the plant journal

SEB Society for Experimental Biology

The Plant Journal (2013) **74,** 160–173 doi: 10.1111/tpj.12104

TECHNICAL ADVANCE/RESOURCE

Development of an integrated transcript sequence database and a gene expression atlas for gene discovery and analysis in switchgrass (*Panicum virgatum* L.)

Ji-Yi Zhang[1,2], Yi-Ching Lee[1,2], Ivone Torres-Jerez[1], Mingyi Wang[1], Yanbin Yin[2,3], Wen-Chi Chou[2,3], Ji He[1], Hui Shen[1,2], Avinash C. Srivastava[1,2], Christa Pennacchio[4], Erika Lindquist[4], Jane Grimwood[4,5], Jeremy Schmutz[4,5], Ying Xu[2,3], Manoj Sharma[6,7], Rita Sharma[6,7], Laura E. Bartley[6,7,†], Pamela C. Ronald[6,7], Malay C. Saha[2,8], Richard A. Dixon[1,2], Yuhong Tang[1,2] and Michael K. Udvardi[1,2,*]

[1]*Plant Biology Division, The Samuel Roberts Noble Foundation, Ardmore, OK 73401, USA,*
[2]*Department of Energy, BioEnergy Science Center (BESC), Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA,*
[3]*Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602, USA,*
[4]*Department of Energy, Joint Genome Institute, Walnut Creek, CA 95598, USA,*
[5]*HudsonAlpha Genome Sequencing Center, Huntsville, AL 35806, USA,*
[6]*Departments of Plant Pathology and the Genome Center, University of California, Davis, CA 95616, USA,*
[7]*Department of Energy, The Joint Bioenergy Institute (JBEI), Emeryville, CA 94608, USA, and*
[8]*Forage Improvement Division, The Samuel Roberts Noble Foundation, Ardmore, OK 73401, USA*

Received 24 September 2012; revised 14 December 2012; accepted 20 December 2012; published online 05 January 2013.
*For correspondence (e-mail mudvardi@noble.org).
The author(s) responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Author Guidelines (http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1365-313X/homepage/ForAuthors.html) is Michael K. Udvardi (mudvardi@noble.org)
†Present address: Department of Botany and Microbiology, University of Oklahoma, Norman, OK, 73019, USA.

SUMMARY

Switchgrass (*Panicum virgatum* L.) is a perennial C4 grass with the potential to become a major bioenergy crop. To help realize this potential, a set of RNA-based resources were developed. Expressed sequence tags (ESTs) were generated from two tetraploid switchgrass genotypes, Alamo AP13 and Summer VS16. Over 11.5 million high-quality ESTs were generated with 454 sequencing technology, and an additional 169 079 Sanger sequences were obtained from the 5′ and 3′ ends of 93 312 clones from normalized, full-length-enriched cDNA libraries. AP13 and VS16 ESTs were assembled into 77 854 and 30 524 unique transcripts (unitranscripts), respectively, using the Newbler and pave programs. Published Sanger-ESTs (544 225) from Alamo, Kanlow, and 15 other cultivars were integrated with the AP13 and VS16 assemblies to create a universal switchgrass gene index (PviUT1.2) with 128 058 unitranscripts, which were annotated for function. An Affymetrix cDNA microarray chip (Pvi_cDNAa520831) containing 122 973 probe sets was designed from PviUT1.2 sequences, and used to develop a Gene Expression Atlas for switchgrass (PviGEA). The PviGEA contains quantitative transcript data for all major organ systems of switchgrass throughout development. We developed a web server that enables flexible, multifaceted analyses of PviGEA transcript data. The PviGEA was used to identify representatives of all known genes in the phenylpropanoid–monolignol biosynthesis pathway.

Keywords: *Panicum virgatum*, switchgrass, bioenergy, transcript sequence, transcriptome, genomics, technical advance.

INTRODUCTION

Switchgrass (*Panicum virgatum* L.) is an out-breeding perennial C_4 grass native to North America. It has been used as forage and for soil conservation in the USA and has been targeted as a source of biomass for biofuel production (McLaughlin and Kszos, 2005; Bouton, 2007; Schmer *et al.*, 2008; Yuan *et al.*, 2008; Keshwani and Cheng, 2009). Therefore, breeding and genetic engineering efforts are under way to improve existing cultivars and

Table 1 Switchgrass 454-cDNA libraries and 454 expressed sequence tags (ESTs)

NCBI SRA accession no.	JGI lib code	Tissues	Plant growth stage/conditions	No. of good ESTs	Mean length (bp)
Summer VS16 454 data					
SRX026147	CFBB	Whole shoot	Leaf development	259 106	201
SRX026148	CFBC	Whole root	Leaf development	205 466	222
SRX026149	CFBF	Whole shoot	Stem elongation	194 426	194
SRX026150	CFBG	Whole root	Stem elongation	174 053	190
SRX026151	CFCZ	Whole shoot	Reproductive stage	219 230	189
SRX026153-4	CFFA	Whole root	Reproductive stage	234 107	205
SRX026155-6	CFFB	Panicles including seeds	Reproductive stage	220 933	212
Subtotal				1 507 321	202
Alamo AP13 454 data					
SRX057824	CCXN	Whole shoot	Stem elongation	733 173	202
SRX057825	CCXO	Whole root	Stem elongation	667 612	206
SRX057830	CFXX	Whole shoot	Leaf development	1 236 020	419
SRX057831	CFXY	Whole root	Leaf development	1 214 630	375
SRX057828	CFXW	Whole shoot	Stem elongation	1 357 290	223
SRX057829	CGGO	Whole root	Stem elongation	1 040 192	404
SRX057827	CGFF	Whole shoot	Reproductive stage	547 278	320
SRX057826	CGFC	Whole root	Reproductive stage	998 691	388
SRX057834	CGTX	Panicles including seeds	Reproductive stage	1 096 949	384
SRX057833	CGFI	Whole shoot	Stem elongation 2 w/drought	362 346	213
SRX057832	CGGU	Whole root	Stem elongation 2 w/drought	918 585	337
Subtotal				10 172 766	316
Total				11 680 087	

NCBI SRA, National Center for Biotechnology Information short read archive; JGI, Joint Genome Institute.

germplasm (Lemus *et al.*, 2002; Bouton, 2007; Chuck *et al.*, 2011; Fu *et al.*, 2011; Wang *et al.*, 2011; Xu *et al.*, 2011).

Most switchgrass cultivars are either tetraploid ($2n = 4x = 36$) or octoploid ($2n = 8x = 72$), which are primarily lowland (latitude) or upland types, respectively (Hopkins *et al.*, 1996). The haploid genome (1C) was estimated to be between 1372 and 1666 Mb (Bennett *et al.*, 2000). Cytogenetic analysis, combined with recent genetic mapping suggested that tetraploid switchgrass is probably amphidiploid (allotetraploid; Martinez-Reyna *et al.*, 2001; Okada *et al.*, 2010; Triplett *et al.*, 2012).

Development of genomic resources, including gene/transcript sequences, single nucleotide polymorphism (SNP) markers, and DNA microarrays for genome-wide expression studies that have been accomplished so far with Arabidopsis (Schmid *et al.*, 2005), *Medicago truncatula* (Benedito *et al.*, 2008), rice (Wang *et al.*, 2010), and maize (Sekhon *et al.*, 2011), is essential for basic and applied research on switchgrass. Partial sequencing of mRNA to produce expressed sequence tags (ESTs) is the most direct and efficient means of generating information about coding regions of the genome. Previously, about half a million ESTs from 18 different cultivars were deposited in public databases, with the majority (346 752) from the cultivar Kanlow (Tobias *et al.*, 2005, 2008). Past efforts to assemble switchgrass ESTs into longer, non-redundant transcripts used sequences from multiple genotypes and cultivars, which complicated the assembly process (PlantGDB-assembled Unique Transcript (PUT), http://www.plantgdb.org/prj/ESTCluster/progress.php; DFCI Switchgrass Gene Index, http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=switchgrass). To complement and extend previous work, we chose single genotypes of two tetraploid switchgrass varieties, namely Alamo AP13 (a lowland cultivar) and Summer VS16 (an upland cultivar), which have been used to generate a mapping population (Missaoui *et al.*, 2005). For the work described here, they were propagated vegetatively (clonally) to minimize sequence complexity and facilitate EST assembly and SNP discovery. Objectives achieved in the work described here include: assembly of full-length or partial mRNA sequences for the majority of functional genes in switchgrass for gene discovery, Affymetrix chip design for switchgrass transcriptomics and development of a gene expression atlas with quantitative transcript data for most genes in all major organs. These resources and a web server hosting all these data sets provide a solid foundation for functional genomics and breeding in switchgrass.

