Final Report DE-FG02-08ER64702
Translational Genomics for the Improvement of Switchgrass

Nicholas C. Carpita, PI; Maureen C. McCann, Co-PI
Purdue University

This report contains no proprietary information and is authorized for distribution

Executive Summary

Our objectives were to apply bioinformatics and high throughput sequencing technologies to identify and classify the genes involved in cell wall formation in maize and switchgrass. Targets for genetic modification were to be identified and cell wall materials isolated and assayed for enhanced performance in bioprocessing. We annotated and assembled over 750 maize genes into gene families predicted to function in cell wall biogenesis. Comparative genomics of maize, rice, and Arabidopsis sequences revealed differences in gene family structure. In addition, differences in expression between gene family members of Arabidopsis, maize and rice underscored the need for a grass-specific genetic model for functional analyses. A forward screen of mature leaves of field-grown maize lines by near-infrared spectroscopy yielded several dozen lines with heritable spectroscopic phenotypes, several of which near-infrared (*nir*) mutants had altered carbohydrate-lignin compositions. Our contributions to the maize genome sequencing effort built on knowledge of copy number variation showing that uneven gene losses between duplicated regions were involved in returning an ancient allotetraploid to a genetically diploid state. For example, although about 25% of all duplicated genes remain genome-wide, all of the cellulose synthase (CesA) homologs were retained. We showed that guaiacyl and syringyl lignin in lignocellulosic cell-wall materials from stems demonstrate a two-fold natural variation in content across a population of maize Intermated B73 x Mo7 (IBM) recombinant inbred lines, a maize Association Panel of 282 inbreds and landraces, and three populations of the maize Nested Association Mapping (NAM) recombinant inbred lines grown in three years. We then defined quantitative trait loci (QTL) for stem lignin content measured using pyrolysis molecular-beam mass spectrometry, and glucose and xylose yield measured using an enzymatic hydrolysis assay. Among five multi-year QTL for lignin abundance, two for 4-vinylphenol abundance, and four for glucose and/or xylose yield, not a single QTL for aromatic abundance and sugar yield was shared. A genome-wide association study (GWAS) for lignin abundance and sugar yield of the 282-member maize Association Panel provided candidate genes in the eleven QTL and showed that many other alleles impacting these traits exist in the broader pool of maize genetic diversity. The maize B73 and Mo17 genotypes exhibited surprisingly large differences in gene expression in developing stem tissues, suggesting certain regulatory elements can significantly enhance activity of biomass synthesis pathways. Candidate genes, identified by GWAS or by differential expression, include genes of cell-wall metabolism, transcription factors associated with vascularization and fiber formation, and components of cellular signaling pathways. Our work provides new insights and strategies beyond modification of lignin to enhance yields of biofuels from genetically tailored biomass.

I. Identification of cell-wall genes of grasses.

Our first objective was to functionally annotate the genes related to cell wall biology for switchgrass, based on homology to maize and rice sequences, and augmented with gene families that are currently of unknown function but are implicated in cell wall development. However, because at project start the switchgrass genome sequencing was still in its infancy and is still overcoming problems with genetic diversity inherent in outcrossing populations, we were directed by DOE to focus on maize comparative genomics.

