Comparative Genomics in Switchgrass Using 61,585 High-Quality Expressed Sequence Tags

Christian M. Tobias,* Gautam Sarath, Paul Twigg, Erika Lindquist, Jasmyn Pangilinan, Bryan W. Penning, Kerry Barry, Maureen C. McCann, Nicholas C. Carpita, and Gerard R. Lazo

Abstract

The development of genomic resources for switchgrass (*Panicum virgatum* L.), a perennial NAD$^+$-malic enzyme type C_4 grass, is required to enable molecular breeding and biotechnological approaches for improving its value as a forage and bioenergy crop. Expressed sequence tag (EST) sequencing is one method that can quickly sample gene inventories and produce data suitable for marker development or analysis of tissue-specific patterns of expression. Toward this goal, three cDNA libraries from callus, crown, and seedling tissues of 'Kanlow' switchgrass were end-sequenced to generate a total of 61,585 high-quality ESTs from 36,565 separate clones. Seventy-three percent of the assembled consensus sequences could be aligned with the sorghum [*Sorghum bicolor* (L.) Moench] genome at a *E*-value of $<1 \times 10^{-20}$, indicating a high degree of similarity. Sixty-five percent of the ESTs matched with gene ontology molecular terms, and 3.3% of the sequences were matched with genes that play potential roles in cell-wall biogenesis. The representation in the three libraries of gene families known to be associated with C_4 photosynthesis, cellulose and β-glucan synthesis, phenylpropanoid biosynthesis, and peroxidase activity indicated likely roles for individual family members. Pairwise comparisons of synonymous codon substitutions were used to assess genome sequence diversity and indicated an overall similarity between the two genome copies present in the tetraploid. Identification of EST–simple sequence repeat markers and amplification on two individual parents of a mapping population yielded an average of 2.18 amplicons per individual, and 35% of the markers produced fragment length polymorphisms.

Published in The Plant Genome 1:111–124. Published 21 Nov. 2008.
doi: 10.3835/plantgenome2008.08.0003

677 S. Segoe Rd., Madison, WI 53711 USA
An open-access publication

SWITCHGRASS (*PANICUM VIRGATUM* L.) is a warm-season C_4 perennial grass, native to North America, and is widely grown for summer grazing and soil conservation. It is a self-incompatible, allogamous species with genome constitution varying from diploid to decaploid (Nielsen, 1944; Talbert et al., 1983). Switchgrass cultivars and ecotypes are classified as either lowland or upland types (Moser and Vogel, 1995). Lowland types are usually tetraploid with genetic composition of $2n = 4x = 36$ with a DNA content of approximately 3 pg. Most of the upland types are either hexaploid ($2n = 6x = 54$) or octaploid ($2n = 8x = 72$) with octaploid DNA content of 5.9 to 6.2 pg $2C^{-1}$ (Hopkins et al., 1996; Lu et al., 1998).

C.M. Tobias and G.R. Lazo, USDA-ARS, Western Regional Research Center, Genomics and Gene Discovery Unit, 800 Buchanan St., Albany, CA; G. Sarath, USDA-ARS, Grain, Forage, and Bioenergy Research Unit, Keim Hall, East Campus, Univ. of Nebraska, Lincoln, NE, 68583; P. Twigg, Dep. of Biology, Univ. of Nebraska, 905 W. 25th St., Bruner Hall, Kearney, NE 68849; E. Lindquist, J. Pangilinan, and K. Barry, Lawrence Berkeley National Lab., 1 Cyclotron Rd., Mail Stop PGF, Berkeley, CA 94720; B.W. Penning and M.C. McCann, Dep. of Biological Sciences, Purdue Univ., West Lafayette, IN 47907; N.C. Carpita, Dep. of Botany & Plant Pathology, Purdue Univ., West Lafayette, IN 47907. Sequencing conducted at and supported through the Community Sequencing Program, DOE-Joint Genome Institute CSP-776898. All sequences have been submitted to the dbEST division of Genbank Accessions FE597478–FE659062. Received 18 Aug. 2008. *Corresponding author (christian.tobias@ars.usda.gov).

Abbreviations: 2,4-D, 2,4-dichlorophenoxyacetic acid; AlaAT, alanine aminotransferase; AspAT, aspartate aminotransferase; CAD, cinnamyl alcohol dehydrogenase; CAH, carbonic anhydrase; C3H, cinnamate 3-hydroxylase; C4H, cinnamate 4-hydroxylase; EST, expressed sequence tag; F5H, ferulate 5-hydroxylase; GO, gene ontology; Ka, synonymous substitution rate; Ks, nonsynonymous substitution rate; NAD$^+$-me, NAD$^+$-malic enzyme; NADP$^+$-me, NADP$^+$-malic enzyme; NCBI, National Center for Biotechnology Information; PAL, phenylalanine ammonia lyase; PEPC, phosphoenol pyruvate carboxylase; PPCK, PEPC kinase; PPDK, pyruvate orthophosphate dikinase; RFLP, restriction fragment length polymorphism; SSR, simple sequence repeat.

Lowland types are tall and robust and possess a bunch-type growth habit. The upland types are shorter and finer, possess longer rhizomes, and spread more readily. Several cultivars from both lowland and upland ecotypes have been developed as forages.

Because switchgrass is the focus of continuing intensive efforts to produce renewable energy economically from lignocellulosic biomass, many steps involved in its production, collection, pretreatment, and processing are undergoing optimization. Genetic improvements are likely to come directly from manipulation of qualities that also will improve its value as forage, such as plant architecture for improved photosynthetic capacity, stress tolerance, optimized cell-wall structure, and digestibility. Other characteristics, such as palatability or alkaloid content, are not important for energy crops. The relative values of these traits are extremely difficult to predict because of advances in conversion technology, cropping systems, valuation of coproducts, and other variables, but high-yielding and better-quality germplasm are being developed from adapted cultivars and ecotypes of switchgrass using conventional selection and breeding techniques (Moser and Vogel, 1995).

Genetic progress may also come through marker-assisted breeding techniques and biotechnological approaches. Molecular mapping, genetic fingerprinting, determination of population structure, and diversity analysis are now routine in many crops but are just beginning to be used in switchgrass breeding efforts. To date, studies in switchgrass have been limited to chloroplast DNA variation (Hultquist et al., 1996), randomly amplified polymorphic DNA (RAPD) diversity among different populations (Gunter et al., 1996), and an investigation into genome organization using restriction fragment length polymorphism (RFLP) markers (Missaoui et al., 2005). Reproducible marker systems amenable to low-cost, high-throughput genotyping that are appropriate for polyploids are needed in switchgrass. However, development of these markers depends on availability of sequence data in large quantities. The objective of this work is to generate and sequence large numbers of expressed sequence tags (ESTs) to facilitate marker development and mining of gene sequence data. Expressed sequence tags derived from morphologically, phenotypically, or genetically distinct sources of cDNA provide inventories of genes and alleles that can subsequently be analyzed in a myriad of ways. These inventories can include simple sequence repeat (EST–SSR) markers, which have become a marker class of choice because they are highly polymorphic, codominant, abundant in genomes, and reproducible and have high rates of transferability across species (Saha et al., 2004; Thiel et al., 2003). These inventories can also contain functional orthologs of genes with recognized functions in model species and, in particular, other members of the *Poaceae*.

In this article, we describe the production and subsequent analysis of 61,585 EST sequences from three different cDNA libraries. These sequences were assembled with existing data to produce minimally redundant consensus sequences that were used for genomewide comparisons to sorghum [*Sorghum bicolor* (L.) Moench]. The representation of several gene families that are directly involved in cell-wall biogenesis and C_4 photosynthesis has been analyzed in detail to provide insight on functional differentiation among individual family members, and several gene families critical for photosynthetic efficiency, stress tolerance, and cell-wall structure are described. Finally, an EST-SSR marker set has been developed that will be useful for future genetic analyses.

MATERIALS AND METHODS

Plant Growth Conditions and cDNA Library Construction

'Kanlow' switchgrass plants were raised from seed in a greenhouse at the University of Nebraska, Lincoln, under a 16/8-h day/night (~26–30°C/~22–26°C) growth regimen, using supplemental lighting from halide lamps (200 mol photons m^{-2} s^{-1}). Soil mixture consisted of 40% Canadian peat, 40% coarse vermiculite, 15% masonry sand, and 5% screened topsoil, amended with 4.4 kg m^{-3} Waukesha fine lime. Plants were watered biweekly with a nutrient solution containing 200 mg kg^{-1} N and with tap water as needed otherwise (Sarath et al., 2007).

Callus was grown from mature caryopsis that had been surface sterilized and placed on callus initiation media with maltose as a carbon source (Somleva et al., 2002). Cultures were maintained in the dark at 28°C. The tissue contained a variety of callus morphologies. Seedling tissue was collected from 4-d-old plants that were germinated in a growth chamber on filter paper. Crown tissue was collected from 6- to 8-wk-old plants that had started to tiller and that were cleaned of roots and tillers. Plant tissues were flash-frozen in liquid nitrogen and stored at −80°C until used for RNA isolation.

Total RNA was extracted from crown, seedling, and callus tissue using the Concert Plant RNA reagent (Invitrogen, Carlsbad, CA). Messenger RNA was purified using the FastTrack 2.0 mRNA isolation system. First strand cDNA synthesis was primed with a *Not*I-oligo(T) adaptor primer followed by second strand synthesis using the Superscript Plasmid System for cDNA Synthesis (Invitrogen). The resulting cDNA was ligated to *Sal*I adapters, digested with *Not*I, size selected, and cloned into the pSPORT1 cloning vector before transformation of Ultramax DH5αFT chemically competent *Escherichia coli* (Invitrogen).