RESULTS AND DISCUSSION

Generation of ESTs by 454 sequencing

Over 1.58 million ESTs were generated from seven Summer VS16 cDNA libraries derived from roots, shoots, and panicles with developing seeds harvested at different stages of development, using 454/Roche pyrosequencing technology and a 454 GS FLX machine with standard reagents. After sequence trimming and culling, 1 507 321

Table 2 Normalized full-length (FL) cDNA libraries and Sanger expressed sequence tags (ESTs)

NCBI dbEST accession no.	JGI code	Tissues and conditions	Clones	Total no. of ESTs	No. of good ESTs	Average length (bp)	Diversity (%)	% of FL-cDNA-clones
JG800964–JG836623	CFNU	Aerial tissues at multiple stages without specific treatment	19 968	39 936	35 660	669	67.6	14.3
JG772376–JG800963	CFNT	Underground tissues at multiple stages without specific treatment	15 744	31 487	28 588	658	68.4	14.7
N/A	EXTA	Pooled RNAs from 32 samples including all possible tissues, some with abiotic stresses	57 600	115 200	104 831	612	48.3	9.3
Total/pooled		93 312	186 623	169 079	646	39.8	11.3	

NCBI dbEST, National Center for Biotechnology Information Database of Expressed Sequence Tags.

ESTs remained (Table 1, Figure S1 in Supporting Information). The average length of an EST was 202 bp.

For Alamo AP13, 11 cDNA libraries were constructed from various organs and developmental stages, including two from drought-stressed plants (Table 1). Libraries were sequenced with 454 GS FLX machines, using either the standard protocol for the first two libraries, which produced about 200 bp read lengths, or the Titanium protocol for the remaining nine libraries, which produced about 400 bp sequences. Over 10 million trimmed ESTs were retained after culling for this genotype.

Analysis of ESTs derived from axenically grown switchgrass revealed that fewer than 0.02% of sequences were derived from microbial contaminants. In contrast, nearly 1% of ESTs in an initial shoot library derived from plants grown in a greenhouse appeared to come from fungal contaminants (Table S1). This guided our decision to use clones propagated initially *in vitro*, with antibiotics and fungicide added to the medium (i.e. axenic plants; Figure S2).

Full-length-enriched cDNA libraries and Sanger-EST generation

Three normalized, full-length-enriched cDNA libraries were constructed from multiple tissues and developmental stages of Alamo AP13 plants grown under optimal or stress conditions (see Experimental Procedures; Table 2). A total of 93 312 cDNA clones were sequenced from both the 5′ and 3′ ends using Sanger sequencing technology, resulting in 169 079 ESTs with an average length of 646 bp, after trimming and culling of sequences. Putative full-length cDNA clones were identified by querying a reference set of coding sequences of C4 grasses, including foxtail millet (*Setaria italica*), sorghum (*Sorghum bicolor*), and maize (*Zea mays*). Alignments of the 5′-end of each clone and the presence of a putative start codon were used to determine that 11.3% of the clones were potentially full length (Table 2).

Sequence assembly and analysis

Assembly of switchgrass ESTs to reconstruct accurate, full-length transcripts is complicated by the heterozygous nature of this out-breeding species and by the sequence similarity between homeologous genes from the ‘A’ and ‘B’ genomes of allotetraploids (Chaudhary *et al.*, 2009). Assembly is further complicated when sequences of multiple genotypes and varieties are combined. To avoid the latter complication, we assembled separately ESTs from Alamo AP13 and Summer VS16. Later, we combined the AP13 and VS16 assemblies with those of other varieties in the public domain.

Several *de novo* assembly programs were compared using the 454-EST data of VS16, including Newbler (454 Life Sciences, http://www.454.com/), mira (Chevreux *et al.*, 2004), tgicl (Pertea *et al.*, 2003), and pave (Soderlund *et al.*, 2009), all with default settings (Figure 1a). The resulting sequence assemblies were deposited in a webserver (http://switchgrassgenomics.noble.org/) for evaluation and comparison purposes. A brief summary of the assemblies is provided in Table S2. Finally, we chose Newbler with the cDNA option (version 2.3) to assemble all 454-ESTs generated in this project, for two main reasons. First, Newbler was specifically designed and optimized for assembling sequences generated by the 454 technology and performs better in this regard than cap3, mira, SeqMan, and clc (Kumar and Blaxter, 2010). Second, tgicl and pave were extremely slow and required enormous computer memory. In fact, we were unable to assemble the 11 million AP13 ESTs with either pave or tgicl in a reasonable time (less than a month) with our computer resources (500 GB memory).

Using Newbler with default settings, including a minimum of 40 bases overlap and 90% sequence identity for assembly, the 1.5 million VS16 454-ESTs were assembled into 34 430 contigs/isotigs and 192 515 singletons. Conceptually, an isotig represents a single splice variant of the primary transcript, which belongs to an isogroup that represents a summation of alternative splice variants from the same gene. An automated quality-control (QC) step was introduced to identify low-quality contigs/isotigs and singletons resulting from Newbler assembly and to discard low-quality sequences and break apart assemblies containing such sequences. Briefly, a PERL script was written to scan sequences with a 10-base sliding window and calculate an average sequence quality score (QS) for each 10-base sequence. If the QS of any 10-base window dropped to 30 or below, the assembly was

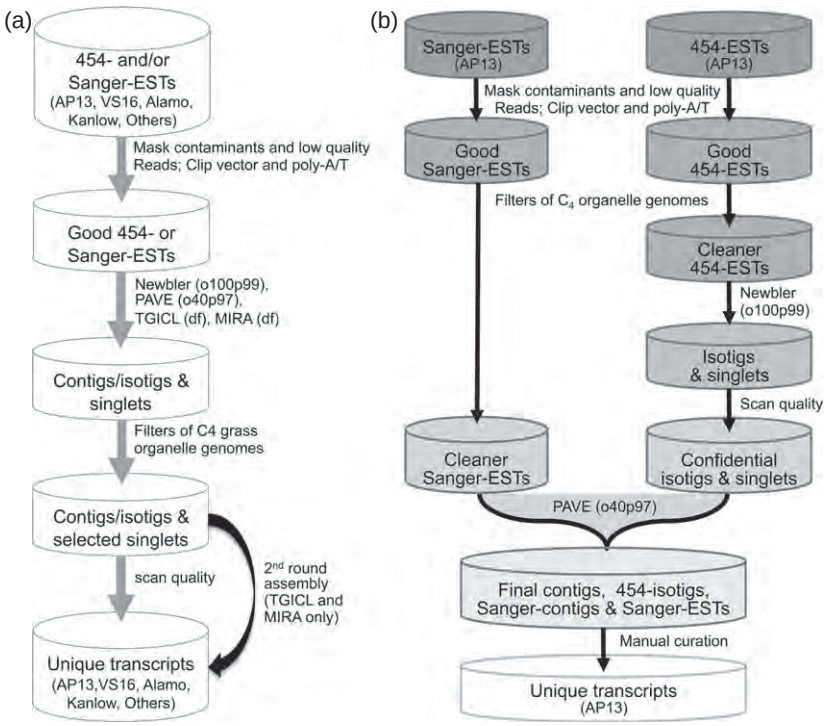


Figure 1. One-step (a) and two-step (b) sequence assembly approaches.
Different assembly programs were tested and compared using 454 and/or Sanger expressed sequence tags (ESTs) in a one-step approach (a). Based on comparison among assembly results, NEWBLER with high stringency was chosen to generate a final assembly of 454-ESTs of VS16, while PAVE with moderate stringency was used to assemble the Sanger-ESTs of Alamo, Kanlow, and other genotypes. The AP13 unitranscripts reported here were assembled using a two-step approach (b), in which 454-ESTs were assembled using NEWBLER with high stringency and Sanger-ESTs were subsequently integrated with the 454-assemblies and singlets using PAVE. o, minimum sequence overlap length (bp); p, percentage of minimum identity required for assembly.

broken apart. Any singleton with QS below 30 was removed from the EST dataset entirely. After applying this QC step, assembly of VS16 454-ESTs using NEWBLER default settings resulted in 32 764 high-quality contigs/isotigs (Figure 2a). The median mismatch in contig/isotig consensus sequences was 2.33% (Table S3), which is close to the estimated 2.5 or 3.8% sequence divergence between the two subgenomes of switchgrass and much higher than the 0.36 or 0.43% variation amongst alleles determined from comparisons of coding (CD) or untranslated (UTR) sequences, respectively (JS, unpublished data). In other words, using NEWBLER with default settings probably resulted in over-assembly of 454-EST sequences.