Sequences of annotated cell wall-related genes of Arabidopsis and their paralogs were used in a conventional BLAST search to find the putative orthologs and paralogous sequences in rice and, subsequently, the homologous sequences in maize. Until very recently, the maize sequence was largely unannotated and incomplete; thus, computer algorithms used to find common sequences based on keywords or full-length sequences, may have missed many relevant genes. In parallel with the maize genome-sequencing project (Schnable et al., 2009), we obtained a more complete view of many cell wall gene families in maize. Dendrograms of cell wall gene families of Arabidopsis, maize, and rice were developed for each species individually and in combination. About 60% of the Arabidopsis genome is annotated with respect to predicted function of its protein products, and ours (http://cellwall.genomics.purdue.edu) and others' web sites have assembled gene families based on known functions. Our characterizations of gene families were consistent with those of the Carbohydrate-Active Enzyme database (http://www.cazy.org/) assembly of families of glycosyl transferases (GTs), glycosyl hydrolases (GHs), and other carbohydrate-metabolizing enzymes. There are 91 gene families of evolutionarily distinct GTs and 112 GHs, with Arabidopsis and rice genes populating 40 and 34 of them, respectively. Although the total number of GTs is higher in rice than in Arabidopsis, 550 versus 445, the numbers of GHs are about the same, 419 for rice and 403 for Arabidopsis. Only a few groups within families of cell wall-related genes have similar numbers of members for Arabidopsis and grasses (Penning et al., 2009). Comparative genomics showed differences in the number of members of a family, in numbers of members of a single group of a family, and in the presence of new family groups (or loss of family groups). Particularly interesting is the five-group family GT31 in Arabidopsis, for which galactosyl transferase functions are associated with both N-glycan formation and galactan backbone formation of AGPs. When grass sequences are included, a new group F emerges, which contains a single Arabidopsis gene with weak homology to genes in groups A and B. Five pairs of rice and maize genes in group F appear after the Arabidopsis-grass divergence (Fig. 1).

Gene family structure is likely a consequence of duplication and divergence in the genomes since the last common ancestor, resulting in splitting of a single gene function between paralogs (subfunctionalization), new function in a duplicate gene (neofunctionalization), or a combination of both events (subneofunctionalization). As we describe below, EST/cDNA data support differential expression of each of the paralogs, indicating their neofunctionalization and subfunctionalization after the tetraploidization event. Genes that encode the enzymes for cell wall biogenesis are assembled in families common to all plants. However, publication of the maize genome sequence (Schnable et al., 2009) has allowed a comparative genomics analysis with rice and Arabidopsis genome sequences that highlights distinctions in the family substructure for cell wall-

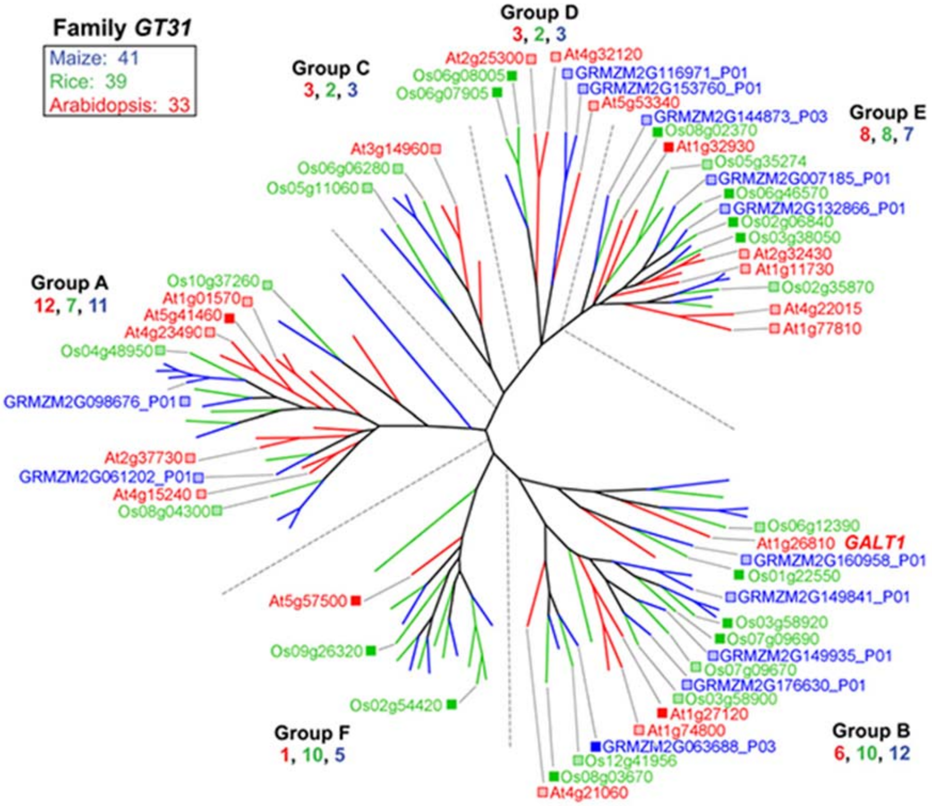

Figure 1. Genes of family GT31. Inclusion of grass sequences forms six subgroups, including the grass-dominated Group F. Genes of family GT31. For accession numbers of all genes in this family, see http://cellwall.genomics.purdue.edu/families/2-3-5/.