Automated DNA Sequencing

The libraries were plated and individual colonies were robotically picked and arrayed into 384 well plates for long-term storage. Sequencing was performed at the USDOE's Joint Genome Institute using their standard dye-terminator sequencing protocols. Individual sequences from each library clone were collected in both the forward and reverse directions, and pairing information was used for subsequent sequence processing steps.

Raw EST sequences were trimmed for vector and adaptor/linker sequences using a tool that searches for common sequence patterns at the ends of the sequences. Insertless clones were identified and removed using the following criteria: 200 bases or more of vector sequence obtained from the 5′ end of the EST or less than 200 bases of vector sequence followed by less than 100 bases of nonvector sequence.

Expressed sequence tags were then trimmed for quality using a sliding window trimmer (window = 11 bases). Once the average quality score in the window was below the threshold (Q15), the EST was split and the longest remaining sequence segment was retained as the trimmed EST. Expressed sequence tag sequences with less than 100 bases of high-quality sequence were screened from the data set. Expressed sequence tags were evaluated for the presence of a polyA or T tail near the ends; if found, the EST was flagged and the polyA/T tail was trimmed off. Additionally, sister ESTs or end pair reads were categorized as follows: if one EST was insertless or a contaminant, by default, the second sister was then categorized with the same category. Alternatively, each sister read was treated separately for the categories: low complexity and low quality.

Sequence Assembly

Sequences were then compared against selected databases containing vector, *E. coli*, bacteriophage λ, chloroplast, mitochondria, or rRNA sequences using the BlastN algorithm (Altschul et al., 1990). Those that aligned with an *E*-value < 10^{-20} were eliminated. Databases included Univec (National Center for Biotechnology Information [NCBI]), *E. coli* (GenBank), mitochondria, plastid (GenBank), and rRNA (GenBank). Expressed sequence tags were then clustered using malign, a kmer-based alignment tool that clusters ESTs on the basis of sequence overlap (kmer = 16, seed length requirement = 32, alignment ID ≥ 98%). Clusters of ESTs were further merged on the basis of sister reads using double linkage. Double linkage requires the presence of two or more matching sister ESTs in each cluster to be merged. Expressed sequence tag clusters were then assembled using CAP3 defaults to form consensus sequences (Huang and Madan, 1999). Clusters may have more than one consensus sequence for various reasons (clone has long insert, clones are splice variants, or consensus sequences are erroneously not assembled). Cluster singlets are clusters of one EST, whereas CAP3 singlets are ESTs that joined a cluster with other ESTs but on assembly by CAP3 a single EST was used to make a separate consensus sequence from the others. Expressed sequence tags from each separate cDNA library were clustered and assembled separately, and subsequently, the entire set of ESTs for all three cDNA libraries and existing switchgrass ESTs available from the dbEST division of Genbank were assembled.

Sequence Analysis

The blastx algorithm was used to match ESTs to EBI UniProt Release 4.2 (Swiss-Prot Rel. 46.2, TrEMBL Rel. 29.2) (Apweiler et al., 2004). The UniProt gene associations to gene ontology (GO) terms were obtained from the Gene Ontology Consortium Website (http://www.geneontology.org) using provided association tables. The gene ontology database available (March 2005) was also used as reference, and data were uploaded and parsed from a MySQL database containing the information presented for GO molecular, biological, and cellular category levels 1 to 4, where each level represents a greater level of detail. Expressed sequence tags that were not annotated with GO terms were further matched to NCBI nonredundant database entries. Those entries that did not match this pool were further compared to NCBI dbEST entries.

Assembled EST sequences were compared to cell-wall gene families from rice (*Oryza sativa* L.) and *Arabidopsis* annotated at the Purdue University Cell Wall Genomics Website (http://cellwall.genomics.purdue.edu). Sequences were compared with tblastn against a local database and hits with *E*-values < 1×10^{-20} and scores of greater than 100 were kept.

Cluster and TreeView (Eisen et al., 1998) were used for visualizing EST representation. Before clustering the data was centered and normalized such that the mean values in each row and column were 0. Data were then analyzed by average linkage clustering while weighting both the genes and arrays (options–cutoff 0.1–exponent 1.0–Euclidian_distance).

Sorghum genome assembly Sbi1.4 (http://www.phytozome.net) was used for alignment with switchgrass EST data. Genome data were used to create a blast database and then queried with EST sequence data. The output was parsed to identify significant hits with an *E*-value of < 1×10^{-20}. Values were then sorted and binned into 1-Mb regions for creating the data files for the figure. Expressed sequence tags and consensus sequences were counted multiple times when they had multiple hits to regions separated by greater than 50 kb.

EST–SSR Screening

Simple sequence repeats containing sequences and primer sequences were taken from the output of BatchPrimer3 (You et al., 2008; http://wheat.pw.usda.gov/demos/BatchPrimer3/). Forward primers were appended at the 5′ end with M13 to allow indirect labeling reactions (Boutin-Ganache et al., 2001). Reverse primers were appended with a variable number of bases at the 5′ end to match the consensus GTTTV (Brownstein et al., 1996). This sequence is found to promote nontemplated (A) addition and facilitated subsequent genotyping. Products were amplified in 5- to 10-μL polymerase chain reactions, polyethylene glycol 8000 precipitated, and sized on an ABI3730xl using homemade markers labeled with PET (Applied Biosystems, Foster City, CA; Symonds and Lloyd, 2004). Markers were genotyped with Genemapper v. 3.7 (Applied Biosystems).

Duplicated Gene Families

The consensus sequences were compared against the sorghum genome's likely coding sequences, which were obtained from http://www.phytozome.net at the USDOE

(Sbi1.4). The best hits were identified, and the corresponding sorghum peptides were used as guide sequences for input to GeneWise v. 2.2 (Birney et al., 2004) with the parameters (–subs 0.01–indel 0.01–quiet–both–alg 333–pep–cdna). The largest predicted open reading frame was checked for the presence of in-frame stop codons and used for codon-based comparisons. A total of 13,710 potential coding sequences and corresponding conceptual translations were produced from the dataset and used for codon-based alignments between consensus sequences to measure relative sequence divergence related to recent and more distant gene duplication events. Sequences with aligned length greater than 300 nucleotides and with at least 40% nucleotide identity were considered to be pairs. These pairs were codon-aligned using ClustalW's default settings (Thompson et al., 1994), and pairwise rates of substitutions at synonymous sites were measured using the YN00 model in phylogenetic analysis by maximum likelihood (PAML) (Yang and Nielsen, 2000). Gaps and ambiguous codons were not included in the analyses.

RESULTS

Sequence Assembly and Composition

Sequence sampling of gene libraries from diverse plant tissues and isolated under a variety of conditions can increase the representation of individual genes as any one gene may be expressed at significant levels in a small subset of tissues and treatments. In an earlier EST sequencing effort, 11,990 ESTs from a total of four different libraries were sampled (Tobias et al., 2005). This sampling resulted in 7810 tentatively unique genes after assembly. For this study, much deeper sequencing of two of these libraries and a third library from the cultivar Kanlow was conducted.

Sequencing was performed at the Joint Genome Institute. In total, 13,440 clones were arrayed from each of the three cDNA libraries created from callus tissue (CBYX), crown tissue (CBYY), and whole seedling tissue (CBYZ). Sequencing was attempted in the forward and reverse directions of each clone. The overall pass rate for sequencing each library was 75.29, 73.37, and 80.45% for CBYX, CBYY, and CBYZ, respectively, and 61,585 sequences along with the corresponding trace files were deposited in the dbEST division of Genbank as accessions FE597478 to FE659062. These reads were then clustered by library; the results of this clustering showing library overlap are given in Fig. 1. When clustered together with existing EST sequences, a nonredundant sequence assembly was produced that included a total of 12,829 clusters, 7532 singletons, and 27,329 different consensus sequences. Several consensus sequences were possible from a single cluster due to nonoverlapping reads from the same clone(s), possible splice variants, chimerism, and the inclusion of paralogous or homeologous sequences within a cluster. Together, these observations explain the approximately 2.1-fold difference in numbers of clusters and consensus sequences. These assembled consensus sequences are available (see Supplementary Fig. S1).

Expressed sequence tag clusters consisting of greater than five individual ESTs and representing all available EST data were clustered by library representation using hierarchical clustering available in Cluster 3.0 (Eisen et al., 1998). The expression by library of each cluster for the three libraries that were sequenced and the additional libraries from previous work were represented graphically and provide an overall view of EST representation by library, showing that each library contained unique sets of transcripts (Fig. 1a). Even taking into account the availability of far fewer sequences from the leaf library, the gene spectrum represented was strikingly different from the seedling library, although both contained photosynthetic tissue. A slightly different representation of the EST assembly data in Fig. 1b shows library overlap as numbers of consensus sequences containing ESTs from each library as well as singletons. The clustering data was plotted on the basis of the number of reads per cluster and established a log relationship between frequency of reads and the number of clusters (Fig. 2a). The largest cluster contained 1090 ESTs and was annotated as a chloroplastic chlorophyll a/b binding protein.