To avoid false mergers of homeologous sequences, we applied more stringent assembly parameters to NEWBLER and increased the minimum sequence overlap to 100 bases and minimum sequence identity to 99%, which increased the length of the shortest contig/isotig to 110 bp and decreased the longest from 6011 to 5001 bp for VS16 454-EST assembly (Figure 2a,b). The total number of VS16 contigs/isotigs decreased slightly from 32 764 to 30 524, using the more stringent assembly parameters followed by the QC step described above. More importantly, the quality of the resulting contigs/isotigs improved substantially, as judged by reduced numbers of mismatches among assembled ESTs (Figure 2c,d). The median degree of mismatch between sequences assembled at high stringency fell to 0.84% (Table S3), which was substantially less than the estimated variation between homeologs and closer to the estimated allelic variation (see above). Therefore, use of the more stringent assembly parameters largely avoided mergers of homeologous gene sequences while bringing together sequences of alleles, which was our objective. (For a comparison of assembly results obtained with a range of sequence mismatch cut-offs see Table S3.)

To assemble the 11.5 million 454-ESTs and the 169 079 Sanger ESTs of Alamo AP13, we employed a two-step strategy, using NEWBLER with high-stringency parameters to assemble the 454-ESTs and PAVE to assemble the resulting 454 contigs/isotigs and singletons together with the Sanger ESTs (Figure 1b). This strategy took advantage of the computationally efficient NEWBLER program to assemble the large number of 454-ESTs and the computationally more demanding PAVE program to generate accurate assemblies of multiple types of sequence (i.e. 454 and Sanger sequences). PAVE had the added advantage of being able to integrate paired-end (5′ and 3′) sequence information from the Sanger reads. Assembly of the 11.5 million 454-ESTs using NEWBLER with stringent parameters followed by the QC step resulted in 102 178 contigs/isotigs and 481 395 high-quality singletons. PAVE, with moderate stringency settings (see Experimental Procedures), was then used to assemble these contigs/isotigs and singletons with the Sanger ESTs (Figure 1b). PAVE assembly of AP13 EST sequences resulted in 77 854 unique transcript sequences (unitranscripts, or UTs), with an average length of 1162 bases, consisting of Sanger-454 sequence hybrids, Sanger contigs, 454-contigs/isotigs, and Sanger singleton sequences. Those 454-sequences that remained singletons after PAVE assembly were not included in the AP13 UT set (Figure 1b). Of

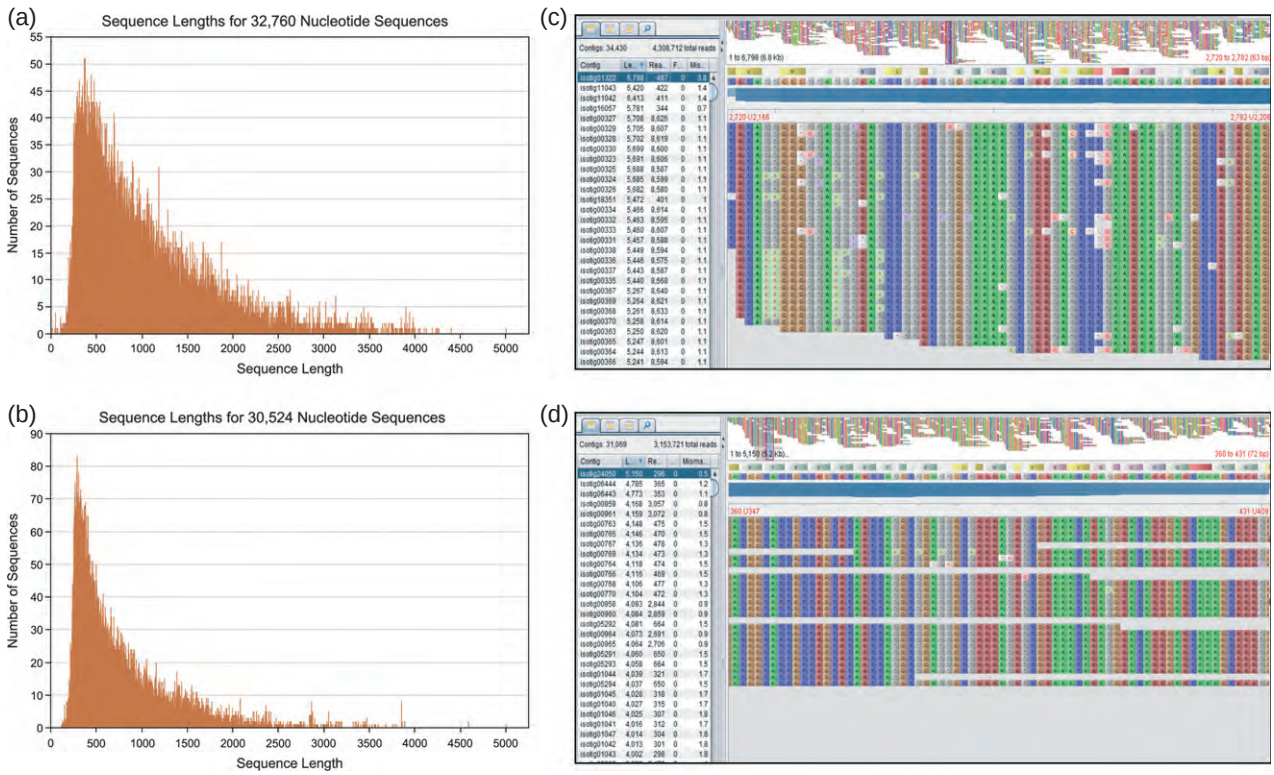


Figure 2. Length distribution of Summer VS16 isotig/contigs and snapshots of representative assemblies.
NEWBLER assembly of 454-expressed sequence tags from VS16 generated fewer and, on average, longer unitranscripts with default parameters (a) than with stringent parameters (b). However, the median mismatch between assembled sequences was 2.19% when default parameters were used (c as a representative), compared with 0.84% with stringent parameters (d as a representative). For a better comparison, sequences longer than 5100 bp were not displayed in (a).

the 34 043 non-redundant paired-end Sanger sequences of AP13, 6704 did not overlap. Integration of 454-sequences bridged 5482 of these 'hanging' paired-end sequences, leaving 1222 unbridged. Interestingly, the number of UTs was approximately twice the number of genes in a typical diploid grass genome – 25 532 protein-coding genes for *Brachypodium* (The International Brachypodium Initiative, 2010), 34 496 for sorghum (Paterson *et al.*, 2009), 40 598 for foxtail millet (http://www.phytozome.net/foxtailmillet.php, version 7), 44 805 gene models (http://rapdb.dna.affrc.go.jp/gene/statistics.html) (Tanaka *et al.*, 2008) or 55 986 gene loci (http://rice.plantbiology.msu.edu/index.shtml) (Ouyang *et al.*, 2007) for rice, indicating that the assembly process resolved transcripts for the 'A' and 'B' genomes of this allotetraploid to some degree. To test this idea, we determined the number of AP13 UTs that best matched each foxtail millet gene, using BLAST, and plotted the distribution of number of switchgrass matches per foxtail millet gene (Figure S3). Only 8.7% of foxtail millet genes matched a single AP13 UT. The largest percentage of foxtail millet genes was represented by two AP13 transcripts, which presumably reflected the diploid versus tetraploid nature of the two genomes. However, many foxtail millet genes were the best BLAST hits of three or more AP13 UTs, which in many cases do not correspond to full-length transcripts.