related genes of the grasses. In some instances, homologous sequences indicate potential orthologous functions, whereas in others, a clear expansion of groups is observed that suggests the evolution of novel functions specific to the type II-walled grasses or to the type I-walled Arabidopsis.

II. Secondary wall-related gene expression

Our second objective was to classify the maize cell wall gene family members that are highly expressed during primary and secondary wall formation. We used RNAseq to chart secondary wall gene expression in rind tissues of field-grown elongating internodes of maize inbred B73, and in greenhouse grown inbreds B73 and Mo17 (Penning et al. 2014a). Tiling arrays had demonstrated that B73, Mo17, and many other inbreds exhibit significant copy-number variation as a result of gene loss following genome duplication, such that inbreds may have two, one, or even zero copies of an ancestral gene at any locus (Springer et al., 2009). We found a remarkable variation of expression between inbreds B73 and Mo17 irrespective of gene dosage.

From preliminary transcriptome analyses we concluded that potential orthologs cannot be identified solely by identifying sequences with the highest sequence similarities with Arabidopsis (Penning et al. 2009). We deep-sequenced RNA transcripts from internodes of greenhouse-grown Mo17 and B73 plants at a developmental stage where secondary wall formation predominates. Consistent with findings of others on expression in primary root tissues (Paschold et al. 2012), we found nearly 70% of the genes of Mo17 and B73 exhibited about a 2-fold difference in expression in internode rind tissues, with 30% exhibiting 5-fold differences, 16% exhibiting 10-fold differences, and about 2% exhibiting over 100-fold differences (Penning et al. 2014a). In some instances, but not all, the fold-difference is a result of the absence of an allele in one of the parental lines. Deletions can be confined to a single gene or to large multi-gene deletions. For example, Chromosome 6 contains three closely spaced deletions totaling 2.1 Mb and corresponding to 56 genes found only in B73 (Fig. 2). Differential expression of GTs and GHs is expected to contribute to natural variation in cell wall composition and architecture revealed in analysis of the IBM lines, the Association Panel, and the NAM populations. Integrating several genetic tools such as GWAS and differential expression will prove useful to resolve those genes for which expression levels contribute to phenotype from candidates for which SNP/INDEL differences indicate more discrete targets for breeding.

