-coumarate:CoA ligase (4CL), hydroxycinnamoyl-CoA transferase (HCT), \( p \)-coumarate 3-hydroxylase (C3H), caffeoyl-CoA O-methyltransferase (CCoAOMT), cinnamoyl-CoA reductase (CCR), ferulate 5-hydroxylase (F5H), and caffeic acid O-methyltransferase (COMT).

Most genes responsible for the monolignol biosynthesis belong to gene families, with probable expression of the different members in different tissues, at different stages, and/or driving the biosynthesis of different monomers (Raes et al. 2003). Based on data of Guillaumie et al. (2007a), at least 37 genes are likely involved in monolignol biosynthesis in maize, while only 10 enzymatic steps are needed to convert phenylalanine into monolignols. In each family, the gene number is variable and ranges from one to seven (Guillaumie et al. 2007a), but most likely all members are not yet known in these families in maize. For example, 22 Arabidopsis homologues of the \( HCT \) gene were found in maize by Penning et al. (2009), while only two were found in Guillaumie’s work (2007a).

Laccases and peroxidases are responsible for formation of lignin macromolecules by
oxidative polymerization of three monolignol radicals (Boerjan et al. 2003; Boudet 2003). In the model species Arabidopsis, 17 laccase genes have been annotated (www.Arabidopsis.org) also with different expression of the different members in different tissues, at different stages (McCaig et al. 2005; Cai et al. 2006). In maize, 14 putative laccase genes were identified by homologous sequence and/or conserved laccase motif search (Andersen et al. 2009) and 13 peroxidase genes were annotated in the maize genetic database (http://www.maizegdb.org). Advancement of genomics and biochemistry, forward and reverse genetics approaches will help to find more genes involved in cell wall formation (as well as lignification). The cell wall is a complex matrix and hundreds of genes in many different pathways are involved in its formation. It was estimated that 10% of all plant genes, are responsible for cell wall formation (cellwall.genomics.purdue.edu; Carpita et al. 2001; Yong et al. 2005). In maize, over 700 putative maize cell wall genes have been identified by a maize cell wall database (http://www.polebio.scvs-tlse.fr/MAIZEWALL/index.html). The accomplishment of sequencing B73 genome (Schnable et al. 2009) is great step toward understanding functions of genes and will facilitate discovering mutants of any gene of interest by comparative genomics. Due to the close evolutionary relationship of grasses and their syntenic organization of genomes, maize is a good model species to transfer functions of genes into other grass species such as in bioenergy grasses.

Cell wall mutants are a valuable genetic resource to study functions of cell wall related genes. Based on fourier transform infrared microspectroscopy (Chen et al. 1998), near-infrared (NIR) spectroscopy (Buyck and Thomas 2001) and other spectroscopic approaches, many cell wall biogenesis-related genes have been identified and their information were stored in a cell wall genomics site (cellwall.genomics.purdue.edu) for Arabidopsis, maize, and rice. Recently, Gille et
al. (2009) proposed a forward chemical genetic approach using hydrolases to screen cell wall mutants with altered cross-linking glycan-xyloglucan in Arabidopsis. Applying NIR spectra to 2200 segregating F2 families of the UniformMu population, Penning et al. (2009) found 39 cell wall mutants and confirmed that at least six of these mutants had altered carbohydrate-lignin interactions by pyrolysis-molecular beam mass spectrometry. And two of the 39 mutants had low lignin phenotypes.

In maize, there are four historically known cell wall mutants-brown midrib mutant (\textit{bm1, bm2, bm3, and bm4}), which show brown pigmentation on the lignified tissue of leaves and/or stems resulting from alteration of cell wall composition around the five leave stage after planting (Barrière et al. 2004b). They have reduced lignin content and altered polymer lignin subunits (Vermerris et al., 2002; Barrière et al. 2004b) and their underlying genes were mapped to maize bins 5.04, 1.11, 4.05, and 9.07 (Vermerris 2009). Characterization of these genes will advance our knowledge of cell wall lignification. However, only the \textit{bm3} gene has been functionally characterized to date. \textit{Bm3} phenotype is caused by structural mutations in the \textit{caffeic acid O-methyltransferase (COMT)} gene (Vignols et al. 1995; Morrow et al. 1997). These mutations lead to reduced COMT activity, which finally results in the reduction of sinapyl residues and elevated 5-hydroxyconiferyl alcohol in the lignin (Lapierre et al. 1988; Provan et al. 1997; Suzuki et al. 1997; Marita et al. 2003). Although Halpin et al. (1998) mapped cinnamyl alcohol dehydrogenase gene (CAD) near \textit{bm1} locus, and found at least one of \textit{CAD} gene showing reduced expression which would explain elevated level of coniferaldehyde in the lignin (Halpin et al. 1998), Vermerris (2009) thought the maize gene orthologous to Arabidopsis ARGONAUTE (\textit{AGO}) gene is the prime candidate gene for \textit{bm1}. The basis of his arguments are: a large set of phenylpropanoid and monolignol biosynthetic genes showing reduced expression in near-
isogenic $bm1$ mutant besides four $CAD$ gene, so do several regulatory genes, e.g. two $MYB$ genes, one $LIM$ gene, and maize $AGO$ gene (Guillaumie et al. 2007b). In addition, Arabidopsis $AGO$ proteins were involved in microRNA-mediated post-transcriptional gene silencing; Moreover, maize $AGO$ gene was mapped to 1 Mb upstream of the $CAD$ gene isolated by Happin et al. (1998) in $bm1$ mutant. Up to now, no information about $bm2$ and $bm4$ gene is available.

Besides four genetically mapped brown midrib mutations, ten new brown midrib mutants which are called as $bm*C$ to $bm*L$ have been discovered (http://maizecoop.cropsci.uiuc.edu/stockcat.php). Previous allelism tests discovered three new brown midrib genes, but these new $bm*$ mutations were not crossed against each other to test whether they belonged to a gene or different genes (Haney et al. 2008). By allele testing the available 10 spontaneous brown midrib mutants against with $bm1$-$4$, we confirmed six new $bm$ mutants, and allelism test between them revealed three new $bm$ genes, which were designated as $bm5$ (Ali et al. 2010), $bm6$ (Ali et al. 2010), and $bm7$ in our lab. These new $bm$ mutants provide additional genetic resources for basic research to understand the cell wall formation, and for forage and biofuels crop breeding.

**Pleiotropic effects of cell wall lignification related genes**

Besides biosynthesis of lignin monomers, the monolignol biosynthetic pathway is involved in biosynthesis of salicylates, coumarins, hydroxycinnamic amides, pigments, UV light protectants, antioxidants, flavonoids, isoflavonoids, anthocyanins, tannins, among others (Dixon and Paiva, 1995; Zabala et al. 2006). Jone (1984) concluded that phenylpropanoid compounds are involved in controlling plant development, growth, xylogenesis, and flowering. For example, chalcone and naringenin, two intermediates in the phenylpropanoid metabolism in plants, inhibit $4CL$ activity (Yun et al. 2001) and suppress the growth of at least 20 annual plant species including maize
(Chen et al. 2004). Flavonoids prevent UV irradiation damages, and accumulated flavonoids could result in suppressed auxin transport and thus reduced plant size (Besseau et al. 2007). The salicylic acid is an important hormone signal and its expression level has substantial effects on pathogen infection response, stress threat, plant fitness, and optimal photosynthesis (Clarke et al. 2004; Heidel et al. 2004; Scott et al. 2004; Martinez et al. 2005; Mateo et al. 2006). Lignin is thought to play important role in plant growth and development, because it is constituents of cell wall and enhances plant rigidity, confers resistance to pathogens and mechanical stress, and is involved in water and nutrition transport (Lewis and Yamamoto 1990; Whetten and Sederoff 1995; Douglas 1996). Therefore, genes involved in cell wall lignification also have effects on other plant development activities, in other words, these genes have pleiotropic effects.

Alterations of lignin biosynthesis in those natural mutants and transgenic plants are often accompanied by remarkable inhibition of plant growth. Generally, reduced lignin content results in weaker stalks, reduced stover and grain yield, and delayed maturity (Pedersen 2005). In maize, brown-midrib (bm) mutants show a decreased lignin content and increased cell-wall digestibility (Barrière and Argillier 1993), but maize bm lines or hybrids show reduced vigor during vegetative growth, a high incidence of stalk breakage at maturity, and decreased grain and stover yield (Miller and Geadelmann 1983 a and b; Ballard et al. 2001; Cox et al. 2001). In other plants, for example tobacco (Pincon et al., 2001; Elkind et al., 1990), Arabidopsis (Derikvand et al. 2008; Franke et al. 2002; Franke et al. 2002; Besseau et al. 2007; Do et al. 2007; Schilmiller et al. 2009), and poplar (Leplé et al. 2007), down regulation or knock outs of monolignol genes result in reduced plant size. Studying T-DNA insertion mutants of 14 laccase genes in Arabidopsis, Andersen et al. (unpublished data) found two mutants- LAC15 and LAC17- with changed monolignol ratios and semi-dwarf or dwarf phenotype, respectively. And T-DNA lines
of Lac17 showed pale seeds, consistent with results of Pourcel et al. (2005). These facts have forced geneticists and breeders to take pleiotropic effect into account, especially on biomass yield, when trying to modifying cell wall by manipulating cell wall lignification related genes.

Background also plays an important role in regulation of the pleiotropic effects. Shi et al. (2006) studied the expression profiles of bm3 mutants by microarray and gene expression subtractive hybridization. They found that up/down regulation of some gene expression as a result of bm3 was affected by the genetic background, while some genes showed the same pattern of regulation regardless of background. Ralph et al. (2004) reported that correlations between lignin content and forage quality can be variable according to the genetic background. Even though, reduced lignin content often resulted in suppressed plant growth, some studies demonstrated different results. After divergent selection for fiber concentration in maize, Wolf et al. (1993) found only weak and inconsistent correlations between lignin content and various agronomic traits. Weller et al. (1985) found no yield difference between bm3 and wildtype isogenic lines. He et al. (2003) developed O-methyltransferase down-regulated maize with a 17% decrease in lignin content, increased digestibility, without effect on dry matter yield. As a result, regulating factors in the background are expected to impact the expression of many genes involved in cell wall lignification and thus their pleiotropic effects. Therefore, searching for new cell wall related genes should not rule out regulatory genes.
References


Biochem Biotechnol 158: 164-179


means of a forward chemical genetic approach using hydrolases. Proc Natl Acad Sci 106: 14699-146704


BMC Genet 5: 19


Lapierre C, Monties B, and Rolando C (1988) A new type of constitutive unit in lignins from the
corn bm3 mutant. C R Acad Sci 3: 723-728


Lauer J (1995) Corn germplasms for silage uses. Field crops 28.5-4


Monties BL (1991) Plant cell walls as fibrous lignocellulosic composites Relations with lignin structure and function, Anim Feed Sci Technol, 32 159–175


Pedersen JF, Vogel KP, Funnell DL: Impact of reduced lignin on plant fitness.. Crop Sci 2005, 45:


characterization of the lignification toolbox in arabidopsis. Plant physiol 133: 1051-1071


wall composition and plant development in a dose dependent manner. Heredity 88: 450-457


CHAPTER 2. MOLECULAR BASIS OF TRAIT CORRELATIONS

Yongsheng Chen\textsuperscript{1,2,3}, Thomas Lübberstedt\textsuperscript{1,4}

1: Department of Agronomy, Iowa State University, Ames, IA 50011-1010, USA
2: Interdepartmental Genetics Graduate Program, Iowa State University, Ames, IA, 50011, USA
3: Primary researcher and author
4: Author for correspondence

Modified from the paper published in Trends in Plant Science
Reprinted with permission of Trends in Plant Science. 2010, 15: 454-461

Abstract

Trait correlations are common phenomena in biology. Plant breeders need to consider trait correlations to either improve correlated traits simultaneously or to reduce undesirable side-effects when improving only one of the correlated traits. Pleiotropy or close linkage are the two major reasons for genetic trait correlations and often confounded at the level of quantitative trait loci (QTL) or genes. With the progress of genetic and genomic approaches, discrimination of intragenic linkage from true pleiotropy (see definition of linkage and true pleiotropy in Table 1) is increasingly possible. This will substantially impact breeding strategies and will be helpful to understand the nature of trait correlations.

Trait correlations and plant breeding

Complex traits do not evolve independently but show correlations during adaptation to new environments (Darwin 1859; Armbruster and Schwaegerle 1996; Bidart-Bouzat et al. 2004; McGlothlin and Ketterson 2008). Thus, breeders have to consider correlated traits in variety development. When the goal is to simultaneously increase the trait values of two negatively correlated characters, the negative correlation is undesirable from a breeder’s perspective.
Examples include negative correlations between grain yield and protein content in maize (*Zea mays* L.) (Duvick and Cassman 1999), durum wheat (*Triticum durum* L.) (Rharrabti et al. 2001), and soybean (*Glycine max.* L.) (Rotundo et al. 2009), as well as between cell wall digestibility and biomass yield for forage maize (Chen et al. 2010). Similarly, positive correlations can be undesirable, if the goal is to increase the trait value for one and to decrease it for the other trait. Such undesirable positive correlations exist for high oil and beta-glucan concentrations in oats (*Avena sativa* L.) used for either milling or feeding (Yan and Frégeau-Reid 2008). However, trait correlations can also be valuable and exploited to improve a trait, which is difficult or expensive to determine, by indirect selection for a correlated trait. For example, the anthesis-silking interval (ASI) has been advocated as an indirect selection criterion for improving grain yield under drought stress (Magorokosho et al. 2003).

Genetic trait correlations can be due to pleiotropy or linkage disequilibrium. The predominant genetic basis of trait correlations is controversial. According to the authors of Ref. (Falconer 1960), pleiotropy is more likely to be the reason for trait associations than linkage. By contrast, the authors of Ref. (Mather and Jinks 1971) advocated that associations between characters are more likely due to linkage. Progress of genetic and genomic approaches increasingly enables the discrimination of linkage from pleiotropy, not only at the gene but also at the sequence polymorphism level. All so far reported cases of genic pleiotropy (see definition of pleiotropy in Table 1) in animals (flies, mice, and humans) were due to intragenic linkage of quantitative trait polymorphisms (see definition of QTP and intragenic QTP in Table 1) (Mackay 2009). Our three objectives are to review (i) methods for dissection of genetic trait correlations, (ii) successful examples of discrimination between intragenic linkage and true pleiotropy with impact on trait correlations in plants, and (iii) to discuss the consequences of either true
pleiotropy or linkage of QTP underlying trait correlations in plant breeding.

**Methods to study the molecular basis of genetic trait correlations**

Dissection of pleiotropic quantitative trait loci (QTL) (see definition of QTL in Table 1) into quantitative trait genes (see definition of QTG in Table 1) and finally into quantitative trait polymorphisms (QTP), contribute to understanding the molecular basis of trait correlations (Figure 1a). QTL mapping studies dealing with two or more traits often detected various pleiotropic QTL (Sibov et al. 2003; Zygier et al. 2005; Hall et al. 2006; Clark et al. 2006; Dudley et al. 2007; Zhang et al. 2008; Chen et al. 2008; Wang et al. 2008). When investigating 1120 trait pairs, the authors of Ref. (Gardner et al. 2007) found a trend that significantly correlated traits (average $r_G=0.46$, see see definition of QTP of $r_G$ in Table 1) shared a higher number of common QTL than traits without significant genetic correlations. The more tightly two traits were correlated, the more QTL they had in common. However, QTL usually span 10-20 centiMorgan intervals and are characterized by a limited genetic resolution (Doerger 2002; Holland 2007) and genetic background effects due to segregating regions throughout the genome. Near isogenic lines (NILs, see definition of NILs in Table 1) obtained after several generations of backcrossing are useful to eliminate “genetic noise” for studying the effect of a single genome region or gene on multiple traits. Based on NILs with or without an introgression segment (see definition in Table 1) on chromosome 1 from wild tomato *Lycopersicon hirsutum* acc, the authors of Ref. (Monforte and Tanksley 2000) demonstrated that distinct loci in this pleiotropic region control total yield and fruit shape. More recently, the pleiotropy of a chromosome segment in rice (*Oryza sativa* L.) conferring non-race-specific resistance to rice blast and poor flavor has been dissected to close gene linkage using NILs (Fukuohda et al. 2009). The number of recombination events and thus genetic resolution in target regions can be substantially increased by fine mapping. For
example, based on 2849 F$_2$ progeny in diploid wheat (*Triticum monococcum* L.), the vernalization gene VRN2 was delimited to a region including three linked genes (Yan et al. 2004). However, whereas traditional QTL mapping approaches can resolve linked genes as causes of pleiotropic QTL, excessively large populations would be required to resolve QTP within genes.

Candidate gene based approaches complement forward genetic methods to further dissect trait correlations. In addition to positional candidates, further candidate genes for correlated traits might be based on physiological, biochemical or functional knowledge, or on transcriptional connectivity (Ayroles et al. 2009). The effect of a gene on correlated traits can be revealed by transforming the expressed dominant allele into recessive plants (complementation approach: Figure 1a). For example, tobacco mosaic virus resistance controlled by a single dominant gene was associated with reduced yield and quality in earlier studies (Holmes et al. 1938; Chaplin et al. 1966; Chaplin and Mann 1978). By transforming an isolated resistance gene to a susceptible tobacco (*Nicotiana tabacum* L.) genotype, increased resistance did not reduce yield and quality (Lewis et al. 2007). The authors of Ref. (Lewis et al. 2007) concluded that linkage drag (see definition in Table 1) was the basis of the earlier reported trait associations. However, this complementation approach is restricted to resolving trait correlations at the gene level.

Genic pleiotropy has been revealed by various mutants, providing *in vivo* evidence (Figure 1a). Examples are the *brown midrib* 1 and 3 genes in maize, which affect both lignin composition and content as well as plant fitness traits, such as plant height, flowering time, and biomass yield (Pedersen et al. 2005). However, comparing knock-out/-down and wildtype alleles has the limitation that the whole gene is considered as pleiotropic unit. Furthermore, linkage drag as molecular basis of trait correlations cannot be excluded, when comparing isogenic lines.

Association mapping, “Targeting Induced Local Lesions IN Genomes” (TILLING), and gene
replacement allow not only dissection of trait correlations at the gene level but also discrimination between true pleiotropy and intragenic linkage of QTP causing trait correlations. Association mapping takes advantage of accumulated historic recombination events in natural populations and is promising for identifying causative polymorphisms for complex traits (Risch and Merikangas 1996; Nordborg and Tavaré 2002), which depends on the extent of linkage disequilibrium (LD; see definition of LD in Table 1) in the materials studied. LD might persist over long distances in self-pollinated species like wheat (Zhu et al. 2008), while it often decays within genes of interest in diverse populations of maize (Remington 2001). Association studies can be candidate gene-based or conducted at a genome wide scale. Genome wide association analysis is currently hampered by the limited availability of high density marker systems in crops, but has been successfully used in Arabidopsis thaliana (Atwell et al. 2010). Following the pioneering candidate gene association study of Dwarf8 (Thornsberry et al. 2001) in maize, this approach has been applied to various crop species including rice, wheat, sorghum (Sorghum bicolor L.), barley (Hordeum vulgare L.), and soybean (Glycine max L.) (Zhu et al. 2008). Association mapping will benefit from next generation sequencing technology (Varshney 2009), which facilitates resequencing and large-scale marker development. In addition, projects like the maize nested association mapping (Yu et al. 2008) and 1001 genomes project for Arabidopsis thaliana (Weigel and Mott 2009) will lead to accumulation of further characterized pleiotropic genes and QTP. However, due to the risk of false positives or negatives in association mapping studies (Yu and Buckler 2006), further validation of pleiotropic QTP in vivo is required.

TILLING provides allelic series of point and short insertion deletion mutations in target genes (Barkley and Wang 2008). It has two advantages compared to most other methods employed in functional genomics. First, it can be applied to any species including those, where
transformation or an endogenous transposon system is not available. Second, the effects of mutations are demonstrated *in vivo*. Although this technique has the potential to resolve trait correlations at both the gene and sequence polymorphism level, low mutation rates limit its application for resolving trait correlations at sequence level. Since the number of mutations per target gene reported in TILLING literature ranges from 1 (Greene et al. 2003) to 106 (Slade et al. 2005), it is unlikely to find mutations for any site of interest within a given gene without substantially increasing the number of lines currently used in TILLING collections. The number of lines required to provide a mutation at any site within a gene depends on the mutation rate. The mutation rate in TILLING populations ranges from 2.1/kb to 1/Mb (Barkley and Wang 2008). Thus, 1425 to 2,995,731 lines are required to find at least one mutation at a specific site with 95% certainty (calculated as (1-p)^n≤5%, where n is the number of lines used for mutant screening, p is the average mutation rate at a site). However, an increased mutation rate decreases the viability of lines in TILLING populations. In addition, the number of mutations per line increases in this case, requiring additional “purification” of target site by backcrossing. In contrast, TILLING is a good resource to find truncation (knock-out) mutations in a gene to validate the hypothesis of genic pleiotropy. As 5% of all mutations result in truncation mutations (Greene et al. 2003), on average only 10 mutant lines are required to find one truncation mutation in a 1kb region with a mutation rate of 2.1/kb (calculated by n*2.1*0.05=1; n is the number of lines used for mutant screening).

More recently, efficient methods for gene specific modification (Townsend et al. 2009) and replacement (Shukla et al. 2009) have become available in plants by taking advantage of Zinc-finger nucleases (ZFNs). ZFNs are based on the fusion of zinc-finger-based DNA recognition modules and a *FokI* nuclease domain. The zinc-finger-based DNA recognition modules are
responsible for specific binding, and the FokI nuclease domain for non-specific cleavage after dimerization (Kim et al. 1996). DNA motif specific ZFNs can induce cleavage at a specific site in double-stranded DNA (Porteus and Carroll 2005) and invoke the DNA repair machinery. Specific mutations can be created by non-homologous endjoining, while endogenous gene replacement can be achieved by homologous recombination (HR) after delivery of both ZFNs as well as engineered DNA for the target locus. Combined with high throughput DNA sequencing, this method can be used for targeting any gene (Townsend et al. 2009). With a high frequency of HR rate of about 2% of transformed cells (Townsend et al. 2009), this approach provides a promising way to knock out genes or validate QTP identified by association mapping in vivo.

**Genetic causes of trait correlations**

Two options have been proposed, how a gene might confer genic pleiotropy (Dudley et al. 2005; Van De Peppel and Holstege 2005; He and Zhang 2006). One is a gene with a single function, but involved in multiple biological processes (Figure 2a). The other is a gene with multiple functions that contribute to different traits (Figure 2b). Resolving genic pleiotropy at the level of sequence polymorphisms, a gene with single function (Figure 2a) is most likely due to a pleiotropic QTP, whereas intragenic linkage of QTP seems more likely the underlying cause of genic pleiotropy in case of a gene with multiple functions. An example for a gene with multiple functions is the OmBBBD (*Oryza minuta* bifunctional nuclease in basal defense response) gene from wild rice (*Oryza minuta* J.). Its protein has both RNase and DNase nuclease activities, at least in vitro (You et al. 2010).

Genetic correlations between traits can also be due to correlated physiological functions, where the first trait leads to the second trait or the second trait depends on the first trait (Figure 2c). An example is provided by the monolignol biosynthetic pathway. Mutations in genes
affecting cell wall lignification, such as maize brown midrib genes, generally decrease lignin content. Reduced lignification affects plant stability. Similarly, some traits are hierarchically structured. For example maize grain yield can be divided into the yield component traits - average grain weight and number of grains per hectare. In this example, genes affecting grain yield component traits are likely to show genic pleiotropy with grain yield (Figure 2c).

Another cause of genic pleiotropy is the trade-off between parallel biosynthetic pathways (Figure 2d). An example is the mutation of HCT (hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase) in Arabidopsis, which is responsible for biosynthesis of shikimate and quinate esters of p-coumaric acid and finally for guaiacyl and syringyl monomers, redirecting metabolic flux from monomer biosynthesis branch to flavonoid biosynthesis branch. The resulting accumulation of flavonoids blocks the auxin transport and thus results in dwarf phenotypes (Besseau et al. 2007). Accordingly, a mutation of this gene leads to a pleiotropic effect with reduced lignin content and plant height (Besseau et al. 2007). The above mentioned two causes of genic pleiotropy are most likely due to true pleiotropy.

Finally, genetic correlations can also be due to physical linkage (Figure 2e) and association of non-linked genes. However, only physical linkage based correlations are stable over many generations in absence of genetic drift and selection. Physical linkage of functionally unrelated genes might result from gene duplication, or other chromosome rearrangements. In procaryotes functionally related genes tend to be organized in operons and thus show close physical linkage, whereas in plant genomes this is generally not the case. However, gene duplications with subsequent differentiation of gene function are abundant. One example for duplications followed by differentiation of gene functions is for the two genes encoding Isopimaradiene synthase (\textit{PaIso}) and Levopimaradiene/abietadiene synthase (\textit{PaLAS}) of Norway spruce (\textit{Picea abies}),
respectively. These two genes, involved in distinct defense responses, are two duplicated genes with only 9% amino acid difference (Keeling et al. 2008). Duplication is thought to contribute to evolution of phenotypic novelty in plants (Flagel and Wendel 2009).

**Examples for resolution of intragenic polymorphisms as causes of genic pleiotropy in plants**

A still limited number of studies succeeded in resolving genic pleiotropy at the sequence level, such as for the *Dwarf8* gene in maize. *Dwarf8* affects both flowering time and plant height, as shown by mutagenesis and QTL studies (Peng et al. 1999; Koester et al. 1993; Schön and Melchinger 1994). Mutant studies showed that polymorphisms in the DELLA domain affect plant height (Peng et al. 1999). Nine polymorphisms across the *Dwarf8* gene including one polymorphism in the promoter region were associated with flowering time (Thornsberry et al. 2001), but none of those was located in the DELLA domain. These polymorphisms were not associated with plant height (Schön and Melchinger 1994). One of the above mentioned nine mutations near the SH2 like domain was validated to be flowering related in a larger population (Camus-Kulandaivelu et al. 2006). Moreover, an insertion/deletion (indel) (see definition of indel in Table 1) polymorphism in the *Dwarf8* promoter was associated with plant height but not flowering time (Andersen et al. 2005). It has recently been reported that an amino acid insertion in the VHYNP domain also leads to a maize dwarf phenotype but less severe than other DELLA domain located mutations (Cassani et al. 2009). The *Slender1* gene of barley is orthologous to *Dwarf8* of maize. Three mutants in *Slender1* have been characterized (Chandler et al. 2002). Among them, *Sln1d* has a non-conservative amino acid substitution in the DELLA domain, and shows a dominant dwarf phenotype similar to the *Arabidopsis GAI* mutant (homologue of *Dwarf8*) which has a 17-amino acid deletion in the DELLA domain (Peng et al. 1997). By
contrast, Sln1c with a 17-amino acid deletion in the COOH-terminal is a recessive mutant and showed rapid growth compared with wild type plants. The phenotypes of these two mutants indicate that the polymorphisms resulting in the mutation of Sln1d and polymorphisms in the 17-amino acid deletion region within the COOH-terminal of Sln1c might be QTP for plant height. Taken together, these results suggest that genic pleiotropy of Dwarf8 regarding plant height and flowering time are due to intragenic linkage of QTP. Probably the flowering time and plant height causative polymorphisms in the promoter are responsible for the expression of Dwarf8 at different developmental stages, while polymorphisms within the coding region affect different domains of the Dwarf8 protein. Since Dwarf8 codes for a transcriptional regulator, some mutations might affect DNA binding specificities for only one but not the other trait.

Opaque2 and CyPPDK1 are two functionally related genes that regulate maize kernel traits such as starch, lysine, and protein content (Schmidt et al. 1992; Cord Neto et al. 1995; Kemper 1999; Manicacci et al. 2009). Association analysis (Manicacci et al. 2009) revealed a SNP each in CyPPDK1 (C817) and Opaque2 (O3988), affecting multiple kernel related traits, such as the protein/starch ratio (Table 2). Each gene also contains other candidate QTP, which exclusively affect a single kernel trait. In conclusion, pleiotropic and linked QTP are the molecular basis for the effects of CyPPDK1 and Opaque2 on multiple kernel traits.

The lycopene epsilon cyclase (LcyE) gene affects both, the α- and β-carotene biosynthetic pathways, which are two parallel branches bifurcating after lycopene biosynthesis (Harjes et al. 2008). In the carotenoid biosynthetic pathway, LcyE is the first enzyme of the α-carotene biosynthetic branch. Biosynthesis of β-carotenes is performed in a separate branch of this pathway, which does not involve LcyE. However, four linked QTP in the LcyE gene were responsible for 58% of variation in both branches possibly due to flux redirection among both
Association studies for genes involved in monolignol biosynthesis coding for enzymes such as *COMT* (Caffeic acid *O*-methyltransferase) revealed independent polymorphisms responsible for cell wall digestibility and biomass yield (Chen et al. 2010). Thus, the most likely reason for pleiotropic effects of mutant genes in isogenic line comparisons (Pedersen et al. 2005) was intragenic linkage. Only one indel polymorphism in a gene coding for 4CL1 (4-coumarate:CoA ligase) showed a pleiotropic effect on both cell wall digestibility and biomass yield traits, whereas all other significantly associated polymorphisms affected only one of the two trait groups (Chen et al. 2010). Therefore, the effect of the majority of monolignol biosynthetic genes on either biomass yield or cell wall digestibility traits was due to intragenic linkage of QTP.

Despite the limited number of dissected pleiotropic genes in plants (Table 2), it is obvious that intragenic linked QTP underlying genic pleiotropy are not rare. A pleiotropic QTP is often linked with other QTP, thus both intragenic linkage of QTP and true pleiotropy can simultaneously contribute to genic pleiotropy. Because most genes participate in more than one single pathway or network, they are expected to influence various traits. Pleiotropy of a QTP indicates that a particular domain is crucial for gene function. Linkage of QTP in a pleiotropic gene indicates involvement of distinct domains in different functions or pathways.

In maize, more often than expected large indels are responsible for allelic differences between individuals (Fu and Dooner 2002; Song and Messing 2003; Brunner et al. 2006). Allelic genome regions might differ in the presence or absence of retroelements and complete genes. Retroelements can be induced (Pouteau et al. 1991; Hirochika et al. 1996; Carlos et al. 2001) and might impact the expression of neighboring genes (Llave et al. 2002; Kashkush et al. 2003; Schramke and Allshire 2003). In the case of presence or absence of complete genes, it is
impossible to resolve causative polymorphisms for different traits below the gene level, as the gene itself is the smallest genetic unit. This increases the chance for occurrence of non-resolvable pleiotropic effects on trait correlations in maize. It is currently unknown, if the large extent of major allelic differences is unique to maize, or also present in other plant species.

**Consequences for plant breeding**

Complex traits and their correlations are controlled by multiple genes and their interactions, as well as environmental factors. However, there might be a small number of QTL or genes contributing significantly to trait correlations. Over multiple generations genic pleiotropy and tight gene linkage are most consistently contributing to trait correlations. An example is the rice Ghd7 locus with strong effects on the number of days to heading, number of grains per panicle, plant height, and stem growth (Xue et al. 2008). Understanding the molecular basis of pleiotropic QTL would be valuable for breeding. Generally, if a locus or gene has major effects on two traits, it can be due to (1) pleiotropic QTP or (2) linked QTP, and the effect of the pleiotropic or linked QTP alleles can be (a) desirable or (b) undesirable with regard to their effect on the two traits (Figure 1b).

In scenario 1a (Figure 1b), a pleiotropic QTP allele with desirable effects on two correlated traits, is straightforward from a plant breeding perspective. A functional marker (Andersen and Lübberstedt 2003) derived from this QTP, or closely linked markers might be useful to ensure fixation in breeding materials. By contrast, a pleiotropic QTP with desirable effect on one trait (T1) but undesirable effect on the second trait (T2) (scenario 1b of Figure 1b) requires identification of favorable alleles at other loci to compensate for negative effects of the pleiotropic QTP allele on trait T2. Knowledge of true pleiotropy will help to avoid efforts to break linkage by searching for recombinants. Moreover, given polymorphic alleles at the
pleiotropic QTP site, a breeder can select an allele with desirable effect on the second trait (T2) but undesirable effect on first trait (T1), if it is easier to compensate for the first trait (T1).

Close intragenic linkage of desirable QTP alleles in coupling (scenario 2a of Figure 1b) is comparable to scenario 1a. Intragenic recombination is highly unlikely. In the case of intragenic linkage of desirable QTP alleles in repulsion, breeders can exploit materials with low LD such as genetic resources to identify haplotypes combining favorable QTP alleles for both traits. Alternatively, breeders might create optimal haplotypes in vitro to replace its endogenous homolog by homologous recombination mediated gene replacement or by exploiting appropriately sized TILLING populations.

In the case of the pleiotropic Ghd7 gene: if true pleiotropy contributes to all four above mentioned traits, and breeders want to cultivate a dual-purpose rice (grain for food, straw for lignocellulosic ethanol production) for locations with long growing period (85 days to heading), then a pleiotropic QTP confers desirable trait correlations. A functional marker could be developed, to transfer the optimal allele from “Minghui 63” into other breeding materials. By contrast, true pleiotropy would cause problems, if breeders want to reduce plant height to avoid lodging, or develop a cultivar for locations with short growing period (64 days to heading). Breeders would have to find other genetic factors to compensate for the undesirable correlations caused by this pleiotropic QTP. If genic pleiotropy of the Ghd7 gene is due to intragenic linkage of QTP for different traits, breeders could maintain the “Minghui 63” allele by marker-aided selection of favorite QTP haplotype, given the “Minghui 63” allele combines all desirable QTP alleles. Otherwise, they can screen large germplasm collections, such as the collection from the International Rice Genebank at IRRI (Jackson 1997), to identify a favorable haplotype combining all desirable QTP alleles.
Acknowledgement

Yongsheng Chen was supported by the Interdepartmental Genetics Graduate Program as well as the RF Baker Center for Plant Breeding at Iowa State University. We would like to thank Dr. Susanne C Brink and four anonymous reviewers for their constructive advice to improve this manuscript.
References


Chaplin JF, Mann TJ (1978) Evaluation of tobacco mosaic resistance factor transferred from
barley to flue-cured tobacco. J Hered 69:175–178

Chaplin JF, Matzinger DF, Mann TJ (1966) Influence of the homozygous and heterozygous mosaic-resistance factor on quantitative character of flue-cured tobacco. Tob Sci 10:81–84


Duvick DN and Cassman KG (1999) Post-green revolution trends in yield potential of temperate


Gardner KM and Latta RG (2007) Shared quantitative trait loci underlying the genetic correlation between continuous traits. Mol Ecol 16: 4195-4209


in Lycopene Epsilon Cyclase tapped for maize biofortification. Science 319: 330-333
involved in mutations induced by tissue culture. Proc Natl Acad Sci USA 93: 7783-7788
He X and Zhang J (2006) Toward a molecular understanding of pleiotropy. Genetics 173:1885-
1891
Holland JB (2007) Genetic architecture of complex traits in plants. Curr Opin Plan Biol 10: 156-
161
Holmes FO (1938) Inheritance of resistance to tobacco-mosaic disease in tobacco.
Phytopathology 28: 553–561
Genebank at IRRI. Plant Mol Biol 35: 61-67
the expression of adjacent genes in wheat. Nat Genet 33:102–106
diterpene synthases involved in conifer defense. Proc Natl Acad Sci USA 105:1085-1090
control of lysine-degrading activities in developing maize endosperm. Plant Cell 11:1981-
1994
Kim YG, Cha J, Chandrasegaran S (1996) Hybrid restriction enzymes: zinc finger fusions to Fok
I cleavage domain. Proc Natl Acad Sci USA 93:1156–1160
Koester RP, Sisco PH, Stuber CW (1993) Identification of quantitative trait loci controlling days
to flowering and plant height in two near-isogenic lines of maize. Crop Sci 33:1209–1216


Monforte AJ and Tanksley SD (2000) Fine mapping of a quantitative trait locus (QTL) from Lycopersicon hirsutum chromosome 1 affecting fruit characteristics and agronomic traits: breaking linkage among QTLs affecting different traits and dissection of heterosis for yield. Theor Appl Genet 100: 471-479


Yan L, Loukoianov A, Blechl A, Tranquilli G, Ramakrishna W, SanMiguel P, Bennetzen JL,


### Table 1. Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Insertion / deletion polymorphism (indel)</strong></td>
<td>Allelic DNA variation due to one or several base pairs of insertions or deletions.</td>
</tr>
<tr>
<td><strong>Intergenic linkage of QTP</strong></td>
<td>Linkage of QTP residing in different genes, each gene affecting a different trait. Linked intergenic QTP contribute to trait correlations.</td>
</tr>
<tr>
<td><strong>Intragenic linkage of QTP</strong></td>
<td>Linkage of QTP within a pleiotropic gene, each QTP affecting a different trait. Linked intragenic QTP contribute to trait correlations.</td>
</tr>
<tr>
<td><strong>Introgression segment</strong></td>
<td>A chromosome segment being introduced from a donor to a recipient line.</td>
</tr>
<tr>
<td><strong>Linkage</strong></td>
<td>Colocalization of sequences on the same chromosome.</td>
</tr>
<tr>
<td><strong>Linkage disequilibrium (LD)</strong></td>
<td>Non-random association of alleles at two or more different polymorphic loci no matter they are physical linked or not.</td>
</tr>
<tr>
<td><strong>Linkage drag</strong></td>
<td>When introducing a genome segment into a recipient genome by, e.g., backcrossing, gene(s) closely located with the target gene tend to be co-transmitted due to linkage.</td>
</tr>
<tr>
<td><strong>Near isogenic lines (NILs)</strong></td>
<td>A series of lines which are identical except for one or a few genetic loci or introgression segments.</td>
</tr>
<tr>
<td><strong>Pleiotropy</strong></td>
<td>We discriminate between genic and true pleiotropy. Genic pleiotropy refers to a gene affecting two or more traits simultaneously. Since in this case genes represent the smallest genetic unit, genic pleiotropy can be caused by intragenic linkage of QTP, each affecting different traits, or true pleiotropy. True pleiotropy is due to a QTP affecting two or more traits.</td>
</tr>
<tr>
<td><strong>Quantitative trait locus (QTL)</strong></td>
<td>A genomic region, which usually consists of multiple</td>
</tr>
</tbody>
</table>
genes, contributing to genetic variation of a quantitative trait.