For quality control, orientation checking, and curation of UTs, Sanger-EST reads of Alamo AP13 generated in this project and of other varieties from the public domain were mapped to the VS16 and AP13 UTs using GMAP (Wu and Watanabe, 2005). Visual examination of the mapping results in GBrowser showed excellent consistency between EST reads and UTs, with a few exceptions dominated by overly assembled sequences, mainly due to vector sequences remaining after clipping. Using the 3′ Sanger reads as the main reference, misassembled sequences were manually reassembled and questionable sequences were removed from the assemblies. Assemblies that contained obvious alternative splice forms were broken into multiple transcripts manually.

We have used the AP13 UT assemblies to clone 91 partial or full-length cDNAs with a 97.8% success rate, and the resulting clone sequences were, on average, 99.4% identical to the corresponding PviUTs.

Development of an integrated switchgrass UT database

To develop an integrated set of UT sequences from multiple switchgrass varieties with minimal redundancy, we

captured all publicly available switchgrass Sanger-EST sequences (Table S4) and assembled them in a variety-centric way before comparing assemblies and eliminating redundancy. Assembly of 58 251 Alamo sequences from multiple genotypes (Tobias *et al.*, 2008; Srivastava *et al.*, 2010), using PAVE with moderate stringency settings resulted in 18 936 UTs, while assembly of 346 752 Kanlow sequences resulted in 56 660 UTs. Assembly of 141 242 sequences from 15 other varieties resulted in 39 823 UTs. The 77 854 AP13 UTs were then compared with those of Alamo, then Kanlow, then the multiple varieties, and finally the UTs of VS16 with the removal of shorter, redundant putative allelic sequences at each step (Figure S4 and see Experimental Procedures for details). As a result, the majority of AP13 UTs (69 793) were retained in the 'non-redundant' dataset along with 10 017 Alamo, 34 412 Kanlow, 2221 VS16 and 13 058 UTs from the other varieties (Table S4). The resulting integrated switchgrass UT (PviUT) set consisted of 128 058 hypothetical transcript sequences with an average length of 1154 bp. The number of PviUTs is substantially larger than the number of genes expected in the allotetraploid switchgrass genome. Lack of full-length assemblies of some gene transcripts, i.e. two or more assemblies covering different, non-overlapping parts of a transcript, may account for some of this 'redundancy'. Assemblies representing alternative splice forms of mRNA also contributed to the surfeit of UTs, particularly when NEWBLER was used for assembly. Finally, sequence divergence between the various genotypes may have resulted in the inclusion of multiple, redundant UTs representing orthologs in the different varieties.

We chose to use axenically grown switchgrass plants for the majority of sequencing work to avoid microbial sequence contamination. As a result, these EST libraries may not have captured transcripts of genes induced by contact with microbes. On the other hand, nearly half of the PviUTs were generated from cDNA sequence of other accessions that were not axenically grown, so microbe-elicited switchgrass genes should be represented in the PviUT database. In fact, subsequent analysis of PviUTs derived from these other accessions indicates that 0.012, 0.256, and 0.792% of PviUTs may be of viral, fungal, or bacterial/organellar origin, respectively.

Although it is not yet possible to determine what fraction of all switchgrass genes are represented by the PviUTs, due to the lack of a complete genome sequence, we attempted to estimate this by determining the fraction of foxtail millet, sorghum, and maize genes that matched at least one PviUT, using BLASTN with an *E*-value cut-off of 1.0×10^{-5}. Around 90% of the genes in these three species matched one or more of the PviUTs (Table S5). Thus, we surmise that the current set of PviUTs represents the vast majority of expressed genes in switchgrass. Based on comparisons with the coding sequences of the three reference grass species that had an average length of coding sequence of 1251, 1271, and 1076 bp in foxtail millet, sorghum, and maize, respectively, between 58.5 and 61.9% of PviUTs appear to contain full-length open reading frames (Table S5).

Similarity between switchgrass and foxtail millet sequences was assessed by comparing PviUTs against annotated transcript sequences of foxtail millet using BLASTN. Over 55% of the PviUTs matched at least one foxtail millet gene transcript with an *E*-value of 1.0×10^{-100} or lower (Figure S5). Approximately 20% of PviUTs matched foxtail millet sequences with *E*-values above 1.0×10^{-100} and below 1.0×10^{-10}. About 20% of PviUTs had little or no sequence similarity to foxtail millet (*E*-value above 0.10).

Prior to annotating the PviUT data, all sequences were oriented from 5′ to 3′ using directional information from Sanger-EST clones if available or information from BLASTX alignments of PviUT sequences to non-redundant protein sequences of the National Center for Biotechnology Information (NCBI). Some sequences (11 142 or 8.7% of the total) could not be oriented with either method. These were analyzed and annotated in both 'forward' and 'reverse' orientations, increasing the total number of switchgrass UTs to 139 200. Among the UTs, 91 617 (65.8%, cutoff *E*-value = 0.01) encoded putative proteins with at least one conserved domain, according to NCBI's Conserved Domain Database (CDD, http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) (Marchler-Bauer *et al.*, 2003, 2011; Figure S6). A total of 7677 different CDD domains were identified in PviUT proteins. The Panther classification system (Abrouk *et al.*, 2010; http://www.pantherdb.org/) was also used to annotate PviUT proteins, of which 72 650 were assigned (*E*-value <0.001) one of 3277 Panther category IDs. Using the UniProtKB database, we got hits for 109 208 (78.5%) PviUTs, and 72 579 (52.1%) were assigned at least one gene ontology (GO) term. The PviUTs were also annotated based on their homology (*E*-value cutoff of 0.01) to gene models of Arabidopsis, rice, and sorghum. KEGG (http://www.genome.jp/kegg) annotations were also performed on PviUT proteins. All of this annotation information is included in the PviUT database (http://switchgrassgenomics.noble.org/). Approximately 30% of PviUTs remained unannotated after the analyses described above. A fraction of these may represent genes unique to switchgrass and/or its close relatives. By comparing all PviUT sequences with the protein sequences of Arabidopsis, rice, and all 31 plant and algal species available at NCBI, using BLASTX with an *E*-value cutoff of 1.0×10^{-5}, we found 48 819 (35.0%), 34 764 (25.0%), and 26 918 (19.3%) PviUTs with no matching sequences in the given species. Grass-, C_4 grass-, or switchgrass-specific gene transcripts are likely to be found amongst these sequences.

Development of a switchgrass gene expression atlas (PviGEA)

We commissioned the design and production of an Affymetrix cDNA chip (Pvi_cDNAa520831) to facilitate switchgrass transcriptome analysis and to develop a reference gene expression atlas (GEA) for this species. The cDNA chip contains a total of 122 973 probe sets, including 122 868 probe sets corresponding to 110 208 PviUT sequences. In addition to the regular control features provided by Affymetrix, there are 68 probe sets for labeling controls and 20 *Brachypodium*-specific probe sets used as negative controls. Bearing in mind the fact that diploid grass genomes sequenced to date, including *Brachypodium distachyon*, foxtail millet, maize, rice, and sorghum, contain between 32 000 and 63 000 protein-coding transcripts (PHYTOZOME v7.0, http://www.phytozome.net/) and that the majority of PviUTs are derived from tetraploid plants, the 110 208 PviUTs with matching probe sets on the cDNA chip probably represent the majority of genes in switchgrass. The chip is produced by Affymetrix and is available to the public.