Additional mapping of all available ESTs to levels 1 to 4 molecular, cellular, and biological categories of GO terms are available from Tobias (n.d.). This annotation resulted in a total of 48,341 (65.9%) EST sequences that could be mapped to GO molecular terms, 46,065 (62.8%) that could be mapped to GO biological terms, and 42,143 (57.5%) that could be mapped to GO cellular terms; in total, encompassing 51,293 ESTs. Of the remaining EST sequences, 14,833 matched existing sequences in the *Poaceae* at a *E*-value $< 1 \times 10^{-20}$.

EST–SSR Marker Development

For the purposes of developing EST–SSR markers, the consensus sequences were screened for SSRs using BatchPrimer 3.0 to identify di- and trinucleotide microsatellite-containing coding sequences. Figure 2b shows the relative abundance and lengths of these microsatellite classes.

The most abundant trinucleotide classes were GC-rich; consistent with findings in other grasses and reflecting codon bias. Of the 2817 di- and trinucleotide repeats greater than 15 nucleotides identified in the consensus sequences, 698 matched the set of 530 conserved grass EST–SSRs found in barley (*Hordeum vulgare* L.), maize (*Zea mays* L.), rice, sorghum, and wheat (*Triticum aestivum* L.) (Kantety et al., 2002), indicating these sequences were well represented in switchgrass. A total of 1780 primer pairs were designed to the EST–SSR using an optimal product size of 200 bp and calculated primer melting point (T_m) of 60°C. These primers were then screened against two parents of a mapping population (Tobias et al.. unpublished data). The results of this screening process give an indication of the ability of the markers to discriminate homeologous loci in polyploid switchgrass, and have been used to determine which markers to screen against the full population. A total of 830 (46%) of the primer pairs reliably amplified and the mean number of

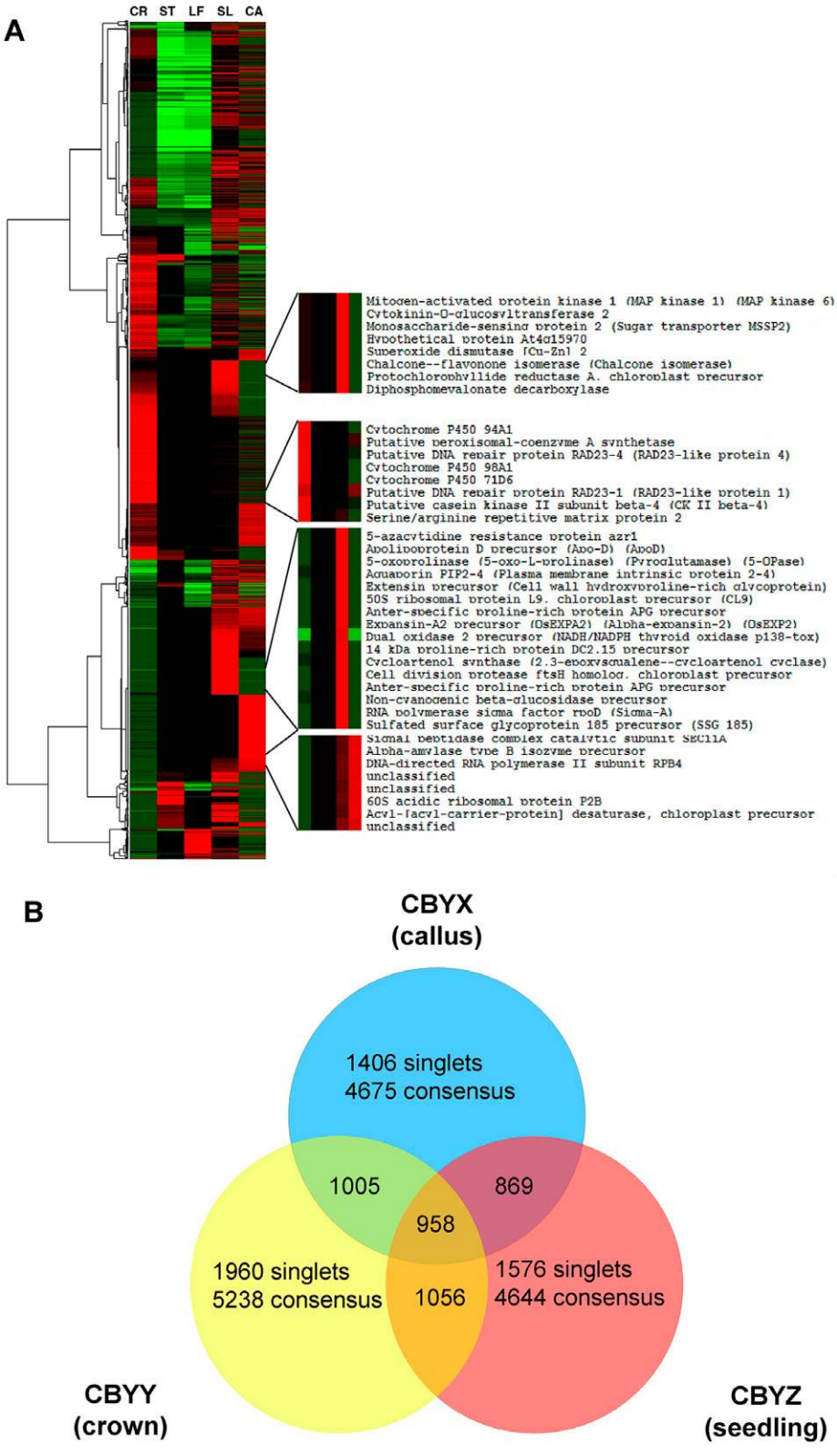

Figure 1. Transcript abundance by library. (A) Cluster analysis was performed on filtered expressed sequence tag (EST) clusters to consider only those clusters represented by greater than five ESTs. Individual columns represent each library included in the sequence assembly. A dendrogram on the left indicates similarity of EST expression profile based on average linkage clustering. The color intensity scale indicates the relative quantity of ESTs in each cluster, with green representing underrepresentation and red representing relative abundance. Black indicates an intermediate values and not underrepresentation. Crown library (CR), stem library (ST), leaf library (LF), seedling library (SD), and callus library (CA) are indicated. Descriptions shown in detail regions are one-line descriptions of best hit to Swissprot database when queried with Blastx (Altschul et al., 1990). Clusters containing more than five ESTs but with no significant hits to the database are listed as unclassified. (B) Library overlap was evaluated for CBYX-Z libraries and the numbers of consensus sequences as well as ESTs that remained unassembled from each library are shown in a Venn diagram.

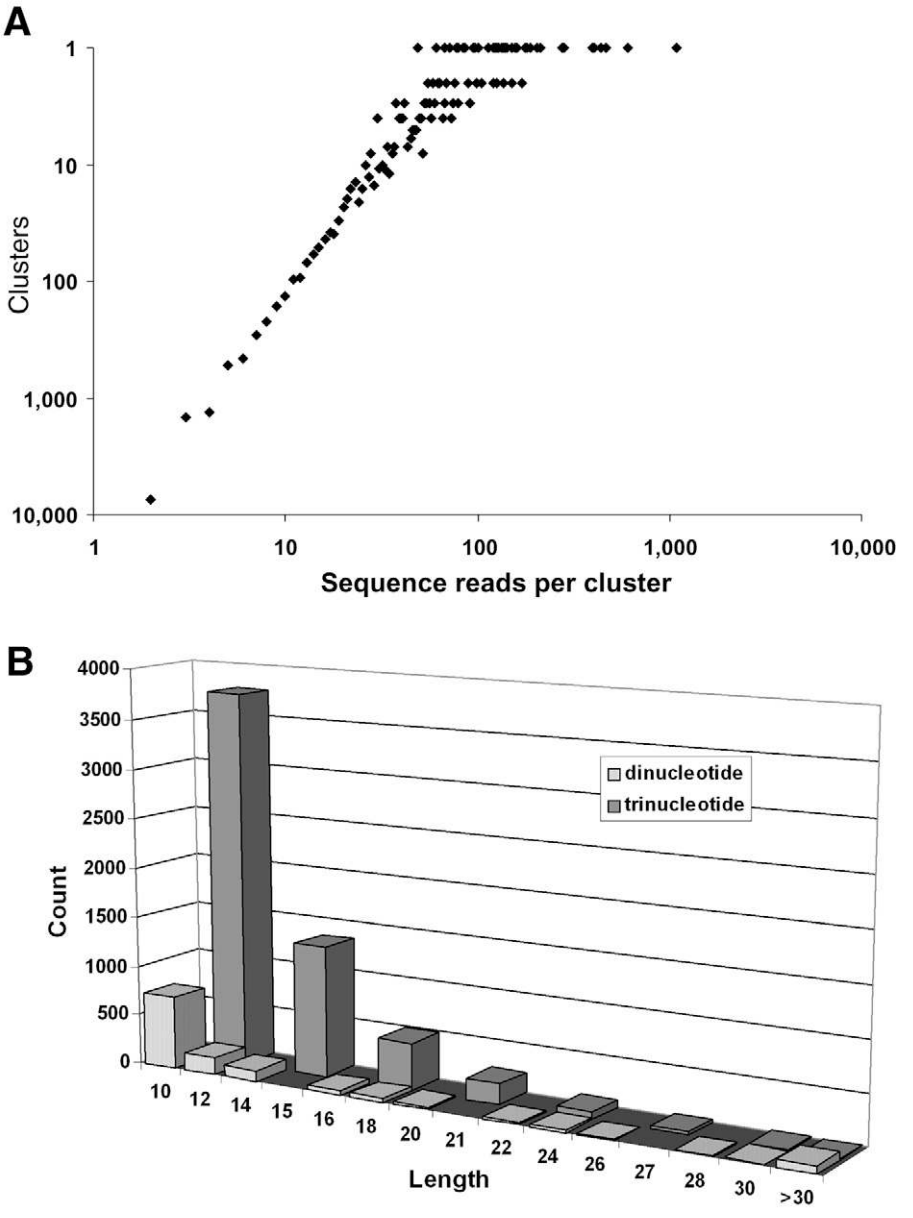

Figure 2. Expressed sequence tag (EST) representation in clusters and numbers of simple sequence repeats (SSRs) found in consensus sequences. The numbers of ESTs per cluster were plotted on a logarithmic scale and are shown in (A). Consensus sequences were screened against BatchPrimer3 (You et al., 2008) to identify SSRs. (B) Di- and trinucleotide repeats were analyzed and their frequency plotted as a function of repeat number.