**Quantitative trait nucleotide (QTN):** a nucleotide polymorphism affecting a quantitative trait.

**Quantitative trait polymorphism (QTP):** SNP or indel polymorphism affecting a quantitative trait.

$r_G$: genetic correlation.

**Single nucleotide polymorphism (SNP):** refers to DNA variation between allelic sequences occurring at a single nucleotide site.
Table 2. Resolution of pleiotropic genes into underlying QTP in maize

<table>
<thead>
<tr>
<th>Gene</th>
<th>Affected traits</th>
<th>QTP</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dwarf8</td>
<td>Plant height</td>
<td>DELLA domain</td>
<td>Peng et al. 1999</td>
</tr>
<tr>
<td></td>
<td>Plant height</td>
<td>promoter 2-bp indel</td>
<td>Andersen et al. 2005</td>
</tr>
<tr>
<td></td>
<td>Flowering time</td>
<td>nine polymorphisms(^a)</td>
<td>Thornsberry et al. 2001, Camus-Kulandaivelu et al. 2006, Andersen et al. 2005</td>
</tr>
<tr>
<td>ae1</td>
<td>Pasting temperature</td>
<td>Ae1-1509(G/A) SNP</td>
<td>Wilson et al. 2004</td>
</tr>
<tr>
<td></td>
<td>Kernel amylose content</td>
<td>Ae1-1689(T/C) SNP</td>
<td>Wilson et al. 2004</td>
</tr>
<tr>
<td>bt2</td>
<td>Kernel oil and protein level</td>
<td>Bt2-925(C/T) SNP(^b)</td>
<td>Wilson et al. 2004</td>
</tr>
<tr>
<td>zag1l</td>
<td>Female ear length, Leaf number, Number of fruitcases</td>
<td>zag1l.1 SNP(^b)</td>
<td>Weber et al. 2008</td>
</tr>
<tr>
<td></td>
<td>Leaf number, Percentage of nondisarticulating fruitcases</td>
<td>PZD00020.3 SNP(^b)</td>
<td>Weber et al. 2008</td>
</tr>
<tr>
<td></td>
<td>Percentage of nondisarticulating fruitcases</td>
<td>PZD00021.2 SNP</td>
<td>Weber et al. 2008</td>
</tr>
<tr>
<td></td>
<td>Tassel branch number</td>
<td>PZD00019.1 SNP</td>
<td>Weber et al. 2008</td>
</tr>
<tr>
<td></td>
<td>Leaf number</td>
<td>PZD00020.4 SNP</td>
<td>Weber et al. 2008</td>
</tr>
<tr>
<td>te1</td>
<td>Female ear length, Number of fruitcases</td>
<td>te1.3 SNP(^b)</td>
<td>Weber et al. 2008</td>
</tr>
<tr>
<td></td>
<td>Female ear length, Number of fruitcases</td>
<td>PZD00006.1 SNP(^b)</td>
<td>Weber et al. 2008</td>
</tr>
<tr>
<td></td>
<td>Number of fruitcases</td>
<td>PZD00008.3 SNP</td>
<td>Weber et al. 2008</td>
</tr>
<tr>
<td>tb1</td>
<td>Lateral inflorescence branch number, Tassel branch number</td>
<td>tb1.18 SNP(^b)</td>
<td>Weber et al. 2008</td>
</tr>
<tr>
<td></td>
<td>Lateral inflorescence branch number, Branch length</td>
<td>tb1.19 SNP(^b)</td>
<td>Weber et al. 2008, Weber et al. 2007</td>
</tr>
<tr>
<td></td>
<td>Branch length</td>
<td>tb1.20 SNP</td>
<td>Weber et al. 2007</td>
</tr>
<tr>
<td></td>
<td>Branch length</td>
<td>tb1.15 SNP</td>
<td>Weber et al. 2007</td>
</tr>
<tr>
<td></td>
<td>Female ear length, Number of fruitcases</td>
<td>tb1.5 SNP</td>
<td>Weber et al. 2007</td>
</tr>
<tr>
<td></td>
<td>Female ear length, Number of fruitcases</td>
<td>tb1.8 SNP</td>
<td>Weber et al. 2007</td>
</tr>
<tr>
<td>ZmGI</td>
<td>Leaf number, Tassel branch number</td>
<td>PZB00049.2 SNP(^b)</td>
<td>Weber et al. 2007</td>
</tr>
<tr>
<td></td>
<td>Tassel branch number</td>
<td>PZB00049.7 SNP</td>
<td>Weber et al. 2007</td>
</tr>
<tr>
<td></td>
<td>Total protein weight, Total soluble amino acids, etc.</td>
<td>C817 SNP(^b)</td>
<td>Manicacci et al. 2009</td>
</tr>
<tr>
<td></td>
<td>Protein/starch ratio, single amino acid concentrations, etc.</td>
<td>C1060 SNP(^b)</td>
<td>Manicacci et al. 2009</td>
</tr>
<tr>
<td></td>
<td>Threonine concentration in kernel</td>
<td>CIDP33 SNP</td>
<td>Manicacci et al. 2009</td>
</tr>
<tr>
<td></td>
<td>Kernel dry matter content, Glc+Fru concentration in kernel</td>
<td>C1060 SNP(^b)</td>
<td>Manicacci et al. 2009</td>
</tr>
<tr>
<td></td>
<td>Ashes, Starch weight, Protein/starch ratio, Lysine Concentration in kernel</td>
<td>P3988 SNP(^b)</td>
<td>Manicacci et al. 2009</td>
</tr>
<tr>
<td></td>
<td>Endosperm to kernel in percentage of dry matter, Endosperm weight, etc.</td>
<td>OP1496, OP1539-2 SNP</td>
<td>Manicacci et al. 2009</td>
</tr>
<tr>
<td></td>
<td>Endosperm to kernel in percentage of dry matter</td>
<td>OP1600 SNP</td>
<td>Manicacci et al. 2009</td>
</tr>
<tr>
<td></td>
<td>Neutral detergent fiber, In vitro digestibility of organic matter, days from sowing to silking</td>
<td>810 indel(^b)</td>
<td>Andersen et al. 2008, Chen et al. 2010</td>
</tr>
<tr>
<td></td>
<td>Days from sowing to silking</td>
<td>454 indel</td>
<td>Chen et al. 2010</td>
</tr>
<tr>
<td>FSH</td>
<td>Neutral detergent fiber</td>
<td>5-6 SNP(^e)</td>
<td>Andersen et al. 2008</td>
</tr>
<tr>
<td></td>
<td>Digestibility of neutral detergent fiber</td>
<td>610 SNP</td>
<td>Andersen et al. 2008</td>
</tr>
<tr>
<td></td>
<td>Neutral detergent fiber</td>
<td>817 SNP</td>
<td>Andersen et al. 2008</td>
</tr>
<tr>
<td></td>
<td>Days from sowing to silking</td>
<td>65 SNP</td>
<td>Chen et al. 2010</td>
</tr>
<tr>
<td></td>
<td>In vitro digestibility of organic matter, Digestibility of neutral detergent fiber</td>
<td>294 SNP</td>
<td>Andersen et al. 2008</td>
</tr>
<tr>
<td>C3H</td>
<td>Digestibility of neutral detergent fiber</td>
<td>1667,1674 SNP</td>
<td>Brenner et al. 2010</td>
</tr>
<tr>
<td></td>
<td>Days from sowing to silking</td>
<td>2358 indel</td>
<td>Chen et al. 2010</td>
</tr>
<tr>
<td></td>
<td>In vitro digestibility of organic matter</td>
<td>2358 indel</td>
<td>Chen et al. 2010</td>
</tr>
<tr>
<td></td>
<td>Dry matter content, Cell wall digestibility</td>
<td>75 indel(^b)</td>
<td>Chen et al. 2016, Guillet-Claude et al. 2004</td>
</tr>
<tr>
<td></td>
<td>Plant height, Days from sowing to silking, Dry matter content</td>
<td>663 indel(^b)</td>
<td>Chen et al. 2010</td>
</tr>
<tr>
<td></td>
<td>Water soluble carbohydrates</td>
<td>344 indel</td>
<td>Brenner et al. 2010</td>
</tr>
<tr>
<td></td>
<td>Neutral detergent fiber, In vitro digestibility of organic matter</td>
<td>890 SNP(^b)</td>
<td>Brenner et al. 2010</td>
</tr>
<tr>
<td></td>
<td>Neutral detergent fiber</td>
<td>670 SNP</td>
<td>Brenner et al. 2010</td>
</tr>
<tr>
<td>PAL</td>
<td>Days from sowing to silking</td>
<td>947-1655(^d)</td>
<td>Chen et al. 2010</td>
</tr>
<tr>
<td></td>
<td>In vitro digestibility of organic matter</td>
<td>2086 indel</td>
<td>Andersen et al. 2007</td>
</tr>
</tbody>
</table>
The nine polymorphisms include MITE185 indel, C/G677 SNP, 702 indel, 1044 indel, 1633C/T SNP, 1964 indel, 3472 indel, 2490G/A SNP, 3570 C/G SNP.

Pleiotropic QTP

Two polymorphisms with LD=1

17 polymorphisms with LD=1
Figures

**Figure 1.** Dissection of genetic trait correlations and its impact on plant breeding

(a). Dissection of genetic trait correlations

Quantitative genetic methods, including QTL mapping and whole genome association mapping, are promising to identify QTL for each trait as well as pleiotropic QTL contributing to correlated traits. Pleiotropic QTL are resolved at the gene level into either genic pleiotropy or linked genes by either in vitro or in vivo methods. Genic pleiotropy can be further resolved into either true pleiotropy or intragenic linkage of QTP, each contributing to a different trait.

The blue arrows indicate the potential of each method in dissecting genetic trait correlations at the level of QTL, genes, or sequence polymorphisms. T1 and T2 represent trait 1 and trait 2. Ellipses represent QTL. An ellipse with T1/T2 on its top or bottom indicates that this QTL contributes to the respective trait(s). Rectangles represent genes. Letters in rectangles represent nucleotides.

*: Targeting Induced Local Lesions IN Genomes

(b). Breeding strategies

Genic pleiotropy can be due to (1) pleiotropic or (2) intragenic linked QTP, and the effect of the pleiotropic or linked QTP can be (a) desirable or (b) undesirable with regard to their effects on the two traits (T1 and T2 in Figure 1a). In scenario 1a, the nucleotide “T” is favorable for improving two traits simultaneously. In scenario 1b, allele "G" from another gene needs to compensate for negative effects of nucleotide "T". These two QTP alleles "T" and "G" can be used to improve the two traits or at least improve trait 1 without significant side effect on trait 2. In scenario 2a, two beneficial nucleotides "A" and "C" can be used for simultaneous improvement of two traits. In scenario 2b, the beneficial nucleotide "A" is selected and the beneficial or neutral allele "T" at the second QTP site needs to be identified.

+: indicates the nucleotide is beneficial for a trait; -: indicates the nucleotide is undesirable; 0: indicates the nucleotide is neutral for the trait(s) of interest. Blue + or - represents the effect of a nucleotide on the first trait (T1 in Figure 1a); red + or - or 0 represents the effect of a nucleotide on the second trait (T2 in Figure 1a).

**Figure 2.** Molecular basis of genetic correlations at the gene level
Genic pleiotropy (a-d) and physical linkage (e) affect genetic correlations. Genic pleiotropy can be a result of the following four events: (a) a single gene with a single molecular function is involved in multiple molecular processes; (b) a gene has multiple functions and each contributes to different traits; (c) a gene is responsible for a trait which leads to or partly contributes to another trait; (d) variation in a gene responsible for a biosynthetic branch results in metabolic flux redirection between two biosynthesis branches; (e) Physical linkage of genes. Each gene contributes to a different trait.

Star in (d) represents the gene whose variation leads to metabolic flux redirection. Black dashed arrow in (d) designates metabolic flux redirection. Colored dashed arrows in (d) indicate increase (up) and decrease (down) of metabolic flux. Black dot in (e) represents centromere.
a. Dissecting genetic trait correlations

Quantitative genetic methods

Significant genetic correlations

In vivo methods

QTL mapping

Chromosome 1 (T1)

Chromosome 2 (T2)

Inter-chromosomal linkage

Genome wide association

Candidate gene association

T1 Genomic pleiotropy

T2 True pleiotropy

T1 Linked genes

T2 Sequence Polymorphism level

Intragenic linkage of QTP

RNAi

T-DNA mutants

Transposon mutants

TILLING

gene replacement

b. Breeding strategy

1a

Desirable QTP

T

1b

T + G

2a

A + C

2b

A + T

Figure 1
Molecular processes 1, 2, and 3

Figure 2

CHAPTER 3. IDENTIFICATION OF NOVEL BROWN-MIDRIB (BM) GENES IN MAIZE BY TESTS OF ALLELISM

Farhad Ali\textsuperscript{1,6}, Paul Scott\textsuperscript{2,5}, Jehan Bakht\textsuperscript{1}, Yongsheng Chen\textsuperscript{3}, Thomas Lübberstedt\textsuperscript{4}

\textsuperscript{1}: Institute of Biotechnology and Genetic Engineering (IBGE), Khyber Pukhtunkhwa Agricultural University, Peshawar 25000, Pakistan
\textsuperscript{2}: 1407 Agronomy Hall ISU, Ames, Iowa 50011, USA
\textsuperscript{3}: 1529 Agronomy Hall ISU, Ames, Iowa 50011, USA
\textsuperscript{4}: 1204 Agronomy Hall ISU, Ames, Iowa 50011, USA
\textsuperscript{5}: Author for correspondence
\textsuperscript{6}: Primary researcher and author

Modified from the paper published in Plant Breeding. Here we added \textit{bm7} to the published paper. Reprinted with permission of Plant Breeding. 2010, 129: 724-726

Abstract

\textit{Brown midrib (bm)} mutations are known to affect cell wall digestibility by altering the quantity and composition of lignins in cell walls, resulting in higher ethanol yield and increased cell wall digestibility. So far, four \textit{bm} genes (\textit{bm1}, \textit{bm2}, \textit{bm3}, and \textit{bm4}) were identified and mapped in maize, the last one (\textit{bm4}) in 1947. In this study, thirteen spontaneous mutations (\textit{bm*}A-M) resulting in the appearance of brown midribs were crossed with \textit{bm1-4} for tests of allelism. From these tests, we report three new \textit{bm} mutants \textit{bm5} (\textit{bm*F}), \textit{bm6} (\textit{bm*J}), and \textit{bm7} (\textit{bm*I}) while other \textit{bm*} lines were either found allelic to \textit{bm1-4} or to one of the \textit{bm*} lines.
Introduction and results

Maize (Zea mays) is one of the most important cereal and fodder crops world-wide. Maize fulfills 15-20% of the daily calorie needs of people in more than 20 countries in the world (Dowswell et al. 1996). Due to high biomass accumulation, maize stover is also a good source for forage and bio-fuel production.

Globally energy security is one of the major concerns besides food security. Due to the non renewable nature of traditional sources of energy (oil, coal, natural gas), efforts are being made to find other avenues for alternative sources of energy, which are both clean and low cost. Ethanol has emerged to be one of the most important alternate energy resources, especially in USA and Brazil, which are two of the major ethanol producing countries. Global ethanol production in 2009 is expected to be 17.5 billion gallons (World’s Ethanol Production Forecast 2008–2012) which is exclusively based on grain, while efforts are being made to utilize lignocellulosic materials for production of ethanol that will not only be more cost efficient, but in addition not inflate prices of food commodities.

One of the major constraints in ethanol production is the conversion efficiency of biomass into ethanol. The goal is to increase the total amount of available digestible sugars, which is limited by cell-wall lignification. Brown midrib (bm) mutants have been shown to increase glucose yield, while reducing the lignin content of cell walls, yielding more ethanol per unit of biomass (Dien et al. 2009). In addition, bm mutants are known to increase the cell wall digestibility by altering lignin composition and reducing the lignin content of cell walls, thus improving the nutritional value of silage in maize and Sudangrass, when used as a fodder crop (Cherney et al. 1991; Campbell and Sederoff 1996; Casler et al. 2003).

The first bm mutant was observed in dent corn at Saint Paul, Minnesota in 1924 (Lauer and
So far, four *bm* mutations have been characterized: *bm1* (Jorgenson 1931), *bm2* (Burnham and Brink 1932), *bm3* (Emerson et al. 1935), and *bm4* (Burnham, 1947). These mutants show reddish brown pigmentation in leaf midribs. Besides leaves this reddish brown pigmentation is also visible in stalk pith after the plant has attained five expanded leaves (Lauer and Coors 1997; Barrière et al. 2004). Coloring eventually disappears on leaves, but remains in the stalk (Lauer and Coors 1997). Biochemical comparison of these mutants with respective wild types showed not only reduced lignin content in *bm* mutants, but also changes in composition of lignin and cell wall digestibility (Barrière and Argillier 1993). Out of the four *bm* mutants (*bm1*-4), *bm3* strongly affects the phenotypic appearance and is reported to improve cell wall digestibility by 16% in comparison to isogenic lines (Barrière and Argillier 1993). The *bm2* mutation affects tissue specific patterns of lignification (Vermerris and Boon 2001) and has a similar phenotype as that of *bm4* (Marita et al. 2003). The *bm1* mutation results from differential expression of the cinnamyl alcohol dehydrogenase (CAD) gene (Halpin et al. 1998) while the *bm3* mutation results from structural changes in the caffeic acid O-methyltransferase (COMT) gene (Vignols et al. 1995). *bm* mutations are also known in sorghum, sudangrass, and pearl millet (Lauer and Coors 1997).

Beside the four known *bm* mutations (i.e., *bm1*, *bm2*, *bm3*, and *bm4*), additional *bm* mutations have been identified, which were not further characterized and named *bm* mutants. Currently there are thirteen such mutant lines (Maize Genetic Stock Center Designations 5803A-M; abbreviated here *bm*A-M) available in the Maize Stock Centre (Maize Genetic Coop Centre, http://maizecoop.cropsci.uiuc.edu/). These spontaneous mutations were not previously tested against the already known *bm1*-4 mutations by tests of allelism to examine, whether these mutations are allelic to the already identified *bm1*-4 mutations, or these are mutations in different
The main aim of the current study was to test these \( \text{bm}^* \) (\( \text{bm}^*\text{A-M} \)) mutations by tests of allelism against the known \( \text{bm} \) mutants (i.e., \( \text{bm}1-4 \)) to identify possible new \( \text{bm} \) genes.

The procedure used for the test of allelism was the same used by Koncz et al. (1990). Midrib color in maize leaf is a qualitative trait where green/white midrib (wild type) is dominant over brown midrubs of \( \text{bm} \) mutants. Thus, a cross of a homozygous \( \text{bm}^* \) line with any homozygous line carrying \( \text{bm}1, 2, 3, \) or \( 4 \) results in either 100% wild type (green midrib) plants, or 100% brown midrib plants indicating that a particular \( \text{bm}^* \) mutation allelic to \( \text{bm}1, 2, 3, \) or \( 4 \). For most of the mutations, results of the tests of allelism were clear-cut, but for mutations with weaker phenotypes (\( \text{bm}^*\text{E} \) and \( \text{bm}^*\text{H} \)), tests were replicated various times for validation. Seed received for \( \text{bm}^*\text{M} \) was segregating for the brown midrib phenotype. Therefore, only plants showing \( \text{bm} \) phenotype were used for the test of allelism. After testing all \( \text{bm}^* \) mutations against \( \text{bm}1-4 \), tests of allelism were performed among \( \text{bm}^* \) mutations, which differed from \( \text{bm}1-4 \).

A total of thirteen novel \( \text{bm}^* \) mutant lines were provided by the Maize Genetic Stock Centre. Seed from these mutant lines was sown in two rows in the field. Crosses were made between \( \text{bm}^* \) and \( \text{bm}1-4 \) lines for tests of allelism. Due to difference in timing to maturity, crosses were made in both directions using both \( \text{bm}^* \) lines and \( \text{bm}1-4 \) as male and female plants to carry out all desired crosses at the Agronomy farm of Iowa State University, USA. Those \( \text{bm}^* \) lines which were identified as non-allelic to \( \text{bm}1-4 \) were crossed amongst themselves to test, whether these mutants are different from one another or allelic. For each cross, three pollinations were made and seed was harvested separately. After drying, seed for each cross was kept in separate bags. A total of twenty kernels (seven kernels from each bag per cross) from all crosses were sown in pots in the greenhouse. A minimum of fifteen plants per cross were used for scoring. Plants were allowed to grow until appearance of the brown midrib phenotype or maturity, whichever came
earlier. The data for Table 1 were collected over a period of two years in the field and greenhouse and tests were replicated in independent trials. Some of the tests of allelism were performed independently by collaborating scientists, Sarah Hake (USDA, personal communication) and Paul Scott (earlier results obtained prior to this study), and the results were generally in agreement with our findings. In case of discrepancies, additional tests of allelism were performed.

\( bm^A \) and \( bm^B \) have radically different phenotypes than all other \( bm \) mutations (Table 1). They were, therefore, excluded from further analyses. \( bm^C, bm^D \) and \( bm^L \) were allelic to \( bm^1 \). All F1 plants of these mutants with \( bm^1 \) resulted in appearance of brown midribs (Table 1). \( bm^K \) was allelic to \( bm^3 \), and \( bm^M \) to \( bm^2 \). \( bm^E, bm^F, bm^G, bm^H \) and \( bm^J \) were non-allelic to \( bm^1-4 \), and were, therefore, crossed among each other for additional tests of allelism to determine the number of independent and novel \( bm \) mutations (Table 2).

The candidate mutations (i.e., \( bm^E, bm^G, bm^H, \) and \( bm^J \)) for novel \( bm \) genes were crossed pair-wise, to ascertain whether these are allelic or distinct \( bm \) mutants (Table 2). These results clearly identified three novel \( bm \) genes. We designate \( bm^F, bm^J, \) and \( bm^I \) as \( bm^5, bm^6, \) and \( bm^7 \) while \( bm^E, bm^G, \) and \( bm^H \) were all found allelic to \( bm^F \).

Four \( bm \) mutations (i.e., \( bm^1-4 \)) have been known for long time. These mutant lines were introgressed into the same background to evaluate their effect on different agronomic parameters. These mutants have shown a drastic effect on cell wall digestibility by altering lignin content and composition resulting in higher yield of total fermentable sugars (Lechtenberg et al.1972; Barrière and Argillier 1993; Vermerris and Boon 2001). According to US Department of Energy, a 1% increase in total available fermentable sugar will result in 2.21 x 10^9 more gallons of ethanol for the same amount of biomass (Department of Energy). Double mutants were previously used and have shown an additive effect on total available fermentable sugars. For the
*bm1bm3* double mutant, total glucose yield obtained after enzymatic saccharification was twice as high as that of the A619 isogenic parent (Vermerris et al. 2007). These three new mutants may prove to be a valuable resource for development of highly convertible maize stover. Further characterization of the novel brown midrib mutants as isogenic line contrasts as well as by genetic mapping are ongoing.

**Acknowledgements**

We acknowledge close collaboration with Dr. Sarah Hake (Adjunct Professor and Center Director; Plant Gene Expression Center USDA-ARS) for this project. We acknowledge Maize Genetics Cooperation Stock Centre for providing seeds of all *bm* lines and Elizabeth Bovenmyer (department of Agronomy, ISU) for her help and support in the field and green house. We acknowledge HEC (Higher Education Commission of Pakistan) IRSIP program for sponsoring Fahad Ali’s visit to ISU, USA. Yongsheng Chen was supported by the Interdepartmental Genetics Graduate Program as well as the RF Baker Center for Plant Breeding at Iowa State University.

**Authors’ Contributions**

FA is the primary research and author, was involved in F₁ material development and F₁ phenotype evaluation, and wrote this manuscript (chapter 30 which includes *bm5*, *bm6*, and *bm7*) as well as the one published in plant breeding which includes only *bm5* and *bm6*. SP provided mutant stocks, and coordinated this project. JB helped AF for manuscript writing. YC was involved in F₁ material development and F₁ phenotype evaluation. LT conceived the idea, and together with FA wrote the manuscript.
References


Emerson RA, Beadle GW, Fraser AC (1935) A summary of linkage studies in maize. Cornell


Lechtenberg VL, Muller LD, Bauman LF, Rhykerd CL, Barnes RF (1972) Laboratory and in vitro evaluation of inbred and F2 populations of brown-midrib mutants of Zea mays L. Agron J 64: 657


World’s Ethanol Production Forecast 2008-2012
Table 1. Results of tests of allelism between \textit{bm*} (\textit{bm*A-M}) mutations and \textit{bm1-4}

<table>
<thead>
<tr>
<th></th>
<th>\textit{bm1}</th>
<th>\textit{bm2}</th>
<th>\textit{bm3}</th>
<th>\textit{bm4}</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{bm*A}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Shootless, not studied</td>
</tr>
<tr>
<td>\textit{bm*B}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bladeless, not studied</td>
</tr>
<tr>
<td>\textit{bm*C}</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>\textit{bm} 1</td>
</tr>
<tr>
<td>\textit{bm*D}</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>\textit{bm} 1</td>
</tr>
<tr>
<td>\textit{bm*E}</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Not \textit{bm1-4}</td>
</tr>
<tr>
<td>\textit{bm*F}</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>\textit{bm} 5</td>
</tr>
<tr>
<td>\textit{bm*G}</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Not \textit{bm1-4}</td>
</tr>
<tr>
<td>\textit{bm*H}</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Not \textit{bm1-4}</td>
</tr>
<tr>
<td>\textit{bm*I}</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Not \textit{bm1-4}</td>
</tr>
<tr>
<td>\textit{bm*J}</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Not \textit{bm1-4}</td>
</tr>
<tr>
<td>\textit{bm*K}</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>\textit{bm} 3</td>
</tr>
<tr>
<td>\textit{bm*L}</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>\textit{bm} 1</td>
</tr>
<tr>
<td>\textit{bm*M}</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>\textit{bm} 2</td>
</tr>
</tbody>
</table>

+  Allelic  
-  Non-Allelic
Table 2. Tests of allelism among *bm* mutants different from *bm*1-4

<table>
<thead>
<tr>
<th></th>
<th><em>bm</em>E</th>
<th><em>bm</em>F</th>
<th><em>bm</em>G</th>
<th><em>bm</em>H</th>
<th><em>bm</em>I</th>
<th><em>bm</em>J</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>bm</em>F</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><em>bm</em>5</td>
</tr>
<tr>
<td><em>bm</em>G</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><em>bm</em>5</td>
</tr>
<tr>
<td><em>bm</em>H</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td><em>bm</em>5</td>
</tr>
<tr>
<td><em>bm</em>I</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td><em>bm</em>7</td>
</tr>
<tr>
<td><em>bm</em>J</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td><em>bm</em>6</td>
</tr>
</tbody>
</table>

+ Allelic
- Non-allelic
CHAPTER 4. GENETIC AND PHYSICAL FINE MAPPING OF THE NOVEL BROWN MIDRIB GENE BM6 IN MAZIE TO A 180KB REGION ON CHROMOSOME 2

Yongsheng Chen¹,²,⁶, Hongjun Liu¹,³, Farhad Ali⁴, M. Paul Scott⁵, Qing Ji¹, Ursula Karoline Frei¹, Thomas Lübberstedt¹,⁷

¹: Department of Agronomy, Iowa State University, Ames, USA, 50011
²: Interdepartmental Genetics Program, Iowa State University, Ames, USA, 50011
³: Maize Research Institute, Sichuan Agricultural University, Ya’an, China, 625014
⁴: Institute of Biotechnology and Genetic Engineering (IBGE), Khyber Pukhtunkhwa Agricultural University, Peshawar, Pakistan, 25000
⁵: USDA-ARS, Corn Insects and Crop Genetics Research Unit, Ames Iowa, 50011
⁶: Primary researcher and author
⁷: Author for correspondence

A paper to be submitted to Theor Appl Genet

Abstract

Brown midrib mutants in maize are known to be associated with reduced lignin content and increased cell wall digestibility, which leads to better forage quality and higher efficiency of cellulosic biomass conversion into ethanol. Four well known brown midrib mutants, named bm1-4, were identified several decades ago. Additional recessive brown midrib mutants have been identified by allelism tests and designated as bm5, bm6, and bm7. In this study, we determined that bm6 increases cell wall digestibility and decreases plant height. bm6 was assigned by quantitative bulked segregant analysis (BSA) onto the short arm of Chromosome 2 and the map location was confirmed in a small segregating F₂ population derived from crossing B73 and a bm6 mutant line. Subsequently, 960 brown midrib individuals were selected from the
same but larger F$_2$ population for fine mapping. With newly developed markers in the target region, the $bm6$ gene was assigned to a 180kb interval flanked by markers SSR_308337 and SSR_488638. In this region, 10 gene models are predicted in the maize B73 sequence. Analysis of these 10 genes as well as genes in the syntenic rice region revealed that four of them are promising candidate genes for $bm6$. Our study provided tightly linked markers for marker-assisted selection in relation to forage quality and lignocellulosic ethanol conversion of stover, will facilitate isolation of the underlying gene, and advance our understanding of brown midrib gene functions.

**Introduction**

Maize has been cultivated mainly for grain production for thousands of years. It is also used for forage by harvesting the whole above ground maize plant a few weeks before maturity and ensiling. Compared with other forage crops, maize forage has several merits. Its quality is relatively consistent and its yield and energy content are high (Lauer 1995). Moreover, the labor and time required for cultivation and harvesting are lower than for many forage crops, which greatly reduces the cost per unit of dry matter (Lauer 1995). Cell wall digestibility is the key factor determining forage quality (Andrieuet et al. 1993; Barrière et al. 2003, 2004b). In Europe, maize is widely grown as forage crop with ~4.6 Mha surface area coverage (Barrière et al. 2004a). In the past decades, breeding efforts in Europe substantially improved in whole plant yield with ~ 0.1t/ha increase per year overall. However, cell wall digestibility substantially decreased, which resulted in a reduced feeding value of elite maize hybrids (Barrière et al. 2005). The USA has the largest maize forage production in the world (Lauer et al. 2001). Despite improvements of 0.13 to 0.16 t/ha per year of forage yield since 1930, there was no improvement for cell wall digestibility (Lauer et al. 2001). Therefore, improvement of cell wall digestibility
deserves consideration in future forage breeding.

Currently, interest in lignocellulosic fuel is increasing. Fuel ethanol, especially the lignocellulosic ethanol, is an attractive alternative to petroleum oil, because it burns cleanly, is renewable, and has very large biomass feedstock supplies (Farrell et al. 2006; Rathin et al. 2011). Ethanol is mainly produced from maize grain or sugarcane at present (Vermerris et al. 2007; Demain, 2009). Grain-derived ethanol will not be sufficient to satisfy future demand for biofuels. For example, the US government has set a goal to produce 36 billion gallons of biofuel with about 44% as cellulosic ethanol and 14% as biomass-based diesel by 2022 (Energy Independence and Security Act 2007). Due to wide geographic adaption and availability, maize stover will likely substantially contribute to cellulosic ethanol production. Currently, cellulosic ethanol is feasible but not economically competitive compared with gasoline and grain-derived ethanol. In order to make cellulosic ethanol economically more competitive, improvement in the quality of biofeedstock is required along with advancement in bioprocessing and availability of effective enzymes and microorganisms for breakdown of polymeric carbohydrates and fermentation (Wyman, 2007).

Cell wall digestibility is highly correlated with the quality of biofeedstocks for the purpose of either forage (Barrière et al. 2003, 2004b) or cellulosic ethanol production (Lorenz et al. 2009). Lignin limits access of cellulolytic enzymes or rumen microorganisms to cellulose and hemicelluloses (Moore and Jung, 2001). Many studies have shown that lignin content is negatively correlated with either in vitro or in vivo cell wall digestibility, for example, studies conducted by Riboulet et al. (2008), Casler and Jung (2006), and Guo et al. (2001). Lignin structure, for example the ratio of syringyl (S) to guaiacyl (G) units, is also reported to affect cell wall digestibility (Grabber et al. 2004; Fontaine et al. 2003). In studying a 104 silage samples,
Taboada et al. (2010) reported significant negative correlations between in vivo dry matter digestibility and either G (r=-0.79) or S unit (r=-0.76), but a positive correlation between in vivo digestibility and the G/S ratio(r=0.57).

In maize, there are four well known mutants, known as brown midrib mutants 1, 2, 3, and 4. They are characterized by reduced lignin content and altered composition as well as increased cell wall digestibility (Barrière et al. 2004b). For example, in vivo neutral detergent fiber digestibility of hybrids with homozygous \( bm3 \) alleles is 9% higher than those with wildtype \( Bm3 \) gene(s) (Barrière et al. 2004a). \( bm3 \) mutations were caused by structural mutations of the caffeic acid O-methyltransferase (COMT) gene (Vignols et al. 1995; Morrow et al. 1997). This gene is responsible for the biosynthesis of the syringyl- \( (S-) \) unit using 5-OH coniferyl alcohol as substrate. The defect of the \( COMT \) gene in \( bm3 \) mutant results in a substantial reduction of total lignin content by 25-40% (Barrière et al. 2004b). Reduced content of sinapyl residues, thus an increased G/S ratio, and elevated level of 5-hydroxyconiferyl alcohol content in lignin were also observed in \( bm3 \) plants (Chabbert et al. 1994a; Provan et al. 1997; Marita et al. 2003). \( bm1 \) was associated with reduced activity of cinnamyl alcohol dehydrogenase 2 (CAD2) gene (Halpin et al, 1998). \( CAD \) is responsible for converting hydroxyl cinnamal aldehydes into alcohols in monolignol biosynthesis (Guillaumie et al. 2007b). Thus reduced CAD activity resulted in accumulated aldehyde and reduced monolignol units in the lignin (Barrière et al 2004b). However, it is still not clear whether \( CAD2 \) is the underlying gene of \( bm1 \) (Guillaumie et al. 2007a). Genes underlying \( bm2 \) and \( bm4 \) have not been isolated and their impact on lignification has not yet been resolved.

13 additional brown midrib mutants named \( bm^*A \) to \( bm^*L \) are listed in the Maize Genetics Stock Centre (http://maizecoop.cropsci.uiuc.edu/stockcat.php). By tests of allelism of these
brown midrib mutants, three new *bm* loci designated as *bm* 5, *bm* 6 (Ali et al. 2010), and *bm* 7 (unpublished data) were identified. So far, no studies have been conducted on these 3 new brown midrib mutations. Characterization of these new mutants on cell-wall lignification will advance our understanding of the change in phenolic compound profiles and their impact on cell wall digestibility, and might ultimately provide genetic tools for forage or bioenergy maize breeding. The objectives of this study were to (1) study the effect of *bm* 6 on cell wall digestibility and plant height, (2) determine the genetic location of *bm* 6, and (3) delimit *bm* 6 to a physical contig to facilitate map-based isolation of *bm* 6.

**Materials and methods**

**Mapping populations**

The F$_2$ mapping population was produced by crossing *bm* 6 (*bm* J) and inbred line B73. We choose B73 because it is one parent of the IBM population, and its genome has been sequenced (Schnable et al. 2009). This will facilitate fine mapping and candidate gene prediction. The *bm* 6 stock used to cross with B73 for creation of mapping population was derived from stock 5803J *bm**-86-87-8875-6* from maize genetics cooperation stock center (http://maizecoop.cropsci.uiuc.edu/). A small F$_2$ population consisting of 181 plants was grown in the greenhouse in the winter of 2009 for validating BSA results and recessive inheritance. A large F$_2$ population with ~5000 plants was grown at the Agronomy farm of Iowa State University during the summer of 2010 for fine mapping. 25 seeds were sown per row in 200 rows. When the brown midrib phenotype could be observed around the sixth leaf-stage in the field, F$_2$ plants with brown midrib phenotype were marked. One week later, the phenotype was re-evaluated. Only leaf tissues of plants with obvious brown midrib phenotype in both phenotype evaluations were collected for DNA extraction.
Investigation of pleiotropic effects of the *bm6* gene

Four plots were randomly selected from the large F2 populations, each consisting of three adjacent rows. The reason for selecting three rows is to ensure a sufficient number of brown midrib plants in a plot. The plant height of all the plants in the four plots was measured from soil level to flag leaf. The effect of the *bm6* gene on plant height was analyzed by mixed model $y_{ij}=\mu+G_i +e_{ij}$ with unbalanced data, because the genotypic effect was only significant, when the effects of genotype, plot, and plot-by-genotype interaction were fitted into the full model. Genotype effect $G_i$ (i=1 to 2) and experimental error $e_{ij}$ (j=1 to 47, depending on the number of plants per plot) were considered as fixed and random variables, respectively. $\mu$ was the overall mean. In order to investigate the digestibility of neutral detergent fiber (DNDF), nine mutant and nine wild-type plants were randomly selected from each plot and stover was harvested to form mutant and wild-type bulks from each plot. Thus, in total four bulked mutant and four bulked wild-type stover samples were analyzed. The stover samples were dried and ground to pass through a 1mm sieve. Quality analysis was performed with near infrared reflectance spectroscopy (NIRS) based on a calibration of 43 samples (Brenner et al. 2011) including one of the eight samples. Wet lab measurement of neutral detergent fiber (ND) for the 43 samples was done according to the protocol available on the ANKOM website (http://www.ankom.com). Finally, neutral detergent fiber (NDF) and In Vitro true digestibility (IVTD) (48 hour rumen fermentation followed by NDF procedure) were predicted for our 8 samples. DNDF was calculated by 1-(IVNDF/NDF). A student’s t-test was used to test the effect of the *bm6* mutation on NDF and DNDF.