To build a switchgrass GEA, we began by measuring transcript levels in all major organ systems at one or more stages of development from germination to flowering: seed germination at 1, 2, 3, and 4 days post-imbibition (Figure S7); whole roots and whole shoots at vegetative stages V1–V5 (Moore *et al.*, 1991); and inflorescence development from the floral rachis meristem stage through to panicle emergence (Figure S8). Because of our interest in lignin biosynthesis, we also harvested different portions of developing internodes (internode 4 of tillers at the stem elongation stage 4; Shen *et al.*, 2009). Internodes were dissected into five equal segments along the main axis and segments one, three, and five were subjected to transcriptome analysis. Vascular bundle tissue, dissected from the middle segment of internode 3, was also analyzed (Figure S9; see Table S6 for an overview of all samples). For each organ or tissue sample, three independent biological replicates were analyzed. All Affymetrix hybridization data were normalized using the robust multi-array average (RMA) procedure provided with Expression Console (Irizarry *et al.*, 2003).

Hybridization data quality was assessed by comparing normalized signals of all probe sets between biological replicates, using Pearson correlation analysis. Correlation coefficients between biological replicates typically exceeded 0.99. Based on analysis of signal to noise of this switchgrass chip, together with other Affymetrix chips used in this laboratory, a normalized transcript level of 32 ($\log_2 = 5$, Figure S10) was set as the threshold below which a gene was determined to be *not* expressed. Consistent with this, values obtained from probe sets with 'absent' signals, as determined by Affymetrix software, exhibited a $\log_2$ normal distribution with an upper boundary around 5. Using this threshold, all 20 negative control probe sets designed upon *Brachypodium*-specific gene sequences detected no transcripts when hybridized to switchgrass nucleic acids (Table S7).

Transcripts were detected for approximately two-thirds of all genes represented by probe sets (79 382) on the cDNA chip in one or more of the organs or tissue samples (Figure S11). Nearly half of these expressed genes, corresponding to 35 000 probe sets, were active in all organs and tissues assayed. Interestingly, the number of active genes was similar in the different organs, corresponding to 56 000–58 000 probe sets (Figure S11). However, the subsets of genes expressed in different organs and tissues differed. For example, genes expressed in roots or roots and stem internodes but not in leaves or flowers included many transporters and transport-related genes possibly involved in plant mineral nutrition, based on GOslim classification (Hu *et al.*, 2008). Quantitative differences in transcript levels between organs revealed a dynamic transcriptome in switchgrass. The majority of genes in switchgrass were subject to transcriptional or post-transcriptional regulation that altered steady-state transcript levels during development. The coefficient of variance (CV) ranged from 6.0 to 392% for expressed genes, with an average of 53% (Figure 3). For comparison, the average CV between the three biological replicates for all genes and all organs was 12.5%. The CV analysis identified a set of stably expressed genes, transcripts of which changed little during development. Approximately 4010 genes exhibited a CV of <15% (Figure 3). Applying additional filters to this subset of genes, including a minimum transcript level of 100 and a transcript ratio <2 when comparing the highest transcript level in any organ with the lowest level in any organ for a given gene, yielded 250 genes, many of which encode proteins with predicted functions (Table S8). Transcript levels of these genes ranged from around 100 to over 20 000. These stably expressed genes can serve as reference genes for normalizing transcript levels of other switchgrass genes prior to comparative gene expression analysis, as in other species (Czechowski *et al.*, 2005; Benedito *et al.*, 2008). Amongst these genes are homologs of common reference genes, such as housekeeping genes encoding ubiquitin (UBI) isoforms and glyceraldehyde-3-phosphate dehydrogenase (GAPDH). Other potential reference genes include genes for cytochrome *c* oxidase and UBI-conjugating enzymes (transcript levels >10 000), histone 3 and double-stranded DNA-binding protein (1000> transcript level <10 000), and DNA polymerase III, topoisomerase II-associated protein, and nuclear histone acetyltransferase (Table S8).

The lack of hybridization of AP13 RNA to one-third of the probe sets may reflect any of the following: very low or no expression of some genes under the growth

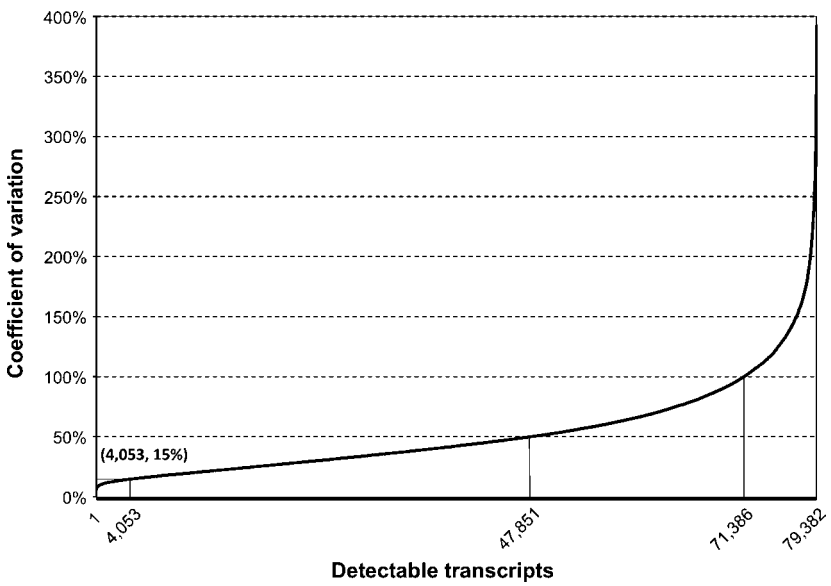


Figure 3. Distribution of coefficients of variation (CV) for transcript levels of all genes expressed during switchgrass development.
The CVs of normalized expression level among 24 organ samples were plotted from low to high for each probe set. Very stable transcript levels amongst all organs were exhibited by 5.1% of genes (represented by 4053 probe sets; CV < 15%), while 12.6% of genes were quite highly regulated and exhibited a CV of 100% or greater.

conditions used; differences in sequence between AP13 and the other varieties used to design probe sets (37% of PviUTs were derived from non-Alamo varieties); the presence of genes in other varieties that are absent in AP13; and the presence of sense and antisense probe sets for over 2000 genes of unknown 5′–3′ orientation (only half of these should detect transcripts). Furthermore, probe sets designed from PviUTs that represent the extreme 5′-end of transcripts may be less effective at detecting labelled probes derived from the 3′ ends of RNA. Normally, probe sets are designed based on the 3′-end of transcripts when these are known. As shown in Table S7, 72.9% of the probe sets derived from AP13 sequences detected AP13 transcripts in our experiments, while only 36.1% of probe sets derived from 'other' pooled genotypes, which included mainly upland varieties, detected AP13 transcripts. Of the 15 393 probe sets that were designed for AP13 contigs of unknown strandedness, less than half (6356 or 41.3%) detected transcripts, as would be expected given that half are not complementary to actual transcripts. In contrast, 72.9% of all AP13-derived probe sets detected transcripts from the switchgrass tissues analyzed. To test whether probe sets designed from PviUTs representing the extreme 5′-end of transcripts were less effective at detecting transcripts than probe sets designed from the 3′-end, we used foxtail millet genes as a reference to separate full-length and partial UT assemblies of AP13. Of the 10 724 probe sets derived from 5′-end partial AP13 transcripts (missing entire 3′-UTRs), only 5730 (53.4%) detected transcripts in the switchgrass tissues analyzed. In contrast, transcripts were detected by 10 424 (77.1%) of the 13 536 probe sets designed from partial sequences representing the 3′-end of transcripts (missing entire 5′-UTRs).

Similarity between transcriptomes of different organs and tissues was assessed by Pearson correlation analysis, taking into account all genes expressed in at least one organ. The resulting correlation matrix revealed clustering of functionally related organs (Figure 4). Roots at different developmental stages (vegetative stages V1–V5 and elongation stage E4) and the crown (E4) clustered together; germinating seeds/seedlings (24–96 h) clustered together; inflorescences at different stages (S1–S4) clustered together and clustered weakly with germinating seeds/seedlings; shoots at various stages of development (V1–V3) clustered together strongly, and slightly less so with leaf blades and sheaths (E4). Stem nodes and internode segments (E4) clustered together strongly, and clustered weakly with shoots, leaf blades, and sheaths (Figure 4).