amplicons per individual was 2.18. A total of 298 were found to be polymorphic between the two parents. The informativeness of the markers differed greatly in the population. One primer pair produced 8 products (4 in each parent), 1 pair produced 7, 2 pairs produced 6, and 10 primer pairs produced 5 products. The EST–SSRs that amplified reliably are cataloged along with primer sequences, predicted product size, and best match to the sorghum genome (see Supplementary Fig. S2)

Mapping Sequences to Sorghum Genome

Sorghum was used as the basis for further comparison as it is currently the only member of the *Panicoideae* subfamily with a sequenced genome and shared a common ancestry with switchgrass between 15 and 25 million years ago (Grass Phylogeny Working Group, 2001). We took the non-redundant sequence assemblies as well as the ESTs and assigned them to one or more positions on the sorghum genome in 1-Mb intervals based on blast similarity scores. The number of ESTs and consensus sequences were then plotted along the chromosomes and major superclusters that were not assembled into chromosomes. Clear biases against the central portion of each chromosome in the pericentric regions are observed (Fig. 3). This bias extended to the superclusters that were not assembled into chromosomes but for which there is sparse EST representation. The EST–SSRs were also localized on the sorghum genome and were found to be evenly represented across the 10 chromosomes. A total of 19,898 or 73% of the consensus sequences matched genome sequences with an *E*-value of $<1 \times 10^{-20}$. A total of 3641 of these aligned best to the same or overlapping regions.

Duplicated Gene Families

Switchgrass open reading frames and peptides predicted by Genewise using guide peptide sequences from sorghum are provided (Supplementary Fig. S3, Supplementary Fig. S4). The coding sequences are not full length in many cases and do not always begin with an initiation codon. These peptides were then used for pairwise estimates of synonymous substitution rates.

For comparison, the same comparisons were used for sorghum coding sequences and for paralogous or orthologous pairs between sorghum and switchgrass (Fig. 4a). In switchgrass, an initial peak representing sequence pairs with a nonsynonymous substitution rate (Ks) value of 0 but differing at other nonsynonymous sites is evident. The main peak shows a slight shift from 0.03 to 0.07 relative to sorghum This shift could result from greater allelic heterozygosity due to sampling multiple genotypes, the outcrossing nature of switchgrass, its polyploidy, or to differences in the rate of the "molecular clock" that might cause sequences to diverge more rapidly in switchgrass. Ancient polyploidization has been detected in EST data, but duplication in the more recent past may not be

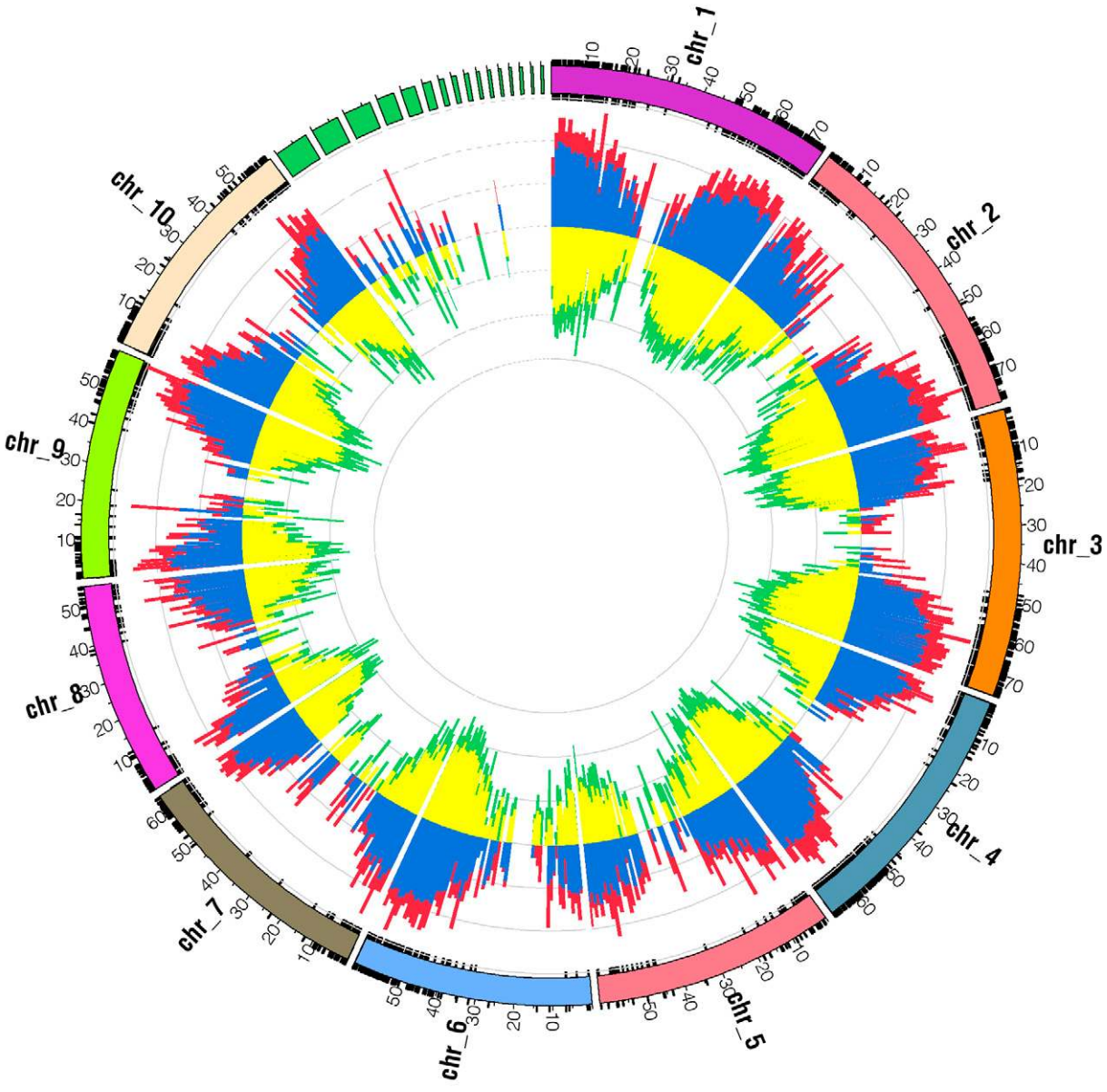

Figure 3. Switchgrass expressed sequence tags (ESTs) mapped to the sorghum genome. The number of ESTs that produced significant alignments to the indicated sorghum chromosome are plotted for each megabase interval. Totals were log transformed to improve visibility, and each radial axis line represents one log interval. Individual sorghum chromosomes (multicolored) and additional superclusters (green) that did not assemble and which represent over 95% of the available sorghum sequence in total are included. The EST frequency (red) and consensus sequence frequency (blue) on the plus strand are shown oriented in the outward direction, while the EST frequency (green) and consensus sequence frequency (yellow) on the minus strand are shown oriented in the inward direction. Positions of EST–SSRs are indicated by black ticks on the surface of the ideogram. Diagram was prepared using Circos (http://mkweb.bcgsc.ca/circos/).

resolved because of low sequence divergence. Ratios of synonymous to nonsynonymous codon substitution in a large majority of sequence pairs, averaged over all sites, was less than 1, supporting the conclusion that negative or purifying selection is acting at most loci (Fig. 4b).

Cell Wall–Related Sequences

Genomewise analysis of sequences related to the biogenesis of the cell wall has been undertaken for several model plant systems (http://cellwall.genomics.purdue.edu). Using EST sequence similarity to previously annotated *Arabidopsis* and rice genes, we focused specifically on cellulose and β-glucan biosynthesis, phenylpropanoid biosynthesis, and peroxidases, which were well represented in the data and are involved directly in determining cell-wall structure. Results of this analysis include previously reported ESTs from a stem and leaf library (Table 1).

Genes in the cellulose synthase (*CesA*) super family, including true CesAs and several groups of cellulose-synthase-like (*Csl*) genes, were detected in moderate abundance. From relative numbers of reads, transcripts of *CesA*s were more abundant than the transcripts from *Csl* genes (Table 1). The three *CesA*s associated with primary wall synthesis, as well as other minor *CesA*s, were more highly represented than the *CesA*s associated with secondary wall formation, with the seedling library

having highest representation. *CslD* and *CslF* family members showed higher abundance in crown tissue, whereas CslE family members showed higher abundance in seedling tissue.

Gene families involved in the phenylpropanoid pathway and, specifically, sequences in the monolignol pathway were analyzed and subcategorized based on predicted role. These families were relatively underrepresented in the callus library but were well represented in crown and seedling libraries. The most abundant categories included the cytochrome P450 hydroxylases *coumarate 3-hydroxylase* (*C3H*), *cinnamate 4-hydroxylase* (*C4H*), and *ferulate 5-hydroxylase* (*F5H*). These were followed by *phenylalanine ammonia lyase* (*PAL*). All of these groups had their highest percentage representation in the stem library (Table 1).