DNA extraction

Leaf tissue was harvested after two phenotypic evaluations. The harvested leaves were freeze
dried for 24 hours. After drying, two steel beads were added into each sample and samples were ground with the Geno/Grinder 2000 (BT&C, Inc. NJ, USA) at 700 strokes/min for 3 minutes. DNA extraction was conducted according to the CTAB protocol used at the Plant Genomics Center of Iowa State University (http://schnablelab.plantgenomics.iastate.edu/docs/resources/protocols/pdf/96wellformat.2010.06.23.pdf).

**Mapping of bm6 to Chromosome 2 by quantitative bulked segregant analysis**

1038 ISU-SNP (http://magi.plantgenomics.iastate.edu/cgi-bin/snp/index.cgi) markers were used to genotype mutant and wild-type F2 bulks with about 20 plants in each bulk at the Plant Genomics Center of Iowa State University. The data were analyzed according to the quantitative BSA principle described by Liu et al. (2010). Briefly, polymorphic codominant markers are expected to have two SNP alleles and their ratio should be equal between two pools at loci that are not linked to bm6. If a marker is linked with bm6 gene, the B73 allele is reduced in the mutant pool and the ratio of B73/mutant allele is lower than "1", but enriched in the wild-type pool. For validation, several simple sequence repeat (SSR) markers near candidate regions were tested for polymorphisms and polymorphic markers were used to genotype the small F2 population consisting of 181 plants.

**Marker development**

After initial mapping of bm6, nearby anchored SSR and insertion deletion polymorphism (IDP) markers were used to delimit the gene to a physical genome region. When public SSR and IDP markers were exhausted within the target region, the B73 sequence between the two flanking markers umc2245 and umc2363 was used for development of new SSR markers. The interval between the two flanking markers spans ~890,000 bp in the B73 sequence
(http://www.maizesequence.org/blast). The ~890,000 bp sequence was subjected to MISA (http://pgrc.ipk-gatersleben.de/misa/) for SSR motif identification with default parameter settings. After SSR motifs were identified, about 250 bp of sequence at each side of every motif were extracted and the resulting ~500bp sequences were stored in Fasta format. The extracted sequences were blasted against the Zea repeat database (download from http://plantrepeats.plantbiology.msu.edu/downloads.html) by local alignment search tool (download from http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download). Only the sequences that did not contain known repeat sequences were used for SSR primer design. Primer design was done in Primer 3 (http://fokker.wi.mit.edu/primer3/input.htm). Default settings in Primer 3 were used, except that GC content was adjusted to 40-60% and a bracket was used to include the SSR motif, which ensures that the amplified fragment includes the SSR motif. For the first cycle of primer design we tried to find one SSR marker per ~50,000 bp within the 890 kb region. Additional primers were designed for further fine mapping, after the first cycle of primers restricted bm6 to a smaller region.

IDP and SSR marker data collection and analysis

PCR products of SSRs and IDPs were separated on 3% metaphor or 1% agarose gels depending on the product size differences between the parental alleles. B73 alleles were scored as “1”, while the PCR band from the bm6 stock was marked as “2”. Heterozygous plants with both “1” and “2” bands were recorded as “1/2”. For dominant markers, only those showing a band in B73 but not in bm6 were used for genotyping. In this case, this type of bands was scored as “3”.

The genotyped portion of the F2 population for fine mapping comprised of plants exclusively
showing brown midrib phenotypes, as these are most informative (Ingvarsdson et al. 2010). Markers showing polymorphisms between two parents will fall into either of the two below mentioned cases: (1) the frequency of heterozygotes identified by a marker is significantly lower than 0.5, which indicates the marker is linked with the \( bm6 \) gene; (2) the frequency of heterozygotes identified by a marker is agreement with 0.5 statistically, which indicates that the marker is not linked with the underlying gene. In the case of a marker co-segregating with the \( bm6 \) gene, it will not detect any heterozygotes in the mapping population assuming the phenotyping is 100% accurate. If a linked marker still identifies brown midrib plants in the mapping population showing the genotypes “1”, “1/2”, or “3”, this marker is not cosegregating with \( bm6 \). umc2245 is closer to the end of short arm of Chr.2 than umc2363, so we call the plants with B73 allele at umc2245 “left recombinants” which define the left border of \( bm6 \). Markers located between umc2245 and \( bm6 \) will detect a fraction of the same recombinants identified by umc2245 (with 1, 1/2, 3 genotypes). Similarly, we call the plants with B73 allele at umc2363 “right recombinants”. Markers located between \( bm6 \) and umc2363 will only detect a fraction of the recombinants identified by umc2363. These two groups of recombinants finally define the genetic interval, where \( bm6 \) must be located based on B73 sequence. In addition, the genetic positions of markers in the target region as well as \( bm6 \) were calculated using JoinMap 4 (Van Ooijen 2006).

**Candidate gene prediction**

After \( bm6 \) was mapped between SSR_308337 and SSR_488638 markers, the start and end position of the physical interval defined by these two markers were obtained by blasting the left and right primer sequences of SSR_308337 and SSR_488638 on maize sequence, respectively (www. Maizesquence.org). Subsequently, the view of this physical interval was obtained by
entering its start and end position on maize sequence, where the working gene sets in both directions are shown. Additionally, synteny to rice was displayed by the “synteny” function on the same website. The graphical display of synteny between rice and maize was produced by SyMAP (Soderlund et al. 2006).

Results

Recessive inheritance of the brown midrib phenotype caused by the bm6 allele

In the test of allelism, F1 plants from crossing bm6 to bm1-4 all showed a non brown midrib phenotype (Ali et al. 2010). In the small mapping population of 181 F2 plants derived from crossing B73 and the bm6 stock, 45 plants showed the brown midrib phenotype. The ratio of brown to green midrib plants is in agreement with a 1:3 segregation (p=0.97), and is in agreement with recessive inheritance of the brown midrib phenotype caused by bm6.

Effect of bm6 on plant height and cell wall digestibility

In total, 11, 15, 9, and 11 mutant plants, and 25, 26, 26, and 30 wild-type plants were identified in those 4 randomly selected “plots”. The least square mean of plant height for wild-type F2 plants is 182.1 cm, while the least square mean of plant height for mutant F2 plants was 171.4 cm, significantly lower than that of wild-type (p<0.01) (Figure 1A). The student’s t-test revealed that there was no significant difference for average NDF between mutant and wild-type bulked samples (p=0.46), while average DNDF was higher in mutant compared to wild-type bulks (Figure 1B) (p=0.07).

Quantitative bulked segregant analysis

The 1038 ISU-SNP markers evenly spread across the 10 maize chromosomes spanning 1736.8 cM in length on ISU.ibm Map (Fu et al. 2006). The marker numbers per chromosome varied from 60 on Chr. 9 to 180 on Chr. 1 (Table 1). In the 2066 assays (1038×2), 1810 had allele
peak areas larger than (or equal to) 15, which resulted in 87.6% base calling. On each chromosome, the base calling rate ranged from 82.1% on Chr. 2 to 94.2% on Chr.9. The quantitative BSA analysis identified six markers on Chr. 2 with strong deviation from 1 for the ratio of “B73 to bm6 stock allele” in wild-type pool to “B73 to bm6 stock allele” in mutant pool (wt:mut ratio). The wt:mut ratio for these six markers ranged from 3.22 to 17.09. Although two markers on Chr.8 also had wt:mut ratios deviating from 1, the ratios were 3.36 and 4.11, which was much lower than those of markers 107679W46 and 38801W30 on Chr.2, whose wt:mut ratios were 13.23 and 17.09, respectively (Table 2). Therefore, we inferred bm6 to be located on Chr.2 linked to those six markers showing most highly biased wt:mut ratios. These six markers spanned a 43 cM region (Table 2), which corresponded to bin 2.01 and 2.02. In order to validate this result, two polymorphic SSR markers, umc1165 and umc2363, which are located in bins 2.01 and 2.02, respectively, were used to genotype the small F2 mapping population consisting of 181 plants. umc2363 showed co-segregation with the mutation, while 5 recombinants were found between umc1165 and bm6. The results confirmed that the bm6 gene is on Chr.2 close to umc2363.

**Fine mapping of bm6 using publicly available markers**

All public SSR markers in bins 2.00, 2.01, and 2.02 were used to test for polymorphisms between B73 and the bm6 stock. The polymorphic markers showing clear bands were used to genotype 192 F2 individuals with brown midrib phenotype from the large F2 mapping population. As all plants in the mapping population included only brown midrib individuals, the polymorphic markers were indicated to be linked with bm6 if the detected heterozygote frequency were significantly lower than 0.5. Mmc0111 identified 50 heterozygotes out of 192 brown midrib individuals, which indicates linkage between mmc0111 and bm6. umc1552, umc2363, and
umc1165 detected only 29, 8, and 8 recombinants, respectively. The genetic distance between *bm6* and mmc0111, umc1552, and umc2363 were 14, 8.5, 2.4 cM (Figure 1). umc2363 and umc1165 co-segregated in these 192 brown phenotypic individuals. umc2245 detected 9 heterozygotes, which were different from those identified by mmc0111. Therefore, umc2245 was located on the other side of *bm6* as mmc0111. Seven pairs of IDP markers were available between umc2245 and umc1165. Two of them were polymorphic but were dominant. Fortunately, PCR amplified bands appearing in B73, so that they could be used for mapping. IDP4732, IDP7712, and umc2245 detected recombination events on the other side compared to mmc0111 (and umc1552, umc2363, umc1165), and were placed 2.3, 1.8, and 1.8 cM relative to the *bm6* gene (Figure 1). IDP7712 and umc2245 co-segregated in these 192 brown phenotypic individuals. Taking together, *bm6* was anchored in a 4.2cm interval delimited by umc2363 (umc1165) and umc2245 (Figure 1).

**Development of new SSR markers to saturate the target region**

Within the ~890,000bp region between umc2245 and umc2363, 99 SSR motifs were identified by MISA, and 20 of them were located in repeat sequences. In the first cycle, we tried to find one SSR marker per ~50 kb, but some regions were highly repetitive. Finally, we designed 14 pairs of primers, among which seven detected polymorphisms between both parents (Table 4). After the target region was further reduced in size, the second cycle of primer design provided eight pairs of primers, but only two of them showed polymorphisms. In total, nine polymorphic SSR markers were obtained from 22 pairs of primers tested, including dominant marker SSR-447208.

**Fine mapping of the *bm6* gene**

umc2363 and umc1165 were used to screen recombinants within the larger set of 960 mutant
F₂ plants. umc2245 detected 20 heterozygotes from 960 plants. These 20 plants were called left recombinants L1 to L20 (Table 5), indicating the recombination occurring on the left side of the bm6 gene. umc2236 detected 19 heterozygotes. These were called right recombinants R1 to R19, which means recombination occurred right of the bm6 gene. All nine newly developed polymorphic markers were used to genotype these 39 recombinants and one control which showed “2” at both umc2245 and umc2363. The dominant marker SSR_447208 was not assigned to the target region, because it detected 31 heterozygotes from these 39 recombinants, more than those identified by either umc2245 or umc2363. Although SSR_66888 and SSR_864814 were polymorphic co-dominant markers, they detected only two genotypes (either the B73 band or the bm6 band) rather than three genotypes in those 39 recombinants. They could still be used for mapping as they indicated, whether the B73 allele was present or absent. Finally, eight primers were assigned to the target region confined by umc2245 and umc2363, as they detected the same number or fewer recombinants identified by these two flanking markers. B73 segments expanded to SSR_308337 for the left recombinants of L11-L20, and to SSR_488638 for the right recombinants of R16-20 (Table 3). Thus, the closest markers flanking bm6 from our study were SSR_308337 and SSR_488638, which is about 180kb in length referring to B73 sequence. The relative genetic position of bm6 to both flanking markers is about 0.5 cM to SSR_308337 and 0.3 cM to SSR_488638 based on the information from 960 F₂ plants, respectively (Figure 1).

In the fine mapping process, another type of “recombinant” was detected. Six plants were heterozygous at umc2245 (two plants) or at both umc2245 and umc2363 loci (four plants). However, the B73 fragment spans from umc2245 to SSR_674176 for all these six plants. This result was in conflict with the above mentioned mapping results. If these are brown midrib plants,
there should be no B73 allele in the region between SSR_308337 and SSR_488638. We suspect that these “recombinants” resulted from phenotyping and/or sampling errors. This would reflect a phenotyping/sampling error rate of 6/(960*4)=0.00156, which is low given that phenotyping was done on single plants. Alternatively, there were double recombination events occurred between SSR_308337 and SSR_488638 within these six plants, but they were not captured as no markers were located within the 180kb region. Anyway, these six plants were considered as suspicious plants and were excluded from the mapping process.

**Candidate genes**

In the target region between markers SSR_308337 and SSR_488638, 10 gene models are predicted in maize B73 sequence (Version 2.0, [www.maizesequence.org](http://www.maizesequence.org)), including a putative GTP cyclohydrolase I 1 gene (GRMZM2G062420), a putative WUS1 gene (GRMZM2G047448), and eight putative uncharacterized or novel genes (Table 6). Among the eight putative uncharacterized or novel genes, one contains a PHD finger-like domain (GRMZM2G047018), indicating it might be a transcription factor. The remaining seven genes do not have any indicated function. However, a comparison of the target region with rice revealed good synteny with a region on Chr. 4 of rice. Seven out of the 10 genes in the target region have homologous genes in rice. Most showed synteny with Chr. 4 except for GRMZM2G342107, with a rice gene hit on Chr.5. Six out of the identified seven homologous genes in rice have gene function annotation, which was used to infer the function of their orthologous counterparts in maize. Based on the information from rice, the genes GRMZM2G046968 might be a zinc finger family gene. The rice homologous counterpart of GRMZM2G342107 was annotated as no apical meristem (NAM) protein-like encoding gene. GRMZ2G062396 is homologous to the naringenin 2-oxoglutarate 3-dioxygenasease encoding gene (*F3H*). The rice homologous counterpart of
GRMZM2G348909 is an IQ calmodulin-binding motif family gene, which might interact with cell wall related kinase WAK2 (Rohila et al. 2010).

**Discussion**

Brown midrib mutants that are characterized by brown-reddish coloration of the vascular tissue in the leaf blade and sheath are reported to have altered cell wall composition (Chabbert et al. 1994a, 1994b; Marita et al. 2003; Barrière et al. 2004). They are also reported to have favorable properties for forage (Barrière and Argillier et al. 1993) and cellulosic ethanol production due to reduced lignin content (Lorenz et al. 2009), especially bm3. Only four mutants known as bm1-4 were characterized in maize until recently. More recently, additional bm mutants in maize were described. Haney et al. (2008) first reported bm*F as bm5. By allelism testing all available brown midrib mutants from the maize stock centre including bm1-4, we confirmed bm5, and identified bm6 (Ali et al. 2010), and bm7 (unpublished data). The allelism tests also revealed that the four mutants known as bm*E, bm*F, bm*G, and bm*F have mutations in the same gene underlying bm5. Only one bm* allele was identified for bm6 (Ali et al. 2010) and bm7, respectively. The well known bm1-4 mutants are mapped on bins 5.04, 1.11, 4.05, and 9.07, respectively. This study mapped bm6 to Chr. 2, which confirmed bm6 as a novel brown midrib gene. Furthermore, the fine mapping of bm6 into a 180kb region will enable final isolation of the underlying gene.

bm1-4 in maize and bmr6, bmr12, bmr18 in sorghum are characterized by reduced lignin content and altered lignin composition, but they often have negative effects on plant fitness, although genetic background and environment might confound their effect on fitness (Pederson et al. 2005). Importantly, bm3 in maize significantly reduces plant grain and/or forage yield (Miller et al. 1983; Miller and Geadelmann 1983; Ballard et al. 2001; Cox et al. 2001), and plant
height (Lee and Brewbaker 1984). *bmr-6* in sorghum also reduces plant height (Casler et al. 2003). F$_2$ plants homozygous for the *bm6* mutation were significantly shorter than wild-type F$_2$ plants in our study. Average DNDF of mutant bulks was 1.4% higher than that of wild-type bulks. In addition, DNDF was always higher in mutant than in wild-type bulks in each of the four randomly selected plots (Figure 1C). However, the difference of average DNDF between mutant and wild-type bulks was just suggestive significant (p=0.07), which might be due to the small sample size. Moreover, wild-type bulks included 2/3 heterozygous (*Bm6bm6*) and 1/3 homozygous plants (*Bm6Bm6*). If the genetic effect of *bm6* gene is additive rather than dominant, clear results might only be obtained by comparing isogenic lines carrying the homozygous *bm6* alleles and that with homozygous wild-type alleles. Thus, the suggestive significance at p=0.07 encourages studying the effect of *bm6* on DNDF and other cell-wall properties by isogenic lines.

Our mapping result provided tightly linked markers, which will speed up development of near isogenic lines. Comparing near isogenic lines with and without *bm6* will help to evaluate the effect of *bm6* on cell wall digestibility traits and agronomic related traits, such as plant height, flowering time, lodging, grain, and forage yield. As near isogenic lines cannot distinguish between genic pleiotropy and tight genic linkage (Chen and Lubberstedt 2010), knocking out the candidate gene by RNAi, will help to discriminate pleiotropy and linkage at the gene level.

The mechanism underlying the brown coloration in the vascular tissue is still not clear. However, Sattler et al. (2010) proposed that color change in the lignin-rich tissue is a good indication of disturbance of monolignol biosynthesis in C4 grasses. In our study, the *bm6* gene was mapped into a 180 kb region in bin 2.01. Nearby this region, some monolignol biosynthesis genes are located, for example *CAD6/SAD* (Ac2155994.3_FG038) in bin2.02, *PAL2a* (AC213314.3_FG039) and *PAL2b* (AC213314.3_FG037) in bin2.03 (Barrière et al. 2009).
our target region, 10 gene models are predicted in B73 sequence. Among these 10 genes, four are promising candidate genes. GRMZ2G062396 is homologous to the Naringenin2-oxoglutarate 3-dioxygenase gene, which is also called flavanone 3-hydroxylase (F3H) gene. GRMZ2G06396 contains the Oxoglutarate/Fe-dep Oxygenase domain and could oxidize phenols or other products in secondary metabolism. Flavonoids as well as lignin are synthesized from phenylpropanoid pathway. P-Coumaryl CoA which is the product of the 4-coumarate:CoA ligase (4CL) enzyme in the phenylpropanoid pathway is used for either lignin monolignol or flavonoid biosynthesis. F3H is co-expressed with other flavonoid enzymes involved in anthocyanin biosynthesis and responsible for plant colorations (Han et al. 2010). In Arabidopsis, genes in the anthocyanin biosynthesis pathway are co-regulated by MYB11 and MYB12 (Stracke et al. 2007). Interestingly, the lignin biosynthesis genes are also co-regulated by MYB transcriptional factors (Zhao et al. 2011). Additionally, lignin biosynthesis is carbon consuming and cross-talking with other physiological processes (Zhao et al. 2011). For example, the reduction of the flux of carbon into lignin resulted in accumulation of flavonoid due to a mutation in the HCT gene (Besseau et al. 2007). Thus, it is reasonable to speculate that a mutation in F3H involved in flavonoid biosynthesis will affect lignin biosynthesis due to either synchronous regulation of flavonoid and lignin biosynthesis or a trade-off of resources between the two biosynthesis branches. Naringenin, substrate for F3H in the flavonoid biosynthesis pathway, inhibits the activity of 4CL in vitro in several species such as petunia (Ranjeva et al. 1976), loblolly pine (Voo et al. 1995), maize, rice, and Echinochloaoryzicol (Deng et al. 2004). Addition of naringenin to rice produced browning leaf tips and decreased coniferyl alcohol as well as the lignin content by about 10% (Deng et al. 2004). In vitro application of naringenin could also inhibit the growth of maize, rice, Echinochloaoryzicol (Deng et al. 2004). Mutation in F3H might lead to accumulation of
naringenin, which will suppress plant growth. In our mapping population, we randomly selected 153 F$_2$ plants. The average plant height of mutant plants (46 plants) was significantly lower than that of wild-type plants (107 plants).

Lignin biosynthesis is highly coordinated by transcription factors (Reviewed by Zhong and Ye 2009; and Zhao and Dixon 2011). Most of the monolignol biosynthesis genes, except $C4H$, $COMT$, and $F5H$, have AC elements in their promoters, which are bound by MYB transcription factors (Raes et al. 2003). Even the three genes which do not have AC elements are also directly or indirectly regulated by lignin-specific MYB transcription factors (Zhou et al. 2009; Zhao et al. 2010). Investigating the differential gene expression by $bm1$-4 isogenic lines, Guillaumie et al. (2007) identified other transcription factors and regulatory genes besides MYB factors showing differential expression between wild-type and mutant isogenic lines. Among them were three Zinc finger like protein coding genes. Thus GRMZM2G046968 gene is another candidate for $bm6$, since its syntenic counterpart in rice is a Zinc finger gene.

Gene GRMZM2G348909 is homologous with a rice IQ calmodulin-binding motif family gene. Wall-associated kinase ($WAK$) is involved in regulation of cell elongation (Lally et al. 2001) and expansion (Wagner and Kohorn 2001). This regulation requires a calcium-dependent binding protein to form the signaling complex (Decreux and Messiaen 2005). In Arabidopsis, the expression of $WAK1$ was more often seen in vascular cells with high levels of expression in older leaves than younger ones (Wagner and Kohorn 2001). $WAK1$ and $WAK2$ could be both wound-induced to express even in adult leaves which have stopped expansion (Wagner and Kohorn 2001). Wounding in plants often stimulates accumulation of lignin around the wound (Hawkins and Boudet 1996). In rice, a $WAK2$-like protein was reported to interact with a calmodulin-binding motif protein (Rohila et al. 2006). Taking together, the putative IQ calmodulin-binding
motif gene in our target region might be involved in lignin regulation by interacting with WAK proteins.

Another interesting gene in our target region is GRMZM2G047488 which is homologous to the WUSCHEL1(WUS1) gene in Arabidopsis. There are two copies of WUS genes in maize (NCBI), which contain homeobox domains. In Arabidopsis, there are seven KNOX homeobox genes. BP (BREVIPEDICELLUS) is one of the seven KNOX homeobox genes and it determines not only internode patterning, but also regulates lignin biosynthesis (Mele et al. 2003). bp mutants resulted in increased lignin, while over-expression of BP resulted in decreased lignin (Mele et al. 2003). However, no similarity between WUS1 and BP was found. Further study these 10 genes in our target region, especially the above mentioned four genes, will help us to isolate the bm6 gene.

Acknowledgements

We would like to thank Elizabeth Bovenmyer, Janine Comstock, and Alice Wang for help in field. We appreciate the generosity of Dr Sarah Hake (Adjunct Professor and Center Director; Plant Gene Expression Center USDA-ARS) and Dr. Erick Vollbrecht (professor in Department of Genetics, Development and Cell biology at Iowa State University) to provide the BSA data for mapping bm6. Yongsheng Chen was supported by the Interdepartmental Genetics Graduate Program as well as RF Baker Center for Plant Breeding at Iowa State University. Hongjun Liu is a visiting student at ISU, supported by China Scholarship Council.

Authors’ contributions

YC is the primary research and author, did the mapping work, analyzed the data, and wrote this manuscript. HL was involved in the fine mapping process. FA did the initial mapping in a small F2 population. PS provides the mutant stock and helped in the manuscript writing. QJ
helped the syntenic analysis. UKF helped to design the new SSR markers. TL coordinated this project and together with YC wrote this manuscript.
References


Rohila JS, Chen M, Chen S, Chen J, Cerny R, Dardick C, Canlas P, Xu X, Gribskov M, Kanrar S,


<table>
<thead>
<tr>
<th></th>
<th>Chr1</th>
<th>Chr2</th>
<th>Chr3</th>
<th>Chr4</th>
<th>Chr5</th>
<th>Chr6</th>
<th>Chr7</th>
<th>Chr8</th>
<th>Chr9</th>
<th>Chr10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Marker</td>
<td>180</td>
<td>109</td>
<td>133</td>
<td>91</td>
<td>108</td>
<td>109</td>
<td>87</td>
<td>79</td>
<td>60</td>
<td>82</td>
<td>1038</td>
</tr>
<tr>
<td>Data calling (%)</td>
<td>85.00</td>
<td>82.10</td>
<td>85.30</td>
<td>90.70</td>
<td>89.40</td>
<td>86.20</td>
<td>89.90</td>
<td>94.20</td>
<td>90.20</td>
<td>87.60</td>
<td></td>
</tr>
<tr>
<td>Coverage(cM)</td>
<td>279.1</td>
<td>213.2</td>
<td>229.8</td>
<td>143.4</td>
<td>141.8</td>
<td>129.6</td>
<td>140.2</td>
<td>168.2</td>
<td>144.8</td>
<td>146.7</td>
<td>1736.8</td>
</tr>
</tbody>
</table>

Data calling % is the percentage of markers with successful base calling. Coverage is the genetic length covered by the markers on each chromosome.
Table 2. Markers closely linked with \textit{bmt6} identified by quantitative BSA

<table>
<thead>
<tr>
<th>Marker</th>
<th>Mutant Pool</th>
<th>Wild Pool</th>
<th>wt:mut</th>
<th>Chr</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Area 1</td>
<td>Area 2</td>
<td>B73/mut ratio</td>
<td>Area 1</td>
<td>Area 2</td>
</tr>
<tr>
<td>107679W46</td>
<td>32.32</td>
<td>5.25</td>
<td>0.16</td>
<td>10.26</td>
<td>28.51</td>
</tr>
<tr>
<td>38801W30</td>
<td>62.19</td>
<td>14.74</td>
<td>0.24</td>
<td>11.54</td>
<td>36.18</td>
</tr>
<tr>
<td>17275W26</td>
<td>17.69</td>
<td>17.37</td>
<td>0.98</td>
<td>8.99</td>
<td>27.44</td>
</tr>
<tr>
<td>38725W3</td>
<td>27.82</td>
<td>8.11</td>
<td>0.29</td>
<td>20.40</td>
<td>27.05</td>
</tr>
<tr>
<td>27015W38</td>
<td>17.80</td>
<td>12.67</td>
<td>0.71</td>
<td>10.78</td>
<td>24.68</td>
</tr>
<tr>
<td>93178W36</td>
<td>16.00</td>
<td>7.56</td>
<td>0.47</td>
<td>11.29</td>
<td>19.32</td>
</tr>
</tbody>
</table>

The B73: mutant ratio (B73/mut ratio) was calculated by Area 2 divided by Area 1. The wt:mut means the ratio of B73/mut ratio in wild type pool divided by that in mutant pool, which would be much higher than “1”, if the marker is linked with \textit{bmt6}. 
Table 3. Mapping of *bm6* by publicly available markers to a region flanked by umc2245 and umc2363

<table>
<thead>
<tr>
<th>Recombinants</th>
<th>IDP4732 (35.32)</th>
<th>umc2245 (30.9)</th>
<th>IDP7712 (36.4)</th>
<th>umc1165 (47.4)</th>
<th>umc2363 (54.15)</th>
<th>umc1552 (75.93)</th>
<th>mnc0111 (93.3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1-3E</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>P1-3D</td>
<td>1</td>
<td>1/2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>P1-6C</td>
<td>1</td>
<td>1/2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>P1-8C</td>
<td>1</td>
<td>1/2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>P2-1H</td>
<td>1</td>
<td>1/2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>P2-3E</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>P2-4H</td>
<td>1</td>
<td>1/2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>P2-6A</td>
<td>1</td>
<td>1/2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>P2-12C</td>
<td>1</td>
<td>1/2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>P1-3A</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
</tr>
<tr>
<td>P1-4D</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
</tr>
<tr>
<td>P1-12H</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1/2</td>
</tr>
<tr>
<td>P2-5C</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
</tr>
<tr>
<td>P2-6E</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
</tr>
<tr>
<td>P2-9F</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P2-10H</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
</tr>
<tr>
<td>P2-12H</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
</tr>
<tr>
<td>P1-2B</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1/2</td>
<td>1/2</td>
</tr>
<tr>
<td>P2-2A</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1/2</td>
<td>1/2</td>
</tr>
<tr>
<td>P1-1H</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1/2</td>
</tr>
<tr>
<td>P2-2C</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1/2</td>
</tr>
<tr>
<td># of Rec.</td>
<td>9</td>
<td>7</td>
<td>7(0)</td>
<td>8(1)</td>
<td>8(1)</td>
<td>29(2)</td>
<td>50(2)</td>
</tr>
</tbody>
</table>

All these markers were used to genotype 192 brown F₂ plants.

G: Numbers in brackets indicate the genetic positions on IBM2 2008 Neighbor Frame 2, except umc1552 whose position is inferred from IBM2 2008 Neighbors 2. IDP4732 was expected between umc2235 and umc2363, but should be upstream of umc2245. Because IDP4732 and IDP7712 were dominant markers, only two kinds of genotypes were detected.

# of Rec: numbers of recombinants being identified in the 192 brown F₂ plants. The numbers in the bracket were recombinants showing only B73 bands. Each marker detected far less than 50 heterozogotes (expected heterozogotes is 90.5) at p=0.01 level.

“1” represents bands from of B73,
“2” represents bands from of *bm6*.
“1/2” represents heterozygotes.
### Table 4. Primers of newly developed SSR markers and their product size in B73 within the *bm6* target region

<table>
<thead>
<tr>
<th>Primer</th>
<th>Motif</th>
<th>Forward</th>
<th>Reverse</th>
<th>Size(bp)</th>
<th>Poly.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSR_66888</td>
<td>TG(7)</td>
<td>TGGCAGCTGATAGTCGTCCAGA</td>
<td>AGCAAGGCCCTGTTATTGGA</td>
<td>101</td>
<td>Y</td>
</tr>
<tr>
<td>SSR_102465</td>
<td>AT(6)</td>
<td>GCGAGTGAATAAACACGAGAA</td>
<td>AAAACCGTGACACCTTGCAAT</td>
<td>147</td>
<td>N</td>
</tr>
<tr>
<td>SSR_157932</td>
<td>CAA(5)</td>
<td>AAGAACATTTGAGCTTGTACAGA</td>
<td>CGAGTGATCGATCTGCAGTG</td>
<td>168</td>
<td>N</td>
</tr>
<tr>
<td>SSR_220551</td>
<td>TCG(5)</td>
<td>CCTCGGAATCCTACACATGA</td>
<td>CCCTCAGTCCACTCCCACCTATG</td>
<td>181</td>
<td>Y</td>
</tr>
<tr>
<td>SSR_263186</td>
<td>TG(10)</td>
<td>CGAGGATTCGAGTCTCAGTGGT</td>
<td>AGGCGAGGAGAAGACGGAGTGGTT</td>
<td>162</td>
<td>N</td>
</tr>
<tr>
<td>SSR_308337</td>
<td>TA(6)</td>
<td>TTCTGTGTTCATCTAGCGACTT</td>
<td>TGTTGCGAGTGTCCTCCATTG</td>
<td>179</td>
<td>Y</td>
</tr>
<tr>
<td>SSR_347455</td>
<td>CAAG(4)</td>
<td>CGTGCCTGACAGAAAGCTG</td>
<td>CATCGGTTAGGTTCCACGAC</td>
<td>178</td>
<td>N</td>
</tr>
<tr>
<td>SSR_472348</td>
<td>TCG(9)</td>
<td>GTGCAAGCAGGAGCTTAA</td>
<td>AAAGACAAAGCCAAATGGG</td>
<td>153</td>
<td>N</td>
</tr>
<tr>
<td>SSR_492215</td>
<td>CAG(7)</td>
<td>CTGCGAGGAGGACCTTT</td>
<td>TCTGATCCAGGGCATTATG</td>
<td>158</td>
<td>Y</td>
</tr>
<tr>
<td>SSR_524924</td>
<td>CGCT(4)</td>
<td>AAGCGACTCCTCAGCTCTCA</td>
<td>GGTCTTTGCTGCTGCTTTC</td>
<td>196</td>
<td>Y</td>
</tr>
<tr>
<td>SSR_641746</td>
<td>CGC(5)</td>
<td>ATACCCGTCATCGCTCATCT</td>
<td>CGAAGGCAGGAGGAAATGAC</td>
<td>199</td>
<td>Y</td>
</tr>
<tr>
<td>SSR_720602</td>
<td>TG(6)</td>
<td>ATCTGCTGGTCTCAGCGATT</td>
<td>AAGACCCCCGCTACACAG</td>
<td>127</td>
<td>N</td>
</tr>
<tr>
<td>SSR_798730</td>
<td>CGT(9)</td>
<td>TACCGTCCACATCGGATTCT</td>
<td>TAGCTCCTGCTCCTAAACA</td>
<td>176</td>
<td>N</td>
</tr>
<tr>
<td>SSR_864814</td>
<td>TA(6)</td>
<td>CTCTTCACACACCATCTCTAAAGACAC</td>
<td>GACGCTCTCCAAACATAATGACA</td>
<td>198</td>
<td>Y</td>
</tr>
<tr>
<td>SSR_32574</td>
<td>AG(9)</td>
<td>TGGAGGAATAGCGAGGAGA</td>
<td>CCGCCTTTGTTTTGTTT</td>
<td>111</td>
<td>N</td>
</tr>
<tr>
<td>SSR_312186</td>
<td>TAT(6)</td>
<td>CTCTGCTCGTCTGTTFFAC</td>
<td>AGAGCGCAAGCAAACACAG</td>
<td>191</td>
<td>N</td>
</tr>
<tr>
<td>SSR_330561</td>
<td>GAT(5)</td>
<td>ATGGTGTCATGTGATGCTAAA</td>
<td>CGATGAAATGCGGAGTGAATG</td>
<td>195</td>
<td>N</td>
</tr>
<tr>
<td>SSR_347315</td>
<td>GT(6)</td>
<td>AGGACCAGGGCCACCACATTAG</td>
<td>GGAAGATGGCGCCTGCTTGA</td>
<td>159</td>
<td>N</td>
</tr>
<tr>
<td>SSR_447208</td>
<td>TCG(5)</td>
<td>GAAGACAGCTGAGCCTGAGC</td>
<td>GCTCGAGATGAGTGGAGTA</td>
<td>231</td>
<td>Y</td>
</tr>
<tr>
<td>SSR_488638</td>
<td>GC(6)</td>
<td>AGGCAACTCCTCTGCTGTTG</td>
<td>CATGATGGCCCACTCCTT</td>
<td>233</td>
<td>Y</td>
</tr>
<tr>
<td>SSR_516108</td>
<td>CAT(5)</td>
<td>GAGCCCCGACATCTCTCTTCT</td>
<td>AAGACCTCCTGCCTGCTAAA</td>
<td>145</td>
<td>N</td>
</tr>
</tbody>
</table>

The numbers in the names indicate the relative position in the ~890000bp sequence. The first 14 primers are designed from the first cycle, the remaining are from the second cycle. **Y**: detected polymorphisms between two parents. **N**: no polymorphisms were detected between two parents.
Table 5. Fine mapping of the *bm6* gene to a 180kb region between *ys308837* and *ys488638* by newly developed SSR markers

<table>
<thead>
<tr>
<th>Recombinants</th>
<th>umc2245</th>
<th>ys6888</th>
<th>ys220551</th>
<th>ys308337</th>
<th>ys488638</th>
<th>ys492215</th>
<th>ys524924</th>
<th>ys674176</th>
<th>ys864814</th>
<th>umc2363</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>R1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
</tr>
<tr>
<td>R2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
</tr>
<tr>
<td>R3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>R4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>R5</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
</tr>
<tr>
<td>R6</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
</tr>
<tr>
<td>R7</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
</tr>
<tr>
<td>R8</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
</tr>
<tr>
<td>R9</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
</tr>
<tr>
<td>R10</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
</tr>
<tr>
<td>R11</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
<td>1//2</td>
</tr>
<tr>
<td>R12</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
<td>1//2</td>
</tr>
<tr>
<td>R13</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
<td>1//2</td>
<td>1//2</td>
<td>1//2</td>
</tr>
<tr>
<td>R14</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
<td>1//2</td>
<td>1//2</td>
<td>1//2</td>
</tr>
<tr>
<td>R15</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
<td>1//2</td>
<td>1//2</td>
<td>1//2</td>
</tr>
<tr>
<td>R16</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
<td>1//2</td>
<td>1//2</td>
<td>1//2</td>
</tr>
<tr>
<td>R17</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
<td>1//2</td>
<td>1//2</td>
<td>1//2</td>
</tr>
<tr>
<td>R18</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
<td>1//2</td>
<td>1//2</td>
<td>1//2</td>
</tr>
<tr>
<td>R19</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1//2</td>
<td>1//2</td>
<td>1//2</td>
<td>1//2</td>
</tr>
<tr>
<td>L1</td>
<td>1//2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L2</td>
<td>1//2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L3</td>
<td>1//2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L4</td>
<td>1//2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L5</td>
<td>1//2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L6</td>
<td>1//2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L7</td>
<td>1//2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L8</td>
<td>1//2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L9</td>
<td>1//2</td>
<td>1</td>
<td>1//2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L10</td>
<td>1//2</td>
<td>1</td>
<td>1//2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L11</td>
<td>1//2</td>
<td>1</td>
<td>1//2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L12</td>
<td>1//2</td>
<td>1</td>
<td>1//2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L13</td>
<td>1//2</td>
<td>1</td>
<td>1//2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L14</td>
<td>1//2</td>
<td>1</td>
<td>1//2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L15</td>
<td>1//2</td>
<td>1</td>
<td>1//2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L16</td>
<td>1//2</td>
<td>1</td>
<td>1//2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L17</td>
<td>1//2</td>
<td>1</td>
<td>1//2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L18</td>
<td>1//2</td>
<td>1</td>
<td>1//2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L19</td>
<td>1//2</td>
<td>1</td>
<td>1//2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L20</td>
<td>1//2</td>
<td>1</td>
<td>1//2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

“1” represents bands from of B73, “2” represents bands from of *bm6*, “1//2” represents heterozygotes, “0” represents missing data.
Table 6. Candidate genes for *bm6* in maize and synteny to rice

<table>
<thead>
<tr>
<th>Genes ID</th>
<th>Location</th>
<th>Gene Annotation</th>
<th>Location</th>
<th>Gene Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRMZM2G049068</td>
<td>2:3306664-3402539</td>
<td>putative uncharacterized protein</td>
<td>LOC_Os04g56730</td>
<td>4:33666503 zinc finger C-x8-C-x5-C-x3-H type family protein, expressed</td>
</tr>
<tr>
<td>GRMZM2G047018</td>
<td>2:3303037-3399002</td>
<td>PHD finger-like domain-containing protein</td>
<td>LOC_Os04g56740</td>
<td>4:33752555 homeobox domain containing protein, expressed</td>
</tr>
<tr>
<td>GRMZM2G047448</td>
<td>2:3387694-3389171</td>
<td>WUS1 protein</td>
<td>LOC_Os04g56780</td>
<td>4:33876315 homeobox domain containing protein, expressed</td>
</tr>
<tr>
<td>GRMZM2G048582</td>
<td>2:3433019-3433400</td>
<td>Novel</td>
<td>LOC_Os04g56790</td>
<td>4:34330406 GTP cyclohydrolase I, putative, expressed</td>
</tr>
<tr>
<td>GRMZM2G062420</td>
<td>2:3542475-3547749</td>
<td>GTP cyclohydrolase I 1</td>
<td>LOC_Os04g56710</td>
<td>4:35620046 GTP cyclohydrolase I 1, putative, expressed</td>
</tr>
<tr>
<td>GRMZM2G062306</td>
<td>2:3556786-3558729</td>
<td>Putative uncharacterized protein</td>
<td>LOC_Os04g56700</td>
<td>4:35624233 naringenin, 3-oxoglutarate 3-dioxygenase, putative, expressed</td>
</tr>
<tr>
<td>GRMZM2G047321</td>
<td>2:338845-3393264</td>
<td>Novel</td>
<td>LOC_Os04g56770</td>
<td>4:33671620 conserved hypothetical protein</td>
</tr>
<tr>
<td>GRMZM2G048069</td>
<td>2:340613-3414955</td>
<td>Putative uncharacterized protein</td>
<td>LOC_Os04g56740</td>
<td>4:3465-6020 IQ esteroligin binding motif family protein, putative, expressed</td>
</tr>
<tr>
<td>GRMZM2G0511859</td>
<td>2:3420867-3421201</td>
<td>Novel</td>
<td>LOC_Os04g56780</td>
<td>5.27846521 no apical meristem protein</td>
</tr>
<tr>
<td>GRMZM2G041017</td>
<td>2:343908-3439887</td>
<td>Novel</td>
<td>LOC_Os04g56750</td>
<td>5.27842657 no apical meristem protein</td>
</tr>
</tbody>
</table>
Figures

Figure 1. Pleiotropic effect of bm6 on plant height and DNDF
A: the effect of bm6 gene on PHT (plant height). It reduces plant height by 10.7 cm (p<0.01). B: the effect of bm6 gene on DNDF (digestibility of neutral detergent fiber). It increase DNDF by 1.4 (p=0.07). C: DNDF of 4 mutant and 4 wild-type bulked samples.