Genes that are expressed specifically in an organ or tissue, and genes that are expressed at substantially higher levels in some organs/tissues than in others can provide insight into specialized processes at work in these organs/tissues. We sought to identify such genes by hierarchical clustering of global transcriptomic data (Figure 5). Using a transcript ratio of >2 as a filter, we found 4940 genes to be more highly expressed in roots than in any other organ of switchgrass (Figure 5). Likewise, 3372 genes were more highly expressed in inflorescences (stages S1–S4) than any other organ; 2766 genes were more highly expressed in germinating seeds; and 7595 genes were more highly expressed in the green organs, shoots (V1–V3), leaf blades and sheaths (E4), stem nodes and internode segments (E4), which grouped together in Figure 5. After sorting these 'marker' genes into Gene Ontology (GO) categories using Plant_GOslim ancestor terms (http://www.ebi.ac.uk/QuickGO/GMultiTerm#tab=edit-terms), we found that genes

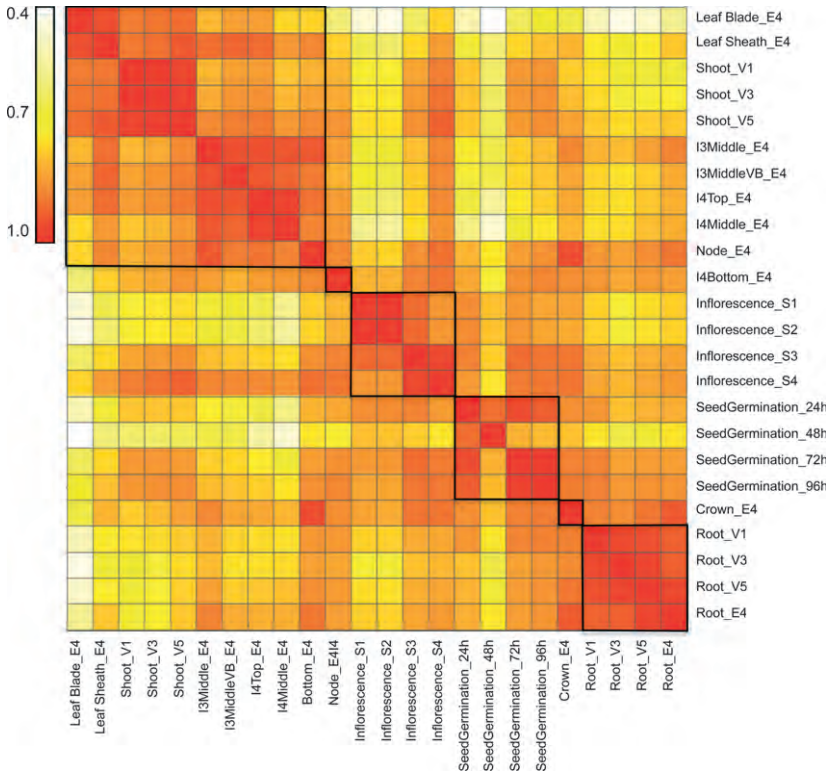
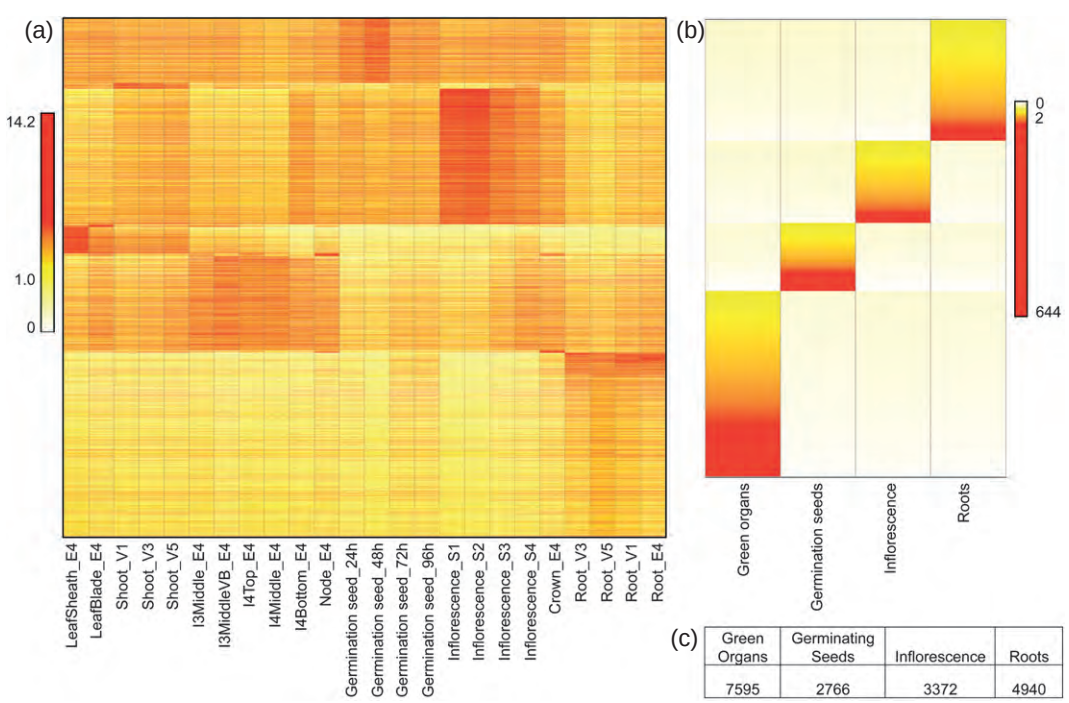


Figure 4. Comparison among transcriptomes of various switchgrass organs.
Pair-wise Pearson correlation coefficients (CCs) were calculated from transcript levels of all expressed genes for all organs. Average CCs were used for hierarchical clustering with the unweighted average (UPGMA) method. The color scale indicates the degree of correlation between transcriptomes of each pair of organs.



Green Organs	Germinating Seeds	Inflorescence	Roots
7595	2766	3372	4940

Figure 5. Organ-enhanced gene expression.
Genes with transcript levels at least twice as high in the indicated organ than in any other organs, which corresponded to 13 962 probe sets, were selected and clustered (a). The color scale in (a) indicates the $\log_2$-transformed ratio of transcript level in this organ to the highest level in any other organ for each gene.
To identify organ-specific genes, the same approach was applied to data accumulated from multiple samples of green organs, germinating seeds, inflorescences, or roots (b). The color scale in (b) indicates the ratio of transcript level in this organ group to the highest level in any others for each gene.
The total number of probe sets representing 'organ-specific' genes in each group is indicated in (c).

involved in carbohydrate metabolism (GO 0005975), transport (GO 0005215), and binding (GO 0005488) were over-represented in roots; genes for DNA metabolism (GO 0006259) and cell communication (GO 0007257) were over-represented in inflorescences; and genes for nucleobase, nucleoside, nucleotide, and nucleic acid metabolism (GO 0006139) and binding (GO 0005488) were over-represented during seed germination (Table S9). These subsets of developmentally regulated genes will be a useful starting point for research into biological processes that define the form and function of specific organs or groups of organs.

Identification of lignin biosynthesis genes

An important goal of switchgrass genomics research is to generate knowledge of lignin biosynthesis, which will facilitate rational approaches to reduce cell wall lignification and the natural recalcitrance of biomass to enzymatic digestion for ethanol production (Simmons *et al.*, 2010). Genes encoding 10 key enzymes involved phenylpropanoid–monolignol biosynthesis have been identified in plant species (Guo *et al.*, 2001; Boerjan *et al.*, 2003; Vanholme *et al.*, 2010). Using annotated phenylpropanoid–monolignol biosynthesis genes of Arabidopsis as query sequences, we identified 324 homologous PviUT sequences, via BLASTX search (*E*-value <0.01). Most of these (298 genes) had been annotated during establishment of the PviUT database. They included genes encoding phenylalanine ammonia-lyase (PAL, 35 sequences), coumaroyl shikimate 3′-hydroxylase (C3′H, 10), cinnamate 4-hydroxylase (C4H, 7), ferulate 5-hydroxylase (F5H, 5), 4-coumarate:CoA ligase (4CL, 58), cinnamoyl CoA reductases (CCR, 87), hydroxycinnamoyl CoA:shikimate hydroxycinnamoyl transferase (HCT, 45), caffeoyl-CoA 3-*O*-methyltransferase (CCoAMOT, 19), caffeic acid 3-*O*-methyltransferase (COMT, 16), and cinnamyl alcohol dehydrogenase (CAD, 42). The corresponding protein sequences were further scrutinized by phylogenetic analyses together with known proteins from