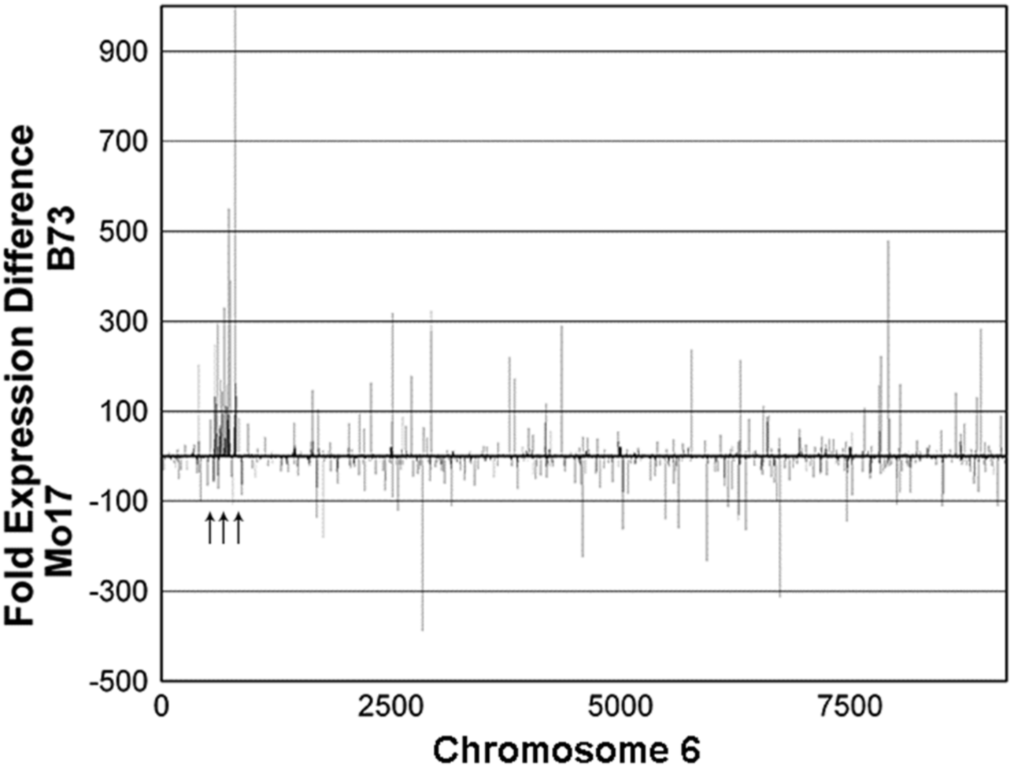

Figure 2. Differential expression of B73 and Mo17 alleles on Chromosome 6 represented by fold-change. B73 is positive fold-change, and Mo17 is negative fold-change. Note three closely spaced deletions in Mo17 (arrows).

III. Classification of genetic variation

As proof of concept, our third objective was to test mutants and transgenic lines of maize with enhanced yield of glucose and xylose in saccharification trials to identify genes that impact structure and degradability of non-cellulosic polysaccharides, and to analyze the function of a gene family of unknown function that had cellulose deficiencies when mutated.

In collaboration with scientists at NREL, we validated Pyrolysis Molecular-Beam Mass Spectrometry (PyMBMS) as a high-throughput method for relative abundances of guaiacyl and syringyl lignin in lignocellulosic cell-wall materials from stems of the IBM population (Penning et al. 2014b). Variations of up to two-fold across the population in phenylpropanoid abundance were observed. Several histochemical and quantitative biochemical assays were used to validate the mass spectrometric data for lignin, hydroxycinnamic acids, and crystalline cellulosic and non-cellulosic glucans and xylans. Pentose from xylans and hexose from cellulosic and non-cellulosic glucans also varied substantially across the population, but abundances of diagnostic fragments for these monosaccharides were not well correlated with the abundance of cell-wall polysaccharides under the conditions we employed. We demonstrated PyMBMS to be a valid high-throughput screen suitable for analysis of lignin abundance in large populations of bioenergy grasses.

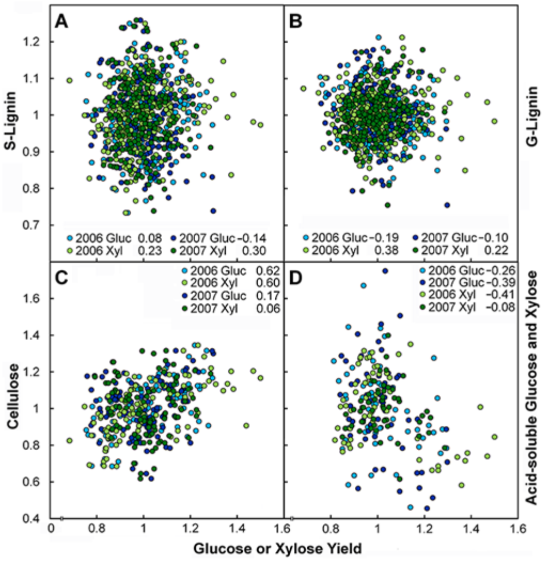

Figure 3. Comparison of lignin and cell-wall polysaccharide abundance with saccharification yield in the IBM population. Syringyl (**A**) and guaiacyl (**B**) lignin calculated from PyMBMS *m/z* ion abundances are compared to relative Glc and Xyl release in standardized saccharification yield assays conducted in two years. Syringyl lignin is the sum of the relative abundance of mass fragments *m/z* 154, *m/z* 167, *m/z* 168, and *m/z* 194; guaiacyl lignin calculated from PyMBMS *m/z* ion abundances *m/z* 124, *m/z* 137, *m/z* 138, and *m/z* 151. Cellulose content (mg/mg cell wall) (**C**), and cell-wall acid-soluble Glc and Xyl (mg/mg cell wall) (**D**) are compared to Glc and Xyl release in standardized saccharification yield assays. Year, sugar, and Pearson's correlation coefficients are inset. All values are mean-normalized relative abundances.