Figure 2. Genetic map of bm6 and its synteny with rice
Left genetic map is constructed by public available SSR and IDP markers, where the bm6 gene was mapped between umc2245 and umc2363. The right genetic map was constructed with newly developed markers, where the bm6 was delineated between YS_308337 and YS_488638. The blue bar is 184kb region of chromosome 2 of maize between YS_308337 and YS_488638. The green bar is 791.5 kb of chromosome 4 of rice. The lines between blue and green bars indicate the synteny of genes between two species.
Figure 1
Figure 2
CHAPTER 5. POLYMORPHISMS IN O-METHYLTRANSFERASE GENES ARE ASSOCIATED WITH STOVER CELL WALL DIGESTIBILITY IN EUROPEAN MAIZE (ZEA MAYS L.)

Everton A Brenner\textsuperscript{1,7}, Imad Zein\textsuperscript{2}, Yongsheng Chen\textsuperscript{1}, Jeppe R Andersen\textsuperscript{3}, Gerhard Wenzel\textsuperscript{2}, Milena Ouzunova\textsuperscript{4}, Joachim Eder\textsuperscript{5}, Birte Darnhofer\textsuperscript{5}, Uschi Fri\textsuperscript{1}, Yves Barrière\textsuperscript{6}, Thomas Lübbeistedt\textsuperscript{1,8}

\textsuperscript{1}: Department of Agronomy, Iowa State University, Agronomy Hall, Ames, IA, 50011, USA
\textsuperscript{2}: Department of Agronomy and Plant Breeding, Technical University of Munich, Am Hochanger 2, 85354 Freising-Weihenstephan, Germany
\textsuperscript{3}: Department of Genetics and Biotechnology, University of Aarhus, Research Center, Flakkebjerg, 4200 Slagelse, Denmark
\textsuperscript{4}: KWS Saat AG, Grimsehlstr 31, 37555 Einbeck, Germany
\textsuperscript{5}: Bavarian State Research Center for Agriculture, Vöttinger Str. 38, 85354 Freising-Weihenstephan, Germany
\textsuperscript{6}: Unite’ de Ge’ne’tique et d’Ame’lioration des Plantes Fourrage’res, INRA, Route de Saintes, 86600 Lusignan, France
\textsuperscript{7}: Primary researcher and author
\textsuperscript{8}: Author for correspondence

Modified from the paper published in BMC Plant Biology 2010, 10: 27
Abstract

**Background:** OMT (O-methyltransferase) genes are involved in lignin biosynthesis, which relates to stover cell wall digestibility. Reduced lignin content is an important determinant of both forage quality and ethanol conversion efficiency of maize stover.

**Results:** Variation in genomic sequences coding for *COMT, CCoAOMT1*, and *CCoAOMT2* was analyzed in relation to stover cell wall digestibility for a panel of 40 European forage maize inbred lines, and re-analyzed for a panel of 34 lines from a published French study. Different methodologies for association analysis were performed and compared. Across association methodologies, a total number of 25, 12, 1, 6 *COMT* polymorphic sites were significantly associated with DNDF, OMD, NDF, and WSC, respectively. Association analysis for *CCoAOMT1* and *CCoAOMT2* identified substantially fewer polymorphic sites (3 and 2, respectively) associated with the investigated traits. Our re-analysis on the 34 lines from a published French dataset identified 14 polymorphic sites significantly associated with cell wall digestibility, two of them were consistent with our study. Promising polymorphisms putatively causally associated with variability of cell wall digestibility were inferred from the total number of significantly associated SNPs/Indels.

**Conclusions:** Several polymorphic sites for three O-methyltransferase loci were associated with stover cell wall digestibility. All three tested genes seem to be involved in controlling DNDF, in particular *COMT*. Thus, considerable variation among *Bm3* wildtype alleles can be exploited for improving cell-wall digestibility. Target sites for functional markers were identified enabling development of efficient marker-based selection strategies.

**Background**

Stover cell-wall digestibility has long been shown to be crucial for forage quality, and more
recently this trait is getting more attention in relation to biofuel production. Research into bioethanol has grown significantly as a response to global warming and increasing prices of fossil fuels. The conversion of lignocellulosic biomass into fermentable sugars by the addition of enzymes has long been recognized as an alternative to the existing starch-based ethanol production (Gray et al. 2006; Hahn-Hägerdal et al. 2006; Lin and Tanaka 2006). Reduced lignin content improves cell wall digestibility due to increased accessibility of cellulose and hemicelluloses by enzymatic procedures, enabling better ethanol conversion efficiency (U.S. Department of Energy (DOE) 2006). Lignins are phenolic polymers resulting from three monolignols: \( p \)-coumaryl, coniferyl, and sinapyl alcohol. These monolignols derive \( p \)-hydroxyphenyl H, guaiacyl G, and syringyl S phenylpropanoid units, respectively, which polymerize by oxidation to form lignins (Boerjan et al. 2003). In maize, lignins are predominantly comprised of G and S units (37.5 and 60.0 \%, respectively) (Lapierre 1993), with a low, but noticeable content in H units which is nearly five times higher than in dicotyledonous plants.

Lignin is synthesized by the phenylpropanoid pathway (Humphreys and Chapple 2002). Phenylalanine ammonia lyase (\( PAL \)) catalyzes the first step by removing ammonia from L-Phe to produce \( p \)-coumaric acid, followed by a series of enzymatic steps involving cinnamate 4-hydroxylase (\( C4H \)), 4-coumarate:CoA ligase (\( 4CL \)), hydroxycinnamoyl-CoA transferase (\( HCT \)), \( p \)-coumarate 3-hydroxylase (\( C3H \)), caffeoyl-CoA \( O \)-methyltransferase (\( CCoAOMT \)), cinnamoyl-CoA reductase (\( CCR \)), ferulate 5-hydroxylase (\( F5H \)), caffeic acid \( O \)-methyltransferase (\( COMT \)), and cinnamyl alcohol dehydrogenase (\( CAD \)) catalyzing the biosynthesis of monolignols. Genes encoding these enzymes controlling the phenylpropanoid pathway in maize have been cloned (Civardi et al. 1999; Collazo et al. 1992; Gardiner et al. 2004; Guillaumie et al. 2007; Halpin et

Four brown midrib mutations in maize (bm1, bm2, bm3, and bm4) are associated with alterations of and reductions in lignin content in stover (Barrière and Argillier 1993). The bm1 mutation decreases cinnamyl alcohol dehydrogenase (CAD) activity (Halpin et al. 1998), while the bm3 mutation results in reduced caffeic acid O-methyltransferase (COMT) activity (Vignols et al. 1995). The genetic events underlying bm2 and bm4 mutations are unknown.

Several studies reported improvement in the digestibility of corn silage in ruminant feeding of brown midrib mutants, especially bm3 materials (Muller et al. 1972; Gallais et al. 1980; Colenbrander et al. 1972; Colenbrander et al. 1973; Colenbrander et al. 1975; Keith et al. 1981; Frenchick et al. 1976). However, negative effects of brown midrib genotypes have also been reported in relation to agricultural fitness, such as reduced grain and stover yield and stalk breakage (Miller et al. 1983; Lee and Brewbaker 1984; Cox and Cherney 2001). Reduced grain yield, combined with other negative effects caused by the bm3 mutation is so significant that it might be difficult to produce superior maize bm3 hybrids in terms of agricultural fitness (Miller et al. 1983; Lee and Brewbaker 1984). However, the same authors report large genetic variation among bm3 lines, and the genotype specificity among bm3 lines in relation to yield reduction Miller et al. 1983; Lee and Brewbaker 1984; Gentinetta et al. 1990).

Although many studies investigated phenotypic effects in relation to down regulation of lignin genes for maize and other species (Marita et al. 2001; Jouanin et al. 2000; Jouanin et al. 2004; Guo et al. 2001; Ralph et al. 2001; Van Doorsselaere et al. 1995; Piquemal et al. 2002), intragenic variability of OMT genes involved in lignin synthesis have not been studied to the same extent. Ultimately, characterized QTN (Quantitative Trait Nucleotide, see definition in Table 1) or QTINDEL (Quantitative Trait Insertion-Deletion) polymorphisms within these genes
would allow development of Functional Markers (FM) (Andersen and Lübberstedt 2003). Prerequisite for distinguishing the effects of different intragenic polymorphisms is low LD (linkage disequilibrium) in the population(s) analyzed. Nucleotide diversity, LD, and associations to forage quality traits have been studied for several genes involved in the phenylpropanoid pathway (Andersen et al. 2007; Fontaine and Barrière 2003; Guillet-Claude et al. 2004; Zein et al. 2007). While LD decreased rapidly within few hundred bp in the COMT and CCoAOMT2 coding genes (Guillet-Claude et al. 2004; Zein et al. 2007), LD persisted over a thousand bp in the CCoAOMT1 gene (Guillet-Claude et al. 2004).

In this study, variation in genomic sequences coding for COMT, CCoAOMT1, and CCoAOMT2 was analyzed in relation to stover cell wall digestibility for the same panel of 40 European forage maize inbred lines investigated by Andersen et al. (Andersen et al. 2007; Andersen et al. 2008) for other “lignin genes” (Experiment 1). In addition, data published by Guillet-Claude et al. (2004) in an association analysis of the three O-methyltransferase genes in relation to cell wall digestibility were re-analyzed (Experiment 2), based on a different statistical approach by using General Linear Model, and including population structure. The objectives were, (1) to identify associations between individual polymorphisms and four forage quality traits determined for maize stover in Exp. 1, (2) to evaluate the impact of sequence alignment parameters on the outcome of association studies, (3) to re-analyze data of Exp. 2 under consideration of population structure, and (4) to compare findings between both experiments for these three jointly analyzed genes.

**Methods**

**Plant Materials and phenotypic Analyses**

The 40 lines within Exp. 1 represent a broad range of Central European forage maize
germplasm, and were extremes within a larger collection of >300 maize inbreds with respect to stover cell-wall digestibility (unpublished data). Thirty-five lines originated from the current breeding program of KWS Saat AG and five lines were from the public domain (AS01, AS02, AS03, AS39, and AS40, identical to F7, F2, EP1, F288, and F4, respectively).

The inbred lines were evaluated in Grucking (sandy loam) in 2002, 2003, and 2004, and in Bernburg (sandy loam) in 2003 and 2004. The experiments consisted of 49 entries in a 7 × 7 lattice design with two replications. Nine entries from this design were consisting of not sequenced checks (such as knock-out mutants of brown midrib loci with known high cell wall digestibility). Therefore, only 40 lines were analyzed for cell wall properties and sequenced. The single row plots were 0.75 m apart and 3 m long with a total of 20 plants. The ears were manually removed and the stover was chopped 50 days after flowering. Approximately 1 kg of stover was collected and dried at 40 °C. The stover was ground to pass through a 1 mm sieve. Quality analyses were performed with near infrared reflectance spectroscopy (NIRS) based on previous calibrations on the data of 300 inbred lines (unpublished results). Four traits were analyzed: Water Soluble Carbohydrate (WSC), Organic Matter Digestibility (OMD), Neutral Detergent Fiber (NDF) and digestible Neutral Detergent Fiber (DNDF). DNDF was estimated by DNDF = 100 – (IVDMD – (100 – NDF))/NDF based on Goering and Van Soest, 1970 (Goering and Van Soest 1970). Since the investigated traits were highly heritable (Guillet-Claude et al. 2004; Barrière et al. 2003), and “trait x location” interactions were not significant, replications and locations were averaged, so that each entry was represented by mean values of the cell wall traits. The results of phenotypic analysis were published previously (Andersen et al. 2007). The heritabilities of NDF and DNDF were 86.5% and 92.2%, respectively (Andersen et al. 2008).

Thirty-four inbred lines were employed in Exp. 2, including public lines and ecotypes both
from Europe and U.S., covering substantial variation in forage quality. The lines were evaluated in three different years: 2006, 2008, and 2009 (unpublished data of INRA Lusignan). Values of stover *in vitro* cell wall digestibility was estimated by $DNDF = (100 \times (ES - (100 - NDF)) / NDF)$ (Struik 1983), based on enzymatic solubility (ES) of Aufrere and Michalet-Doreau (Aufrere and Michalet-Doreau 1983).

**DNA isolation, PCR amplification, and DNA sequencing**

The inbred lines from Exp. 1 were grown in the greenhouse for DNA isolation. Leaves were harvested three weeks after germination, and DNA was extracted using the Maxi CTAB method (Saghai-Marooft et al. 1984). Polymerase chain reaction (PCR) primers were developed based on maize mRNA sequences identified in GenBank by employing BLAST (Altschul et al. 1990) for three genes: *COMT*, *CCoAOMT1* and *CCoAOMT2*. PCR reactions contained 20 ng genomic DNA, primers (200 nM), dNTPs (200 μM), 1 M Betain and 2 units of Taq polymerase (Peqlab, Erlangen, Germany), in a total reaction volume of 50 μl. A touchdown PCR program was: denaturation at 95°C for 2 min, 15 amplification cycles: 45 sec at 95°C; 45 sec at 68°C (minus 0.5°C per cycle), 2 min at 72°C, followed by 24 amplification cycles: 45 sec at 95°C; 45 sec at 60°C, 2 min at 72°C, and a final extension step at 72°C for 10 min.

Products were separated by gel electrophoresis on 1.5% agarose gels, stained with ethidium bromide and photographed using an eagle eye apparatus (Herolab, Wiesloch, Germany). Amplicons were purified using QiaQuick spin columns (Qiagen, Valencia, USA) according to the manufacturer instructions, and directly sequenced using internal sequence specific primers and the Big Dye1.1 dye-terminator sequencing kit on an ABI 377 (PE Biosystems, Foster City, USA). Electropherograms of overlapping sequencing fragments were manually edited using the software Sequence Navigator version 1.1, from PE Biosystems.
Full alignments were built for \textit{CCoAOMT1} and \textit{CCoAOMT2} genes by using default settings of the CLUSTAL W program (Thompson et al. 1994). Several Indels were present among the different \textit{COMT} alleles in Exp. 1 (Zein et al. 2007). Thus, four different alignment parameters were set in CLUSTAL W (Thompson et al. 1994) to validate the consistency of polymorphic sites. The first alignment was based on default parameters, additional alignments were constructed by using different parameter settings in relation to gap penalty. The exon-intron structure of the three O-methyltransferase genes was estimated by alignments of genomic to mRNA sequences.

In Exp. 2, primer pairs for \textit{CCoAOMT1} and \textit{CCoAOMT2} were designed based on published cDNA sequences (accession numbers AJ242980 and AJ242981), respectively. For both genes, fragments of about 1.2 to 1.3 kb were amplified, encompassing the 5’-UTR and the complete coding region (Guillet-Claude et al. 2004). Since the \textit{COMT} promoter sequence was not available in databases, a walking-PCR procedure was performed to amplify the 5’-flanking region. Sequencing was performed for each PCR fragment in both directions by Isoprim (Toulouse, France) and MWG-Biotech (Ebersberg, Germany). The sequences containing singletons were checked by re-amplifying genomic DNA and partially re-sequencing the appropriate alleles. Contigs were constructed using SeqWeb (GCG Wisconsin Package). Sequences were aligned using CLUSTAL W (Thompson et al. 1994).

**Population structure and association analysis**

Lines evaluated in Exp. 1 were genotyped with 101 simple sequence repeat (SSRs) markers providing an even coverage of the maize genome. Population structure was estimated from the SSR data by the \textit{Structure} 2.0 software (Pritchard et al. 2000; Falush et al. 2003). The Q matrix estimating membership coefficients for each individual in each subpopulation was produced. A
burn-in length of 50,000 followed by 50,000 iterations was applied. The Admixture model was applied with independent allele frequencies.

Association between polymorphisms and mean phenotypic values were performed by the General Linear Model (GLM) analyses in TASSEL. The Q matrix produced by Structure was included as covariate in the analysis to control for populations structure. The polymorphisms were determined as significant for p-adj_Marker (based on 10,000 permutations) equal to 0.05 or less. p-adj_Marker is a permutation based experiment-wise error rate which controls the error rate over all the markers tested.

The same parameters were applied to perform the Logistic Regression in TASSEL. In this association analysis a logistic regression ratio test is used to evaluate associations involving quantitative traits while controlling for population structure (Thornsberry et al. 2001). The trait values permuted relative to the fixed haplotypes were recalculated for 10,000 permutations.

Associations were also tested by the Mixed Linear Model (MLM) in TASSEL (Yu et al. 2006). The MLM accounts not only for overall population structure (Q), but also finer scale relative kinship (K). Loiselle kinship coefficients (Loiselle et al. 1995) between lines (a K matrix) were estimated by the SPAGeDI software (Hardy and Vekemans 2002) based on the SSR data mentioned above. Negative values between two individuals in the K matrix were set to zero. The Bonferroni Step-down correction (Holm 1979) was applied to correct for multiple testing as the p-values for the MLM analysis are expressed on a single marker basis in TASSEL.

Significant intron polymorphisms in Exp.1 were analyzed for alterations in motif sequences, using the inbred line W64 as reference sequence (AY323283). Differential splicing was tested by comparing expressed sequences with genomic sequences. The discrimination of favorable and unfavorable alleles at each significant associated polymorphic site at COMT was done by
grouping lines by alleles and averaging DNDF values within the same group.

Guillet-Claude et al. (2004) tested for associations between cell wall digestibility and polymorphic sites of COMT, CCoAOMT1, and CCoAOMT2 by performing Multiple Linear Regression without considering population structure. Our re-analysis included population structure in the GLM analysis of TASSEL. The P-value of all individual polymorphisms (including singletons) was estimated based on 10,000 permutations. Population structure data were obtained from Camus-Kulandaivelu et al. (2006), where five sub-populations were found using 55 SSR loci in Structure (Pritchard et al. 2000; Falush et al. 2003).

Results

Genetic diversity within COMT, CCoAOMT1, and CCoAOMT2

The analysis of overlapping sequences of common inbred lines between Exp.1 and Exp.2 (F2, F4, F7, F288, EP1, and W64) revealed consistency of sequences for the COMT gene across both experiments. Only one SNP each was different among overlapping sequences of F4 and EP1 inbred lines, respectively (Table 1). For CCoAOMT1, F2 and F7 sequences were identical in both experiments, whereas between two (F288) and 13 (EP1) polymorphisms were identified for the other three inbred lines (Table 1). For the CCoAMT2 gene, only two inbred lines were common between both studies. One SNP and one large Indel were identified for F288, whereas sequences perfectly matched for EP1. For each of the three genes, sequences of lines used in both experiments, were in all cases most similar across both experiments: even when comparing the EP1 CCoAOMT1 allele from Exp. 1 with all alleles in Exp. 2, the Exp. 2 EP1 allele was the most similar sequence to Exp. 1 EP1 among all Exp. 2 sequences.

Analysis of haplotype diversity for the 40 inbred lines employed in Exp. 1 revealed 13, 9, and 6 haplotypes for COMT, CCoAOMT1, and CCoAOMT2, respectively (Table 2). The range of
haplotype means for all four stover quality traits was larger for COMT compared to the other two genes (Table 2). A total number of 26, 12, and 14 haplotypes was previously identified by Guillet-Claude et al. (2004), for COMT, CCoAOMT1, and CCoAOMT2, respectively. When sequences from both experiments were jointly analyzed for haplotype numbers, 38 haplotypes were discriminated for COMT, and 15 haplotypes each for CCoAOMT1, and CCoAOMT2 (data not shown).

**Association analysis of COMT in Experiment 1**

Varying alignment lengths and number of polymorphic sites were observed when four different parameter sets were applied for aligning COMT by CLUSTAL W. By GLM based on the first alignment (default settings), 16 polymorphisms were significantly associated with DNDF, whereas alignment settings 2, 3, and 4 identified 13, 14, and 14 polymorphic sites, respectively (Table 3), significantly associated with DNDF. Different gap opening and gap extension penalties lead to different Indel numbers and sizes. Some polymorphisms were observed to change positions, others were created or vanished. Only eight polymorphic sites showing significant associations with DNDF were in common among all four alignments.

All subsequent analyses were performed for default sequence alignment settings, in order to be able to compare findings reported here with previous studies (Andersen et al. 2007; Andersen et al. 2008). In Exp. 1, a total number of 16 COMT polymorphic sites were significantly associated with DNDF using GLM analysis: 9 SNPs and 7 Indels (Table 4). The same sites were identified by MLM analysis, with the exception of one SNP located at 1439 bp and one Indel located at 1638 bp based on the reference sequence AY323283 of Genbank. Logistic regression analysis in TASSEL identified a larger number of significant polymorphic sites in association with DNDF: 16 SNPs and 9 Indels, respectively (Table 4).
No polymorphic sites were observed to be significantly associated with DNDF in the first exon, and most of the polymorphic sites identified for this trait were located in the intron, where 13 SNPs and 9 Indels were detected. In the second exon, three SNPs were significantly associated with DNDF in positions 2103, 2178, and 2185 bp, each leading to amino acid substitutions (Ser/Pro, His/Asp, and Arg/Pro, respectively). Most of the significantly associated polymorphic sites identified for DNDF were in high LD (Fig. 1).

No significant associated SNP or Indel was identified between COMT and NDF, OMD, and WSC by MLM or GLM (Table 4). All significantly associated polymorphic sites for these traits were identified by Logistic Regression. For NDF, a synonymous SNP was identified in the second exon of the COMT gene. Low values of LD were observed between this SNP and other polymorphic sites associated with DNDF, OMD, and WSC (Fig. 1). All polymorphic sites (six SNPs and six Indels) associated with OMD matched positions of DNDF associated polymorphic sites, most of them located in the intron. Only two SNPs (2103 and 2178 bp) were identified in the second exon of the gene, both causing changes in amino acid sequences (Ser/Pro and His/Asp, respectively). All polymorphic sites identified for WSC were located in the intron: five SNPs and one Indel with high LD among each other (Fig. 1). Six singletons in complete LD located at positions 1064, 1070, 1071, 1072, 1073, and 1074 bp were identified by GLM to be significantly associated with all four traits, but not included in Table 4.

Eight significantly associated Indels in the COMT intron affected five motifs representing binding sites for transcription factors RAV1, GAmyb, and DOFs 1, 2, and 3. Tests for differential splicing did not reveal divergent patterns of exon inclusion (data not shown). Across polymorphisms significantly associated with DNDF, lines containing mostly favorable alleles had on average higher DNDF values than lines containing mostly unfavorable alleles (Fig. 2).
However, one line (AS22) with all favorable alleles had a low DNDF value and three lines (AS26, AS29 and AS37) containing only unfavorable alleles had high DNDF values.

**Association analysis of CCoAOMT1 and CCoAOMT2 in Experiment 1**

Association analysis for CCoAOMT1 revealed substantially fewer polymorphic sites associated with the traits investigated as compared to COMT. Two SNPs were identified by GLM: one SNP was associated with NDF and located in the third intron, and a second SNP was associated with both NDF and OMD, located in the fourth intron. One additional SNP identified by Logistic Regression (Table 5) associated with WSC was identified in the fifth exon, leading to amino acid substitution. High LD was observed between SNPs of sites 726 and 944, while low LD was observed between site 1299 and sites 726 and 944 bp (Fig. 1).

Only two Indels showed significant trait associations for CCoAOMT2 by Logistic Regression, and only one was also detected by GLM (Table 5). Both Indels were located in the second intron of the CCoAOMT2 gene, and both were associated with WSC. Both polymorphisms were in complete LD. Intron polymorphisms of CCoAOMT1 and CCoAOMT2 did not cause changes in known motifs, and were not found to cause differential splicing.

**Association analysis of COMT, CCoAOMT1, and CCoAOMT2 in Experiment 2**

Guillet-Claude et al. (2004) identified polymorphic sites of COMT associated with DINAGZ, all located in non-coding regions. No association was found between polymorphisms within CCoOAMTI and DINAGZ, while an 18-bp Indel associated with DINAGZ was identified for CCoAOMT2. DINAGZ and DNDF are both estimates of CWD based on in vitro enzymatic solubility, differing on how the fully digestible part of the non-cell-wall part is computed. The correlation between the two traits is high (nearly 0.90, INRA Lusignan unpublished data). For
this reason, the association results from Guillet-Claude et al. (2004) should be phenotypically comparable to our analysis (using DNDF).

After taking population structure into account, five sub-populations were identified for this line panel by Camus-Kulandaivelu et al. (2006). In total 14 polymorphic sites showed significant associations with DNDF for the three investigated genes (Table 6). Thirteen of those polymorphic sites were singletons, and most of them were located in non-coding regions.

For COMT, three SNPs and two Indels were identified by GLM. The SNP located at position 1948 bp was also identified as significantly associated with DNDF in Exp. 1. This SNP was observed to be in high LD with another polymorphic site at position 1981 bp, not identified in Exp.1, and low LD with the other three trait-associated sites identified in Exp. 2. The two polymorphic sites at positions 749 bp and 1981 bp were also present in Exp.1, but not significantly associated with any of the four traits analyzed. Two polymorphisms located in the promoter (at positions 342 and 659 bp) were not polymorphic in Exp.1.

Contrary to the findings of Guillet-Claude et al. (2004), two SNPs and three Indels showed significant associations between CCoAOMT1 and DNDF. One SNP located at position 944 bp was also found to be significantly associated with NDF and OMD in Exp. 1. This SNP was in high LD with Indels at positions 704 and 734 bp, and in low LD with polymorphic sites at positions 956 and 972 bp. The Indel at positon 734 bp was also observed in Exp.1, but it was not significantly associated with any four traits investigated in Exp. 1. Four SNPs in CCoAOMT2 sequence were significantly associated with DNDF. All polymorphisms were singletons not identified in Exp.1, and three of them were located in non-coding regions.

Discussion

Impact of sequence alignments and association analysis methods on results of association
Changing alignments parameters for gap opening and gap extension costs leads to the creation of alignments with different Indel sizes and frequencies (Pons and Vogler 2006). Different studies investigated consequences of different alignments parameters settings in the outcome of phylogenetic studies (Wheeler 1995; Morrison and Ellis 1997; Xia et al. 2003; Terry and Whiting 2005). However, little is known about the effect of different alignment settings in association analysis. Our results demonstrate that different sets of polymorphisms are identified by association analyses, depending on alignments settings (Table 3). This is especially true for sequences with multiple Indels as observed for COMT, where choosing optimal alignment parameters might be difficult. Comparison of different alignments varying for gap costs may be used to identify the most consistent significantly associated polymorphic sites.

In this study, cell wall traits of two relatively small populations (40 and 34 lines) were tested for associations with two large sets of polymorphic sites (SNPs and Indels). The large number of independent variables (SNPs/Indels) in relation to number of dependent variables (lines/phenotypic data) increases the chances of false positive associations due to overfitting (Piepho and Gauch 2001). However, we believe that high heritability values of the cell wall traits analyzed (Andersen et al. 2008; Barrière et al. 2003) and stringent significance levels estimated by 10,000 permutations limit the rate of false positives (Bernardo 2004).

The control of systematic differences due to population structure and kinship has also been suggested to avoid spurious associations and for greatest statistical power in association studies (Yu et al. 2006; Freedman et al. 2004). In this study, two of our models (GLM and Logistic Regression) were controlling for population structure, and the mixed model (MLM) approach was controlling for both population structure and kinship. Most of the lines evaluated in this
study could be clearly identified as Flint or Dent lines (see Figure five: Andersen et al. 2007).

Across the investigated genes in this study, most of the significantly associated polymorphisms were detected by Logistic Regression, followed by GLM and MLM. A few previous studies compared results of association analyses from these methods. Andersen et al. (2005) in an attempt to validate associations of the Dwarf8 locus identified by Thornsberry et al. (2001), identified a larger number of polymorphisms associated with flowering time when using GLM (without control of population structure) as compared to Logistic Regression (with control of population structure). According to the authors, GLM identified more significant polymorphisms, because variation of causal alleles was closely correlated with population structure. In corroboration with our results, Andersen et al. (2008) observed a larger number of significant polymorphisms by GLM analysis accounting for population structure as compared to MLM accounting for both population structure and kinship corrections. Fewer significant polymorphisms identified by MLM are expected, as MLM also corrects for kinship, and significant polymorphisms identified in GLM but not in MLM might thus reflect identity of polymorphisms by descent. Investigating three QTL in 277 diverse maize inbred lines, Yu et al. (2006) concluded that the model accounting for both population structure (Q) and relative kinship (K) gave the best approximation to the cumulative distribution of p-values when compared to models containing only K, only Q and a simple model (in which family structure is ignored).

Across the three genes and four traits within Exp. 1, 13 polymorphisms were consistent across the three methods, while one polymorphism was identified by both GLM and MLM, 3 by both GLM and Logistic Regression. No significant polymorphic sites were consistent between both MLM and Logistic Regression. Logistic Regression was the method detecting the highest
number of significant polymorphisms in our study. With this method, permutation tests are performed for individual markers, not controlling for experiment-wise error rates. Consequently, a larger number of false positives are expected for this method of analysis when compared to analyses that control for experiment-wise error, like GLM in our study. Logistic Regression (and GLM in one occasion) revealed significant associations between the three OMT genes with WSC, with none of these three genes being involved in biosynthesis of soluble carbohydrates (Tables 4 and 5). However, we were able to identify one SNP in Exp.1 by Logistic Regression (but not by the other two methods) that was consistent with Exp. 2 (SNP at position 1948 of the COMT gene, Table 6). Moreover, many polymorphisms identified by Logistic Regression were also identified by GLM and MLM when considering p-values between 0.05 and 0.10, especially for associations between COMT with NDF, OMD, and WSC (data not shown). It can be inferred that these QTN/Indels are at the threshold of being detectable, and even small changes in statistical methods or parameters might lead to (non-) significance of these sites. Therefore, significant polymorphisms identified by association analyses, need to be validated by independent experiments (ongoing).

Consistency of association results across experiments

Two SNPs were identified as significantly associated in both datasets (Exp.1 and Exp.2). The SNP at position 1948 bp in the COMT intron was significantly associated with DNDF. At this C/T polymorphic site, a sequence motif for a DOF3 zinc finger transcription factor was identified for the allele with the T base present. The second SNP consistent across experiments was located in the site 944 bp at the fourth intron of CCoAOMT1. This SNP did not cause motif or splicing alterations, but it was significantly associated with both NDF and OMD in Exp.1. The consistency of these two SNPs across experiments makes them excellent candidates for
development of FM.

In addition, three polymorphic sites identified in Exp. 2 were also observed in Exp. 1, but not significantly associated with the investigated traits. The relative low number of common polymorphic sites between Exp. 1 and Exp. 2 can in part be explained by the different regions of genes that each experiment investigated. The overlapping region of each gene common to both studies was not representing the whole gene. Moreover, for some lines overlapping sequences were not completely identical in both studies (Table 1), most likely due to residual heterozygosity as the “same” lines in both studies had in all cases the most similar sequences. In addition, most of the significantly associated polymorphic sites identified in Exp. 2 were singletons, which have a low probability of being identified across datasets. Finally, Exp. 1 and 2 were conducted in quite different environments in Germany and France, respectively, which might explain for the different performance of the “same lines” with regard to CWD in both experiments.

**Genes controlling forage quality**

Most of the significant SNPs/Indels identified in Exp. 1 were identified for COMT (44 significant polymorphisms) while a smaller number of sites was identified for CCoAOMT1 and CCoAOMT2 (three and two significant polymorphic sites, respectively). Guillet-Claude et al. (2004) also observed a larger number of polymorphisms for COMT compared to CCoAOMT1 and CCoAOMT2. In both experiments, the alignment sequence size of COMT was approximately two times larger than CCoAOMT1 and CCoAOMT2, whereas the number of significantly associated polymorphisms for COMT were substantially higher than 2-fold.

Andersen et al. (2008) investigated the same set of 40 lines for six other “lignin genes” in relation to NDF, IVDOM, and DNDF, and identified significant associations for 4CL1, C3H, and
Genes. The DNDF range between haplotypes within these three genes was largest for C3H (10.0%), followed by F5H (7.7%) and 4CL1 (2.3%). COMT haplotypes differed for DNDF by 17.8% (Table 2). Larger differences between haplotypes suggest that COMT is a promising candidate gene to derive markers affecting stover cell-wall digestibility. Moreover, down regulation of COMT in Arabidopsis, poplar, alfalfa and maize affecting the lignin content/structure support the strong effect of this gene on variation for cell wall digestibility (Marita et al. 2001; Jouanin et al. 2000; Jouanin et al. 2004; Guo et al. 2001; Ralph et al. 2001; Van Doorsmelaere et al. 1995; Piquemal et al. 2002).

**Polymorphisms associated with cell wall digestibility**

The association analyses revealed a considerable number of SNPs and Indels associated with cell wall traits for the genes analyzed. Across experiments, methodologies and genes, 11 significant QTNs/Indels were located in coding regions and 52 significant polymorphisms were located in non-coding regions. Some of the significant associations identified in this study may reflect a statistical artifact. The probability of false-positive associations is higher in case of rare alleles and cases where a polymorphic site was only detected with one of the alignments. In addition, several of the identified significant associations can likely be explained by high LD to closely linked causal QTN/QTINDELs. Within an unresolved linkage block, likely only one or few polymorphisms are causative (Figs. 1 and 3). Polymorphisms causing amino acid changes are promising candidates for causative QTN / QTINDELs. The SNPs identified to cause amino acid changes in positions 2178 and 2185 bp of COMT (Table 4) were identified by all statistical methods, being, therefore, strong candidate sites to cause DNDF variation.

Several significantly associated polymorphisms located in non-coding regions have been reported. Andersen et al. (2008) identified two respective SNPs in the F5H gene as significantly
associated with NDF and DNDF. Wilson et al. (2004) identified significant intron polymorphisms associated with kernel composition traits for \(sh1\) and \(sh2\) genes. Polymorphisms located in introns can be “causative” if affecting transcript abundance (expression and stability) of genes (Jeong et al. 2006). The analysis of motif alterations revealed eight significant associated Indels in \(COMT\) introns disrupting the sequence of five motifs, which represent binding sites for transcription factors RAV1, GAmyb, and DOFs 1, 2, and 3. DOF–type transcription factors are well known for their function as transcriptional activators or repressors of tissue-specific gene expression. More recently, they have been identified as putative regulators of lignin biosynthetic genes, based on the microarray analyses of Arabidopsis ectopic lignification mutants (Rogers et al. 2005). Therefore, although polymorphisms in positions 1439, 1920, 1948, 1952, 1953, and 1954 bp are located in the \(COMT\) intron, they might play an important role as cis-regulators of gene expression, causing changes in the lignin content and variation in DNDF values. Intron polymorphisms might also be part of splicing signals that could cause differential splicing encoding structurally and functionally distinct protein products. However, no divergent patterns of exon inclusion were observed for the three genes analyzed (data not shown).