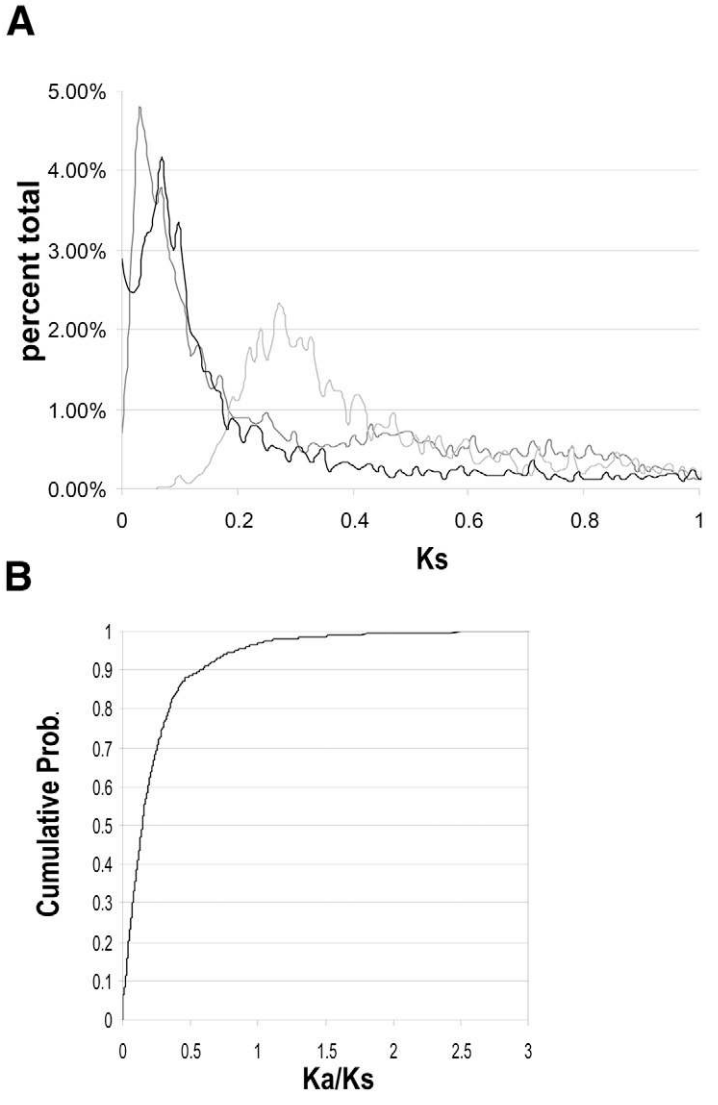

Figure 4. Synonymous substitutions in duplicated genes. Similar pairs of genes were codon aligned and synonymous substitutions (Ks) rates are shown in (A). Rates were plotted as a percentage of all pairs analyzed. A total of 6077 comparisons were made for switchgrass (black line). A total of 12,966 comparisons were made for sorghum (dark-gray line). A total of 4954 switchgrass–sorghum comparisons were made (light-gray line). The cumulative distribution of the nonsynonymous to synonymous substitutions (Ka/Ks) is shown in (B). Values greater than 1 are used as evidence for positive selection acting on one of the gene pairs.

One consensus sequence represented by ESTs from stem tissue had 88% protein identity to sorghum *bmr18* (Bout and Vermerris, 2003) and maize *bm3* (Collazo et al., 1992) caffeic acid/5-hydroxyferulic acid *O*-methyltransferase (COMT) genes. A distinctly different partial sequence represented by EST in crown tissue showed 85 and 90% protein identity to the same sorghum and maize COMT genes. Three distinct switchgrass consensus sequences belonged to a monocot-specific subfamily that shared 94 to 95% protein identity with maize *bm1* cinnamyl alcohol dehydrogenase (CAD) and cloned sorghum CAD sequences. Ten other specific CAD-like sequences were most closely related to identified subfamilies that lack conserved amino acids believed to determine specificity for aromatic alcohols.

We examined the peroxidase gene family in greater detail because of its skewed distribution in the libraries and large number of isoforms of the class III plant-specific peroxidases in both *Arabidopsis* (73 genes) and rice (138 genes). Our analysis detected 23 different switchgrass class III peroxidases encoding products predicted to be full length. These were given individual designations, and the predicted peptides were aligned with representative members of the rice family of peroxidases as well as one representative peroxidase from fern, moss, and liverwort. A dendrogram shows the relationships of the switchgrass peroxidases, including members in six of the eight defined class III peroxidase groups in rice (Fig. 5). PviPrx07 was the only sequence responsible for the abundance of peroxidase ESTs sampled from the callus library. It possessed an N-terminal signal sequence and was predicted to localize to either the vacuole or cell wall by PSORT. Another isoform (PviPrx14) was most closely related to and matched the consensus motif found in peroxidase group V.1, a monocot-specific grouping of unknown function(s) (Passardi et al., 2004).

C_4 Photosynthesis

C_4 photosynthesis contributes to increased water use efficiency and reduced photorespiration. Three different subtypes are defined based on the major decarboxylation enzyme in the bundle sheath; either $NADP^+$-malic enzyme, NAD^+-malic enzyme, or phosphoenol pyruvate carboxykinase. We examined genes associated with C_4 photosynthesis in the EST libraries, their differential library representation, and evidence of specific features that may be associated with C_4 activity of the NAD^+-me subtype present in switchgrass. In maize and sorghum ($NADP^+$-me C_4), the carbonic anhydrase (*cah*) gene family contains C_4 members with duplicated catalytic domains (Wyrich et al., 1998). However, in switchgrass, we only detected genes containing a single catalytic domain. The most abundantly expressed *cah* consensus sequences each contained 40 to 60 ESTs and were nearly exclusive to the seedling (green tissue)

library. We estimated that there were five different genes present based on the number of consensus sequences and analysis of the 5′ untranslated regions.

Most NAD$^+$-me plants are thought to utilize a mitochondrial malic-enzyme. In several studies on plant mitochondria, this enzyme has been demonstrated to have an octomeric structure consisting of distinct α and β subunits, but the subunit composition of the NAD$^+$-me involved in C_4 metabolism is not well characterized. In NADP$^+$-me plants, this function is performed by a chloroplast enzyme consisting of identical subunits. The switchgrass consensus sequences included nine members composed of a total of 23 ESTs that were most closely related to α and β subunits of plant mitochondrial NAD$^+$-me's. Although abundantly recorded in the seedling library, no trend based on EST representation allowed determination of which among these nine were involved in C_4 photosynthesis. Seven additional consensus sequences were related but were not likely involved in C_4 pathways. The consensus sequences also contained one full-length member representing 10 ESTs from the seedling library that was 87% identical at the protein level to the maize chloroplast NADP$^+$-me.

Plastid, mitochondrial, and cytosolic isoforms of aspartate aminotransferase (AspAT) have been described in the closely related C4 NAD$^+$-me plant *Panicum mileaceum*. The major C_4 active isoforms are present in the mitochondria of bundle sheath and the cytosol of mesophyll cells functioning in nitrogen metabolism and for shuttling reduced carbon between cell types (Taniguchi et al., 1995). Closely corresponding sequences were identified in switchgrass. Two mAspAT sequences were 96% identical at the protein level to the *P. mileaceum* mitochondrial isoform. Two additional cAspAT sequences were 95% identical to the cytoplasmic isoform. The cAspAT sequences were also the most highly represented in the seedling library. Four related sequences were 95% identical at the protein level to the plastidic isoform of *P. miliaceum* AspAT, which is present at much lower levels and not believed to play a major role in C_4 metabolism.

Alanine aminotransferase (AlaAT) also serves to shuttle C_3 carbons between the cytosol of the bundle sheath and mesophyll cells. Two consensus sequences in switchgrass comprising 146 ESTs, the large majority of which were derived from the seedling cDNA library, were 97% identical at the protein level to the light regulated AlaAT-2 of *P. mileaceum* and were predicted to be cytoplasmic (Son and Sugiyama, 1992). These also were 91% identical to a maize hypoxia-induced AlaAT (Muench and Good, 1994).

Three closely related sequences similar to phosphoenol pyruvate carboxylase (PEPC) were overrepresented in the seedling library. These displayed 89, 87, and 82% protein identity to the C_4 isoforms of PEPC in sorghum. Pyruvate orthophosphate dikinase (PPDK), which converts ATP + pyruvate to phosphoenol pyruvate in mesophyll cells, displayed 89% protein identity to PPDK from the closely related genus *Echinochloa frumentacea* and was highly represented in the seedling library with two related consensus sequences representing a total of 206 ESTs. A total of 11 ESTs closely matched PEPC kinase (PPCK) from maize and sorghum. However, none of these were derived from the seedling library. Six were from preexisting sequences accessed from a leaf library derived from more mature leaf tissue.

Table 1. Selected gene representation in sequences from all available switchgrass expressed sequence tags (ESTs). Values are numbers of ESTs in each library. Genbank accession numbers of sequences used to prepare this table are listed in Supplementary Fig. S5.