The significant effort to reduce or modify lignin in biomass crops is predicated on the assumption that it is the principal determinant of the recalcitrance of biomass to enzymatic digestion for biofuels production. We defined quantitative trait loci (QTL) in the IBM recombinant inbred maize population using pyrolysis molecular-beam mass spectrometry. Also with colleagues from NREL, we then applied a high-throughput saccharification assay, including a proxy for a steam explosion pretreatment, to determine the accessibility of biomass to a standard Ctec2 cocktail of enzymes to release glucose and xylose (Selig et al., 2011).

We analyzed cell wall structures and saccharification potential of developing stems and stover of maize B73, as control, and our 39-member *nir* mutant collection of cell wall-related mutants available at project start. Although large differences were found in lignin abundance and saccharification yield among the *nir* mutants, we observed no expected inverse correlation. When extended to the natural diversity of well-mapped populations represented in recombinant inbred lines, such as the IBM population, the PyMBMS and saccharification yield assays allow rapid identification of QTL of the genes relevant for biomass improvement. The NAM populations, which are 200 recombinant inbred lines each derived from crosses of B73 with 25 diverse inbreds,

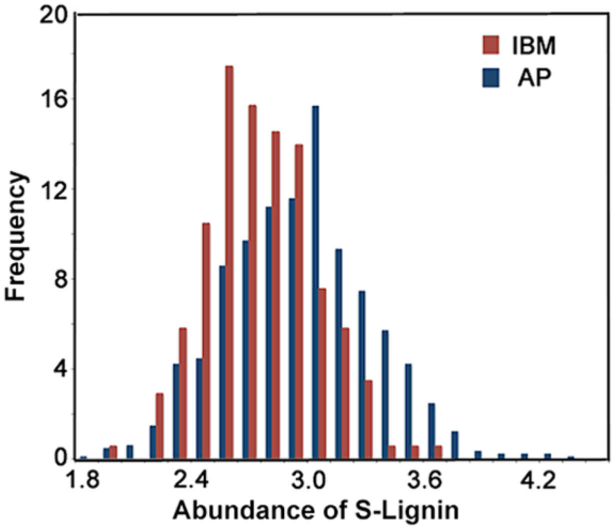

Figure 4. Comparison of frequency distribution of S-lignin abundance determined by PyMBMS in the AP vs. IBM populations.

includes many high-biomass tropical maize inbreds that capture a substantial amount of the existing genetic diversity of the species (http://www.panzea.org). These rich genetic resources greatly facilitate the discovery of biomass-relevant genes. Among five multi-year QTL for lignin abundance, two for 4-vinylphenol abundance, and four for glucose and/or xylose yield, not a single QTL for aromatic abundance and sugar yield was shared. These differences in QTL for lignin and saccharification traits are underscored by the lack of a correlation between lignin abundance and saccahrification across the populations (Fig. 3). Thus, although lignin composition and abundance are critical factors in the recalcitrance of biomass to enzymatic hydrolysis to sugars, our QTL analysis indicates it is not the only factor. Pretreatments above the phase transition temperature for lignin cause lignin to melt and redistribute within biomass. When the over-arching impact of lignin is removed by such pretreatments, other factors involving polysaccharide composition and architecture are revealed as additional determinants of recalcitrance.

A genome-wide association study (GWAS) for lignin abundance and sugar yield of the 282-member maize Association Panel provided candidate genes in the eleven QTL of the B73 and Mo17 parents, but showed that many other alleles impacting these traits exist among this broader pool of maize genetic diversity. This observation is underpinned by the finding that diversity of lignin abundance is greater in the Association Panel than represented by the transgressive segregation of the IBM population (Fig. 4).