Conclusions

Lines containing favorable alleles at putative QTN or QTINDELs were generally, but not always associated with high phenotypic values for cell wall digestibility. As cell wall digestibility is a polygenic inherited quantitative trait, likely unfavorable alleles at other loci affecting this trait are responsible for this finding, either by strong main effects or interaction with \(COMT\). In agreement with these observations, Barrière et al. (2008) identified conventional inbreds with equally high cell wall digestibility as \(bm3\) lines, supporting that other genes can
have strong effects on this trait. Identifying and combining favorable alleles from multiple genes affecting cell wall digestibility is, therefore, mandatory for functional marker-based improvement of stover for forage and biofuel usage.

List of abbreviations


Acknowledgments

We thank KWS Saat AG (Einbeck) and the German ministry for education and science (BMBF) for financial support of the EUREKA project Cerequal. EAB and YC are supported by the RF Baker Center for Plant Breeding at Iowa State University.

Authors’ contributions

EAB prepared the manuscript. EAB, JRA, and YC performed data analysis. IZ carried out allele sequencing. GW contributed to experimental design. BD and JE provided phenotypic data. MO provided the SSR data and together with GW contributed to experimental design. UKF performed sequence alignments. YB provided data from Exp.2 and reviewed the manuscript. TL coordinated the project and together with EAB prepared the manuscript. All authors read and approved the final manuscript.
References


Pollmer WG, Phipps RH:319-339


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


reduced caffeic acid O-methyltransferase activity. Plant Physiol 123:1363-1374


Comparison of lignin deposition in three ectopic lignification mutants. New Phytol 168: 123–140


methyltransferase activity. Plant J 8:855–64


Table 1. Comparison of sequences of common inbred lines in Exp. 1 and Exp. 2

<table>
<thead>
<tr>
<th>Inbred lines</th>
<th>COMT Bp Overlap</th>
<th>COMT Differences</th>
<th>CCoAOMT1 Bp Overlap</th>
<th>CCoAOMT1 Differences</th>
<th>CCoAOMT2 Bp Overlap</th>
<th>CCoAOMT2 Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2</td>
<td>2051</td>
<td>0</td>
<td>1326</td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>F4</td>
<td>1947</td>
<td>1 SNP</td>
<td>1327</td>
<td>5 SNP, 1 Indel</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>F7</td>
<td>2090</td>
<td>0</td>
<td>1249</td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>F288</td>
<td>2052</td>
<td>0</td>
<td>1328</td>
<td>1 SNP, 1 Indel</td>
<td>655</td>
<td>1 SNP, 1 Indel</td>
</tr>
<tr>
<td>EP1</td>
<td>2115</td>
<td>1 SNP</td>
<td>1254</td>
<td>9 SNP, 4 Indel</td>
<td>751</td>
<td>0</td>
</tr>
<tr>
<td>W64</td>
<td>2088</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
**Table 2.** Number of haplotypes based on single nucleotide polymorphisms (SNPs) in the *COMT*, *CCoAOMT1*, and *CCoAOMT2* genes of maize, and minimum, maximum and variance of phenotypic values of lines representing individual haplotypes

<table>
<thead>
<tr>
<th>Gene</th>
<th>No Hap.</th>
<th>WSC Min</th>
<th>WSC Max</th>
<th>WSC Range</th>
<th>NDF Min</th>
<th>NDF Max</th>
<th>NDF Range</th>
<th>OMD Min</th>
<th>OMD Max</th>
<th>OMD Range</th>
<th>DNDF Min</th>
<th>DNDF Max</th>
<th>DNDF Range</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>COMT</em></td>
<td>13</td>
<td>13.69</td>
<td>23.68</td>
<td>9.99</td>
<td>52.92</td>
<td>65.90</td>
<td>12.98</td>
<td>61.55</td>
<td>76.65</td>
<td>15.1</td>
<td>42.18</td>
<td>59.98</td>
<td>17.8</td>
</tr>
<tr>
<td><em>CCoAOMT1</em></td>
<td>9</td>
<td>16.04</td>
<td>23.28</td>
<td>7.24</td>
<td>52.93</td>
<td>61.93</td>
<td>9.00</td>
<td>68.02</td>
<td>75.55</td>
<td>7.53</td>
<td>50.20</td>
<td>59.98</td>
<td>9.78</td>
</tr>
<tr>
<td><em>CCoAOMT2</em></td>
<td>6</td>
<td>13.28</td>
<td>20.69</td>
<td>7.41</td>
<td>54.95</td>
<td>63.03</td>
<td>8.08</td>
<td>71.00</td>
<td>74.46</td>
<td>3.46</td>
<td>54.07</td>
<td>59.98</td>
<td>5.91</td>
</tr>
</tbody>
</table>
Table 3. Significantly associated polymorphic sites of *COMT* with DNDF, identified by GLM, based on alignments resulting from four different parameters settings of CLUSTAL W

<table>
<thead>
<tr>
<th>Site</th>
<th>Alignments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 (2461 bp)</td>
</tr>
<tr>
<td>701</td>
<td>X</td>
</tr>
<tr>
<td>737</td>
<td>X</td>
</tr>
<tr>
<td>781</td>
<td></td>
</tr>
<tr>
<td>782</td>
<td>X</td>
</tr>
<tr>
<td>787</td>
<td>X</td>
</tr>
<tr>
<td>793</td>
<td></td>
</tr>
<tr>
<td>824</td>
<td>X</td>
</tr>
<tr>
<td>845</td>
<td>X</td>
</tr>
<tr>
<td>855</td>
<td>X</td>
</tr>
<tr>
<td>897</td>
<td>X</td>
</tr>
<tr>
<td>962</td>
<td>X</td>
</tr>
<tr>
<td>1010</td>
<td>X</td>
</tr>
<tr>
<td>1059</td>
<td>X</td>
</tr>
<tr>
<td>1240</td>
<td>X</td>
</tr>
<tr>
<td>1351</td>
<td>X</td>
</tr>
<tr>
<td>1352</td>
<td>X</td>
</tr>
<tr>
<td>1354</td>
<td>X</td>
</tr>
<tr>
<td>1356</td>
<td>X</td>
</tr>
<tr>
<td>1388</td>
<td>X</td>
</tr>
<tr>
<td>1417</td>
<td></td>
</tr>
<tr>
<td>1419</td>
<td></td>
</tr>
<tr>
<td>1432</td>
<td></td>
</tr>
<tr>
<td>1562</td>
<td>X</td>
</tr>
<tr>
<td>1612</td>
<td>X</td>
</tr>
<tr>
<td>1619</td>
<td>X</td>
</tr>
</tbody>
</table>

Site numbers denote bp alignment positions of individual SNPs and starting positions of Indels based on reference sequence AY323283 of Genbank.
**Table 4.** Polymorphic sites of *COMT* associated with DNDF identified by GLM, MLM, and Logistic Regression tests for DNDF

<table>
<thead>
<tr>
<th>Site</th>
<th>Snp/Indel</th>
<th>E/I</th>
<th>aa change</th>
<th>GLM</th>
<th>MLM</th>
<th>REG</th>
<th>Other trait*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1233</td>
<td>C-G</td>
<td>I</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>WSC</td>
</tr>
<tr>
<td>1235</td>
<td>A-T</td>
<td>I</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>WSC</td>
</tr>
<tr>
<td>1236</td>
<td>C-G</td>
<td>I</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>WSC</td>
</tr>
<tr>
<td>1240</td>
<td>C-T</td>
<td>I</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>WSC</td>
</tr>
<tr>
<td>1243</td>
<td>10</td>
<td>I</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>WSC</td>
</tr>
<tr>
<td>1261</td>
<td>A-G</td>
<td>I</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>WSC</td>
</tr>
<tr>
<td>1296</td>
<td>C-T</td>
<td>I</td>
<td>–</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>OMD</td>
</tr>
<tr>
<td>1331</td>
<td>C-T</td>
<td>I</td>
<td>–</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>OMD</td>
</tr>
<tr>
<td>1377</td>
<td>A-T</td>
<td>I</td>
<td>–</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>OMD</td>
</tr>
<tr>
<td>1381</td>
<td>7</td>
<td>I</td>
<td>–</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>OMD</td>
</tr>
<tr>
<td>1439</td>
<td>A-C</td>
<td>I</td>
<td>–</td>
<td>X</td>
<td>–</td>
<td>X</td>
<td>–</td>
</tr>
<tr>
<td>1449</td>
<td>4-8</td>
<td>I</td>
<td>–</td>
<td>X</td>
<td>X</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>1547</td>
<td>6</td>
<td>I</td>
<td>–</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>OMD</td>
</tr>
<tr>
<td>1589</td>
<td>A-G</td>
<td>I</td>
<td>–</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>OMD</td>
</tr>
<tr>
<td>1638</td>
<td>1</td>
<td>I</td>
<td>–</td>
<td>X</td>
<td>–</td>
<td>X</td>
<td>OMD</td>
</tr>
<tr>
<td>1811</td>
<td>6</td>
<td>I</td>
<td>–</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>OMD</td>
</tr>
<tr>
<td>1902</td>
<td>1</td>
<td>I</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>X</td>
<td>–</td>
</tr>
<tr>
<td>1907</td>
<td>3</td>
<td>I</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>X</td>
<td>–</td>
</tr>
<tr>
<td>1916</td>
<td>A-G</td>
<td>I</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>X</td>
<td>–</td>
</tr>
<tr>
<td>1917</td>
<td>A-C</td>
<td>I</td>
<td>–</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>–</td>
</tr>
<tr>
<td>1918</td>
<td>A-G</td>
<td>I</td>
<td>–</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>–</td>
</tr>
<tr>
<td>1919</td>
<td>C-T</td>
<td>I</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>X</td>
<td>–</td>
</tr>
<tr>
<td>1920</td>
<td>28</td>
<td>I</td>
<td>–</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>OMD</td>
</tr>
<tr>
<td>1948</td>
<td>C-T</td>
<td>I</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>X</td>
<td>–</td>
</tr>
<tr>
<td>1952</td>
<td>C-T</td>
<td>I</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>X</td>
<td>–</td>
</tr>
<tr>
<td>1953</td>
<td>A-T</td>
<td>I</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>X</td>
<td>–</td>
</tr>
<tr>
<td>1954</td>
<td>77</td>
<td>I</td>
<td>–</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>OMD</td>
</tr>
<tr>
<td>2032</td>
<td>A-C</td>
<td>I</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>X</td>
<td>–</td>
</tr>
<tr>
<td>2103</td>
<td>C-T</td>
<td>E2</td>
<td>Ser-Pro</td>
<td>–</td>
<td>–</td>
<td>X</td>
<td>OMD</td>
</tr>
<tr>
<td>2178</td>
<td>C-G</td>
<td>E2</td>
<td>His - Asp.</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>OMD</td>
</tr>
<tr>
<td>2185</td>
<td>C-G</td>
<td>E2</td>
<td>Arg - Pro</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>–</td>
</tr>
<tr>
<td>2693</td>
<td>A-C</td>
<td>E2</td>
<td>Syn.</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>NDF</td>
</tr>
</tbody>
</table>

*All polymorphic sites associated with WSC, OMD and NDF were identified only by Logistic Regression. E: exon; I: intron; aa: Amino Acid; Syn.: Synonymous substitution
Site numbers denote bp alignment positions of individual SNPs and starting position of Indels based on reference sequence AY323283 of Genbank.
Table 5. Significantly associated polymorphic sites of *CCoAOMT1* and *CCoAOMT2* genes identified by GLM, MLM and Logistic Regression (REG) with NDF, OMD, and WSC, respectively.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Trait</th>
<th>Site</th>
<th>Snp/Indel</th>
<th>E/I</th>
<th>aa change</th>
<th>GLM</th>
<th>MLM</th>
<th>REG</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>CCoAOMT1</em></td>
<td>NDF</td>
<td>726</td>
<td>C-T</td>
<td>I3</td>
<td>–</td>
<td>X</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>NDF</td>
<td>944</td>
<td>C-T</td>
<td>I4</td>
<td>–</td>
<td>X</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>OMD</td>
<td>944</td>
<td>C-T</td>
<td>I4</td>
<td>–</td>
<td>X</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>WSC</td>
<td>1299</td>
<td>C-G</td>
<td>E5</td>
<td>Ala - Gly</td>
<td>–</td>
<td>–</td>
<td>X</td>
</tr>
<tr>
<td><em>CCoAOMT2</em></td>
<td>WSC</td>
<td>404</td>
<td>1</td>
<td>I2</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>WSC</td>
<td>414</td>
<td>4</td>
<td>I2</td>
<td>–</td>
<td>X</td>
<td>–</td>
<td>X</td>
</tr>
</tbody>
</table>

E: exon; I: intron; aa: Amino Acid.
Site numbers denote bp alignment positions of individual SNPs and starting position of Indels. (Based on reference sequences AY323264 and AY279022, for *CCoAOMT1* and *CCoAOMT2*, respectively).
Table 6. Significant polymorphic sites identified in Exp. 2 in comparison to Exp. 1

<table>
<thead>
<tr>
<th>Gene</th>
<th>Exp.2 polymorphisms</th>
<th>Exp.1</th>
<th>Site</th>
<th>Region</th>
<th>SNP/Indel</th>
<th>aa Change</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMT</td>
<td></td>
<td></td>
<td>342</td>
<td>Prom</td>
<td>A-T</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>659*</td>
<td>Prom</td>
<td>1</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>749*</td>
<td>E1</td>
<td>6</td>
<td>–</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1948*</td>
<td>I1</td>
<td>G-T</td>
<td>–</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1981*</td>
<td>I1</td>
<td>C-T</td>
<td>–</td>
<td>0.92</td>
</tr>
<tr>
<td>CCaOMT1</td>
<td></td>
<td></td>
<td>704*</td>
<td>I3</td>
<td>3</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>734*</td>
<td>I3</td>
<td>2</td>
<td>–</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>944*</td>
<td>I4</td>
<td>C-G</td>
<td>–</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>956*</td>
<td>I4</td>
<td>11</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>972*</td>
<td>I4</td>
<td>A-G</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>CCaOMT2</td>
<td></td>
<td></td>
<td>187*</td>
<td>E1</td>
<td>A-G</td>
<td>His - Arg</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>688*</td>
<td>I3</td>
<td>A-G</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>717*</td>
<td>I3</td>
<td>C-T</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>720*</td>
<td>I3</td>
<td>C-T</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

E: exon; I: intron; Prom: promoter; aa: Amino Acid; Syn.: Synonymous substitution; *: singleton
Site numbers denote bp alignment position of individual SNPs and starting position of Indels based on the reference sequences AY323283, AY323264 and AY279022, for COMT, CCaOMT1 and CCaOMT2, respectively.
Figures

Figure 1. Linkage disequilibrium among significantly associated SNPs and Indels estimated by TASSEL, for DNDF, OMD, NDF and WSC of a) COMT, b) CCoAOMT1 in Experiment 1.

Figure 2. Discrimination of favorable and unfavorable COMT alleles in relation to average DNDF for lines in Exp.1.

Figure 3. Linkage disequilibrium among significantly associated SNPs and Indels estimated by TASSEL, for DNDF of a) COMT, b) CCoAOMT1 c) CCoAOMT2 in Experiment 2.
Figure 1

a) COMT

b) CCoAOMT1
Figure 2
Figure 3
CHAPTER 6. POLYMORPHISMS IN MONOLIGNOL BIOSYNTHETIC GENES ARE ASSOCIATED WITH BIOMASS YIELD AND AGRONOMIC TRAITS IN EUROPEAN MAIZE (ZEA MAYS L.)

Yongsheng Chen\textsuperscript{1,2,6}, Imad Zein\textsuperscript{3}, Everton Alen Brenner\textsuperscript{1}, Jeppe Reitan Andersen\textsuperscript{4}, Mathias Landbeck\textsuperscript{5}, Milena Ouzunova\textsuperscript{5}, Thomas Lübbestedt\textsuperscript{1,7}

\textsuperscript{1}: Department of Agronomy, Iowa State University, Ames, Iowa 50011, USA
\textsuperscript{2}: Interdepartmental Genetics Graduate Program, Iowa State University, Ames, Iowa 50011, USA
\textsuperscript{3}: Department of Agronomy and Plant Breeding, Technical University of Munich, Am Hochanger 2, 85354 Freising-Weihenstephan, Germany
\textsuperscript{4}: Department of Genetics and Biotechnology, University of Aarhus, Research Center Flakkebjerg, 4200 Slagelse, Denmark
\textsuperscript{5}: KWS Saat AG, Grimsehlstr. 31, 37555 Einbeck, Germany
\textsuperscript{6}: Primary researcher and author
\textsuperscript{7}: Author for correspondence

Modified from the paper published in BMC Plant Biology 2010, 10: 12

Abstract

**Background:** Reduced lignin content leads to higher cell wall digestibility and, therefore, better forage quality and increased conversion of lignocellulosic biomass into ethanol. However, reduced lignin content might lead to weaker stalks, lodging, and reduced biomass yield. Genes encoding enzymes involved in cell-wall lignification have been shown to influence both cell wall digestibility and yield traits.

**Results:** In this study, associations between monolignol biosynthetic genes and plant height (PHT), days to silking (DTS), dry matter content (DMC), and dry matter yield (DMY) were
identified by using a panel of 39 European elite maize lines. In total, 10 associations were detected between polymorphisms or tight linkage disequilibrium (LD) groups within the COMT, CCoAOMT2, 4CL1, 4CL2, F5H, and PAL genomic fragments, respectively, and the above mentioned traits. The phenotypic variation explained by these polymorphisms or tight LD groups ranged from 6% to 25.8% in our line collection. Only 4CL1 and F5H were found to have polymorphisms associated with both yield and forage quality related characters. However, no pleiotropic polymorphisms affecting both digestibility of neutral detergent fiber (DNDF), and PHT or DMY were discovered, even under less stringent statistical conditions.

**Conclusion:** Due to absence of pleiotropic polymorphisms affecting both forage yield and quality traits, identification of optimal monolignol biosynthetic gene haplotype(s) combining beneficial quantitative trait polymorphism (QTP) alleles for both quality and yield traits appears possible within monolignol biosynthetic genes. This is beneficial to maximize forage and bioethanol yield per unit land area.

**Background**

Elevating the polysaccharide to lignin ratio is one possible approach to improve the quality of biofeedstocks for ethanol conversion (Lorenz et al. 2009). It is believed that cell wall lignin content is negatively correlated with forage digestibility (Ralph et al. 2004) and bioethanol production (Torney et al. 2007). Removing lignin by oxidative pretreatment could significantly increase the release of available sugars in subsequent enzyme hydrolysis compared to the untreated control (Duncan et al. 2009). In maize, a 1% increase in available cellulose is expected to increase the potential ethanol production from 101.6 to 103.3 gallons per dry ton of biomass, as calculated using the U.S. Department of Energy’s Theoretical Ethanol Yield Calculator and Feedstock Composition Database (U.S. Department of Energy). Theoretical maximum ethanol yields from
biomass are highly correlated ($r^2=0.9$) with acid detergent lignin concentration (Isci et al. 2008). According to Lorenz et al. (2009), variation in ethanol yield is driven by glucan convertibility, which is highly correlated with ruminal digestibility and lignin content. Besides the lignin content, other aspects of cell wall lignification like the ratio of syringyl to guaiacyl lignin units affect cell wall digestibility (Grabber et al. 2004; Fontaine et al. 2003) and, therefore, likely ethanol production from biofeedstocks. The syringyl to guaiacyl ratio impacts the efficiency of cell wall hydrolysis in forage sorghums (Corredor et al. 2009). In summary, modification of cell wall lignification is a promising route to improve the quality of bioenergy crops.

However, reduced lignin content can influence the overall plant performance. Generally, reduced lignin content results in weaker stalks, reduced stover and grain yield, and delayed maturity (Pedersen et al. 2005). In maize, *brown-midrib (bm)* mutants show a decreased lignin content and increased cell-wall digestibility (Barrière and Argillier 1993). For instance, lignin content is reduced by one third and cell wall digestibility is increased by 9% in *bm3* lines or hybrids (Barrière et al. 2004). However, maize *bm* lines or hybrids show reduced vigor during vegetative growth, a high incidence of stalk breakage at maturity, and decreased grain and stover yield (Miller and Geadelmann 1983; Miller et al. 1983; Ballard et al. 2001; Cox et al. 2001). Similarly, *bm* hybrids of Sudan grass and sorghum also show reduced dry matter yield (Casler et al. 2003; Beck et al. 2007). Genetically engineered tobacco with reduced *CCoAOMT* (Pinçon et al. 2001) or *PAL* activities (Elkind et al. 1990), poplar with down-regulated *CCR* activity (Leplé et al. 2007), *Arabidopsis* with a mutation in the *CCR1* (Mir Derikvand et al. 2008), *C3H* (Franke et al. 2002), and *C4H* genes (Schilmiller et al. 2009), or with double mutations in the *COMT1* and *CCoAOMT1* genes (Do et al. 2007) showed reduced plant size. By silencing the *HCT* gene in *Arabidopsis*, Besseau et al. (2007) obtained mutants with modified lignin structure as well as
repressed plant growth. Silencing of \textit{HCT} resulted in redirection of the metabolic flux into flavonoids, which suppressed auxin transport.

Decreased lignin content does not necessarily have negative effects on plant growth. After divergent selection for fiber concentration in maize, Wolf et al. (1993) found only weak and inconsistent correlations between lignin content and various agronomic traits. Weller et al. (Weller et al. 1985) found no yield difference between \textit{bm3} and wildtype isolines. He et al. (2003) developed \textit{O}-methyltransferase down-regulated maize with a 17\% decrease in lignin content, increased digestibility, without effect on dry matter yield. In aspen, repression of \textit{4CL} led to a 45\% reduction in lignin content (Hu et al. 1999). While the structural integrity at both the cellular and whole-plant level was not affected, enhanced leaf, root, and stem growth were observed, as well as increased cellulose content (Hu et al. 1999). By simultaneously silencing \textit{HCT} and \textit{CHS} genes, Besseau et al. (2007) obtained normal growing \textit{Arabidopsis} plants with substantially altered lignin composition. In summary, cell-wall lignification is generally, but not always, negatively correlated with biomass yield and other agronomic traits. These correlations can be due to: (1) linkage of genes controlling monolignol biosynthesis and biomass yield, (2) pleiotropy at the level of genes but not QTPs within monolignol biosynthetic genes affecting both groups of traits, and (3) pleiotropic effects of QTP(s) within monolignol biosynthetic genes. The underlying genetic cause(s) for these correlations impact the strategy for breeding of bioenergy crops.

Ten enzymes are involved in converting phenylalanine to monolignols in maize, and the majority is encoded by two or more genes (Guillaumie et al. 2007). Four genes encode PAL proteins in \textit{Arabidopsis}, which catalyze the first step in the phenylpropanoid pathway (Raes et al. 2003). In maize, PAL has both phenylalanine and tyrosine ammonia lyase activity (Rösler et al. 1997) and at least five contigs with \textit{PAL/TAL} annotation were identified (Guillaumie et al. 2007).
The other enzymes involved in biosynthesis of monomers include cinnamate 4-hydroxylase (C4H), 4-coumarate:CoA ligase (4CL), hydroxycinnamoyl-CoA transferase (HCT), p-coumarate 3-hydroxylase (C3H), caffeoyl-CoA O-methyltransferase (CCoAOMT), cinnamoyl-CoA reductase (CCR), ferulate 5-hydroxylase (F5H), caffeic acid O-methyltransferase (COMT), and cinnamyl alcohol dehydrogenase (CAD), with at least two, seven, two, one, five, eight, two, one, and seven sequences were identified, respectively (Guillaumie et al. 2007). Association mapping is a promising approach to identify candidate QTPs for traits of interest (Guillet-Claude et al. 2004; Andersen et al. 2007; Andersen et al. 2008; Thornsberry et al. 2001). The CCoAOMT2 gene is co-localized with a QTL for cell wall digestibility and lignin content (Roussel et al. 2002), and an 18-bp indel in the first exon was found to be associated with cell wall digestibility (Guillet-Claude et al. 2004). In addition, associations have been identified between neutral detergent fiber (NDF) and polymorphisms within PAL, 4CL1, C3H, and F5H genes, between in vitro digestibility of organic matter (IVDOM) and polymorphisms within PAL, 4CL1, and C3H, and between digestibility of neutral detergent fiber (DNDF) and polymorphisms in C3H and F5H genes (Andersen et al. 2007; Andersen et al. 2008). However, genes encoding any of these 10 enzymes have so far not been studied in relation to biomass yield-related traits. In this study, the relationship between 10 monolignol biosynthetic genes belonging to eight enzyme encoding genes or gene families and the biomass yield-related traits: plant height (PHT), days to silking (DTS), dry matter content (DMC), and dry matter yield (DMY) were analyzed. Only one or two gene member(s) of each gene family were amplified. Our objectives were to investigate, (1) whether candidate quantitative trait polymorphisms (QTPs) for these four traits can be identified in monolignol biosynthetic genes, and (2) whether candidate QTPs for biomass yield-related traits and cell wall digestibility traits act pleiotropically by comparing the results of this study.
with results from previous forage trait association studies (Andersen et al. 2007; Andersen et al. 2008; Brenner et al. 2010). The results are discussed with respect to implications for breeding of maize for forage and lignocellulosic ethanol production.

Methods

Plant materials

A panel of 39 European elite inbred lines including 22 Flint and 17 Dent lines used for forage quality studies (Andersen et al. 2007; Andersen et al. 2008; Brenner et al. 2010) were employed in this study. Five lines (AS01=F7, AS02=F2, AS03=EP1, AS39=F288, and AS40=F4) were from the public domain, the remaining inbred lines were provided by KWS Saat AG (Table 1). These lines were selected as extremes with respect to DNDF from a larger set of >300 European lines (unpublished data).

Agronomic trials

Four biomass yield-related traits were evaluated for these 39 lines in Grucking (sandy loam) and Bernburg (sandy loam) in 2002 and 2003, respectively. Field trials were performed as 7*7 lattice design with two replications in each environment. 20 plants were planted per plot in single row plots, 3 m long and 0.75 m apart. Analysis of forage quality related traits: water soluble carbohydrate (WSC), in vitro digestibility of organic matter (IVDOM), neutral detergent fiber (DNF), and digestibility of neutral detergent fiber (DNDF) have previously been reported (Brenner et al. 2008). In our study, four biomass yield-related traits PHT, DTS, DMC, and DMY were analyzed. PHT was measured as distance from soil level to the lowest tassel branch after flowering. DTS was measured as days from sowing to silking. Dry matter content (DMC) of stover (g/kg) (ears were manually removed) was determined 50 days after flowering and dry matter yield (DMY) was measured in tons per hectare.
**Phenotypic data analyses**

Mean values, heritability, and variance components of each biomass yield-related trait and correlations between the above mentioned eight traits were calculated in PLABSTAT version 3A (Utz 2003). Briefly, analyses of variance were performed for each experiment separately. Adjusted entry means and effective error mean squares were used to compute the combined variances and covariances across environments for each trait. The sums of squares for entries were subdivided into variation among inbred lines, environments, interaction between inbred lines and environments, and error. Variance components were computed for lines and environments, considering them as random effects in the statistical model: Phenotype=effects of lines + effects of environments + effects of lines by environment (P=mean+L+E+L×E). F-tests were employed for testing the homogeneity of lines, environments and interactions between lines and environments according to the approximation given by Satterthwaite (Satterthwaite 1946). Heritabilities ($h^2$) for each trait were calculated on an entry-mean basis, and confidence intervals for $h^2$ were obtained according to Knapp et al. (1985). Phenotypic and genotypic correlations between eight traits were calculated by standard procedures (Mode and Robinson 1959).

**DNA extraction, amplification, and sequencing**

Leaves of each of the 39 lines were harvested in the greenhouse three weeks after germination for DNA extraction by the Maxi CTAB method (Saghai-Maroof et al. 1984). Primers for PCR amplification of $C4H$, $4CL1$, $4CL2$, $C3H$, $F5H$, $CAD$, $PAL$, $COMT$, $CCoAMT1$, and $CCoAMT2$, as well as amplification conditions were described elsewhere (Andersen et al. 2007; Andersen et al. 2008; Brenner et al. 2010; Zein et al. 2007). Two overlapping fragments were amplified for $PAL$ and $COMT$ to cover the complete genes, whereas partial gene sequences were obtained for the other genes. Sequences were aligned in CLUSTALW (Thompson et al. 1994) and stored in...
Nexus format for haplotype analysis in DnaSP (Rozas et al. 2003). Only one member of the \textit{PAL}, \textit{C4H}, \textit{F5H}, and \textit{CAD} gene families, respectively, was amplified. The reference sequences used for primer design were L77912, AY104175, AX204869, and AJ005702 (GenBank accession number). Primers for \textit{COMT} and \textit{C3H} were designed based on M73235 and AY17051. Two members of the \textit{CCoAOMT} gene family corresponding to \textit{CCoAOMT1} and \textit{CCoAOMT2} (Civardi et al. 1999) were amplified. Two members of the \textit{4CL} gene family corresponding to the sequences reported by Puigdomenech et al. (2001) were amplified.

**Population structure and association analysis**

101 publicly available simple sequence repeat markers (SSR) (http://www.maizegdb.org/ssr.php), evenly distributed across the whole genome of maize, were employed to genotype the 39 inbred lines. SSR data were used to infer the population structure in Structure 2.0 software (Pritchard et al. 2000; Falush et al. 2003). Individual lines were grouped based on marker profiles by the Bayesian clustering method of Structure 2.0. The membership coefficients for each individual in each subpopulation were calculated with a burn-in length of 50,000 followed by 50,000 iterations and stored in a Q matrix. Inbreds were treated as haploids. Based on these SSR marker data, finer scale relative kinship (K)- Loiselle kinship coefficients (Loiselle et al. 1995) between lines were calculated in SPAGeDi (Hardy and Vekemans 2002). Values on the diagonal of the K matrix were set as 2, and negative values in the matrix indicating that two individuals were less related than randomly chosen individuals (Hardy and Vekemans 2002) were set to 0.

Association analyses were carried out using the general linear model (GLM), and mixed linear model (MLM) in TASSEL 2.01 software (Yu et al. 2006) to test associations between polymorphisms of the 10 monolignol biosynthetic genes and four biomass traits. The threshold
for $P$-values was set to 0.05. In all models, the Q matrix was used to account for overall population structure. 10,000 permutations were used to determine the $P$-value for association of each polymorphism by GLM. The $P$-value adjusted for multiple tests was obtained by a step-down MinP procedure (Ge et al. 2003), implemented in TASSEL. For MLM, the K matrix was included to account for relative kinship between individuals (Yu et al. 2006). Trait associated polymorphisms with $r^2>0.85$ and $D'>0.9$ were assigned to a tight LD group (Mizuta et al. 2006). The phenotypic variation explained by this tight LD group was considered to be equal to the phenotypic variation of that polymorphism with the largest effect in this region. The False Discovery Rate (FDR) was determined to correct for multiple testing by MLM (Benjamini and Hochberg et al. 1995).

Results

Phenotypic data analyses

Mean phenotypic values for individual lines across four environments ranged from 109.3 to 197.1 cm for PHT, 68.5 to 85.7 days for DTS, 23.2% to 36.0% for DMC, and 2.5 to 8.4 t/ha for DMY. Overall mean values were 152.0 cm, 78.6 days, 28.5 %, and 5.3 t/ha, respectively, for these four traits (Table 1). Variance components for genotype and interactions between genotype and environment were significant ($P=0.01$) and variance components for environment were significant ($P=0.01$) for PHT, DTS, and DMC. Heritabilities were 88.0 %, 92.0 %, 85.7 %, and 81.9 % for PHT, DTS, DMC, and DMY, respectively (Table 1). Means of dent lines were significantly higher than means of flint lines for DTS ($P=0.01$), DMY ($P=0.01$), and PHT ($P=0.05$), whereas DMC was not significantly different between dent and flint lines.

PHT was positively correlated with DTS and DMY at both phenotypic and genotypic levels, with phenotypic and genotypic correlation coefficients ranging from $r=0.62$ ($P=0.01$) to 0.69
DNDF was negatively correlated with PHT (phenotypic correlation coefficient \( r_p = -0.45 \), genotypic correlation coefficient \( r_g = -0.47 \), \( P = 0.01 \)), as well as DMY (\( r_g = -0.24 \), \( P = 0.05 \)) (Table 2).

Previous studies reported the haplotype diversity of these ten monolignol biosynthetic genes (Andersen et al. 2007; Andersen et al. 2008; Brenner et al. 2010; Zein et al. 2007). The number of haplotypes ranged from 2 to 12 for the ten monolignol biosynthetic genes (see Additional file 1). \textit{COMT}, \textit{CCoAOMT1}, and \textit{F5H} showed the largest phenotypic ranges among haplotype classes for PHT (121.2-171.4 cm) and DMC (23.2%-30.3%), DMY (4.1-8.4 t/ha), and DTS (72.2-84.1 days), respectively.

**Association analyses**

Association analyses revealed that six genes, coding for \textit{COMT}, \textit{CCoAOMT2}, \textit{4CL1}, \textit{4CL2}, \textit{F5H}, and \textit{PAL} proteins, were associated with at least one of the four biomass yield related traits. 10 associations were identified by GLM when including population structure in the analysis and controlling for multiple testing. Among those, seven were validated by MLM (Tables 3 and 4), which, in addition to population structure, corrects for finer scale relative kinship. However, none of these polymorphisms identified by MLM remained significant after controlling for multiple testing by FDR. At the \textit{PAL} locus a tight LD group containing 17 polymorphisms with \( r^2 = 1 \) was associated with days to silking (DTS). The 39 lines were classified into two groups by this LD group. The lines including AS1-8, 11-22, 24, and 29 were six days earlier than the remaining lines. This LD group explained 7% of the total DTS variation in our population. At the \textit{4CL2} locus, a tight LD group consisting of two SNPs (at position 192 and 217) in complete LD explained 14.3% of the phenotypic variation for PHT. The SNP at position 217 led to an amino acid change. The lines with the TG allele at these two positions were on average 17 cm higher than the lines with the CA allele. At the \textit{CCoAOMT2} locus, three polymorphisms (an indel
starting at position 75, two SNPs at position 144 and 406) were in a tight LD group with \( r^2 > 0.89 \), which explained 23.5% of the phenotypic variation for DMC. Another indel, which starts at position 663 in this locus explained 25.8%, 18.5%, and 10.5% of variation for PHT, DTS, and DMC, respectively. At the 4CL1 locus, two indels (starting at position 454 and 810) were both associated with DTS, and explained 20.2% and 6% of the phenotypic variation, respectively. These two indels both resulted in reading frame shifts, with one of those being a singleton. Lines with an Adenine insertion at position 454 silked on average three days earlier than the remaining lines. The COMT gene has been shown to strongly affect cell wall digestibility and plant growth. However, only one polymorphism was detected for associations with DTS. The indel in the 3’UTR was detected only by GLM and explained 10.3% of the phenotypic variation for DTS. Finally, one trait association was detected at the F5H locus, which was a missense substitution at position 65 and explained 22.4% of the phenotypic variation for DTS.

Pleiotropic polymorphisms affecting biomass yield and forage quality

In order to increase the chance of finding potential pleiotropic QTP affecting both biomass yield related and digestibility traits, associations of monolignol biosynthetic genes (Andersen et al. 2007; Andersen et al. 2008; Brenner et al. 2010) were determined without multiple test adjustment. In our study, two additional trait associations were detected only by MLM, one of which was an association between a synonymous SNP in the COMT gene and PHT, the other one was between a tight LD group in the F5H gene (two SNPs at position 5 and 6 in complete LD) and DMY. Despite of these relaxed statistical test conditions, only two polymorphisms in 10 monolignol biosynthetic genes were associated with both biomass yield-related and cell wall digestibility traits. The indel starting at position 810, resulting in a reading frame shift in the 4CL1 gene, was associated with IVDOM (Andersen et al. 2008) and DTS identified by both
GLM and MLM. It was also associated with NDF identified by GLM (Andersen et al. 2008). The tight LD group with two SNPs in complete LD in the F5H gene, resulting in a substitution from Pro to Arg, was associated with both DMY (by MLM) and NDF (by GLM) (Andersen et al. 2008). In addition, the tight LD group in the PAL gene showing association with DTS in our study was also associated with NDF (Andersen et al. 2007). However, the association between this LD group and NDF was only detected when population structure was not considered. In summary, no pleiotropic polymorphisms associated with DNDF and DMY or PHT were identified.

**Discussion and conclusion**

**Impact of the association analysis method on QTP identification**

Two statistical approaches (GLM and MLM) were employed as in previous association studies for better comparison across quality (Andersen et al. 2007; Andersen et al. 2008; Brenner et al. 2010) and yield related traits (this study). In those former studies, the same line panel, gene sequences, and marker data have been used. Inclusion of both population structure and relative kinship reduces the number of false positive associations compared to including population structure alone (Yu et al. 2006). In the present study, most of the associations identified for biomass yield and other agronomic traits by GLM were also identified by MLM, although none of the associations identified by MLM remained significant after controlling for multiple testing. Therefore, we cannot exclude the possibility that familiar relatedness resulted in false positives. However, this result might also suggest that inclusion of relative kinship information might in some cases mask genuine associations, comparable to likely false negatives of flowering time caused by inclusion of population structure for the Dwarf8 gene in European maize (Andersen et al. 2005). In this example, likely true effects of QTP on flowering time were confounded with
presence of one particular allele set in flint, the other in dent lines.