Family description	Callus	Crown	Seedling	Leaf	Stem	Total no. EST	Total no. consensus
Phenylpropanoid biosynthesis							
4CL	19	26	11	0	8	64	16
C3H F5H & C4H	30	80	28	8	7	153	68
CAD	3	33	16	5	7	64	14
CCoAOMT	4	15	5		4	28	9
CCR	6	44	37	2	3	92	32
COMT	0	24	37	2	6	69	9
HCT	6	31	34	9	9	89	23
PAL	20	58	14	2	20	114	21
Cellulose and glucan biosynthesis							
CesA Primary	9	8	16	1	1	35	10
CesA Secondary	0	4	1	0	0	5	5
CesA Other	9	5	19	0	0	33	10
CslA	2	1	2	0	0	5	5
CslC	0	0	2	0	0	2	1
CslD	2	5	0	0	5	12	5
CslE	0	2	4	0	0	6	6
CslF	2	6	2	0	1	11	4
C4 photosynthesis							
CAH	0	1	269	0	20	290	13
PEPC	7	4	111	8	13	143	14
PPCK	4	1	0	6	0	11	4
ME	13	10	39	0	8	70	19
AlaAT	21	9	143	0	2	175	12
AspAT	20	16	46	2	3	87	12
PPDK	13	2	194	0	9	218	8
Peroxidases	160	124	106	1	9	400	87
Total no. ESTs clustered	23,128	22,155	21,624	2268	4158	73,333	27,329

DISCUSSION

Genetic Marker Development

Although centuries of selection and breeding have been successful in developing improved cultivars in major food crops, very little attention has been paid until recently to the improvement of switchgrass and other perennial herbaceous and woody species that hold the most potential for use in production of biofuels from lignocellulosic biomass. High dry matter yield can be obtained with switchgrass,

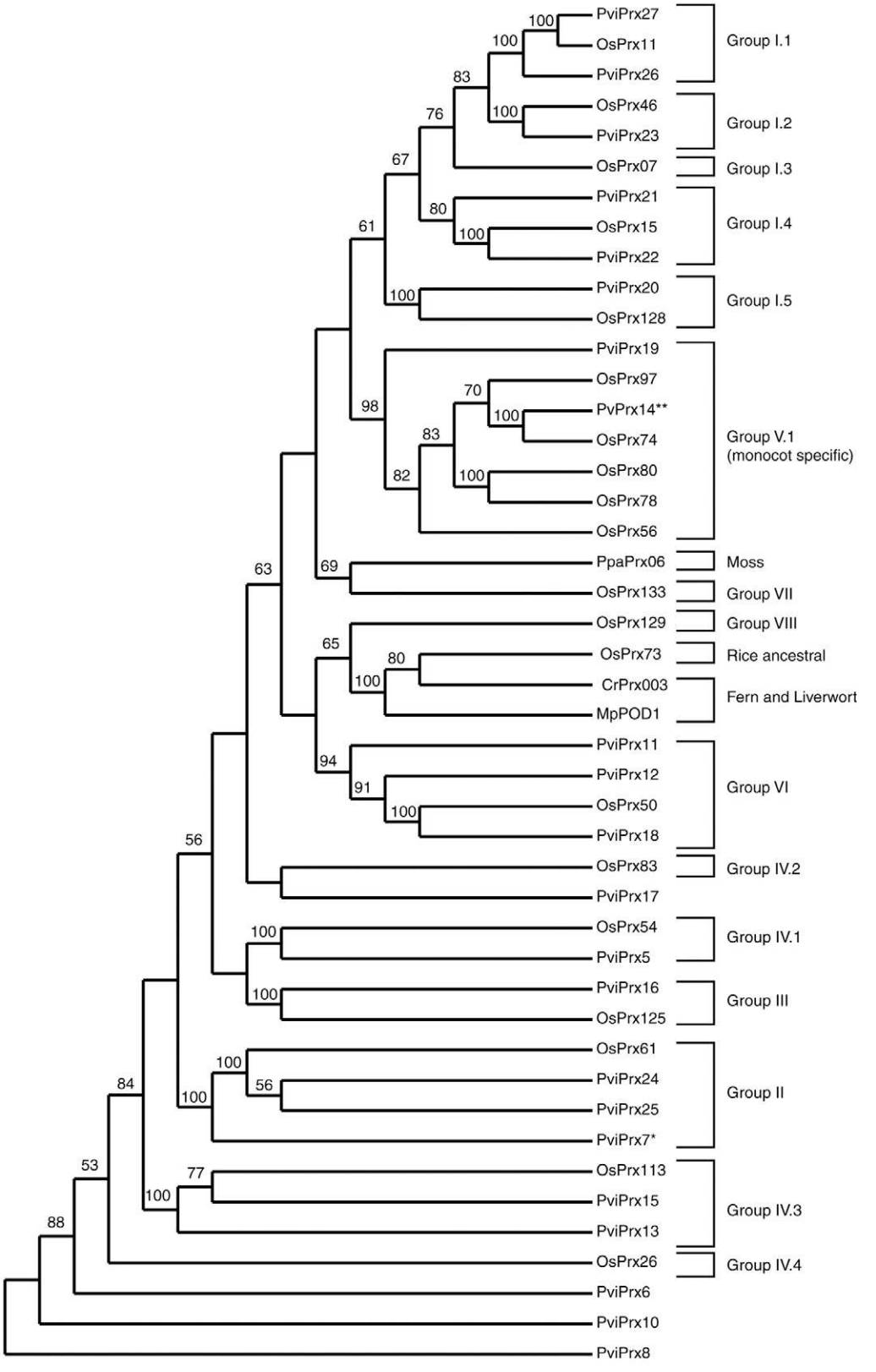

Figure 5. Evolutionary relationship of peroxidase gene family members in switchgrass and rice. Full-length switchgrass (*Panicum virgatum*; PviPrx) peroxidase sequences were obtained by virtual translation, followed by manual verification of start and stop codons in the expressed sequence tags. Switchgrass sequences were aligned with selected rice (*Oryza sativa*; OsPrx) sequences belonging to different groups as assigned by Passardi et al. (2004) and one peroxidase sequence from a liverwort (Marchantia; MpPOD1), moss (Physcomitrella; PpaPRX06), and a fern (Ceratopteris; CrPrx03) using ClustalW (Thompson et al., 1994). Nonswitchgrass peroxidase sequences were obtained from PeroxiBase (http://peroxibase.isb-sib.ch/). Aligned sequences were used to generate a dendrogram using the neighbor-joining method within the Phylip program (Phylogeny Inference Package v. 3.6). Peroxidase sequence differentially expressed in callus is indicated by *. Peroxidase sequence related to monocot specific group is indicated by **.

and net energy yields of approximately 60 GJ ha^{-1} yr^{-1} are possible (Schmer et al., 2008). This translates into ethanol yield estimates of between 4400 and 5600 L ha^{-1} (Sladden et al., 1991), leading the USDOE's Bioenergy Feedstock Development Program to choose switchgrass as a model species (McLaughlin and Walsh, 1998).

The rationale for continuing to use libraries from Kanlow switchgrass is that this high-yielding northern lowland ecotype shows a high degree of phenotypic variation and is incorporated into several breeding programs (Casler et al., 2004; McLaughlin et al., 1999). Further, as much allelic and phenotypic variation in this species is found among cultivars and populations, and even ploidy levels, as between cultivars, populations, or ecoregions (Casler et al., 2007; Gunter et al., 1996; Gunter et al., 2003). Nevertheless, continued EST sequencing efforts with several other cultivars to detect the full extent of intraspecific sequence diversity are ongoing.

The sequencing effort resulted in high-quality data with most ESTs capable of alignment to the sorghum genome. These data can be used for targeted marker development in switchgrass, allowing integration of mapping efforts with data generated from other grass species and for use in candidate gene-based approaches to perform association mapping. Identifying markers that are polymorphic and highly informative in any switchgrass population is complicated by a number of factors because of its allogamous, polyploid nature. Marker informativeness depends greatly on the number of detectable alleles, the presence of shared alleles in the parents, size homoplasy, and the phase of the markers (Luo et al., 2001). In addition, linkage phase is not known in most F_1 mapping populations. Where the ordering of markers is uncertain, reference to a related sequenced genome can sometimes help resolve ambiguity. In this regard, we have provided data on switchgrass microsatellites and positional information using sorghum as a reference. The basis of fragment length polymorphisms at many SSR loci in maize is often not strictly due to differences in repeat number but to insertions or deletions outside of the repeat region (Lia et al., 2007; Matsuoka et al., 2002). It is still unknown to what extent these types of mutations are important within coding SSRs that are under strong selective constraints. The targeting of untranslated regions, or intron-spanning regions that are under weaker selection would also be possible with these EST sequences and would provide additional useful marker data.

Genome Duplication

Codon-based comparisons can be useful for estimating synonymous and nonsynonymous substitution rates and looking at codon usage. These can be used to evaluate gene family evolution through interspecific and intraspecific comparisons, and the relative rates of synonymous versus nonsynonymous substitutions (Ka/Ks) have been widely used as indicators of the strength of selective pressure (Zhang and Yu, 2006). We used this analysis to ascertain the degree of similarity of the two distinct switchgrass genomes. An alternative measure of divergence that only considered the rates of transversion at fourfold degenerate sites (4DTv), gave very similar results.