A pervasive problem in QTL studies is the difficulty in identifying causative genes among the numbers of genes within large physical intervals. GWAS can be used to refine the search by taking advantage of natural genetic variation in SNPs and INDELs among a great many genotypes of a species. Taking account of kinship, GWAS gave several strong candidate genes with p values $\leq 10^{-8}$ within the QTL intervals displaying significant SNP/INDEL support and consistent with allelic variation between Mo17 and B73 (Penning et al. 2014c). For saccharification yield, we identify a Prefoldin gene and three genes of unknown function in QTL1, and a particularly large number in QTL3, with genes encoding three *Scarecrow9-like* and one *Scarecrow14-like* transcription factors, a Thaumatin-like protein, a Rhombin family protein (*BL4*), a Brassinolide-signaling kinase (*BSK1*), a powdery mildew resistance protein, a Cyclin D7, and four genes of unknown function (Fig. 5A). For lignin and 4-vinylphenol, candidate genes with p values $\leq 10^{-8}$ within the QTL intervals displaying SNP/INDEL support are an Inositol phosphoceramide synthase1 and a Harpin-related protein in QTL5, and two Leucine-rich repeat proteins and a protein of unknown function in QTL8 (Fig. 5B).

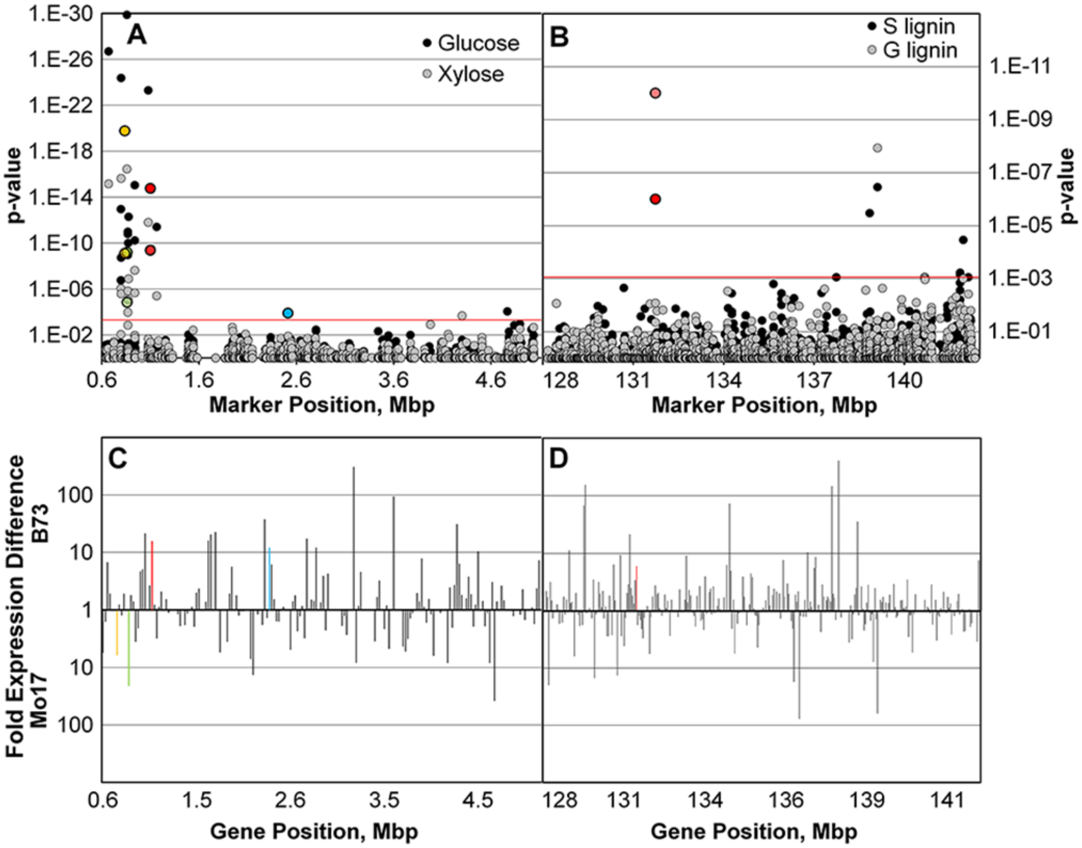

Figure 5. Marker and gene positions as identified by GWAS and differential expression, respectively, in QTL3 and QTL8. Physical map locations in mega basepairs (Mbp) (**A** and **B**) were identified using flanking markers, and a threshold value of 0.001 is indicated by red lines. Differential expression fold-change (**C** and **D**) on shown in logarithmic scale, with higher B73 fold-change upward, and higher Mo17 fold-change downward. GWAS predictions for QTL3 for Glc yield (**A**) are paired with expression fold-differences between B73 and Mo17 parents across the QTL interval (**C**). GWAS predictions for QTL8 for G- and S-lignin (**B**) are paired with expression fold-differences between B73 and Mo17 parents across the QTL interval (**D**). Markers flanking the QTL locations in cM were used to match the closest markers with physical map positions (www.maizegdb.org). Expression analysis (RNAseq) was performed on cDNA populations derived from developing internodes 4 and 5 of greenhouse-grown B73 and Mo17 parents 63 days after planting during peak secondary wall formation. Color codes for symbols in (**A**) and (**C**), all glucose-yield related, are as follows: orange, gene of unknown function; green, *scarecrow9-like* transcription factor (*SCR9-like*); red, *BR-signaling kinase1* (*BSK1*), and blue, *β-xylosidase* (*BXL1*). Dark red symbol (S lignin) and light red symbol (G-lignin) both indicate the same candidate gene of unknown function in (**B**) and (**D**).