**Characterization of polymorphisms associated with biomass yield and agronomic traits**

We compared trait-associated (27) with not-associated polymorphisms (255) within the 10 monolignol biosynthetic genes regarding (i) the distribution among SNPs and indels, and (ii) polymorphisms among coding and non-coding sequences. Based on Chi-square tests, trait-associated polymorphisms for biomass yield-related traits were not preferentially due to either SNPs or indels, and not primarily located in either coding or non-coding gene regions.

Polymorphisms in conserved motifs with impact on protein function or abundance are more likely candidates for causative QTPs (Rebbeck et al. 2004). Within the *PAL* gene in our study, 1 out of 17 polymorphisms in the LD group associated with DTS was located within a possible bipartite RAV1 binding site (MOTIF Search; Kagaya et al. 1999). RAV1 has been suggested as a negative regulator of plant growth and development (Hu et al. 2004). In addition, five polymorphisms in the same LD group were located within Dof-like motifs (MOTIF Search). Dof transcription factors play a critical role in plant growth and development (Yanagisawa et al. 2004). Those six polymorphisms are more likely candidates for causative QTPs, whereas the remaining 11 significant associations within the same LD group are more likely due to linkage. To pinpoint causative polymorphisms, further dissection based on additional alleles at low LD is required. In the *CCoAOMT2* gene, a 40-60 bp indel at position 663 was just six base pairs upstream of a 3’ splicing donor site, spanning a potential “branching site” for splicing. Consequently, this indel might affect splicing and in this way interfere with the mRNA sequence and function of *CCoAOMT2*. Moreover, this indel also spanned part of a bipartite RAV1 binding site (MOTIF Search). Interestingly, this site was associated with three biomass yield-traits. Although LD decay was rapid in *CCoAOMT2*, the indel and two SNPs, which are at positions 75,
144, and 406, respectively, were tightly linked \((r^2>0.89)\). The indel resulted in two amino acid (Asparagine and Glycine) deletions compared with the \(CCoAOMT2\) allele of maize inbred line F2 (NCBI accession number AY279014.1). The other two SNPs were either synonymous or intron located SNPs. Thus, the indel is a more promising candidate QTP compared to the other two SNPs. Two DTS associated polymorphisms in \(4CL1\), which were both single nucleotide indels, led to frame shift mutations. One indel starting at position 810 introduced a premature stop (Andersen et al. 2008). The other indel changed the peptide sequence substantially, since it is located close to the transcription initiation site. In \(4CL2\), two polymorphisms in complete LD were associated with PHT. One of them changed the amino acid sequence and is, therefore, a more likely candidate QTP. In the \(F5H\) gene, Leucine to Proline and Proline to Arginine substitutions, were associated with DTS and DMY, respectively. Both are expected to change protein structure dramatically based on the Blosum-62 substitution matrix (Berg et al. 2007). Proline is very different from other amino acids due to its aliphatic side chain bonded to both nitrogen and \(\alpha\)-carbon atoms. In summary, some of the above mentioned trait associated polymorphisms or LD groups likely change protein sequence and expression dramatically, and are consequently the most likely QTPs affecting agronomic traits. However, future studies with maize populations with very low LD or alternative approaches are required for validation.

**Pleiotropic effects of monolignol biosynthetic genes**

Besides biosynthesis of lignin monomers, the monolignol biosynthetic pathway is involved in biosynthesis of salicylates, coumarins, hydroxycinnamic amides, pigments, UV light protectants, antioxidants, and flavonoids (Dixon and Palva 1995). Jone (1984) concluded that phenylpropanoid compounds are involved in controlling plant development, growth, xylogenesis, and flowering. For example, chalcone and naringenin, two intermediates in the phenylpropanoid
metabolism in plants, inhibit 4CL activity (Yu et al. 2001) and suppress the growth of at least 20 annual plant species including maize (Chen et al. 2004). Moreover, mutants in genes coding for C3H, C4H, PAL, CCoAOMT1, CCR1 and HCT show effects on plant growth (Pinçon et al. 2001; Elkind et al. 1990; Leplé et al. 2007; Mir Derikvand et al. 2008; Franke et al. 2002; Schilmiller et al. 2009; Do et al. 2007; Besseau et al. 2007). This is likely due to redirection of metabolic flux and accumulation of compounds, like naringenin, flavonoids, chalcone, which have the potential to perturb hormone homeostasis and ultimately affect plant growth.

In our study, polymorphisms affecting both biomass yield and cell wall digestibility were identified in six monolignol biosynthetic genes (encoding for COMT, CCoAOMT2, 4CL1, 4CL2, F5H, and PAL). These findings indicate that at least some of the monolignol biosynthetic genes act pleiotropically on both lignin content or composition and biomass yield or other agronomic traits. However, only two polymorphisms, the indel at position 810 in the 4CL1 gene and the LD group with SNPs resulting in substitution from Pro to Arg (Andersen et al. 2008) in the F5H gene, were found to be associated with both biomass yield and cell wall digestibility traits without controlling multiple testing. After controlling multiple testing, only the indel in the 4CL1 gene was associated with both DTS and IVDOM. Thus, the majority of QTPs identified in our study affected only one of the two groups of traits. Intragenic linkage of respective QTPs was more abundant than pleiotropic QTPs. According to our findings, most QTPs for both groups of traits are expected to segregate independently in germplasm with low LD.

Another important implication from our results is, that pleiotropy identified by comparison of wild-type with knock-out alleles, might in several cases turn out to be due to close linkage of intragenic QTPs with effects on different pathways and traits. An example is the well-studied Dwarf8 gene. This gene has been shown to affect plant height, when comparing mutant and wild
type alleles (Peng et al. 1999). However, association analyses with a range of wildtype alleles revealed candidate QTPs for flowering time, but not for plant height (Thornsberry et al. 2001). In Dwarf8, the DELLA domain is thought to affect plant height (Peng et al. 1999), while other polymorphisms affect flowering time. The DELLA domain was conserved in the 92 inbred lines used for an association analysis (Thornsberry et al. 2001). Similarly, previous bm3 mutant studies implied that the COMT coding gene acts pleiotropically on both forage quality and yield characters. However, after adjustment for multiple testing only one polymorphism was associated with DTS in our analysis, whereas eight different polymorphisms were associated with DNDF (Brenner et al. 2010). Since earlier reports on pleiotropy of bm mutations were based on isogenic lines, another explanation might be closely linked genes in introgressed donor segments affecting either quality or yield characters.

**Implications for plant breeding**

Although the genetic correlation between DNDF and DMY was significant (P=0.05), it was very low (r=-0.24) in these 39 inbred lines. Hence, it is very likely that the majority of genes affecting either biomass yield or cell wall digestibility traits are different. Our results support that monolignol biosynthetic genes affect both biomass yield related and cell wall digestibility traits. Intragenic linkage of QTPs was the more frequent cause for “pleiotropy” compared to pleiotropic polymorphisms. No QTP in our study was associated with PHT and DNDF, or DMY and DNDF. Considering these correlations and association data together, we conclude that breeders can employ optimal wildtype alleles for monolignol biosynthetic genes to improve cell-wall digestibility, without penalty on DMY.
List of abbreviations


Acknowledgments

We would like to thank KWS Saat AG (Einbeck) and the German ministry for education and science (BMBF) for financial support of the EUREKA project Cerequal. We thank the constructive advice from three anonymous reviewers. YC and EAB are now supported by the RF Baker Center for Plant Breeding at Iowa State University. YC was also supported by the interdepartmental genetics graduate program at Iowa State University.

Author’s contributions

YC performed the data analysis and prepared the manuscript. IZ carried out allele sequencing. EAB was involved in data analysis. JRA helped with statistical data analysis. MO provided the SSR data. ML has run the field trial. TL coordinated the project and together with YC prepared the manuscript. All authors read and approved the final manuscript.
References


Both caffeoyl Coenzyme A 3-O-methyltransferase 1 and caffeic acid O-methyltransferase 1
are involved in redundant functions for lignin, flavonoids and sinapoyl malate biosynthesis in
*Arabidopsis*. Planta 226:1117-1129

Increased Saccharification Yields from Aspen Biomass Upon Treatment with Enzymatically
Generated Peracetic Acid. Appl Biochem Biotechnol 160:1637-1652

Abnormal plant development and down-regulation of phenylpropanoid biosynthesis in
transgenic tobacco containing a heterologous phenylalanine ammonia-lyase gene. Proc Natl
Acad Sci USA 87:9057-9061

genotype data: linked loci and correlated allele frequencies. Genetics 164:1567-1587

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

Franke R, Humphreys JM, Hemm MR, Denault JW, Ruegger MO, Cusumano JC, Chapple C
(2002) The *Arabidopsis* REF8 gene encodes the 3-hydroxylase of phenylpropanoid
metabolism. Plant J 30:33-45

Test 12:1-77

Grabber JH, Ralph J, Lapierre C, Barrière Y (2004) Genetic and molecular basis of grass cell-


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


Kagaya Y, Ohmiya K, Hattori T (1999) RAV1, a novel DNA-binding protein, binds to bipartite
recognition sequence through two distinct DNA-binding domains uniquely found in higher plants. Nucleic Acids Res 27:470-478


MOTIF Search. Available at: http://motif.genome.jp/


327:467–479


Roussel V, Gibelin C, Fontaine AS, Barrière Y (2002) Genetic analysis in recombinant inbred lines of early dent forage. II. QTL mapping for cell wall constituents and cell wall digestibility from per se value and top cross experiments. Maydica 47: 9-20


Utz HF (2003) PLABSTAT: a computer program for statistical analysis of plant breeding experiments. Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany


Table 1. Phenotypic means, variance components, and heritabilities for four agronomic traits across four environments

<table>
<thead>
<tr>
<th>Line</th>
<th>PHT</th>
<th>DTS</th>
<th>DMC</th>
<th>DMY</th>
</tr>
</thead>
<tbody>
<tr>
<td>F_AS1</td>
<td>109.3</td>
<td>70.8</td>
<td>36.0</td>
<td>3.2</td>
</tr>
<tr>
<td>F_AS2</td>
<td>154.4</td>
<td>75.3</td>
<td>27.0</td>
<td>3.9</td>
</tr>
<tr>
<td>F_AS3</td>
<td>136.4</td>
<td>76.3</td>
<td>25.4</td>
<td>5.6</td>
</tr>
<tr>
<td>F_AS4</td>
<td>150.7</td>
<td>71.7</td>
<td>25.5</td>
<td>4.8</td>
</tr>
<tr>
<td>F_AS5</td>
<td>121.2</td>
<td>71.5</td>
<td>28.8</td>
<td>2.8</td>
</tr>
<tr>
<td>F_AS6</td>
<td>151.7</td>
<td>78.3</td>
<td>29.4</td>
<td>4.4</td>
</tr>
<tr>
<td>F_AS7</td>
<td>141.1</td>
<td>74.0</td>
<td>27.5</td>
<td>4.2</td>
</tr>
<tr>
<td>D_AS8</td>
<td>137.7</td>
<td>74.5</td>
<td>29.9</td>
<td>4.0</td>
</tr>
<tr>
<td>D_AS9</td>
<td>179.6</td>
<td>84.5</td>
<td>27.4</td>
<td>8.1</td>
</tr>
<tr>
<td>D_AS10</td>
<td>168.8</td>
<td>85.3</td>
<td>29.0</td>
<td>7.4</td>
</tr>
<tr>
<td>D_AS11</td>
<td>159.7</td>
<td>82.5</td>
<td>25.4</td>
<td>5.5</td>
</tr>
<tr>
<td>F_AS12</td>
<td>145.2</td>
<td>80.3</td>
<td>30.4</td>
<td>3.7</td>
</tr>
<tr>
<td>F_AS13</td>
<td>150.0</td>
<td>80.7</td>
<td>30.7</td>
<td>4.2</td>
</tr>
<tr>
<td>F_AS14</td>
<td>171.4</td>
<td>79.2</td>
<td>23.2</td>
<td>5.4</td>
</tr>
<tr>
<td>F_AS15</td>
<td>152.4</td>
<td>77.5</td>
<td>29.5</td>
<td>4.1</td>
</tr>
<tr>
<td>F_AS16</td>
<td>128.2</td>
<td>72.5</td>
<td>26.9</td>
<td>4.2</td>
</tr>
<tr>
<td>F_AS17</td>
<td>197.1</td>
<td>77.5</td>
<td>30.7</td>
<td>5.7</td>
</tr>
<tr>
<td>F_AS18</td>
<td>122.6</td>
<td>68.5</td>
<td>32.7</td>
<td>2.5</td>
</tr>
<tr>
<td>F_AS19</td>
<td>159.9</td>
<td>75.8</td>
<td>28.0</td>
<td>4.7</td>
</tr>
<tr>
<td>F_AS20</td>
<td>164.2</td>
<td>77.8</td>
<td>25.9</td>
<td>5.9</td>
</tr>
<tr>
<td>F_AS21</td>
<td>151.4</td>
<td>76.2</td>
<td>26.9</td>
<td>5.0</td>
</tr>
<tr>
<td>F_AS22</td>
<td>144.4</td>
<td>78.0</td>
<td>28.9</td>
<td>6.9</td>
</tr>
<tr>
<td>F_AS23</td>
<td>141.1</td>
<td>77.5</td>
<td>27.0</td>
<td>5.8</td>
</tr>
<tr>
<td>F_AS24</td>
<td>133.6</td>
<td>75.8</td>
<td>24.7</td>
<td>4.8</td>
</tr>
<tr>
<td>D_AS25</td>
<td>150.1</td>
<td>80.0</td>
<td>30.0</td>
<td>4.3</td>
</tr>
<tr>
<td>D_AS26</td>
<td>146.0</td>
<td>81.0</td>
<td>27.9</td>
<td>6.2</td>
</tr>
<tr>
<td>D_AS27</td>
<td>146.5</td>
<td>83.7</td>
<td>26.6</td>
<td>5.4</td>
</tr>
<tr>
<td>D_AS28</td>
<td>158.7</td>
<td>81.2</td>
<td>29.0</td>
<td>4.5</td>
</tr>
<tr>
<td>D_AS29</td>
<td>155.8</td>
<td>80.2</td>
<td>26.0</td>
<td>6.7</td>
</tr>
<tr>
<td>D_AS30</td>
<td>168.7</td>
<td>79.3</td>
<td>34.3</td>
<td>5.1</td>
</tr>
<tr>
<td>D_AS31</td>
<td>161.2</td>
<td>81.3</td>
<td>31.3</td>
<td>5.0</td>
</tr>
<tr>
<td>D_AS32</td>
<td>167.2</td>
<td>83.5</td>
<td>27.6</td>
<td>6.4</td>
</tr>
<tr>
<td>D_AS33</td>
<td>177.8</td>
<td>81.3</td>
<td>28.0</td>
<td>8.4</td>
</tr>
<tr>
<td>D_AS35</td>
<td>155.0</td>
<td>81.7</td>
<td>26.2</td>
<td>7.2</td>
</tr>
<tr>
<td>D_AS36</td>
<td>171.8</td>
<td>85.7</td>
<td>28.5</td>
<td>7.5</td>
</tr>
<tr>
<td>D_AS37</td>
<td>157.1</td>
<td>81.2</td>
<td>31.3</td>
<td>5.3</td>
</tr>
<tr>
<td>F_AS38</td>
<td>154.2</td>
<td>80.8</td>
<td>30.1</td>
<td>6.0</td>
</tr>
<tr>
<td>D_AS39</td>
<td>131.0</td>
<td>80.5</td>
<td>29.5</td>
<td>5.3</td>
</tr>
<tr>
<td>F_AS40</td>
<td>153.7</td>
<td>80.2</td>
<td>29.8</td>
<td>5.9</td>
</tr>
</tbody>
</table>

Phenotypic means
Flint  147.0±18.5  76.2±3.4  28.4±2.9  4.7±1.1
Dent  158.4±13.4  81.6±2.6  28.7±2.3  6.0±1.4
Overall  152.0±17.3  78.6±4.1  28.5±2.6  5.3±1.4
Variance components

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lines</td>
<td>8.3**</td>
<td>12.6**</td>
<td>7.00**</td>
<td>5.5**</td>
</tr>
<tr>
<td>Environments</td>
<td>44.0**</td>
<td>133.3**</td>
<td>30.3**</td>
<td>2.5</td>
</tr>
<tr>
<td>L×E</td>
<td>3.5**</td>
<td>6.2**</td>
<td>2.7**</td>
<td>4.0**</td>
</tr>
<tr>
<td>LSD5</td>
<td>16.8</td>
<td>3.3</td>
<td>2.7</td>
<td>1.64</td>
</tr>
</tbody>
</table>

Heritability% and CI interval

<table>
<thead>
<tr>
<th></th>
<th>88.0</th>
<th>92.0</th>
<th>85.7</th>
<th>81.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>90% CI on ( h^2 )</td>
<td>78.9-92.6</td>
<td>85.7-95.3</td>
<td>74.9-91.3</td>
<td>68.2-88.9</td>
</tr>
</tbody>
</table>

These lines are the same 40 lines (except D_AS34) used by Andersen et al. 2008. Flint- and dent lines are denoted by F_ and D_ prefixes, respectively. PHT: plant height (cm); DTS: days to silking; DMC: dry matter content of stover; DMY: dry matter yield of stover (tons per hectare) LSD5: least significant difference at 5% level between lines CI: confidence interval

*, ** significant at 5% and 1% level, respectively.
**Table 2:** Phenotypic and genotypic (*italics*) correlations between forage quality and yield-related traits

<table>
<thead>
<tr>
<th></th>
<th>DTS</th>
<th>DMC</th>
<th>DMY</th>
<th>WSC</th>
<th>IVDOM</th>
<th>NDF</th>
<th>DNDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHT</td>
<td>0.64**</td>
<td>-0.17</td>
<td>0.64**</td>
<td>0.24</td>
<td>-0.31</td>
<td>0.06</td>
<td>-0.45**</td>
</tr>
<tr>
<td></td>
<td>0.62**</td>
<td>-0.17</td>
<td>0.69**</td>
<td>0.24*</td>
<td>-0.35**</td>
<td>0.11</td>
<td>-0.47**</td>
</tr>
<tr>
<td>DTS</td>
<td>-0.13</td>
<td>0.72**</td>
<td>0.17</td>
<td>-0.17</td>
<td>0.08</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.09</td>
<td>0.78**</td>
<td>0.14</td>
<td>-0.22*</td>
<td>0.17</td>
<td>-0.28*</td>
<td></td>
</tr>
<tr>
<td>DMC</td>
<td></td>
<td></td>
<td>-0.35*</td>
<td>-0.51**</td>
<td>-0.39*</td>
<td>0.61**</td>
<td>-0.15</td>
</tr>
<tr>
<td></td>
<td>-0.37**</td>
<td>-0.56**</td>
<td>-0.40**</td>
<td>0.66**</td>
<td>-0.17*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DMY</td>
<td></td>
<td></td>
<td></td>
<td>0.44**</td>
<td>0.00</td>
<td>-0.23</td>
<td>-0.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.47**</td>
<td>-0.03</td>
<td>-0.23*</td>
<td>-0.24*</td>
</tr>
</tbody>
</table>

** Significant at P=0.01. * Significant at P=0.05.

PHT: plant height in cm
DTS: days from sowing to silking
DMC: % dry matter content of stover
DMY: Dry matter yield of stover in tons per hectare
WSC: Water soluble carbohydrates
IVDOM: *In vitro* digestibility of organic matter
NDF: Neutral detergent fiber
DNDF: Digestibility of neutral detergent fiber
Table 3. Associations between individual polymorphisms or LD groups and biomass yield and agronomic traits. Positions of trait associated polymorphisms are presented according to public reference sequences.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Position</th>
<th>Associated trait</th>
<th>R^2%</th>
<th>Identified by</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAL</td>
<td>947-1655^LD,1</td>
<td>DTS</td>
<td>7.4%</td>
<td>GLM^{*}, MLM^{**}</td>
</tr>
<tr>
<td>4CL2</td>
<td>192, 217^LD,2</td>
<td>PHT</td>
<td>14.3%</td>
<td>GLM^{*}, MLM^{**}</td>
</tr>
<tr>
<td>COMT</td>
<td>2358</td>
<td>DTS</td>
<td>10.3%</td>
<td>GLM^{*}</td>
</tr>
<tr>
<td>F5H</td>
<td>65</td>
<td>DTS</td>
<td>22.4%</td>
<td>GLM^{*}</td>
</tr>
<tr>
<td>4CL1</td>
<td>454</td>
<td>DTS</td>
<td>20.2%</td>
<td>GLM^{*}, MLM^{**}</td>
</tr>
<tr>
<td></td>
<td>810</td>
<td>DTS</td>
<td>6.0%</td>
<td>GLM^{*}, MLM^{**}</td>
</tr>
<tr>
<td>CCoAOMT2</td>
<td>75,144, 406^LD,3</td>
<td>DMC</td>
<td>23.5%</td>
<td>GLM^{*}, MLM^{**}</td>
</tr>
<tr>
<td></td>
<td>663</td>
<td>PHT</td>
<td>25.8%</td>
<td>GLM^{*}</td>
</tr>
<tr>
<td></td>
<td>663</td>
<td>DTS</td>
<td>18.5%</td>
<td>GLM^{*}, MLM^{**}</td>
</tr>
<tr>
<td></td>
<td>663</td>
<td>DMC</td>
<td>10.5%</td>
<td>GLM^{*}, MLM^{**}</td>
</tr>
</tbody>
</table>

^LD linkage group in a linkage block; ^1 the LD group contains 17 polymorphisms in complete LD in PAL and 947 and 1655 are the first and last polymorphisms; ^2 the LD group contains two polymorphisms which are in complete LD; ^3 the LD group contains three polymorphisms with r^2>0.89. Associations identified by GLM were after controlling multiple testing, while those identified by MLM were not. Polymorphism positions are denoted by the position in our alignment; R^2% is the proportion of phenotypic variance explained by the detected polymorphisms/LD group; * and ** mean P<0.05 and P<0.01, respectively. 4CL: 4-coumarate:CoA ligase, C3H: p-coumarate 3-hydroxylase, C4H: cinnamate 4-hydroxylase, CAD: cinnamyl alcohol dehydrogenase, CCoAOMT: caffeoyl-CoA O-methyltransferase, COMT: caffeic acid O-methyltransferase, F5H: ferulate 5-hydroxylase; PHT: plant height in cm; DTS: days from sowing to silking; DMC: % dry matter content of stover; DMY: dry matter yield of stover in tons per hectare.
### Table 4. Polymorphism character and position in reference sequence

<table>
<thead>
<tr>
<th>Gene</th>
<th>Position</th>
<th>Polymorphism</th>
<th>Position in reference sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAL</td>
<td>947</td>
<td>Intron SNP</td>
<td>TTTAGAGACGATCGCAATCCA(C/T)</td>
</tr>
<tr>
<td></td>
<td>1652,1655</td>
<td>Intron SNP</td>
<td>CCTGTGCGCTGGCTCCC(T/C)TT(C/T)</td>
</tr>
<tr>
<td>4CL2</td>
<td>192</td>
<td>synonymous SNP</td>
<td>TACTGCTTCGGGAAGATGGG(T/C)</td>
</tr>
<tr>
<td></td>
<td>217</td>
<td>Val to Ile SNP</td>
<td>TGGCGAGCGGCGTGCTGCCG(T/C)</td>
</tr>
<tr>
<td>COMT</td>
<td>2358</td>
<td>3'UTR indel</td>
<td>CGCCGTCTGCTTCTTT(T/C)</td>
</tr>
<tr>
<td>F5H</td>
<td>65</td>
<td>Leu to Pro SNP</td>
<td>ACGCGCGACAATATCAAGGC(T/C)</td>
</tr>
<tr>
<td>4CL1</td>
<td>454</td>
<td>Exon indel</td>
<td>TGTGTGGGCGGATGCGCGGGG(+/-)</td>
</tr>
<tr>
<td></td>
<td>810</td>
<td>Exon indel</td>
<td>GCTGTGGGAGGATGCGCGCGGGG(+/-)</td>
</tr>
<tr>
<td>CCoAOMT2</td>
<td>75</td>
<td>Exon Indel</td>
<td>CAGGCGAGCGGCAAACGGCA(+/-)</td>
</tr>
<tr>
<td></td>
<td>144</td>
<td>synonymous SNP</td>
<td>CTGCTAAGAGCGACGACCT(C/G)</td>
</tr>
<tr>
<td></td>
<td>406</td>
<td>Intron SNP</td>
<td>ACCGAGATCTGAGAACCACCA(A/G)</td>
</tr>
<tr>
<td></td>
<td>663</td>
<td>Intron indel</td>
<td>CTTAGGATCTTACCACCCC(+/-)</td>
</tr>
</tbody>
</table>

Polymorphism character and position in reference sequence. (N/N): SNP substitution; (+/-): indel; s: singleton; M73235, AY323238, AY279014, AX204867, AX204868, and AX204869 were used as reference sequences for COMT, CCoAOMT1, CCoAOMT2, 4CL1, 4CL2, and F5H respectively. The reference sequences for PAL are the conserved sequences before each polymorphism in our alignment.
**Additional file 1.** Haplotype number, average, minimum, and maximum of biomass yield-related trait values for each monolignol biosynthetic gene. Haplotype numbers and inbred lines included in each haplotype group: see (Andersen et al. 2007) for PAL, (Andersen et al. 2008) for 4CL2, 4CL1, CAD, C3H, C4H, and F5H, and (Brenner et al. 2010) for CCoAOMT1, CCoAOMT2, COMT

<table>
<thead>
<tr>
<th>Gene</th>
<th>Haplotype Num</th>
<th>Haplotype Min</th>
<th>Haplotype Max</th>
<th>Mean</th>
<th>Haplotype Min</th>
<th>Haplotype Max</th>
<th>Mean</th>
<th>Haplotype Min</th>
<th>Haplotype Max</th>
<th>Mean</th>
<th>Haplotype Min</th>
<th>Haplotype Max</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMT</td>
<td>12</td>
<td>121.2</td>
<td>171.4</td>
<td>146.9</td>
<td>71.5</td>
<td>82.4</td>
<td>78.1</td>
<td>23.2</td>
<td>30.3</td>
<td>28.6</td>
<td>2.7</td>
<td>6.0</td>
<td>4.9</td>
</tr>
<tr>
<td>CCoAOMT1</td>
<td>9</td>
<td>131.0</td>
<td>177.8</td>
<td>154.4</td>
<td>75.7</td>
<td>85.8</td>
<td>79.7</td>
<td>26.0</td>
<td>29.9</td>
<td>28.4</td>
<td>4.1</td>
<td>8.4</td>
<td>5.9</td>
</tr>
<tr>
<td>CCoAOMT2</td>
<td>6</td>
<td>131.0</td>
<td>161.2</td>
<td>147.9</td>
<td>74.0</td>
<td>81.3</td>
<td>79.1</td>
<td>27.6</td>
<td>31.2</td>
<td>29.8</td>
<td>4.2</td>
<td>5.9</td>
<td>5.0</td>
</tr>
<tr>
<td>PAL</td>
<td>8</td>
<td>146.5</td>
<td>179.6</td>
<td>156.5</td>
<td>75.7</td>
<td>84.6</td>
<td>80.5</td>
<td>26.5</td>
<td>31.2</td>
<td>28.6</td>
<td>4.2</td>
<td>8.1</td>
<td>5.6</td>
</tr>
<tr>
<td>C4H</td>
<td>4</td>
<td>143.9</td>
<td>164.1</td>
<td>152.1</td>
<td>75.5</td>
<td>80.0</td>
<td>78.3</td>
<td>28.2</td>
<td>28.8</td>
<td>28.6</td>
<td>4.5</td>
<td>5.6</td>
<td>5.1</td>
</tr>
<tr>
<td>4CL1</td>
<td>4</td>
<td>152.5</td>
<td>167.2</td>
<td>158.4</td>
<td>76.2</td>
<td>83.6</td>
<td>4.5</td>
<td>27.6</td>
<td>31.2</td>
<td>3.8</td>
<td>4.9</td>
<td>6.4</td>
<td>5.4</td>
</tr>
<tr>
<td>4CL2</td>
<td>7</td>
<td>142.0</td>
<td>162.4</td>
<td>150.2</td>
<td>75.0</td>
<td>80.6</td>
<td>78.4</td>
<td>26.9</td>
<td>30.6</td>
<td>28.7</td>
<td>3.9</td>
<td>6.0</td>
<td>4.9</td>
</tr>
<tr>
<td>C3H</td>
<td>7</td>
<td>150.7</td>
<td>171.4</td>
<td>159.9</td>
<td>71.7</td>
<td>81.1</td>
<td>79.0</td>
<td>23.2</td>
<td>29.5</td>
<td>26.8</td>
<td>4.5</td>
<td>5.4</td>
<td>5.1</td>
</tr>
<tr>
<td>CAD</td>
<td>2</td>
<td>152.2</td>
<td>155.3</td>
<td>153.7</td>
<td>75.3</td>
<td>79.1</td>
<td>77.2</td>
<td>28.3</td>
<td>28.5</td>
<td>28.4</td>
<td>4.9</td>
<td>5.3</td>
<td>5.1</td>
</tr>
<tr>
<td>F5H</td>
<td>5</td>
<td>133.6</td>
<td>169.4</td>
<td>150.7</td>
<td>72.2</td>
<td>84.1</td>
<td>77.7</td>
<td>24.7</td>
<td>29.2</td>
<td>27.5</td>
<td>3.8</td>
<td>6.2</td>
<td>5.1</td>
</tr>
</tbody>
</table>
CHAPTER 7. EXTENSIVE GENETIC DIVERSITY AND LOW LINKAGE DISEQUILIBRIUM WITHIN THE COMT LOCUS IN GERMPLASM ENHANCEMENT OF MAIZE POPULATIONS

Yongsheng Chen\textsuperscript{1,2,4}, Michel Blanco\textsuperscript{3}, Qing Ji\textsuperscript{1}, Ursula Karoline Frei\textsuperscript{1}, Thomas Lübberstedt\textsuperscript{1,5}

\textsuperscript{1}: Department of Agronomy, Iowa State University, Ames, USA, 50011
\textsuperscript{2}: Interdepartmental Genetics Program, Iowa State University, Ames, USA, 50011
\textsuperscript{3}: USDA-ARS, Plant Introduction Station, Ames, IA, USA, 50011
\textsuperscript{4}: Primary researcher and author
\textsuperscript{5}: Author for correspondence

A paper to be submitted to PLoS Genetics

Abstract

The Caffeic acid 3-\textit{O}-methytransferase (COMT) gene is a prime candidate for cell wall digestibility improvement based on the characterization of brown midrib mutants. We compared the genetic diversity and linkage disequilibrium at this locus (COMT locus) between landraces sampled within the Germplasm Enhancement of Maize (GEM) project and 70 elite lines. In total, we investigated 55 exotic alleles from GEM at the COMT locus, and discovered more than 400 polymorphisms in a 2.2 kb region. The pairwise nucleotide diversity ($\pi$) for the exotic alleles was 0.0172, and much higher than the reported pairwise nucleotide diversity in elite inbred lines ($\pi=0.0047$ to 0.0067). The pairwise nucleotide diversity ($\pi$) for exotic alleles was 0.01705 in the overlapping region (~1949bp) between exotic populations and 70 elite lines. The average number of nucleotide differences between any two randomly selected alleles was 27.69 for exotic COMT alleles, much higher than 18.02 in elite alleles. The non-synonymous to synonymous SNPs was 3:1 in exotic COMT alleles, significantly higher than 1:1 in elite alleles. The LD decay in exotic populations was at least 4 times faster than in elite lines with $r^2>0.1$ persisting only about 100bp. In conclusion, the alleles sampled in the GEM line collection offer a valuable genetic resource to
broaden genetic variation for the COMT gene, and likely any genes, in elite background. Moreover, the low LD makes these materials suitable for high resolution association analyses.

**Introduction**

Maize has been widely used for forage. Each year, ~4.6 and ~4 Mha of maize are grown for forage purposes in Europe and USA, respectively (Barrière et al. 2004a; Lauer et al. 2001). In the past decades, whole plant yield increased substantially in Europe (Barrière et al. 2004a). However, the digestibility of cell walls steadily decreased (Barrière et al. 2005). Lauer et al. (2001) reported no increase of cell wall digestibility in the U.S. Cell wall digestibility is a major trait affecting forage quality (Andrieu et al. 1993; Barrière et al. 2003, 2004b).

Two reasons are blamed for the reduction of cell wall digestibility of maize during the increase of whole plant yield (Barrière et al. 2005). The first is loss of favorable alleles for cell wall digestibility when selecting favorable alleles for standability. The second is genetic drift during breeding for high grain yield, which also resulted in the loss of beneficial alleles for cell wall digestibility. Barrière et al. (2004a, 2005) suggested to exploit genetic resources from old landraces for increasing cell wall digestibility.

Brown midrib mutants are known to affect cell wall digestibility. In maize, there are four well $\texttt{bm}$ mutants ($\texttt{bm1-bm4}$), which are characterized by brown midrib coloration in the vascular leaf tissues and reduced lignin content or altered lignin composition (Barrière et al. 2004b). More recently, three new brown midrib mutants named $\texttt{bm5}$, $\texttt{bm6}$ (Ali et al. 2010), and $\texttt{bm7}$ (Lütherstedt et al., unpublished results) were identified. It was observed that $\texttt{bm5}$ and $\texttt{bm6}$ both have increased digestibility of neutral detergent fiber (DNDF) (Chen et al., unpublished data). Reduced lignin content and altered lignin composition lead to higher cell wall digestibility of $\texttt{bm}$ mutants (Barrière et al. 2004b), and thus better forage quality. As a consequence, $\texttt{bm1-7}$ mutants
are potentially valuable genetic resources for forage breeding. However, *bm1*-4 mutants are often associated with undesirable agronomic traits, such as lodging, reduced growth vigor, grain, and stover yield (Pederson et al. 2005). Among the four *bm* mutants, *bm3* has the strongest effect on both lignin, cell wall digestibility, and agronomic traits (Barrière and Argillier 1993; Barrière et al. 2004b). The *bm3* phenotype is due to structural mutations in the COMT gene coding caffeic acid O-methyl transferase (Vignols et al. 1995; Morrow et al. 1997), which is involved in monolignol biosynthesis. Previous studies based on isogenic lines of *bm3* suggested genic pleiotropy of the *bm3* locus for cell wall digestibility and biomass yield related traits (Lee and Brewbaker 1984; Inoue and Kasuga 1989; Gentinetta et al. 1990). A candidate gene-based association study (Chen et al. 2010) indicated that its genic pleiotropy might be due to intragenic linkage of QTPs. Therefore, an optimal haplotype of the COMT gene which combines favorable quantitative trait polymorphisms (QTPs) for both cell wall digestibility and biomass yield might exist. Three studies have been performed to study the genetic diversity within this locus (Fontaine and Barrière 2003; Guillet-Claude et al. 2004; Zein et al. 2007). However, their studies were all conducted with elite lines. To our knowledge, no report on characterization of the genetic diversity of the COMT locus in exotic landraces is available.

This study aims to exploit the genetic diversity of COMT alleles sampled in a wide range of exotic races from the Germplasm Enhancement of Maize (GEM) project (Pollak 2003). The objectives were to (1) study the genetic diversity and linkage disequilibrium (LD) within the COMT locus in 55 exotic alleles sampled from 41 races, (2) study the genetic diversity and LD of COMT in 70 elite lines combined from Zein and Guillet-Claude’s reports (Zein et al. 2007; Guillet-Claude et al. 2004), (3) compare the genetic diversity and LD between exotic populations and elite lines, (4) study the phylogenetic relationship between the 55 sampled exotic alleles and
70 elite lines, and discuss the impact of our results on forage breeding, and the usefulness of GEM lines for breeding novel elite lines.

**Materials and methods**

**Plant materials**

In order to broaden the germplasm base of commercial hybrids in USA, USDA-ARS, public land grand universities, and private seed companies initiated the Germplasm Enhancement of Maize (GEM) project by collecting exotic maize races across the world, and introducing them into two expired Plant Variety Protection (PVP) elite lines. The aim is to provide a random collection of novel exotic alleles for pre-breeding purposes. This project is sampling the majority of ~250 known maize races. In each race, several plants were selected and backcrossed with the two expired PVP lines (PHB47 and PHZ51), which resulted in several backcross (BC) families per race.

Our goal was to sample exotic COMT alleles from a range of BC families of the GEM project and to sequence them to study their genetic diversity and LD. We compared those with 70 elite lines, which were studied by Guillet-Claude et al. (2004) and Zein et al. (2007). In these two studies, in total 76 entries were studied, including six lines present in both studies.

As we sampled COMT alleles from BC families within the GEM program, half of the BC₁ plants were expected to be homozygous at the COMT locus with two copies of recurrent parent COMT alleles, and the other half to be heterozygous with one COMT allele from an exotics GEM donor. Therefore, the first step was to identify BC₁ plants with COMT alleles different from the recurrent parents (PHB47 and PHZ51). In 2008, 478 BC₁ families from 131 races in BC₁ generation were available. Two plants were randomly sampled from each BC₁ family to screen for exotic COMT alleles. Using insertion and deletion (indel) and simple sequence repeat
(SSR) markers (see below), we identified 252 plants that were heterozygous at the COMT locus, which were backcrossed to the recurrent parents to produce BC\textsubscript{2} families. The 252 BC\textsubscript{2} families along with another 374 BC\textsubscript{1} families within GEM project were grown in 2009 and were screened for exotic COMT alleles. One plant was randomly selected from each BC\textsubscript{1} and BC\textsubscript{2} family for screening. If no heterozygous plant was found in a family, a 2\textsuperscript{nd} plant was picked for genotyping in the \textit{bm3} region. For BC\textsubscript{2} families, random sampling was done for the third time if the first and second randomly selected samples did not find heterozygous plants. In total, 285 heterozygous plants (169 from BC\textsubscript{2} families and 116 from BC\textsubscript{1} families) with one copy of the COMT allele from GEM were identified.