This type of EST data analysis has detected the signature of ancient large-scale duplication events in maize and poplar (*Populus*), but recent polyploidization events, even in species considered to be allopolyploids, such as cotton (*Gossypium* L.) and bread wheat (*Triticum durum* Desf.), can be obscured by the initial peak predicted based on a constant rate of small-scale gene duplication and gene loss (Blanc and Wolfe, 2004; Sterck et al., 2005). In our analysis, we see no clear signature of large-scale recent genome duplication. These events may be undetectable because the two homeologous genomes are not sufficiently diverged from one another and cannot be clearly discriminated from paralogous sequences resulting from local gene duplication and dispersion events. However, in 3641 cases, different consensus aligned best to the same or overlapping genome regions of sorghum. These may be instances of alternative splicing or cases where both genomes are being expressed and are divergent enough to resolve as unique consensus sequences. The ratio of nonsynonymous to synonymous substitutions can be used to detect positive selection at many or single sites within sequences and across lineages. Data from these pairwise calculations support a model of negative selection, but these comparisons can be biased by several factors, including the initial selection of gene pairs and the methods used to estimate rates of substitution.

Genome size estimates of Kanlow are approximately 3.09 to 3.37 pg DNA $nuclei^{-1}$ with $2n = 4x$ complement of 36 chromosomes (Hultquist et al., 1996; Lu et al., 1998). This is close to twice the size of the sorghum genome. Estimates also place the date of divergence of the two species somewhere between 15 and 25 million years ago (Grass Phylogeny Working Group, 2001). These data and the efficient placement of most unique genes of high certainty onto the sorghum genome indicate that sorghum is an adequate model for genome-based comparisons. Estimating the evolutionary time frame for the speciation events that produced the two lineages using molecular data from this study is complicated by the possibility of significantly different rates of sequence evolution across gene lineages and possibly different clock rates that can be affected by factors such as generation time and effective population size (Zhang et al., 2002). Pairwise comparisons based on EST data can be further confounded by the possibility of including paralogous sequences, multiple comparisons of splice variants of the same gene, and sequencing errors that would all tend to overestimate clock-based dates. More studies are required in this area to date the relatively recent evolutionary history in this subfamily.

Gene Inventories

It is estimated that up to 10% of genes in plants are involved in the biogenesis of the cell wall (Yong et al.,

2005). The CesAs comprise at least three major and several minor clades in *Arabidopsis*, rice, and maize (Holland et al., 2000; Vergara and Carpita, 2001). The *CesA*s associated with primary wall cellulose synthesis—that is, orthologs of *AtCesA1*, *AtCesA3*, and *AtCes6*—show higher numbers of ESTs in callus, crown, and seedling tissue. Secondary cell-wall formation would be much more prominent in mature tissues, such as stem, but the libraries probed here may have been too small to adequately represent this difference, given the overall very small number of *CesA* and *Csl* family members detected. Nevertheless, the secondary wall-associated CesAs—that is, the putative orthologs of *AtCesA4, AtCes7,* and *AtCesA8*—are highest in crown, representing initiation of vascular development, and absent in callus. *CslD* is the one family showing more than a single read in stem tissue compared to other cellulose synthase super family members. *CslD* genes function in wall synthesis in growing root hairs in *Arabidopsis* and rice, and although some CslD genes have been shown to have moderate expression in *Arabidopsis* stem tissue, their function is unknown (Favery et al., 2001, Kim et al., 2007). *CslF* genes encode at least one of the catalytic components of the mixed linkage (1→3),(1→4)-β-D-glucan synthase (Burton et al., 2006). The *CslE*s are most closely related to cyanobacterial cellulose synthases by sequence homology but are of unknown function (Nobles and Brown, 2004).

Lignin, a phenylpropanoid, is critical for water conductance, physical strength, and pest resistance but is also problematic as it creates an impediment to efficient cell-wall hydrolysis. A preliminary classification of sequences by library showed that peroxidases and genes involved in phenylpropanoid biosynthesis are extremely well represented in switchgrass. The biosynthesis of lignin-monomers, their polymerization, and insolublization within the cell wall through oxidative coupling are controlled by both these classes of genes. The genes for phenylpropanoid biosynthesis were relatively underrepresented in the callus library while a single peroxidase that was most similar to a barley endosperm-specific and vacuole-localized protein (BP1) upregulated by Black Point infection (*Alternaria* spp. *Cochliobolus sativus*, *Fusarium* spp.; Rasmussen et al., 1997) was highly abundant. Peroxidases play major roles in determining cell-wall architecture and are believed to exert this influence by engendering loss of cell-wall extensibility through crosslinking of cell-wall components, particularly phenolic compounds. There is also a well-established link between auxin levels, embryogenic potential, and peroxidase levels in callus where peroxidase activity can be used as a biomarker for embryogenic potential (Kochba et al., 1977; Thorpe and Gaspar, 1978). We believe that the observed abundance of peroxidase ESTs in callus is likely due to auxin-regulated gene upregulation and the embryogenic potential of the callus that was utilized. Peroxidase family members other than the one that was present in callus are likely to be more important for controlling the extent of crosslinking within the cell wall during development. Close orthologs maize and sorghum CAD and COMT genes that are responsible for the *brown-midrib* mutations (Vermerris et al., 2007) in these species were also identified and may lend themselves to manipulation in switchgrass.

Physiological adaptation of switchgrass to high temperatures and water stress is due in part to C_4 photosynthesis. In contrast to maize and sorghum, C_4 photosynthesis in switchgrass is primarily of the NAD^+-me type (Warner et al., 1987), although the genus contains examples of all three C_4 subgroups. These different subtypes are characterized by metabolic, anatomical, and physiological differences. There appeared to be relatively few structural differences between sorghum, maize, rice, and switchgrass nuclear-encoded C_4 genes, supporting the hypothesis that these genes have acquired new functional specificity based on duplication and acquisition of different spatial and temporal expression patterns. There were in many cases clear biases in library representation associated with specific consensus sequences present in the seedling library, which was partially composed of photosynthetic tissue. The clearest difference in protein structure was the apparent absence in switchgrass of carbonic anhydrase sequences containing duplicated catalytic domains. These are present in both maize and sorghum but absent in rice. This type of duplication would have been readily detected via paired read information and the large number of ESTs in this group. If true, this places the structural duplication events that occurred in the *cah* genes at a point after diversification of the *Panicoideae* subfamily in the lineage that gave rise to the *Andropogoneae*.

Conclusions

These switchgrass EST sequencing efforts underscore the utility of long reads and sequencing of relatively conserved transcribed regions to produce useful gene inventories even in a highly outcrossing species. We have utilized the assembled sequences for the purposes of marker development and were successful in annotating the sequences, aligning them with a reference genome, and in discerning significant gene expression patterns based on EST representation in gene families that are important for cell-wall biogenesis and adaptation to stress. The quality and depth of reads enabled discovery of marked differences in patterns of expression and gene structure that have provided support for functional specialization within the gene families analyzed.

Acknowledgments

The authors would like to acknowledge Humphrey Wanjugi and Jennifer Bragg for critical reading of the manuscript. This work was supported through the community sequencing program of the U.S. Department of Energy, project 776898, through the U.S. Department of Agriculture, Agriculture Research Service CRIS 5325-21000-13 and 5440-21000-028, and supported in part by NIH Grant P20 RR16569 from the BRIN Program of the National Center for Research Resources, a University of Nebraska at Kearney Research Services Council University Research & Creative Activity Grant, and an NSF Plant Genome Research Grant DBI-0217552 (to N.C.C., M.C.C.). Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

References

Altschul, S., W. Gish, W. Miller, E. Myers, and D. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Apweiler, R., A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, et al. 2004. UniProt: The universal protein knowledgebase. Nucleic Acids Res. 32:D115–D119.

Birney, E., M. Clamp, and R. Durbin. 2004. GeneWise and genomewise. Genome Res. 14:988–995.

Blanc, G., and K. Wolfe. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16:1667–1678.

Bout, S., and W. Vermerris. 2003. A candidate-gene approach to clone the sorghum Brown midrib gene encoding caffeic acid O-methyltransferase. Mol. Genet. Genomics 269:205–214.

Boutin-Ganache, I., M. Raposo, M. Raymond, and C. Deschepper. 2001. M13-tailed primers improve the readability and usability of microsatellite analyses performed with two different allele-sizing methods. Biotechniques 31:24–28.

Brownstein, M., J. Carpten, and J. Smith. 1996. Modulation of non-templated nucleotide addition by *Taq* DNA polymerase: Primer modifications that facilitate genotyping. Biotechniques 20:1004–1010.

Burton, R.A., S.M. Wilson, M. Hrmova, A.J. Harvey, N.J. Shirley, B.A. Stone, E.J. Newbigin, A. Bacic, and G.B. Fincher. 2006. Cellulose synthase-like *CslF* genes mediate the synthesis of cell wall (1,3;1,4)-β-D-glucans. Science 311:1940–1942.

Casler, M., C. Stendal, L. Kapich, and K. Vogel. 2007. Genetic diversity, plant adaptation regions, and gene pools for switchgrass. Crop Sci. 47:2261–2273.

Casler, M., K. Vogel, C. Taliaferro, and R. Wynia. 2004. Latitudinal adaptation of switchgrass populations. Crop Sci. 44:293–303.

Collazo, P., L. Montoliu, P. Puigdomenech, and J. Rigau. 1992. Structure and expression of the lignin O-methyltransferase gene from *Zea mays* L. Plant Mol. Biol. 20:857–867.

Eisen, M., P. Spellman, P. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95:14863–14868.

Favery, B., E. Ryan, J. Foreman, P. Linstead, K. Boudonck, M. Steer, P. Shaw, and L. Dolan. 2001. KOJAK encodes a cellulose synthase-like protein required for root hair cell morphogenesis in *Arabidopsis*. Genes Dev. 15:79–89.