Several significant GWAS locations within QTL but with no SNP/INDEL support in Mo17 and B73 indicates that the maize genome diversity holds many more trait-determining genes in the Association Panel than are present in the narrowed diversity represented in these two parents. For example, a xylosidase in QTL3 is a GWAS candidate gene, but has no polymorphisms between B73 and Mo17. While the activity of the β-xylosidase might contribute to a saccharification phenotype in the IBM population, it is a potential target for selective breeding only with other genotypes in the Association Panel. A cluster of *Scarecrow-like9* and *Scarecrow-like14* transcription factor genes provides exceptionally strong candidate genes emerging from the GWAS study (Penning et al. 2014c). In addition to these and genes associated with cell-wall metabolism, candidates include several other transcription factors associated with vascularization and fiber formation, and components of cellular signaling pathways.

In summary, GWAS primarily reveals genes in which SNPs might alter allele strength across a diverse population of genotypes, irrespective of transcript abundance, and differential expression of the two RIL parents provides candidates based on expression level, irrespective of polymorphisms that might affect activity. Genes with no polymorphisms might be expressed differently as an indirect effect, for example, of mutation in a *trans*-acting factor elsewhere in the genome. Integrating several genetic tools such as GWAS and differential expression has proven useful to resolve those genes for which expression levels contribute to phenotype from candidates for which SNP/INDEL differences indicate more discrete targets for breeding.

Additional References

Paschold A, Jia Y, Marcon C, Lund S, Larson NB, Yeh CT, Ossowski S, Lanz C, Nettleton D, Schnable PS, et al. (2012) Complementation contributes to transcriptome complexity in maize (*Zea mays* L.) hybrids relative to their inbred parents. Genome Res 22: 2445-2454

Selig MJ, Tucker MP, Law C, Doeppke C, Himmel ME, Decker SR (2011) High throughput determination of glucan and xylan fractions in lignocelluloses. Biotechnol Lett 33: 961-967

Springer NM, Ying K, Fu Y, Ji TM, Yeh CT, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, et al. (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLOS Genet 5: e1000734

Publications

Penning, B.W., Babcock, N.C., Klimek, J.F., Mosier, N.S., Springer, N.M., Thimmapuram, J., Weil, C.F., McCann, M.C., Carpita, N.C. (2014c) Transcriptome variation during secondary cell wall formation in maize parental inbred lines B73 and Mo17. In preparation.