**Molecular markers for screening COMT alleles from GEM germplasm**

Previous studies demonstrated high levels of genetic variation in the intron region of the COMT gene (Guillet-Claude et al. 2004; Zein et al. 2007). Especially, there was substantial indel variation within the intron region, which could be used to derive indel markers for screening different COMT alleles (Zein et al. 2007). Based a large indel at the 3’ end of the intron, we designed an indel marker for screening exotic COMT alleles different from recurrent parent alleles. The left (forward) primer was 5’-GCCTGGCTGATGCTGATAGA-3’. The right (reverse) primer was 5’-CACCGAGTGGTTCTTCATGC-3’. To increase the chance to find COMT alleles which differ from the recurrent parent COMT alleles, closely linked SSR markers were used in addition. The COMT gene is located in the overlap region of bacterial artificial clones (BAC) AC203909 and AC196475 (www.maizedb.org). AC203909 overlaps with AC205176. More than 20 SSR marker pairs were designed based on B73 sequence from the above three BACs (AC203909, AC196475, and AC205176). The SSR marker AC203909S02 (left primer: 5’-AATTCCATCATTCGCGTACC-3’; right primer: 5’-CGCTTGACAACAGACCACAC-3’)
produced clear bands and a high frequency of heterozygotes. This SSR and the above mentioned indel marker that was located within the intron region of COMT gene were used for screening heterozygotes at the COMT locus in BC families.

**DNA extraction, PCR amplification, and DNA sequencing**

Leaves from the BC families and the two inbred lines PHB47 and PHZ51 used as recurrent parents were harvested for DNA isolation after six weeks of sowing. Harvested leaves were freeze dried for 24 hours. After drying, two steel beads were added to each sample. Samples were ground by a Geno/Grinder 2000 (BT&C, Inc., NJ, USA) at 700 strokes/min for 3 minutes. DNA extraction was conducted according to the CTAB protocol used at the Plant Genomics Center of Iowa State University (http://schnablelab.plantgenomics.iastate.edu/docs/resources/protocols/pdf/96wellformat.2010.06.23.pdf).

For screening of COMT alleles from GEM, DNA amplification was performed in a 15 μl reaction mixture containing 1.5 μl 10× buffer, 1 μL genomic DNA (~100 ng), 0.3 μL of each left and right primer (10 μM), 0.3 μl dNTPs (10 mM), 1.5 μl Mgcl₂(25 mM), 2 μl Taq polymerase (~1 unit), and 8.1 μl purified water. PCR amplification included an initial denaturation step at 94 °C for 2 min, and was followed by 35 amplification cycles: 94 °C for 30 s, 57°C for 30 s, 72°C for 40 s, and a final extension step at 72°C for 10 min. The products were separated in 1% agarose gels, visualized by ethidium bromide staining and photographed by using the FOTO/UV 300 system (Fotodyne InC., WI, USA).

For cloning of COMT alleles, DNA amplification was performed in a 10μl reaction mixture containing 2 μl 5xbuffer, 0.5μl DNA (~50ng), 0.2 μl of dNTPs (10mM), 0.2 μl of each left and right primer (10μM), 0.2μl of 50 GC Advantage2 (BD Biosciences Clontech, Palo Alto, Calif.,
USA), 1µl of GC melt, and 5.7 µl purified water. The left and right primers (Figure 1) are Promoter F: TCCACGCGAGCTGCCACCGTCGCTATCGC-3’ (Guillet-Claude et al. 2004) and Ex2_R: 5’-CGCACATGGCAGAGACAA-3’ (Zein et al. 2007). The PCR program included an initial denaturation step at 94°C for 2 min, eight amplification cycles: 94°C for 30s, 68°C for 45s (-1°C per cycle), 72°C for 4 min, followed by 27 cycles of: 94 °C for 30 s, 60°C for 30 s, 72°C for 4 min, and a final extension step at 72°C for 10 min. PCR products were cloned into pGEM-T vector (Promega, Madison, USA) and transformed into DHα52 competent cells by heat shock for 90 seconds. Transformed DHα52 competent cells were cultivated on solid Lysogeny broth (LB) medium over night. Out of the 285 heterozygotes, 190 were randomly selected for cloning. In addition, the PHB47 and PHZ51 were also sequenced for COMT gene for later sequence comparison to pick alleles from BC families different from those.

When DNA was extracted from BC plants, both recurrent parental and exotic COMT alleles were amplified by PCR and transformed, at an expected ratio of 1:1. One clone of each transformation event was randomly selected and plasmid DNA was extracted by QuickLyseMiniprep Kit (QIAGEN, USA) according to its protocol. Plasmid DNA was sequenced once by the In_R primer 5’GTTGAACGGGATGCCGCCGTC-3’ (Guillet-Claude et al. 2004) (Figure 1) using the Applied Biosystems 3730x DNA Analyzer (AB Applied Biosystems, USA). Intron sequence was compared with its recurrent COMT allele first. A threshold was applied to identify sequences that were most likely different from the recurrent parental allele and not due to sequencing error. We selected alleles that either had one or more indel(s) or at least 5 single nucleotide polymorphism (SNP) (one base pair indels were considered as a SNP) differences compared to the respective recurrent allele sequence (either from PHB47 or PHZ51). If the randomly selected clone happened to be the recurrent parental
allele, we selected another clone from the same transformation for the second or third time for intron sequencing. The total sequenced number of fragments was larger than 190, but some failed or were of poor quality. Finally, we obtained intron sequences for 156 plasmids, which had good sequencing quality. Among these 156 intron sequences, 61 had indels compared with its respective recurrent allele, 15 had at least five SNPs. This number (76+15) was consistent with a 1:1 ratio (recurrent: exotic alleles). In addition, 28 and 50 sequenced alleles had four or fewer SNPs compared with recurrent alleles.

For whole gene sequencing, we selected the samples, which had indels or at least five SNPs within the intron compared with its respective recurrent alleles. We also selected some samples which had only four SNPs. In total, we selected 86 samples for whole gene sequencing. Three primers were used whole gene sequencing, including Ex1_F 5’GGTGAGCCGTCGGCCCCAATAAACCTT-3’, In_F 5’ATGAACCAGGACAAGGGTCCTCATG-3’, and Ex2_R (Guillet-Claude et al. 2004) (Figure 1). Sequencing was repeated 2-3 times for each fragment. Sequencing PCR was performed in 10 µl reaction mixture including 1.75 µl 5 x sequencing buffer, 1 µl BigDye (Applied Biosystems), 0.256 µl of primer (10 µM), 1 µl of plasmid DNA (~50 ng), and 4.97 µl purified water. The PCR cycle consisted of an initial denaturation step at 96 ºC for 2 minutes, and 25 cycles of amplification: 96 ºC for 30 s, 50 ºC for 1 min, and 60 ºC for 4 min. Ultimately, whole gene sequences of 57 exotic alleles were obtained. Six to 10 fragment sequences were obtained for each of these 57 alleles, including 2 or 3 sequences for each of Ex1_F primer, In_F primer, and Ex_R2 primer, and one In_R (Figure 1) primer.

Analysis of sequence data

Sequence assembly and sequence error estimation for exotic COMT alleles was done by using
CodonCode Aligner (V. 3.7.1, Codon Code Corporation, USA). Usually, the sequence quality both at the beginning and end of each sequence is low. Therefore, sequences at both ends were clipped by the clip function implemented in CodonCode Aligner, before whole gene sequence assemblies were performed. Fifty-five of the 57 sequenced exotic alleles had low sequencing errors at a rate of less than one base pair of miscalling across the assembled contig. Moreover, all these 55 alleles were different from corresponding recurrent alleles. To study the genetic diversity at the COMT locus in exotic populations, all assembled sequences of the 55 alleles were used for alignment, which resulted in missing data for some alleles at 5’ or 3’ ends. All alignments were performed in CLUSTALW (Thompson et al. 1994), and resulting alignment files were further analyzed by using DnaSp Version 5.10.01 software (Librado and Rozas 2009).

For comparison between alleles from elite lines and exotic populations, sequence at both ends was removed for comparison of the overlapping region across all COMT alleles. In our study and the report of Zein et al. (2007), the sequenced region covered ~150 bp of the 5’-UTR and ~250 bp 3’ flanking regions besides the two coding regions and the intron. In the study of Guillet-Claude et al. (2004), COMT sequences covered ~1000 bp of the promoter region and 5’-UTR, coding region and intron, but some sequences did not cover the 3’-UTR region, and stopped 18 bp ahead of the stop codon (TGA). In order to precisely compare alleles from exotic and elite lines, the overlapping region between the three studies were extracted. The respective overlapping sequence started 98 bp ahead of start codon (ATG) and stopped 18 bp ahead of the stop codon (TGA), spanning 1851 bp of the B73 COMT allele (GenBank accession number AY323295.1). In conclusion, the alignment of 55 exotic COMT alleles spanned 2141 bp (Table 4), and the alignment of the 70 COMT alleles from Zein et al. (2007) and Guillet-Claude et al. (2004) spanned 2136 bp (Table 4). These two alignments were used for comparison between exotic
populations and elite lines for the 5'-UTR, coding region, and intron. For convenience, the overlapping region between both sets (from 98 bp ahead of the start codon (ATG) to 18 bp ahead of the stop codon) was called overlap region hereafter.

Two estimates of diversity, $\pi$ and $\theta$ were calculated. $\pi$ is the average number of nucleotide differences per site (Nei 1987), calculated by the total number of pairwise nucleotide differences divided by the total number of comparisons. $\theta$ is a neutral mutation parameter with an estimated value $4N_e\mu$, where $N_e$ and $\mu$ are the effective population size and mutation rate per generation (Kimura 1969), respectively. Under the neutral mutation hypothesis that neutral mutations are randomly maintained in the history of evolution, $\theta$ can be unbiased estimated by different estimates of nucleotide diversity, for example, $\pi$, $s/(1+i)$ (where $i=1$ to $n-1$), $\eta/(1+i)$, $(n-1)\eta/n$ and $k$. The variable $s$ is the number of segregating sites, $\eta$ is the total number of mutations, $\eta_s$ is the total number of singletons (polymorphic sites appearing only once in the sample), $k$ is average number of nucleotide differences between sequence pairs, and $n$ is number of nucleotide sequences. Comparison of $\theta$ based on $\pi$ and $s$ lead to Tajima’s D test (Tajima 1987). Comparison of $\theta$ based on $\eta$ and $\eta_s$ leads to Fu and Li’s $D^*$ test, and comparison of $\theta$ based on $\eta_s$ and $k$ results in Fu and Li’s $F^*$ (Fu and Li 1993).

Linkage disequilibrium (LD) was estimated by squared allele-frequency correlations ($r^2$) (Weir 1996) considering only informative parsimony SNP sites (which contain at least two types of nucleotides and each presents at least twice). Significance of LD between sites was tested by Fisher’s exact test implemented in DnaSP. LD decay was visualized by plotting $r^2$ against physical distance as well as depicted by the LD function implemented in Tassel software version 3.0 (Bradbury et al. 2007). For phylogenetic analysis of the alleles from elite lines and exotic populations, a neighbor joining tree was constructed by MGPA software version 5.0 (Tamura et
al. 2011) using default settings. 30 haplotypes discovered from the 70 distinct lines and 55 exotic haplotypes were treated as taxa in tree construction for reducing the tree space. In tree construction, an *Oryza sativa* O-methyltransferase gene was used as outgroup.

The B73 COMT sequence (GenBank accession number AY323295.1) which has annotation of coding regions was used to infer the coding regions of 55 exotic COMT alleles as well the 30 elite haplotype COMT alleles. The predicted coding region DNA sequences were translated into amino acids by the ‘translate’ function in the Sequence Manipulation Suite (http://www.bioinformatics.org/sms2/). The amino acids were aligned in CLUSTALW (Thompson et al. 1994) to reveal the (non-) conservative status of amino acid substitutions. The neighbor joining trees of amino acids were drawn in CLUSTALW (Thompson et al. 1994).

**Results**

**Summary of sequencing exotic COMT alleles**

Fifty-Seven exotic alleles were sequenced and assembled into contigs which spanned from the 5’-UTR (~150bp) to the 3’flanking region (~250bp) of the COMT gene. Two contigs had sequence errors exceeding one base miscall within the COMT gene, and were thus excluded from further genetic diversity analyses. Assembly of the allele from sample no. 3659 resulted in two short contigs in CodonCode Aligner (V. 3.7.1, Codon Code Corporation, USA). The first contig of the allele from sample no. 3659 was 1463 bp in length with an estimated error rate of 4.89×10^{-4} and it spanned 5’-UTR, the first exon, and intron. The second contig had 848 bp with an estimated error rate of 8.09×10^{-4} spanning the second exon. The short overlapping of 22 bp did not allow CondonCode Aligner to assemble them. The contigs were thus manually assembled. All other alleles were assembled into a single contig. In total, 55 alleles from 41 races were included for genetic diversity analyses (Table 1). The sequencing coverage for these 55 alleles
varied from 2.7 to 4.2 times with 5995 to 9202 bp being sequenced for each allele. The contig lengths varied between 2135 to 2342 bp with a sequencing error rate ranging from $4.02 \times 10^{-5}$ to $3.63 \times 10^{-4}$, which translates to 0.09 to 0.82 miscalled base pairs per contig (Table 1). On average, the contig length was 2217 bp with a 3.4-fold sequence coverage and an error rate of $1.16 \times 10^{-4}$.

The 41 races originated from 10 countries in Central or South America (Table 1), with eight races from Peru, seven from Bolivia, seven from Colombia, five from Venezuela, five from Ecuador, four from Mexico, and two from Paraguay. In addition, Guatemala, Cuba, and Chile contributed one race each.

**Sequence variation for the COMT locus in exotic populations**

**Polymorphisms and haplotypes**

The full alignment of 55 exotic alleles spanned 2466 bp including 477 sites with missing data or alignment gaps, which resulted in 1989 sites of net alignment without missing data and gaps. The missing data (for some alleles) were at the two ends of the alignment, spanning 44 and 46 bp of alignment at the 5’ and 3’ end, respectively. Based on the alignment, a total of 438 SNPs were discovered, including 95 parsimony informative SNPs and 343 singletons (Table 2), which lead to one parsimony informative SNP per 20.9 bp (1989/95) or totally one SNP per 4.5 bp (1989/438), on average. These 438 SNPs formed 55 distinct haplotypes, which means each allele is a unique haplotype. Half of the SNPs (219 of 438) were located in the coding regions, the other in non-coding regions. In coding regions, 30 (13.7%) and 189 SNPs (86.3%) were parsimony informative SNPs and singleton SNPs, respectively (Table 2). While 65 (29.7%) and 154 (70.3%) SNPs were informative SNPs and singletons in non-coding regions. Non-coding regions had a significantly lower percentages of singletons and higher percentages of parsimony SNPs than coding regions (p=0.00005). Of the 219 SNPs located in coding regions, 57 (26.03%)
were synonymous SNPs, while 162 (73.97%) were non-synonymous SNPs leading to amino acid replacement.

Excluding the missing data (44 bp at the 5’ end and 46 bp at the 3’ end) at both sides of the alignment, 72 indels were discovered in the alignment (Table 3), which resulted in one indel per 33 bp of alignment on average (2376/72, where 2376 was the number of sites excluding missing sites). However, the indels were predominantly distributed in non-coding regions (p=0.05), where 63 indels were identified in 1281 bp sites of alignment, while only nine indels were identified in 1096 bp sites of alignment for the coding regions. Among those nine indels located in coding regions, eight were one base pair indels and one was a 3bp indel. The eight one base indels resulted in frame shift mutations and/or introduction of stop codons (see details in discussion). The allele from sample no. 3619 had a singleton indel in each of the two exons. Three alleles (from sample no. 3656, no. 3834, and no. 3659) had such indels in the second exon, while the other three alleles (from sample no. 3730, no. 3133, and no. 3720) had it in the first exon. The allele from sample no. 3725 had a 3 bp deletion in the first exon, which leads to loss of an “Ala” amino acid rather than reading frame shift.

**Nucleotide diversity**

Pair-wise nucleotide diversity was highest in the intron region (\(\pi=0.027\)), followed by the 3’flanking (\(\pi= 0.020\)) and coding regions (\(\pi=0.011\)). The 5’-UTR region had the lowest diversity (\(\pi=0.0073\)). The nucleotide diversity (\(\pi\)) of the entire gene was 0.017 (Table 2). The k value, which is the average number of nucleotide differences between any two randomly selected alleles (Tajima 1983), was 34.22 for the entire gene. To get an overview of the nucleotide diversity in different region of this gene, \(\pi\) was calculated in sliding windows of 100 bp using a step size of 10 bp (Figure 2), with alignment gaps not being considered. The highest diversity
was found at the 5’ and 3’ ends of the intron, with π values up to ~0.037 and ~ 0.034.

Decay of linkage disequilibrium and recombination

LD between all (87) parsimony informative SNPs except (8) sites segregating for three nucleotides were estimated. In total, 3741 pair-wise comparisons were obtained and their significance was checked by Fisher’s exact test. 150 pairs showed significance after Bonferroni multiple comparison adjustment. Seventy (46.7%) of the 150 site pairs were less than 100 bp apart, which indicates dramatic rapid LD decay. The plot of r² against physical distance between parsimony informative SNPs indicated that r² ≥ 0.1 could persist only about 100 bp (Figure 3a). The very low level of LD was supported by the 28 recombination events in the 2466 bp alignment. There were four and six recombination events being detected in the first and second exon, and 15 in the intron, and three in the 3’ flanking region. There was one recombination per 88.1 bp of alignment on average.

Selection

Tajima’s D (Tajima 1989), Fu and Li’s D*, and Fu and Li’s F* (Fu and Li 1993) test statics were all lower than -1.87 (Table 2) and were all significant for each region (UTR, coding, intron, and 3’flanking) as well as the entire gene (p=0.05). These indicate that selection occurred within each region as well as the entire COMT gene. These negative values also indicated an excess of low frequency polymorphisms, partly due to the large number of singletons. Regarding indel polymorphisms, the Tajima’s D (Tajima 1989) test was negative for the 5’-UTR, intron, and coding regions, as well as for the entire gene, but only significant for coding regions (p=0.05) (Table 3).

Sequence variation for the COMT locus within the overlap region of elite lines

Polymorphisms and haplotypes
The alignment of the overlap region of the COMT locus for the 70 elite lines spanned 2136 sites with 323 sites with gaps due to indels, which lead to 1860 net sites (numbers before vertical lines in Table 4). According to the alignment, 56 SNPs were discovered, of which two were singleton SNPs (Table 4). This resulted in one parsimony informative SNP per 34.4bp (1860/54) and one (singleton and parsimony informative) SNP per 33.2 bp on average, respectively. These SNPs were predominantly distributed in non-coding regions (p=0.05), with 42 and 14 in non-coding and coding regions, respectively. Six (42.9%) of the coding region located SNPs lead to amino acid changes. In addition, 45 indels were identified in the full alignment, but none located in coding region. With these 56 SNPs, the 70 elite lines were classified into 30 distinct haplotypes (Hap1-Hap30). Hap1, 8, and 19 were found in 15, 9, and 10 elite lines, each of 8 haplotypes (Hap21, 14, 13, 6, 18, 5, 18, and 25) in two to four elite lines, while most haplotypes (19) were found in only one elite line (Figure 5).

**Nucleotide diversity**

The nucleotide diversity was highest in the intron region (\(\pi = 0.021\)), while both coding (\(\pi = 0.0037\)) and 5'-UTR (\(\pi = 0.0026\)) regions had much lower diversity (numbers before vertical lines in Table 4). The overall nucleotide diversity (\(\pi\)) of the overlap region was 0.011. The \(k\) value was 19.65.

**Decay of linkage disequilibrium and recombination**

Fifty two parsimony informative sites (excluding the sites segregating for three or four nucleotides) lead to 1081 site pairs were tested for significance of LD. 319 of these 1081 \(r^2\) values were significant after Bonferroni multiple comparison adjustment. One hundred forty eight (46%) of those 319 site pairs were more than 200 bp apart. A plot of \(r^2\) against physical distance between the informative polymorphic pairs clearly indicated that the LD could persist...
~500 bp with $r^2 \geq 0.1$ (Figure 3b). There were only 14 recombination sites detected in the overlap region among elite lines, with five and nine occurring in the coding and intron regions, respectively. These resulted in one recombination per ~152.6 bp (2136/12) of alignment on average.

**Selection**

For elite COMT alleles, Tajima’s D (Tajima 1989), Fu and Li’s D*, and Fu and Li’s F* (Fu and Li 1993) test statistics were all positive for the overlap region, coding region, and intron, except for Tajima’s D for the 5’-UTR region (numbers before vertical lines in Table 4). The three test statistics were all significant for intron region as well as the entire overlap region ($p=0.05$). Regarding coding regions, only Fu and Li’s D* (Fu and Li 1993) test indicated significant selection ($p=0.05$). The positive test statistic indicated an excess of alleles with intermediate frequencies.

**Comparison of COMT locus at DNA level between exotic and elite alleles**

The findings for the overlap and the entire sequenced region of the 55 exotic alleles were only slightly different for the estimates of polymorphisms, nucleotide diversity, and tests for the neutral model (by Tajima’s D, Fu and Li’s D*, Fu and Li’s F* tests). For comparison, the estimates of the overlap region for the 55 exotic alleles are also listed in Table 4.

Increased genetic variation was observed in exotic alleles compared to elite alleles, which was supported by higher k value (Tajima 1983) in exotic alleles (the average number of nucleotide differences between any two randomly selected alleles). The k value (Tajima 1983) in exotic haplotypes was 27.69, much higher than that (k=18.02) in elite haplotypes.

The distribution of SNPs was significantly different between exotic populations and elite lines ($p=0.00007$). In elite alleles, most (42/56=75%) of the SNPs were located in non-coding regions. In the exotic populations, only (187/402) 46.5% of SNPs were discovered in non-coding regions.
(Table 4). In the coding regions, 14 SNPs were identified within the overlap region in elite alleles. Eight and six of the 14 SNPs resulted in synonymous and non-synonymous substitutions, respectively (Table 4). The ratio of non-synonymous to synonymous was close to 1:1. In contrast, 158 non-synonymous and 57 synonymous SNPs were located in the coding region in exotic alleles, leading to about 3:1 synonymous to non-synonymous ratio, significantly higher than that in elite alleles (p<0.05). Based on the average number of synonymous (266.48 for elite alleles; 258.44 for exotic alleles) and non-synonymous sites (807.52 for elite alleles; 788.56 for exotic alleles) in coding regions, the polymorphism in synonymous (8/266.48=3.0% variation) and non-synonymous sites (46/258.44=17.80% variation) in exotic alleles were about 6 and 24 times higher than those (6/807.52=0.74% and 143/788.56=18.13% variation, respectively) in elite alleles.

In both elite and exotic COMT alleles, the nucleotide diversity had the same trend across the overlap region, with the highest π values in the intron, followed by coding regions and the 5’-UTR (Figure 2). In exotic populations, the π value of the overlap region was 0.017, 1.55 times of that of elite lines (π=0.011). The nucleotide diversity of the non-coding regions (π=0.026) in exotic alleles were only 1.3 times of counterparts (π=0.020) in elite lines. Yet, the nucleotide diversity (π=0.011) of coding regions in exotic populations was 2.97 times of its counterpart (π=0.0037) of elite lines. Different from elite lines, the two exons contributed much to the high nucleotide diversity of the whole gene in exotic alleles. Moreover, each part of the entire gene had higher π in exotic populations than in elite lines (Table 4; Figure 2).

At the COMT locus in exotic populations, the plot of $r^2$ against physical distance for informative polymorphic pairs indicated a much more dramatic LD decay in exotic populations than for elite alleles. In exotic populations, $r^2 \geq 0.1$ could persist ~100bp (Figure 3a), while $r^2 \geq 0.1$
could persist over 500bp for elite populations (Figure 3b). Another way to describe the LD decay was to build the relationship between LD and physical distance by regression analysis (Sokal and Rohlf 1981). The LD decay in exotic populations could be fitted by $Y=0.101-0.0524X$, where $Y$ is the $r^2$ values between sites and $X$ is the distance between sites measured by kb. In elite lines, the LD decay could be fitted by $Y=0.2927-0.2309X$. Hence, the LD decay in exotic alleles was ~4.41 times more rapid than (0.2309/0.0524) that in elite lines per one kb.

**Comparison of predicted COMT amino acids between exotic and elite COMT alleles**

As some amino acids have similar polarities, hydrophobic attributes, and electric charges, the interchange between them might not alter the protein function. Therefore, we dig into the amino acid level to get insight of the conservative/non-conservative amino acid substitutions as well as the protein diversity. Nine of the 55 exotic COMT alleles had indels in coding regions, of which eight were expected have reading frame shifts and were not included in the alignment analysis. The alignment of other 47 exotic predicted (47 haplotype) COMT amino acid chains in CLUSTALW (Thompson et al. 1994) revealed 106 amino acid substitutions in total. 39 (36.8%) and 29 (27.3%) of them were conservative and semi-conservative amino acid substitutions, respectively. The other 38 (35.9%) were non-conservative amino acid substitutions. In addition, most of non-conservative amino acid substitutions were located in the functional domains, with three and 24 within the dimerisation and methyltransferase domain of COMT protein (Figure 5). The 30 elite COMT haplotypes had only six amino acid substitutions, much less that found in the 47 exotic haplotypes. Moreover, there was no non-conservative amino acid substitutions in the 30 elite haplotype amino acids alignment, instead only three conservative and three semi-conservative amino acid substitutions were found (Figure 5). Reduced protein diversity was expected in elite COMT alleles as lower number of non-conservative and semi-conservative
amino acid substitutions than in exotic COMT alleles. Based on the neighbor-joining (Saitou and Nei, 1987) tree of protein sequences, at least 45 classes of COMT proteins might exist (Figure 6). Plus the eight alleles with reading frame shifts, the 55 exotic COMT alleles might be classified into at least 53 classes, while there were only seven classes of 30 elite COMT proteins (Figure 6). These evidences indicated strong selection occurred from landraces to elite lines, which resulted in much less protein variation in COMT locus in elite gene pool.

**Phylogenetic analysis of exotic and elite COMT alleles**

The Phylogenetic tree was constructed by the neighbor-joining (NJ) method (Saitou and Nei, 1987) based on the overlap region of COMT and contained 55 exotic haplotypes and 30 elite haplotypes (Figure 7). The tree had two major clusters with the first major cluster being highlighted (Figure 7) and was called cluster 1. The other cluster, which was not highlighted, was called cluster 2. Besides Hap_5, Hap_6, and sample no. 3798 (dark highlighted), cluster 1 has two major sub-clusters, which were highlighted with green (called cluster 1-1) and yellow color (called cluster 1-2). In cluster 1-2, most (11 out 18) of the elite lines are flint lines and the other seven included (six) five dent lines and one mixed origin line (Noordlander_VCA145). In this sub-cluster, COMT alleles from the majority of elite lines of the Hap_19 (marked by a star in Figure 7) were considered to hold all favorable QTPs for digestibility of neutral detergent fiber (DNDF) in the study of Brenner et al. (2010; 2011), including (F)1, (F)2, (F)04, (F)06, (F)15, (F)22, (F)23, (F)24. (F)16 in Hap_23, and (D)39 in Hap_21 (marked by stars in Figure 7) also combines all favorable QTPs. The exotic alleles in this sub-cluster are from Venezuela (two exotic populations), Paraguay (two exotic populations), Colombia (one exotic population), and Bolivia (one exotic population). Similar to cluster 1-2, most elite lines in cluster1-1 are flint lines (24 out 32). Almost all of the elite COMT alleles with both favorable and unfavorable QTPs for
DNDF (Brenner et al. 2010) were observed in this major cluster, including (F19), (F)20, (F)13, F(12), F(21), (F)07, (F)03,F(17),F(5), (F)18, (F)05,(F40), (D)35, (D)18, (D)31, and (D)27, which were distributed in Hap_8, Hap_13, Hap_14, Hap_15, Hap_18, and Hap_25 (marked by triangles in Figure 7). This sub-cluster included exotic populations mainly from Venezuela (six exotic populations), Mexico (three exotic populations), and Colombia (two exotic populations). In addition, one exotic population was from each of Peru, Colombia, and Chile. In cluster 2, all elite lines are dent lines. In this major cluster, the elite lines in Hap_1, Hap_3, and Hap_4 (marked with by crosses in Figure 7) were considered to hold only unfavorable QTPs for DNDF for the COMT locus (Brenner et al. 2010; 2011). The exotic populations in this cluster included seven populations from each of Colombia and Peru, eight from Bolivia, four from Ecuador, two from Cuba, and one each from of Venezuela, Paraguay, and Mexico.

**Discussion**

**Evaluation of genetic variation of GEM materials**

The goal of GEM (Pollak 2003) is to broaden the genetic base of commercial hybrid corn by incorporation of novel and useful germplasm collected from open pollinated races world-wide. Therefore, genetic variation in GEM is expected to be higher than among elite inbred lines. However, the molecular diversity within GEM has so far not been studied. Zein et al. (2004) reported pair-wise nucleotide diversity 0.00834 of the COMT locus in panel of 40 inbred lines, which were collected based on extreme values for DNDF from over 300 inbred lines used for forage breeding in Europe. Guillet-Claude et al. (2004) studied inbred lines from both USA and Europe, representing broad range variation for cell wall digestibility. As they selected wider germplasm than Zein et al. (2007), they reported pair-wise nucleotide diversity 0.011. The pair-wise nucleotide diversity was reported to be between 0.0047 to 0.0067 for thousands of genes in
maize inbred lines (Ching et al. 2003; Yamasaki et al. 2005; Rafalski et al. 2009). It is obvious, that the genetic variation within COMT even in elite lines is higher than many other genes in maize, as well as seven of other nine reported monolignol genes (C4H, 4CL1, C3H, F5H, CAD, PAL, CCoAOMT1) with π ranging from 0.00049 to 0.0062 (Andersen et al. 2007 and 2008; Guillet-Claude et al. 2004). The nucleotide diversity of other two reported monolignol genes CCoAOMT2 (π=0.0084, Guillet-Claude et al. 2004) and 4CL2 (Andersen et al. 2008) are quite close to that of COMT found in elite lines. However, the genetic variation within COMT randomly selected from GEM (π=0.01705) is much higher than that (π=0.01057) in the combined 70 elite lines (Table 4). This indicates that GEM can substantially broaden genetic variation at the COMT locus and other loci.

For breeders, their interest to broaden genetic diversity is driven by the goal to obtain more phenotypic variation for breeding purposes. Within the COMT locus, the genetic diversity across the whole gene is higher in GEM than that in elite lines, but the extent is different within the gene. The genetic diversity of the intron in elite COMT alleles almost catches up with the genetic diversity of intron in exotic COMT alleles. While, the genetic diversity of coding regions in exotic COMT alleles was 3 times of that in elite lines. As alteration in the coding regions is more likely to change gene expression and/or protein structure, there is an increased chance of altered lignin composition and content among those GEM alleles. Moreover, both the number of non-synonymous SNP (162) and the ratio of non-synonymous SNP to synonymous (3:1) amino acid changes are much higher than those (6 non-synonymous SNPs and 1:1 ratio) in elite lines. Furthermore, much higher of protein variation were expected in exotic COMT alleles as more non-conservative amino acid substitutions (Figure 5) and more protein diversity (Figure 6, Figure 7) in exotic than elite COMT alleles. Based on these we expect more variation at the
protein level in GEM materials, which will lead to a wider variation of cell wall digestibility.

**Mining of potential useful COMT alleles**

The \( \text{bm3} \) phenotype is caused by structural changes in the COMT gene (Vignols et al. 1995; Morrow et al. 1997). So far, three \( \text{bm3} \) mutations have been identified. Vignols et al. (1995) first characterized \( \text{bm3}-1 \) and \( \text{bm3}-2 \). \( \text{bm3}-1 \) has a B5 retrotransposon inserted into the junction of intron and the second exon, which produces a chimeric COMT mRNA with part of wild type COMT mRNA, followed by B5 long terminal repeat (LTR) sequence. As a stop codon is introduced by the B5 LTR, the mRNA of \( \text{bm3}-1 \) is 155 amino acids shorter than the wild type mRNA (364 amino acids). A large deletion, which occurred downstream of and close to the B5 insertion site in \( \text{bm3}-1 \), caused the \( \text{bm3}-2 \) mutation. In \( \text{bm3}-2 \), part of the second intron of the COMT gene is removed. \( \text{bm3}-1 \) produces very low levels of chimeric mRNA, while no mRNA is produced in \( \text{bm3}-2 \), both probably due to post-transcriptional regulation. \( \text{bm3}-3 \) has a larger deletion than \( \text{bm3}-2 \), with the second exon lost (Morrow et al. 1997).

To date, no additional mutations in this commercially important gene have been identified. In our study, we found eight exotic alleles having indels within the coding regions. These indels lead to frame shift and/or introduce stop codons, which finally change peptide content and/or result in premature mRNA. Four alleles (sample no. 3619, no. 3730, no. 3133, and no. 3720) had indels in the first codon and are expected to produce very short premature mRNA. Their peptide lengths are expected to be 130, 11, 60, and 130 amino acids, which are far shorter than the wild type peptide. Moreover, the amino acid sequence of these alleles is dramatically changed due to reading frame shift. There are only 68, 3, 40, and 34 consecutive amino acids identical between these four alleles and the B73 allele. The allele from sample number 3659 had a “T” insertion at the end of the second exon, which destructed the stop codon in the wild type allele and is
expected to produce a longer amino acid chain. The alleles from sample number 3656 had a one basepair deletion in the second exon and introduced a stop codon nearby the deletion, so that the expected protein length is 180 amino acids. The allele from sample number 3834 is expected to produce 318 amino acids. In conclusion, all these eight alleles might produce (pre-) mRNA, resulting in proteins with different length and content. Except for the allele from sample number 3725, which had lost an “Ala” amino acid in the region between two domains, the other seven COMT alleles might produce similar brown midrib phenotypes as \textit{bm3} due to pre-mature mRNA and/or altered peptide content. Yet they might be with different severity on lignin content/composition, cell wall digestibility, and agronomic traits, because the mutations are in different domains. As a consequence, they might be valuable genetic resources to study the function of different domains of COMT and are potential genetic resources for forage breeding. This demonstrates the usefulness of GEM materials for identification of new mutations for genetic and breeding purposes.

**Potential advantage of GEM materials for association analyses**

Candidate gene association analyses are promising to search for causative polymorphisms, provided that the LD in the respective association panel is low to remove the linkage disequilibrium between causative and non-causative polymorphisms. Brenner et al. (2010) identified putative causative polymorphisms for cell wall digestibility within the COMT gene in a panel of 40 elite lines. Although several polymorphisms were associated with DNDF, OMD, NDF, and WSC, LD prevented identification of causative polymorphisms. After we combined the 40 elite lines with 34 lines used by another group (Guillet-Claude et al. 2004), the LD extent in elite lines was still high as LD block with \( r^2 > 0.1 \) could persist about 500bp (Figure 3b). Figure 3b also shows, that there might be three LD blocks \( (r^2 > 0.1) \) within COMT locus in elite
lines, with the largest one spanning about 950 bp. In contrast, the LD extent was estimated as only one fourth of that in elite with LD block ($r^2>0.1$) only persisting ~100bp. Only one LD block exists within COMT in GEM materials which spans about 250 bp (Fig. 3a). This demonstrates the usefulness of GEM for high resolution association analyses.

The nature of genic pleiotropy of monolignol biosynthesis genes was studied by a candidate gene association approach (Chen et al. 2010). Genic pleiotropy of monolignolin biosynthesis genes regarding biomass yield related and cell wall digestibility related traits was most likely to be intragenic linkage of QTPs. Thus, optimal haplotypes which combine QTPs beneficial for both cell wall digestibility and biomass might exist, and can be identified from a large genetic pool. In order to best define how beneficial a haplotype is for biomass yield and cell wall digestibility, it is prerequisite to find the causative polymorphism(s) for each trait. Since the very low LD in GEM would break the linkage between polymorphisms responsible for biomass yield related and cell wall digestibility traits, it is possible to obtain haplotypes with beneficial QTP alleles for both groups of traits. Therefore, our next step is to construct isogenic lines for all the different COMT haplotypes. After we obtain these genetic tools, we will evaluate their effects on cell wall digestibility and other agronomic traits. The in vivo results will help us test our hypothesis of genic intragenic linkage for COMT gene regarding biomass yield and cell wall digestibility, which was proposed based on association analysis (Chen et al. 2010). If this hypothesis is true, there will be modified B47 and Z51 elite lines with improved cell wall digestibility, which could be directly used in hybrid breeding.

**Evolution of the COMT Locus**

COMT is a key enzyme in monolignol biosynthesis. The structural mutations in COMT gene lead to the brown midrib phenotype (Vignols et al. 1995; Morrow et al. 1997) which is associated
with both reduced lignin content and alter lignin composition (Barrière et al. 2004b). Lignin plays a pivotal role in supporting plant standability, transferring water, and resistance to pathogen attack (Zhao and Dixon 2011). All these traits are targets of breeding. In particular, improvement of maize standability contributed a lot to the large increase of maize yield in the past decades (Barrière et al. 2005). For these reasons, it is reasonable to assume that COMT gene has been experiencing selection pressure. All the Tajima’s D (Tajima 1989), Fu and Li's D*, Fu and Fi's F* (Fu and Li 1993) tests indicated significant selection occurred in COMT locus in both exotic and elite COMT alleles (Table 4). Virtually, the combined 70 elite lines were selected to represent large variation of cell wall digestibility, thus it is not unexpected to see selection in elite COMT locus. Selection will lead to reduced genetic diversity. Therefore, the pair-wise nucleotide diversity in elite COMT alleles only maintained 62.0% diversity in sampled COMT alleles from GEM which collected exotic open pollinated races originally.