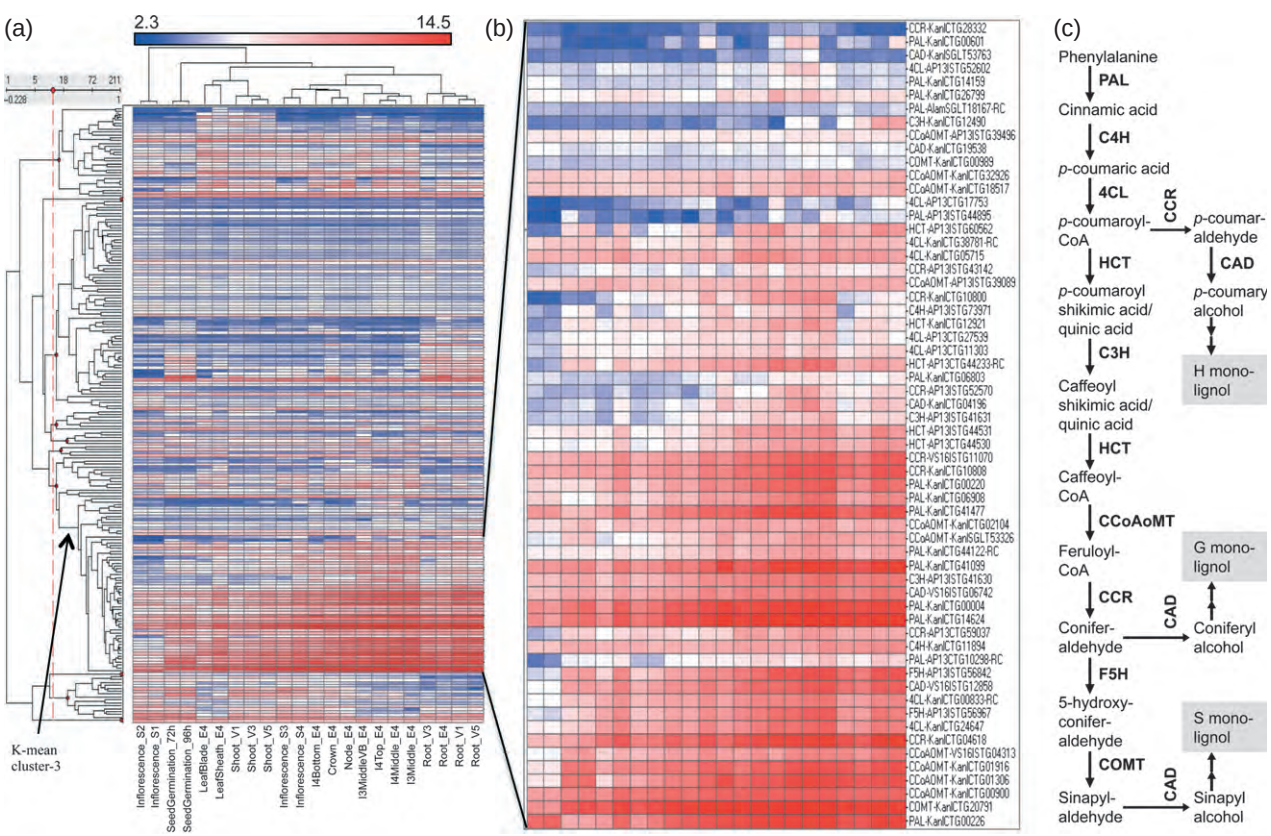


Figure 6. Hierarchical clustering of phenylpropanoid–monolignol biosynthesis pathway genes.
Transcript levels of the putative lignin biosynthesis pathway genes in various organs at multiple stages of plant development were subjected to hierarchical clustering analysis. Both genes and organs were clustered using the Pearson correlation method (a).
A cluster of 60 integrated switchgrass unitranscripts (*K*-means cluster no. 3 of 8) was implicated in lignin biosynthesis based on the correlation between tissue lignification and gene expression (b). This cluster contains all 10 known enzyme-encoding genes of the lignin pathway. The color scale indicates $\log_2$-transformed transcript level.
The phenylpropanoid–monolignol biosynthesis pathway and enzyme nomenclature in (c) is based on published work (Simmons *et al.*, 2010; Vanholme *et al.*, 2010).

Arabidopsis and maize, and outlier switchgrass proteins were discarded. As a result, we annotated 283 putative phenylpropanoid–monolignol biosynthesis genes in switchgrass (Table S10).

We used the switchgrass GEA to identify which of the putative phenylpropanoid–monolignol biosynthesis genes are likely to play key roles in lignin production. Transcripts of 211 (74.6%) of these genes were detected in at least one organ. Hierarchical clustering of these genes based on their transcript levels revealed a cluster of 60 genes that was expressed in all lignified organs/tissues (Figure 6). Interestingly, transcript levels of many of these were lowest in inflorescence meristems and in germinating seedlings, which exhibit little lignification. The highest expression of these genes was found in highly lignified organs, especially roots and stems. Furthermore, the expression of most of these 60 genes increased during maturation of different organs, including inflorescences of S1–S4, and shoots of V1–V5. Lignin content is correlated with tissue and organ maturation, particularly in stems (Shen *et al.*, 2009). Therefore, the developmental regulation of this set of 60 genes is consistent with roles in monolignol biosynthesis. Notably, this set contains representatives of a full complement of genes required for phenylpropanoid–monolignol biosynthesis (Table S11). Of these, one of the eight 4*CL* genes, both *COMT* genes, and one of the six *CAD* genes have been proven to be involved in switchgrass lignin biosynthesis (Fu *et al.*, 2011; Saathoff *et al.*, 2011; Xu *et al.*, 2011). Thus, the PviGEA serves to identify genes involved in important biological processes in switchgrass.

To facilitate exploration of the PviGEA we have developed a web server that enables flexible, multifaceted analyses of transcript data and provides a range of additional information about genes, including annotation that helps users formulate hypotheses about gene function. Transcript data can be accessed with an Affymetrix probe identification number, DNA sequence, gene name, functional description in natural language, putative functional domains, GO and KEGG annotation terms, and annotation based on BLASTX results against UniProt. Flexible tools to select a subset of experiments and to visualize and compare expression profiles of multiple genes have been implemented. Data can be downloaded in tabular form for use by common analytical and visualization software. The web server will be updated regularly with new gene expression data and annotations. Importantly, the architecture of PviGEA enables it to handle RNA-seq data also, which means that it can serve as a 'one-stop-shop' for switchgrass transcriptomics. To this end, we plan to import additional switchgrass transcriptome data into PviGEA, as they become available. The PviGEA server is accessible at http://switchgrassgenomics.noble.org/.

EXPERIMENTAL PROCEDURES

Plant materials and propagation

To minimize microbial contamination of plant samples destined for sequencing, an axenic, *in vitro* plant propagation protocol was used (Alexandrova *et al.*, 1996). Antibiotic (100 mg L^{-1} Timentin) and fungicide (10 mg L^{-1} nystatin) were included in MS-based media (Murashige and Skoog, 1962), for shoot initiation from node, shoot growth, and rooting (Figure S2). After 2 months (for Summer VS16) or 4 months (for Alamo AP13) of *in vitro* culture, rooted plants were transferred to pre-autoclaved MetroMix 300 substrate (Sungro® Horticulture, http://www.sungro.com/) and grown in a walk-in growth chamber at 30/26°C day/night temperature with a 16-h photoperiod (250 μM m^{-2} sec^{-1}). Half-strength modified Hoagland basal salt mixture (Hoagland and Arnon, 1950) was used once a week as fertilizer.

Organs were harvested at six developmental stages, including leaf development (VLD: V2), stem elongation (STE: E2 and E4), and reproductive phases (REP: R2, S2, and S6) (Moore *et al.*, 1991).