Penning, B.W., Sykes, R.W., Babcock, N.C., Dugard, C.K., Held, M.A., Klimek, J.F., Shreve, J., Fowler, M., Gamblin, D., Ziebell, A., Davis, M., Decker, S.R., Filley, T.R., Mosier, N.S., Springer, N.M., Thimmapuram, J., Weil, C.F., McCann, M.C., Carpita, N.C. (2014b) Genes impacting rates of enzymatic digestion of sugars from maize biomass are independent of those for phenylpropanoid abundance. *In revision*

Penning, B.W., Sykes, R.W., Babcock, N.C., Dugard, C.K., Klimek, J.F., Gamblin, D., Davis, M., Filley, T.R., Mosier, N.S., Weil, C.F., McCann, M.C., Carpita, N.C. (2014a) Validation of PyMBMS as a high-throughput screen for lignin abundance in lignocellulosic biomass of grasses. *Bioenergy. Res., in press* [DOI 10.1007/s12155-014-9410-3]

Carpita, N.C., McCann, M.C. (2010) The maize mixed-linkage (1→3),(1→4)-□-D-glucan polysaccharide is synthesized at the Golgi membrane. *Plant Physiol.* **153**, 1362-1371

Schnable, P.S., et al. [158 authors] (2009) The B73 maize genome: complexity, diversity and dynamics. *Science* **326**, 1112-1115

Penning, B., Hunter, C.T., Tayengwa, R., Eveland, E., Dugard, C.K., Olek, A. Vermerris, W., Koch, K.E., McCarty, D.R., Davis, M., Thomas, S.R., McCann, M.C., Carpita,

N.C. (2009) Genetic resources for maize cell wall biology. *Plant Physiol.* **151,** 1703-1728

Invited lectures, seminars and other presentations given at conferences, educational or research institutions

March 2013	“*Genes impacting lignocellulose abundance and saccharification yield are different*”, symposium talk, annual meeting of the British Sustainable Bioenergy Centre, Crewe, U.K.
August 2012	“*Gene discovery in maize for secondary wall biogenesis*”, seminar at the National Renewable Energy Lab (NREL), Golden, CO
March 2012	“*Defining expression networks in maize for secondary wall biogenesis*”, plenary speaker, 55th Maize Genetics Meeting, Portland, OR
October 2011	“*Capturing the genetic diversity of maize for the improvement of energy grasses*”, symposium lecture at the annual SACNAS Convention, San Jose, CA
May 2011	“*Mechanisms of biosynthesis of cellulose, the mixed-linkage β-D-glucan, and other (1→4)-β-D-glycans*”, symposium lecture at the Brazilian Biochemistry and Molecular Biology Sympoisum, Foz do Iguaçu, Brazil
March 2011	“*Capturing the genetic diversity of maize for the improvement of energy grasses*”, banquet speaker, Atlantic Section of ASPB, College Park, MD
March 2011	“*Capturing the genetic diversity of maize for the improvement of energy grasses*”, keynote lecture, Midwest Section of ASPB, West Lafayette, IN
August 2010	“*Capturing the diversity of maize for the improvement of energy grasses*” 2nd Pan-American Congress on Plants and Bioenergy, São Pedro, Brazil.
May 2010	“*Improvement of grasses as lignocellulosic bioenergy crops*” 5th Frontiers in Bioenergy Conference, Purdue University
February 2010	“*Capturing the diversity of maize for the improvement of energy grasses*” Seminar in Plant Biology Colloquium Series, University of Minnesota
January 2010	“*Translational genomics for the improvement of switchgrass*” Plant and Animal Genome Meeting, San Diego, CA.
November 2009	“*Capturing the genetic diversity of maize for improvement of bioenergy grasses*.” XIII National Congress of Plant Molecular Biology and 6th Mexico-US, Guanajuato, Mexico.
September 2009	“*Translational Genomics for the improvement of switchgrass”*. DOE GTL PIs Meeting, Washington, DC.
February 2009	“*Maize: A genetic model for improvement of energy grasses*”, Seminar, Interdisciplinary Plant Biology Group, University of Missouri, Columbia, MO