At the DNA level, much higher frequency of singletons was observed in exotic COMT alleles. In contrast only two singleton SNPs were discovered in elite lines. This indicates that most of the singleton SNPs might be lost during inbreeding from the open pollinated races. Part of the lost singleton SNPs might be favorable alleles for cell wall digestibility improvement, since inbreeding mostly focused on yield in past breeding practice. The higher frequency of singleton SNPs in exotic COMT alleles could be due to the wide collection of isolated landraces. During domestication, mutations specific to a geographical region might have advantages regarding fitness and become fixed through positive natural selection. On the other hand, it is also possible for deleterious mutations being fixed in isolated populations if they are introduced into new areas and followed by rapid population expansion and breeding (Buckler and Thornsberry 2002), which has been observed in Arabidopsis (Purugganan and Suddith 1999; Kawabe et al. 2000;
Kuittinen and Aguade 2000) and might also occurred in maize. Similarly, the factors that random drift and bottle neck effect during introducing populations into new areas and followed breeding activities based only the introduced materials could also result in alleles specific to a certain geographic areas. Since the collected races are open pollinated, the singleton are mostly in heterozygous status and not selected against. The mutations and alleles which are specific to a certain geographical region most likely appear as singletons when they are aligned with the alleles from other geographical regions. Since our materials are sampled from races covering a wide geographical area, high level of singletons were expected. On contrast, developing inbred lines from open pollinated races lead to homozygous fate of alleles. The singleton SNPs are prone to be lost during in breeding for inbred line development as if they are deleterious sites. At the level of predicted amino acids, no non-conservative amino acid substitutions were found in the 30 elite haplotypes, in contrast, there were plenty of such substitutions in exotic COMT alleles. These again supported the inbreeding practices has kicked out the most protein variation in COMT and it is a target under human selection.

Acknowledgement

We would like to thank Elizabeth Bovenmyer for help in leaf tissue collecting, Janine Comstock for help in plasmid extraction. We greatly appreciate Dr. Randy Shoemaker allowing us to use facilities in his lab for sequencing. We also thank Rebecca Nolan for technical support during the sequencing. This project is supported by Baker Center for Plant Breeding at Iowa State University.

Authors’ Contributions

CY is the primary research and author, did field and lab work to get the sequence data, analyzed the data, and wrote this manuscript. BM provided plant materials and helped in the
manuscript writing. JQ helped to screen out the heterzyotes. FUK helped to design SSR markers.

LT conceived the idea, and together with CY designed the experiment and wrote the manuscript.
References


(2004b) Genetic and molecular basis of grass cell wall biosynthesis and degradability. II. Lessons from brown-midrib mutants. C R Biol 327:847-860


Fu YX and Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133: 693-709

Weir BS (1996) Genetic data analysis II. Sinauer, Sunderland


Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics, 61:893-903


Pollak LM (2003) The history and success of the public-private project on germplasm enhancement of maize (GEM). Advances in agronomy 78: 45-87

Purugganan MD, Suddith JI (1999) Molecular population genetics of floral homeotic loci: departures from the equilibrium-neutral model at the APETALA3 and PISTILLATA genes of Arabidopsis thaliana. Genetics, 151: 839-848


Weir BS (1996) Genetic data analysis II. Sinauer, Sunderland

Yamasaki M, Tenaillon MI, Bi IV, Schroeder SG, Sanchez-Villeda H, Doebley JF, Gaut BS, McMullen MD (2005) A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell 17: 2859-2872


<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3641</td>
<td>BC</td>
<td>PHB47</td>
<td>CAMELIA CHI411</td>
<td>Chile</td>
<td>8523</td>
<td>2270</td>
<td>3.8</td>
<td>1.14E-04</td>
<td>0.26</td>
</tr>
<tr>
<td>3675</td>
<td>BC</td>
<td>PHB47</td>
<td>SAN MARCENO # [GU1724][INIA]</td>
<td>Guatemala</td>
<td>8412</td>
<td>2174</td>
<td>3.9</td>
<td>1.50E-04</td>
<td>0.33</td>
</tr>
<tr>
<td>3713</td>
<td>BC</td>
<td>PHZ51</td>
<td>CHANDELLE CUB68 CI</td>
<td>Cuba</td>
<td>7664</td>
<td>2170</td>
<td>3.5</td>
<td>1.32E-04</td>
<td>0.29</td>
</tr>
<tr>
<td>3714</td>
<td>BC</td>
<td>PHZ51</td>
<td>CHANDELLE CUB68 CI</td>
<td>Cuba</td>
<td>8800</td>
<td>2179</td>
<td>4.1</td>
<td>6.92E-05</td>
<td>0.15</td>
</tr>
<tr>
<td>3139</td>
<td>BC</td>
<td>PHB47</td>
<td>Avati Moroti Guapi - PAG 139</td>
<td>Paraguay</td>
<td>6132</td>
<td>2156</td>
<td>2.8</td>
<td>1.08E-04</td>
<td>0.23</td>
</tr>
<tr>
<td>3140</td>
<td>BC</td>
<td>PHZ51</td>
<td>Avati Moroti Guapi - PAG 139</td>
<td>Paraguay</td>
<td>6354</td>
<td>2174</td>
<td>2.9</td>
<td>1.85E-04</td>
<td>0.40</td>
</tr>
<tr>
<td>3145</td>
<td>BC</td>
<td>PHB47</td>
<td>Avati Moroti Ti - PAG 124 B</td>
<td>Paraguay</td>
<td>5874</td>
<td>2182</td>
<td>2.7</td>
<td>1.31E-04</td>
<td>0.29</td>
</tr>
<tr>
<td>3646</td>
<td>BC</td>
<td>PHB47</td>
<td>GORDO [CHI131][CIMYT]</td>
<td>Mexico</td>
<td>8692</td>
<td>2218</td>
<td>3.9</td>
<td>1.03E-04</td>
<td>0.23</td>
</tr>
<tr>
<td>3705</td>
<td>BC</td>
<td>PHZ51</td>
<td>ELOTES OCCIDENT DGO236</td>
<td>Mexico</td>
<td>8490</td>
<td>2178</td>
<td>3.9</td>
<td>6.01E-05</td>
<td>0.13</td>
</tr>
<tr>
<td>3707</td>
<td>BC</td>
<td>PHZ51</td>
<td>DULCILLO DEL NO SON57</td>
<td>Mexico</td>
<td>6450</td>
<td>2231</td>
<td>2.9</td>
<td>1.35E-04</td>
<td>0.30</td>
</tr>
<tr>
<td>3709</td>
<td>BC</td>
<td>PHZ51</td>
<td>BOFO DGO123</td>
<td>Mexico</td>
<td>8349</td>
<td>2173</td>
<td>3.8</td>
<td>8.73E-05</td>
<td>0.19</td>
</tr>
<tr>
<td>3683</td>
<td>BC</td>
<td>PHB47</td>
<td>UCHIM ECU681</td>
<td>Ecuador</td>
<td>8290</td>
<td>2257</td>
<td>3.7</td>
<td>7.87E-05</td>
<td>0.18</td>
</tr>
<tr>
<td>3685</td>
<td>BC</td>
<td>PHB47</td>
<td>ENANO GIGANTE ECU</td>
<td>Ecuador</td>
<td>7920</td>
<td>2173</td>
<td>3.6</td>
<td>1.46E-04</td>
<td>0.32</td>
</tr>
<tr>
<td>3792</td>
<td>BC</td>
<td>PHZ51</td>
<td>CANGUIL GR. ECU447</td>
<td>Ecuador</td>
<td>6176</td>
<td>2173</td>
<td>2.8</td>
<td>1.25E-04</td>
<td>0.27</td>
</tr>
<tr>
<td>3798</td>
<td>BC</td>
<td>PHZ51</td>
<td>TUSILLA ECU569 ICA</td>
<td>Ecuador</td>
<td>8987</td>
<td>2168</td>
<td>4.1</td>
<td>7.92E-05</td>
<td>0.17</td>
</tr>
<tr>
<td>3805</td>
<td>BC</td>
<td>PHZ51</td>
<td>CANDELA ECU699 ICA</td>
<td>Ecuador</td>
<td>5515</td>
<td>2153</td>
<td>2.6</td>
<td>8.73E-05</td>
<td>0.19</td>
</tr>
<tr>
<td>3176</td>
<td>BC</td>
<td>PHB47</td>
<td>Canilla - VEN 693</td>
<td>Venezuela</td>
<td>6439</td>
<td>2192</td>
<td>2.9</td>
<td>7.29E-05</td>
<td>0.16</td>
</tr>
<tr>
<td>3199</td>
<td>BC</td>
<td>PHZ51</td>
<td>Chandelier - VEN 409</td>
<td>Venezuela</td>
<td>6493</td>
<td>2158</td>
<td>3</td>
<td>1.31E-04</td>
<td>0.28</td>
</tr>
<tr>
<td>3243</td>
<td>BC</td>
<td>PHB47</td>
<td>Costeño - Antioquia 394</td>
<td>Venezuela</td>
<td>5903</td>
<td>2145</td>
<td>2.8</td>
<td>1.23E-04</td>
<td>0.26</td>
</tr>
<tr>
<td>3599</td>
<td>BC</td>
<td>PHB47</td>
<td>ARAGUITO VEN678</td>
<td>Venezuela</td>
<td>7441</td>
<td>2185</td>
<td>3.4</td>
<td>7.56E-05</td>
<td>0.17</td>
</tr>
<tr>
<td>3600</td>
<td>BC</td>
<td>PHB47</td>
<td>ARAGUITO VEN678</td>
<td>Venezuela</td>
<td>5995</td>
<td>2167</td>
<td>2.8</td>
<td>1.80E-04</td>
<td>0.39</td>
</tr>
<tr>
<td>3652</td>
<td>BC</td>
<td>PHB47</td>
<td>COSTENO [VEN775][ICA]</td>
<td>Venezuela</td>
<td>9079</td>
<td>2299</td>
<td>4</td>
<td>1.02E-04</td>
<td>0.23</td>
</tr>
<tr>
<td>3653</td>
<td>BC</td>
<td>PHB47</td>
<td>COSTENO [VEN775][ICA]</td>
<td>Venezuela</td>
<td>8890</td>
<td>2328</td>
<td>3.8</td>
<td>1.64E-04</td>
<td>0.38</td>
</tr>
<tr>
<td>3654</td>
<td>BC</td>
<td>PHB47</td>
<td>COSTENO [VEN775][ICA]</td>
<td>Venezuela</td>
<td>7029</td>
<td>2342</td>
<td>3</td>
<td>1.40E-04</td>
<td>0.33</td>
</tr>
<tr>
<td>3656</td>
<td>BC</td>
<td>PHB47</td>
<td>COSTENO [VEN775][ICA]</td>
<td>Venezuela</td>
<td>8109</td>
<td>2333</td>
<td>3.5</td>
<td>1.32E-04</td>
<td>0.31</td>
</tr>
<tr>
<td>3659b</td>
<td>BC</td>
<td>PHB47</td>
<td>COSTENO [VEN775][ICA]</td>
<td>Venezuela</td>
<td>6138</td>
<td>2288</td>
<td>2.7</td>
<td>4.89E-04; 8.09E-04</td>
<td>0.69*; 0.72^a</td>
</tr>
<tr>
<td>3239</td>
<td>BC</td>
<td>PHZ51</td>
<td>Coroico blanco - BOV 406</td>
<td>Bolivia</td>
<td>6437</td>
<td>2164</td>
<td>3</td>
<td>1.32E-04</td>
<td>0.29</td>
</tr>
<tr>
<td>3260</td>
<td>BC</td>
<td>PHZ51</td>
<td>Cubano dentado - BOV 585</td>
<td>Bolivia</td>
<td>6320</td>
<td>2169</td>
<td>2.9</td>
<td>1.03E-04</td>
<td>0.22</td>
</tr>
<tr>
<td>3632</td>
<td>BC</td>
<td>PHB47</td>
<td>PISAN BOV344</td>
<td>Bolivia</td>
<td>7644</td>
<td>2260</td>
<td>3.4</td>
<td>6.34E-05</td>
<td>0.14</td>
</tr>
<tr>
<td>3635</td>
<td>BC</td>
<td>PHB47</td>
<td>PISAN BOV344</td>
<td>Bolivia</td>
<td>8985</td>
<td>2255</td>
<td>4</td>
<td>9.82E-05</td>
<td>0.22</td>
</tr>
<tr>
<td>3751</td>
<td>BC</td>
<td>PHZ51</td>
<td>PISAN BOV344</td>
<td>Bolivia</td>
<td>8205</td>
<td>2182</td>
<td>3.8</td>
<td>9.06E-05</td>
<td>0.20</td>
</tr>
<tr>
<td>3636</td>
<td>BC</td>
<td>PHB47</td>
<td>KARAPAMPA BOV978</td>
<td>Bolivia</td>
<td>7354</td>
<td>2278</td>
<td>3.2</td>
<td>6.63E-05</td>
<td>0.15</td>
</tr>
<tr>
<td>3747</td>
<td>BC</td>
<td>PHZ51</td>
<td>KCELO BOV948</td>
<td>Bolivia</td>
<td>9130</td>
<td>2178</td>
<td>4.2</td>
<td>5.83E-05</td>
<td>0.13</td>
</tr>
<tr>
<td>3765</td>
<td>BC</td>
<td>PHZ51</td>
<td>POJ CHICO BOV800</td>
<td>Bolivia</td>
<td>6483</td>
<td>2154</td>
<td>3</td>
<td>7.86E-05</td>
<td>0.17</td>
</tr>
<tr>
<td>3819</td>
<td>BC</td>
<td>PHZ51</td>
<td>COROICO BOV1064</td>
<td>Bolivia</td>
<td>6414</td>
<td>2173</td>
<td>3</td>
<td>1.04E-04</td>
<td>0.23</td>
</tr>
<tr>
<td>3610</td>
<td>BC</td>
<td>PHB47</td>
<td>YUCATAN TOL389 ICA</td>
<td>Colombia</td>
<td>9143</td>
<td>2252</td>
<td>4.1</td>
<td>1.51E-04</td>
<td>0.34</td>
</tr>
<tr>
<td>3613</td>
<td>BC</td>
<td>PHB47</td>
<td>YUCATAN TOL389 ICA</td>
<td>Colombia</td>
<td>8583</td>
<td>2247</td>
<td>3.8</td>
<td>1.77E-04</td>
<td>0.40</td>
</tr>
<tr>
<td>3614</td>
<td>BC</td>
<td>PHB47</td>
<td>YUCATAN TOL389 ICA</td>
<td>Colombia</td>
<td>8057</td>
<td>2135</td>
<td>3.8</td>
<td>4.85E-05</td>
<td>0.10</td>
</tr>
<tr>
<td>3720</td>
<td>BC</td>
<td>PHZ51</td>
<td>YUCATAN TOL389 ICA</td>
<td>Colombia</td>
<td>6859</td>
<td>2276</td>
<td>3</td>
<td>8.89E-05</td>
<td>0.20</td>
</tr>
<tr>
<td>Sample</td>
<td>Gen.</td>
<td>Loc.</td>
<td>Population</td>
<td>Coverage</td>
<td>Length (bp)</td>
<td>Error</td>
<td>Exp. wrong</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>------</td>
<td>------------</td>
<td>----------</td>
<td>-------------</td>
<td>-------</td>
<td>------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3725</td>
<td>BC2</td>
<td>PHZ51</td>
<td>CANDELA ECU531</td>
<td>Colombia</td>
<td>8680</td>
<td>2174</td>
<td>4</td>
<td>6.17E-05</td>
<td>0.13</td>
</tr>
<tr>
<td>3602</td>
<td>BC2</td>
<td>PHB47</td>
<td>ANDAQUI CAQ307 FT</td>
<td>Colombia</td>
<td>7621</td>
<td>2259</td>
<td>3.4</td>
<td>1.30E-04</td>
<td>0.29</td>
</tr>
<tr>
<td>3608</td>
<td>BC2</td>
<td>PHB47</td>
<td>MONTANA NAR625</td>
<td>Colombia</td>
<td>6629</td>
<td>2167</td>
<td>3.1</td>
<td>5.68E-05</td>
<td>0.12</td>
</tr>
<tr>
<td>3697</td>
<td>BC2</td>
<td>PHB47</td>
<td>SABANERO CUN342</td>
<td>Colombia</td>
<td>8061</td>
<td>2252</td>
<td>3.6</td>
<td>1.19E-04</td>
<td>0.27</td>
</tr>
<tr>
<td>3784</td>
<td>BC2</td>
<td>PHZ51</td>
<td>CABUYA SAN316 FT C</td>
<td>Colombia</td>
<td>7858</td>
<td>2165</td>
<td>3.6</td>
<td>1.78E-04</td>
<td>0.39</td>
</tr>
<tr>
<td>3829</td>
<td>BC2</td>
<td>PHZ51</td>
<td>CHOCOCENO CHO340</td>
<td>Colombia</td>
<td>6321</td>
<td>2160</td>
<td>2.9</td>
<td>8.14E-05</td>
<td>0.18</td>
</tr>
<tr>
<td>3133</td>
<td>BC1</td>
<td>PHZ51</td>
<td>Arizona - LIB 16</td>
<td>Peru</td>
<td>6240</td>
<td>2257</td>
<td>2.8</td>
<td>3.57E-04</td>
<td>0.81</td>
</tr>
<tr>
<td>3116</td>
<td>BC1</td>
<td>PHZ51</td>
<td>Amarillo Huancabamba - PIU 17B</td>
<td>Peru</td>
<td>6018</td>
<td>2163</td>
<td>2.8</td>
<td>7.76E-05</td>
<td>0.17</td>
</tr>
<tr>
<td>3619</td>
<td>BC2</td>
<td>PHB47</td>
<td>CON PUNT CUZ13</td>
<td>Peru</td>
<td>7463</td>
<td>2250</td>
<td>3.3</td>
<td>3.63E-04</td>
<td>0.82</td>
</tr>
<tr>
<td>3625</td>
<td>BC2</td>
<td>PHB47</td>
<td>PERLA ANC23</td>
<td>Peru</td>
<td>6171</td>
<td>2228</td>
<td>2.8</td>
<td>1.05E-04</td>
<td>0.23</td>
</tr>
<tr>
<td>3730</td>
<td>BC2</td>
<td>PHZ51</td>
<td>CON PN CUZ13</td>
<td>Peru</td>
<td>9202</td>
<td>2178</td>
<td>4.2</td>
<td>6.12E-05</td>
<td>0.13</td>
</tr>
<tr>
<td>3734</td>
<td>BC2</td>
<td>PHZ51</td>
<td>CUZCO CUZ217</td>
<td>Peru</td>
<td>8551</td>
<td>2180</td>
<td>3.9</td>
<td>4.02E-05</td>
<td>0.09</td>
</tr>
<tr>
<td>3735</td>
<td>BC2</td>
<td>PHZ51</td>
<td>CUZCO CUZ217</td>
<td>Peru</td>
<td>8063</td>
<td>2174</td>
<td>3.7</td>
<td>5.11E-05</td>
<td>0.11</td>
</tr>
<tr>
<td>3737</td>
<td>BC2</td>
<td>PHZ51</td>
<td>SG HUANCAY JUN164</td>
<td>Peru</td>
<td>6134</td>
<td>2186</td>
<td>2.8</td>
<td>1.20E-04</td>
<td>0.26</td>
</tr>
<tr>
<td>3740</td>
<td>BC2</td>
<td>PHZ51</td>
<td>SG HUANCAY JUN164</td>
<td>Peru</td>
<td>8958</td>
<td>2171</td>
<td>4.1</td>
<td>9.42E-05</td>
<td>0.20</td>
</tr>
<tr>
<td>3834</td>
<td>BC2</td>
<td>PHZ51</td>
<td>SARCO ANC184</td>
<td>Peru</td>
<td>6349</td>
<td>2167</td>
<td>2.9</td>
<td>1.22E-04</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Min: 5995 | 2135 | 2.7 | 4.02E-05 | 0.09 |
Max: 9202 | 2342 | 4.2 | 3.63E-04 | 0.82 |
Mean: 7517 | 2217 | 3.4 | 1.16E-04 | 0.26 |

2: indicates two contigs being produced in the CodonCode Aligner (V. 3.7.1, Codon Code Corporation, USA). The first contig spans 5’UTR, the first exon, and intron. The second contig spans the second exon.
Gen.: Generation; Rec. parent: recurrent parent; Cov.: coverage of sequencing (times); Est. error: estimated sequencing error of the whole contig; # Exp. wrong: expected wrong base pair calling per contig.
a: estimation for the first contig.
b: estimation for the second contig.
Calculation of min, Max, and mean do not take the allele from the sample 3659 into account.
Table 2. Summary of SNP polymorphisms, genetic diversity estimates, and selection estimates for COMT in exotic populations

<table>
<thead>
<tr>
<th></th>
<th>Entire region</th>
<th>Non-coding regions</th>
<th>Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Alignment sites</td>
<td>Total</td>
<td>5'-UTR</td>
</tr>
<tr>
<td>All alleles(55)</td>
<td>2466</td>
<td>1370</td>
<td>144</td>
</tr>
<tr>
<td>net sites</td>
<td>1989</td>
<td>904</td>
<td>50</td>
</tr>
<tr>
<td>Haplotypes</td>
<td>55</td>
<td>55</td>
<td>11</td>
</tr>
<tr>
<td>SNPs(singl)</td>
<td>438(343)</td>
<td>219(154)</td>
<td>10(10)</td>
</tr>
<tr>
<td>(\pi)</td>
<td>0.0172</td>
<td>0.02474</td>
<td>0.00727</td>
</tr>
<tr>
<td>ThetaNuc</td>
<td>0.05011</td>
<td>0.05488</td>
<td>0.04371</td>
</tr>
<tr>
<td>Tajima's D</td>
<td>(-2.3653^{**})</td>
<td>(-1.9643^*)</td>
<td>(-2.3435^{**})</td>
</tr>
<tr>
<td>Fu and Li's D*</td>
<td>(-5.8496^{**})</td>
<td>(-4.9721^{**})</td>
<td>(-4.8746^{**})</td>
</tr>
<tr>
<td>Fu and Li's F*</td>
<td>(-5.3844^{**})</td>
<td>(-4.5631^{**})</td>
<td>(-4.7608^{**})</td>
</tr>
</tbody>
</table>

UTR: untranslated region

Numbers in the parentheses of “# SNPs (singl)” are numbers of singleton SNPs.

*, \(P < 0.05\); **, \(P < 0.01\); ***, \(P < 0.001\)
Table 3. Summary of indel polymorphisms, their genetic diversity and selection estimates within \textit{COMT} locus in exotic populations

<table>
<thead>
<tr>
<th></th>
<th>Entire</th>
<th>Non-coding</th>
<th>5'-UTR</th>
<th>Intron</th>
<th>3' flanking</th>
<th>Coding</th>
<th>coding1</th>
<th>coding2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alignment sites</td>
<td>2377</td>
<td>1281</td>
<td>100</td>
<td>966</td>
<td>215</td>
<td>1096</td>
<td>422</td>
<td>674</td>
</tr>
<tr>
<td>Indel sites</td>
<td>388</td>
<td>377</td>
<td>50</td>
<td>262</td>
<td>65</td>
<td>11</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Indel events</td>
<td>72</td>
<td>63</td>
<td>6</td>
<td>43</td>
<td>14</td>
<td>9</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>(\pi(i))</td>
<td>0.00627</td>
<td>0.00667</td>
<td>0.0106</td>
<td>0.017</td>
<td>0.0043</td>
<td>0.00022</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tajima's D</td>
<td>-0.18365</td>
<td>-1.20759</td>
<td>0.030834</td>
<td>0.58121</td>
<td>-1.98089</td>
<td>-1.85303*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

UTR: untranslated region; \(\pi(i)\) is diversity of indels.
*: \(P < 0.05\)
**Table 4.** Comparison of polymorphisms, genetic diversity estimates, and selection estimates for *COMT* within the overlap region of COMT locus between inbred lines and exotic populations

<table>
<thead>
<tr>
<th></th>
<th>Entire overlap</th>
<th>Non-coding regions</th>
<th>Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>5'-UTR</td>
<td>Intron</td>
</tr>
<tr>
<td><strong>All alleles(70)(55)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>number of sites</td>
<td>2136(2141)</td>
<td>1062(1066)</td>
<td>98(100)</td>
</tr>
<tr>
<td>net sites</td>
<td>1800(1818)</td>
<td>780(754)</td>
<td>52(50)</td>
</tr>
<tr>
<td>Indel</td>
<td>45(58)</td>
<td>45(49)</td>
<td>5(6)</td>
</tr>
<tr>
<td>SNP(singl)</td>
<td>56(2)/40(2)(315)</td>
<td>42(2)/187(130)</td>
<td>1(0)/10</td>
</tr>
<tr>
<td>Haplotype</td>
<td>30(55)</td>
<td>15(55)</td>
<td>21(0)</td>
</tr>
<tr>
<td>Pi</td>
<td>0.011(0.017)</td>
<td>0.019(0.0264)</td>
<td>0.0025(0.00727)</td>
</tr>
<tr>
<td>Theta&lt;sub&gt;Nuc&lt;/sub&gt;</td>
<td>0.006(0.05037)</td>
<td>0.013(0.0562)</td>
<td>0.00039(0.004371)</td>
</tr>
<tr>
<td>Tajima's D&lt;sup&gt;*&lt;/sup&gt;</td>
<td>2.110(2.3809)**</td>
<td>2.474(1.9499)*</td>
<td>-0.935(2.3435)**</td>
</tr>
<tr>
<td>Fu and Li's D&lt;sup&gt;**&lt;/sup&gt;</td>
<td>1.777(2.854)**</td>
<td>1.378(1.8821)**</td>
<td>0.316(4.8746)**</td>
</tr>
<tr>
<td>Fu and Li's F&lt;sup&gt;**&lt;/sup&gt;</td>
<td>2.828(3.389)**</td>
<td>2.298(4.9206)**</td>
<td>0.274(4.7008)**</td>
</tr>
</tbody>
</table>

UTR: untranslated region

*<sup>a</sup>: mutations caused by the SNPs, which will larger than SNP number if there are SNPs with three or four variant.

*, P < 0.05; **, P < 0.01; ***, P < 0.001

Numbers in the parentheses of “# SNPs(singl)” row are numbers of singleton SNP. Numbers before and after vertical lines are estimations for elite lines and exotic populations, respectively.
Figures

**Figure 1.** Schematic diagram of the *COMT* gene and primers used for cloning and sequencing. The length of arrow which denote primers indicate how long the sequencing could achieve if using that primer in our sequencing, except the primer Promoter F, which is only used for cloning together with Ex2R. The wide boxes represent the exons. The line between two boxes represents the intron. The narrow boxes represent 5’ UTR and 3’ UTR regions.

**Figure 2.** Nucleotide diversity values (π represented on Y axis) in sliding windows (length=100, step size=10).
Red line represents the nucleotide diversity (π) in 55 exotic populations. Green line represents the nucleotide diversity (π) in 70 inbred lines.

**Figure 3.** Plot of $r^2$ (Y-axis) against physical distance (bp) between pair sites along COMT locus (X-axis) in 55 exotic (a) and 70 elite COMT alleles (b).
Only parsimony informative sites were considered. A logarithmic trend was fitted to the data.

**Figure 4.** LD among parsimony informative SNPs across COMT locus in exotic populations (a) and inbred lines (b).
The positions of informative SNPs in the alignment are given. Lower left triangle: P-values derived from Fisher’s exact test. Upper right triangle: $r^2$ values. Values upper and below of the diagonal are $r^2$ and probability values. The largest LD blocks in exotic and elite populations were indicated by vertical green lines in the left.

**Figure 5.** Distribution of conserved, semi-conserved, and non-conserved amino acid substitutions due to SNPs across the whole COMT domain.
The whole COMT protein has 364 AA and the figure is depicted in proportion to the length of two domains. The two domains are domerisation and methytransferase domain, which are about 30 and 240AA in length. The two box stand for the domains, and the lines stand for non-domain regions. The three numbers separated by two slashes represents conserved, semi-conserved, and non-conserved AA caused by SNPs. The numbers up and below the lines are for the 47 exotic and 30 elite COMT haplotypes. Here only 47 exotic COMT haplotypes are aligned because the other eight had indels in coding regions and are expected to produce COMT protein with more or less than 364 AAs.

**Figure 6.** Comparison NJ trees of predicted amino acids between exotic and elite COMT haplotypes.
The eight exotic haplotypes which were predicted to have frame shifting in coding region were not included in the tree construction, thus only 47 AAs of exotic COMT haplotypes were listed in the left figure. Most of the 47 AA were unique and only proteins of COMT alleles from sample 3709, 3652 and 3140 could be classified into the same group. Plus the eight ones with indels in the coding regions which expected to bring frame shifting, the 55 exotic COMT proteins could be classified into 53 groups. The AAs of 30 elite COMT haplotypes could be classified into seven groups, which were indicated by the numbers (1-7) in the right figure.

**Figure 7.** Phylogenetic tree of elite and exotic COMT DNA haplotypes.
The F and D in parentheses indicated either dent or flint pool the inbred line belonging to. The prefix (F) or (D) were directly added to the Arabic numerals (1-42) which were used to distinguish the 42 elite lines in the study of Zein et al. (2007). The names of the other 34 elite lines which were listed in the report of Guillet-Claude et al. (2004) and began with English letter rather than Arabic number were also added with the prefix (F) or (D).

76 entries which were studied by Guillet-Claude et al. (2004) and Zein et al. (2007) actually equal to 70 distinct elite lines. The 55 exotic COMT alleles were grouped into 55 unique haplotypes. 76 entries (70 distinct lines) were classed into 30 haplotypes. To reduce the space of the phylogenetic tree, only the 88 haplotypes were used for the tree construction. The elite lines included in each haplotype were listed after the haplotype number (for example Hap_1). The numbers after haplotype number were the number of elite lines included in that haplotype. Hap_1 includes 15 elites. Besides the eight lines listed in this figure, there were still (D)30, (D)29, (D)26, (D)11, (D)09, (D)10, and (D)28 belonging to this haplotype.

(F)03 is equal to (F)EP1; (D)W64 = (D)W64A; (F)1 = (F)F7; (F)2 = (F)F2; (D)F288 = (D)39; (F)40 = (F)F4.
Figure 1
Figure 2
Figure 3
Figure 4
Figure 5
Figure 6

NJ tree of 47 AA sequences in exotic COMT Haplotypes

NJ tree of 30 AA sequences in elite COMT Haplotypes
Oryza-sativa_(japonica_cultivar-group)-O-methyltransferase

Figure 7
CHAPTER 8. GENERAL CONCLUSIONS

The digestibility of cell wall of stover is major factor affecting the maize stover used for either forage or lignocellulosic ethanol. Lignification of cell walls greatly influences cell wall digestibility. \textit{bm} mutants proved that they are good resources to improve cell wall digestibility and it is very promising to manipulate the genes involved in cell wall lignification. However, genes involved in cell wall lignification act pleiotropically on both cell wall digestibility and biomass yield or agronomic related traits. Therefore, understanding the nature of pleiotropic effects of genes involved in cell wall lignification will help us design appropriate strategies for bioenergy crop breeding via manipulation of cell wall related genes.

In my dissertation, I reviewed the current knowledge and approaches to study the nature of pleiotropy. Basically, statistical and in-vivo methods can be employed to dissect trait correlations at different levels, that is at QTL level, gene level, and DNA polymorphism level. Each method has its own merits, but also has drawbacks. For example, candidate gene association studies can dissect trait correlations at the DNA polymorphism level, but is they are greatly limited by the LD extent in the association panel and results might contain false positives, thus requires in-vivo validation. Therefore, it is necessary to combine different methods to finally understand the pleiotropic effects of genes involved in cell wall lignification.

First, natural mutants were employed for our purpose. There have been 13 uncharacterized brown midrib mutants called \textit{bm} star available in the Maize Stock Center. By allelism test, we found three new \textit{bm} mutants different from the historically well characterized \textit{bm1-4}. These three new \textit{bm} mutants were designated as \textit{bm5}, \textit{bm6}, and \textit{bm7}. \textit{bm5} was represented by \textit{bm*E}, \textit{bm*F}, \textit{bm*G}, and \textit{bm*H}. \textit{bm6} and \textit{bm7} were each only represented once in mutation stocks \textit{bm*J} and \textit{bm*I}, respectively. In addition, we found three new potential allelic mutations (\textit{bm*C}, \textit{bm*D}, \textit{bm*G})
One new potential allelic mutation (\textit{bm}*M) to \textit{bm}2 and (\textit{bm}*K) to \textit{bm}3, respectively.

We found that \textit{bm}6 increases cell wall digestibility and decreases plant height by comparison of mutant bulks with wild-type bulks in a F\textsubscript{2} segregating population, which was developed by crossing B73 and the \textit{bm}6 line (\textit{bm}*J). \textit{bm}6 was then assigned to the short arm of Chromosome 2 by quantitative bulked segregant analysis (BSA) and its location was confirmed in a small segregating F\textsubscript{2} population. Finally, the \textit{bm}6 gene was precisely mapped into a 180kb region based on 960 brown F\textsubscript{2} plants selected from the same but larger segregating F\textsubscript{2} population. There were 10 genes being predicted in the 180kb region in the B73 genome. Based on the annotation information of B73 as well as annotations in the syntenic region of rice, four genes are good candidates for \textit{bm}6, including a putative flavanone 3-hydroxylase (F3H) gene, a putative zinc finger transcription factor gene, a putative IQ calmodulin-binding motif gene, and a putative \textit{WUSCHEL}1 gene. This study provided tightly linked markers for marker-assisted selection in relation to forage quality and lignocellulosic ethanol conversion of stover. It will facilitate isolation of the underlying gene, and advance our understanding of cell wall biosynthesis, formation of brown midrib phenotypes, and pleiotropic effects of \textit{bm}6 gene on biomass yield and cell wall digestibility.

In association analyses between O-methyltransferase genes (COMT, CCoAOMT1, and CCoAOMT2) and cell digestibility related traits, we found a few to several polymorphisms in each gene associated with some of the cell wall digestibility related traits. In particular, 25 polymorphisms in COMT were associated with DNDF, some of which were consistent across association methodologies and association populations. These findings support that O-methyltransferase genes, especially COMT, are good candidates for enhancing cell wall
digestibility. Using the same association panel, we found at least six out of the ten monolignol genes having DNA polymorphisms associated with biomass yield-related or agronomic traits, which supported presence of pleiotropic effects at the gene level. However, most of the polymorphisms showing significance in association analyses were associated with only one group of traits, either cell wall digestibility or biomass yield-related or agronomic traits. Thus pleiotropic effects of monolignol genes were most likely due to intragenic linkage of QTPs. Therefore, optimal haplotypes which combine QTPs beneficial for both cell wall digestibility or biomass yield related/agronomic traits might exist.

In order to validate our hypothesis that the pleiotropic effects of monolignol genes were due to intragenic linkage of QTPs and to evaluate the potential usage of GEM for enhancing cell wall digestibility, we randomly selected 55 exotic COMT alleles from GEM project. We compared their genetic diversity and linkage disequilibrium at this locus (COMT locus) with those in 70 elite lines. We found a larger number of DNA polymorphisms among exotic COMT alleles (with over 400 SNPs) than among elite COMT alleles (with only 56 SNPs). The nucleotide diversity ($\pi$) among exotic COMT alleles was 1.6 times higher than among elite COMT alleles, and much higher than reported nucleotide diversities for several genes among elite inbred lines. In particular, the ratio of non-synonymous to synonymous (3:1 vs 1:1) and non-conservative amino acid substitutions (38 vs 0) in exotic COMT alleles were much higher than in elite COMT alleles. All these facts indicate abundant genetic diversity in GEM materials, which can be employed to greatly enhance the improvement of cell wall digestibility. The LD decay within the COMT locus in exotic populations was at least four times more rapid than for elite lines with $r^2 > 0.1$ persisting only about 100 bp. The low LD extent within COMT locus in exotic populations might help to break the linkage between polymorphisms responsible for biomass yield-related or
agronomic traits and cell wall digestibility traits, which will help us to validate our hypothesis and obtain haplotypes with beneficial QTP alleles for both cell wall digestibility and biomass yield or agronomic traits.
ACKNOWLEDGEMENT

When I am working on the keyboard writing this acknowledgement, I suddenly realize that my PhD study as well as my student career will end soon. Without the help of so many people, I cannot image how hard it would be for me to get my degree. So I would like to take this opportunity to express my gratefulness to all people who have helped me during my PhD study. I apologize if somebody is missed here because there are numerous persons who have helped me.

First, I would like to express my extremely grateful thanks to my advisor Dr. Lübbeastedt. I appreciated him to offer me the opportunity to do the rotation and finally join his lab four years ago. His genius and deep insight on my projects ensured the success of my research. He is not only a wise advisor in my study but also a good mentor in my life, guiding me when I was wondering in the cross road. I will for sure miss the time in his lab. I also want to thank my other POS committee members, Dr. Paul Scott, Dr. William Beavis, Dr. Phill Becraft, and Dr. Nettleton, for their constructive advice on my research. I appreciated the willingness of Dr. Phill Becraft to share mutant resources with me in the beginning of my research, although finally I found the mutants did not fit into my research. I am grateful to Dr. Paul Scott and Dr. Randy Shoemaker for allowing me to use their lab facilities. Many thanks to Ms Rebecca Nolan for her kind help in my sequencing project. The support in both field and lab from Ms Elizabeth Bovenmyer is also greatly appreciated. I am sincerely grateful for Dr. Uschi Frei’s help on bioinformatics in my research. I also would like to thank my officemates, Everton Brenner, Constantin Jansen, Andrea Arias, for their help and inspiring discussion on my study and research. I wish to thank the hard work from Farhad Ali and Hongjiu Liu in the brown midrib project.

Finally, I want to express my great appreciation and love to my family members. Thank my
parents and my wife Qing Ji for they selflessly support and consistent encouragement. It is their love and support that make me not afraid of any defeat in my life. I wish to dedicate my dissertation to my wife Qing Ji and my lovely newborn twin boys, Harrison and David.