Grass Phylogeny Working Group. 2001. Phylogeny and subfamilial classification of the grasses (*Poaceae*). Ann. Mis. Bot. Gard. 88:373–457.

Gunter, L., A. Black, S. Ratnayeke, G. Tuskan, and S. Wullschleger. 2003. Assessment of genetic similarity among 'Alamo' switchgrass seed lots using RAPD markers. Seed Sci. Technol. 31:681–689.

Gunter, L.E., G.A. Tuskan, and S.D. Wullschleger. 1996. Diversity among populations of switchgrass based on RAPD markers. Crop Sci. 36:1017–1022.

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

Hopkins, A., C. Taliaferro, C. Murphy, and D. Christian. 1996. Chromosome number and nuclear DNA content of several switchgrass populations. Crop Sci. 36:1192–1195.

Huang, X., and A. Madan. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9:868–877.

Hultquist, S., K.P. Vogel, D. Lee, K. Arumuganathan, and S. Kaeppler. 1996. Chloroplast DNA and nuclear DNA content variations among cultivars of switchgrass, *Panicum virgatum* L. Crop Sci. 36:1049–1052.

Kantety, R., M. La Rota, D. Matthews, and M. Sorrells. 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum, and wheat. Plant Mol. Biol. 48:501–510.

Kim, C.M., S.H. Park, B. Il Je, S.J. Park, H.L. Piao, M.Y. Eun, L. Dolan, and C.D. Han. 2007. *OsCSLD1*, a *cellulose synthase-like D1* gene, is required for root hair morphogenesis in rice. Plant Physiol. 143:1220–1230.

Kochba, J., S. Lavee, and P. Spiegel-Rey. 1977. Differences in peroxidase activity and isoenzymes in embryogenic and nonembryogenic 'Shamouti' orange ovular callus lines. Plant Cell Physiol. 18:463–467.

Lia, V., M. Bracco, A. Gottlieb, L. Poggio, and V. Confalonieri. 2007. Complex mutational patterns and size homoplasy at maize microsatellite loci. Theor. Appl. Genet. 115:981–991.

Lu, K., S. Kaeppler, K. Vogel, K. Arumuganathan, and D. Lee. 1998. Nuclear DNA content and chromosome numbers in switchgrass. Great Plains Res. 8:269–280.

Luo, Z.W., C.A. Hackett, J.E. Bradshaw, J.W. McNicol, and D. Milbourne. 2001. Construction of a genetic linkage map in tetraploid species using molecular markers. Genetics 157:1369–1385.

Matsuoka, Y., S. Mitchell, S. Kresovich, M. Goodman, and J. Doebley. 2002. Microsatellites in *Zea*- Variability, patterns of mutations, and use for evolutionary studies. Theor. Appl. Genet. 104:436–450.

McLaughlin, S., J. Bouton, D. Bransby, B. Conger, W. Ocumpaugh, D. Parrish, et al. 1999. Progress in developing switchgrass as a bioenergy feedstock. p. 282–298 *In* J. Janick (ed.) Perspectives on new crops and new uses. Am. Soc. Hortic. Sci. Press, Alexandria, VA.

McLaughlin, S., and M. Walsh. 1998. Evaluating environmental consequences of producing herbaceous crops for bioenergy. Biomass Bioenergy 14:317–324.

Missaoui, A., A. Paterson, and J. Bouton. 2005. Investigation of genomic organization in switchgrass (*Panicum virgatum* L.) using DNA markers. Theor. Appl. Genet. 110:1372–1383.

Moser, L., and K.P. Vogel. 1995. Switchgrass, big bluestem, and indiangrass. p. 409–420. *In* R.F. Barnes et al. (ed.). An introduction to grassland agriculture. Iowa State Univ. Press, Ames.

Muench, D.G., and A.G. Good. 1994. Hypoxically inducible barley alanine aminotransferase: CDNA cloning and expression analysis. Plant Mol. Biol. 24:417–427.

Nielsen, E. 1944. Analysis of variation in *Panicum virgatum*. J. Agric. Res. 69:327–353.

Nobles, D.R., and R.M. Brown, Jr. 2004. The pivotal role of cyanobacteria in the evolution of cellulose synthases and cellulose synthase-like proteins. Cellulose 11:437–448.

Passardi, F., D. Longet, C. Penel, and C. Dunand. 2004. The class III peroxidase multigenic family in rice and its evolution in land plants. Phytochemistry 65:1879–1893.

Rasmussen, C., A. Henriksen, A. Abelskov, R. Jensen, S. Rasmussen, J. Hejgaard, et al. 1997. Purification, characterization and stability of barley grain peroxidase BP 1, a new type of plant peroxidase. Physiol. Plant. 100:102–110.

Saha, M., M. Mian, I. Eujayl, J. Zwonitzer, L. Wang, and G. May. 2004. Tall fescue EST-SSR markers with transferability across several grass species. Theor. Appl. Genet. 109:783–791.

Sarath, G., L.M. Baird, K.P. Vogel, and R.B. Mitchell. 2007. Internode structure and cell wall composition in maturing tillers of switchgrass (*Panicum virgatum* L.). Bioresour. Technol. 98:2985–2992.

Schmer, M., K. Vogel, R. Mitchell, and R. Perrin. 2008. Net energy of cellulosic ethanol from switchgrass. Proc. Natl. Acad. Sci. USA 105:464–469.

Sladden, S.E., D.I. Bransby, and G.E. Aiken. 1991. Biomass yield, composition and production costs for 8 switchgrass varieties in Alabama. Biomass Bioenergy 1:119–122.

Somleva, M., Z. Tomaszewski, and B. Conger. 2002. Agrobacterium-mediated genetic transformation of switchgrass. Crop Sci. 42:2080–2087.

Son, D., and T. Sugiyama. 1992. Molecular cloning of an alanine aminotransferase from NAD-malic enzyme type C4 plant *Panicum miliaceum*. Plant Mol. Biol. 20:705–713.

Sterck, L., S. Rombauts, S. Jansson, F. Sterky, P. Rouze, and Y. Van de Peer. 2005. EST data suggest that poplar is an ancient polyploid. New Phytol. 167:165–170.

Symonds, V., and A. Lloyd. 2004. A simple and inexpensive method for producing fluorescently labelled size standard. Mol. Ecol. Notes 4:768–771.

Talbert, L., D. Timothy, J. Burns, J. Rawlings, and R. Moll. 1983. Estimates of genetic parameters in switchgrass. Crop Sci. 23:725–728.

Taniguchi, M., A. Kobe, M. Kato, and T. Sugiyama. 1995. Aspartate aminotransferase isozymes in *Panicum miliaceum* L, an NAD-malic enzyme-type C4 plant: Comparison of enzymatic properties, primary structures, and expression patterns. Arch. Biochem. Biophys. 318:295–306.

Thiel, T., W. Michalek, R. Varshney, and A. Graner. 2003. Exploiting EST databases for the development and characterization of gene-derived

SSR-markers in barley (*Hordeum vulgare* L.). Theor. Appl. Genet. 106:411–422.
Thompson, J., D. Higgins, and T. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.
Thorpe, T., and T. Gaspar. 1978. Changes in isoperoxidase during shoot formation in tobacco callus. In Vitro 14:522–529.
Tobias, C. n.d. EST sequencing of the model grass Panicum virgatum. Available at http://wheat.pw.usda.gov/panicum/GO/ (verified 13 Oct. 2008). USDA-ARS, Albany, CA.
Tobias, C., P. Twigg, D. Hayden, K. Vogel, R. Mitchell, G. Lazo, et al. 2005. Analysis of expressed sequence tags and the identification of associated short tandem repeats in switchgrass. Theor. Appl. Genet. 111:956–964.
Vergara, C.E., and N.C. Carpita. 2001. β-D-Glycan synthases and the *Ces*A gene family: Lessons to be learned from the mixed-linkage (1→3),(1→4)β-D-glucan synthase. Plant Mol. Biol. 47:145–160.
Vermerris, W., A. Saballos, G. Ejeta, N.S. Mosier, M.R. Ladisch, and N.C. Carpita. 2007. Molecular breeding to enhance ethanol production from corn and sorghum stover. Crop Sci. 47:S142–S153.
Warner, D., M. Ku, and G. Edwards. 1987. Photosynthesis, leaf anatomy, and cellular constituents in the polyploid C4 grass *Panicum virgatum*. Plant Physiol. 84:461–466.
Wyrich, R., U. Dressen, S. Brockmann, M. Streubel, C. Chang, D. Qiang, et al. 1998. The molecular basis of C4 photosynthesis in sorghum: Isolation, characterization and RFLP mapping of mesophyll- and bundle-sheath-specific cDNAs obtained by differential screening. Plant Mol. Biol. 37:319–335.
Yang, Z., and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17:32–43.
Yong, W., B. Link, R. O'Malley, J. Tewari, C. Hunter, C. Lu, et al. 2005. Genomics of plant cell wall biogenesis. Planta 221:747–751.
You, F.M., N. Huo, Y.Q. Gu, M. Luo, Y. Ma, D. Hane, et al. 2008. BatchPrimer3: A high throughput web application for PCR and sequencing primer design. BMC Bioinformatics 9:253.
Zhang, L., T. Vision, and B. Gaut. 2002. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. Mol. Biol. Evol. 19:1464–1473.
Zhang, Z., and J. Yu. 2006. Evaluation of six methods for estimating synonymous and nonsynonymous substitution rates. Genomics Proteomics Bioinformatics 4:173–181.