Plant stress treatments

Abiotic stresses were applied to plants at the E2 stage. After withholding water for 5, 10, and 20 days, plants showed mild, moderate, and severe drought stress symptoms. Volumetric water content of the soil/sand (3:1) mixture in pots at harvest times was 10–12, 6–8 and 2–4%, respectively. We isolated RNA from drought-stressed and rewatered plants (24 h post-rewatering) and pooled it into shoot and root RNA samples. Salt stress was applied by adding 500 mM NaCl to the fertilizer solution. Shoots and roots were sampled after 1 and 24 h of salt treatment. Cold stress was applied by sequentially decreasing the growth chamber temperature each day from 30/26°C (day/night) to 24/24°C, 18/18°C, 12/12°C, and 10/10°C on the final day. Heat stress was applied by increasing the temperature each day from 30/26°C to 35/30°C, 39/35°C, and 44/39°C on the final day. Shoots and roots were sampled at the last two cold or heat treatments.

RNA isolation, cDNA synthesis, and sequencing

Total RNA was isolated using a cetyltrimethyl ammonium bromide protocol followed by LiCl purification (Chang *et al.*, 1993) and quality controlled using a bioanalyzer (Agilent 2100, http://www.agilent.com/).

For 454-sequencing, mRNA was prepared by one round of oligo-(dT) purification. First-strand cDNA was synthesized using $dT_{15}VN_2$ primer and SuperScript III enzyme. Double-stranded cDNA was synthesized using the RNA replacement approach with *Escherichia coli* DNA ligase, DNA polymerase I, and RNase H. The resulting cDNA was fragmented by sonication. Adaptor ligation, single-stranded template DNA preparation, immobilization, emulsion PCR, and subsequent sequencing were carried out according to the manufacturer's instructions (454 Life Sciences). The 454-EST sequences were trimmed to remove adaptor, vector, and polyA/T sequences. The ESTs of low quality and low complexity were removed. The remaining sequences were compared with the NCBI nucleotide database to identify and remove non-cellular sequences (any hit to categories including Viroid, Virus, and Unclassified). Plant organelle sequences were identified and removed using the SeqClean (http://compbio.dfci.harvard.edu/tgi/software/) program with sorghum and maize chloroplast and mitochondria genome sequences as references. For AP13, ESTs shorter than 100 bp were also excluded.

Normalized, full-length-enriched cDNA libraries for Sanger sequencing were made from RNA of genotype AP13 from various tissues, developmental stages, and treatments (Table 2). Total RNA was used to synthesize cDNA using the SMART approach (Zhu *et al.*, 2001). Amplified cDNA was normalized using the duplex-specific nuclease normalization method (Zhulidov *et al.*, 2004), followed by size selection to enrich full-length cDNA.

Cloned cDNA was sequenced from both the 5′ and 3′ ends, using pDNR-LIB dir and rev primers, respectively, with an ABI 3730 DNA analyzer. Raw sequences were trimmed to remove vector sequences and low-quality regions. Poly-A or -T tracts near the ends of sequences were also removed. Trimmed ESTs <100 bases in length were set aside. The remaining EST sequences were queried against the GenBank nucleotide database, via BLAST, to identify and remove non-plant and plant-organellar sequences, as described above.

Sequence assembly and analysis

De novo assembly of 454-EST sequences was performed using NEWBLER (version 2.3) with the '-cDNA' option. For stringent assembly, a minimum overlap of 100 bp with at least 99% identity was required to join two sequences. A PERL script was written to scan the quality of the output sequences. A 10-base sliding window was used and the sequence quality score (QS) for each 10-base sequence was calculated. Windows with QS <30 were cut off and removed.

For AP13, Sanger-ESTs and 454-contigs and singletons were assembled using PAVE (Soderlund *et al.*, 2009), with the following parameter settings: SELF_JOIN = 40 97 20p; CLIQUE = 200 97 20; TC1 = 200 97 20; TC2 = 150 97 20; TC3 = 100 97 20 (Figure 1). Other switchgrass Sanger-ESTs were downloaded from NCBI-dbEST, grouped based on genotype, and assembled using PAVE as described above.

Five UT data sets were generated (Figure S3). A universal, low-redundancy switchgrass UT set (PviUT) was produced from the cultivar/accession assemblies in an iterative manner: five subsets were compared pair-wise and sequentially using the BLASTN program. Redundant sequences were identified as sharing 80% overlap and 90% identity and the longest or highest-quality sequence was retained. The PviUT sequences were orientated from 5′- to 3′- using information about insert orientation vectors, the polarity of matching DNA (i.e. 5′–3′) or protein (N- to C-terminal) sequences in GenBank. When orientation couldn't be determined, both forward and reverse complementary sequences were included.

Putative full-length cDNA sequences were identified by comparing them to predicted coding (CDS) and transcript sequences of sorghum (*S. bicolor*), foxtail millet (*S. italica*), and maize (*Z. mays*; http://www.phytozome.net/, released on 5 November 2010). Briefly, 5′-end reads were compared with reference sequences using BLAST with an *E*-value $\leq 1.0 \times 10^{-4}$ to identify homologs. A query start preceding the target start was taken as evidence for a full-length cDNA clone.

Unitranscript deduced protein sequences were annotated based on the presence of conserved domains in the CCD database at NCBI (Marchler-Bauer *et al.*, 2003, 2011; http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), matches to hidden Markov models (HMMs) at PANTHER (Mi *et al.*, 2010; http://www.pantherdb.org/), and results of BLASTP-searches against KEGG pathway genes for Arabidopsis, rice, and sorghum. To annotate PviUTs with GO terms, sequences were blasted against the UniProtKB database (http://www.uniprot.org/help/uniprotkb; *E*-value threshold 10^{-6}). The Plant_GOslim tool at http://www.animalgenome.org/bioinfo/tools/catego/ was used to classify genes with correlated expression patterns (Hu *et al.*, 2008).

Switchgrass cDNA chip design and gene expression atlas

An Affymetrix cDNA-format microarray chip was designed based on PviUT version 1.2 sequences and orientation, primarily using 600 bp at the 3′-end of each PviUT as targets. A 49-format chip design with 11-μm feature size was used. Eleven 25-mer probes were assigned to each PviUT where possible. No mismatch probes were included in the design.

Plant material for the switchgrass gene expression atlas (PviGEA) was derived from seeds of AP13 plants that were pollinated by other Alamo individuals. The RNA was isolated from tissues pooled from at least six individual plants. For the germination time series (Figure S7), seeds were sterilized with 20% commercial bleach prior to germination in the dark at 27 ± 1°C. Seedlings were planted in synthetic soil sunshine professional growing media (Sungro® Horticulture), and grown in a greenhouse with a 16/8 h and 29/24°C day/night cycle. Half-strength Hoagland fertilizer was used once a week. Shoots and roots were harvested separately at multiple vegetative and leaf developmental stages. Four stages of seed germination (24 h after imbibition through to 96 h) and three stages of shoot and root vegetative growth (V1, V3, and V5) were profiled to determine gene expression levels. Inflorescence development was divided into four major stages covering early meristem/primordium initiation (Inflorescence–S1), floret formation (Inflorescence–S2), rachis elongation (Inflorescence–S3), and panicle emergence (Inflorescence–S4) (Figure S8). For spatial gene expression profiling, plants at the stem elongation stage (E4–E5) were used. Whole roots and crowns were sampled separately. Leaf blades, leaf sheaths, nodes, and internode were harvested separately from the E4 stage tillers (Shen *et al.*, 2009). Internode 4 (E4i4) was excised and the top (t, 20% of the total length.), middle (m, 20%), and bottom (b, 20%) fragments were sampled separately. The middle fragment of internode 3 (E4i3 m) was used to isolate the vascular bundle tissue (Figure S9). Sample designation and tissue description are summarized in Table S6.

ACKNOWLEDGEMENTS

We thank Dr Will Nelson of the University of Arizona/Bio5 for help using PAVE. This work was supported by the BioEnergy Science Center, a US Department of Energy Bioenergy Research Center, through the Office of Biological and Environmental Research in the DOE Office of Science. The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under contract no. DE-AC02-05CH11231.

ACCESSION NUMBERS

All 454-ESTs obtained in this study have been submitted to NCBI's short read archive (SRA, accession numbers listed in Table 1) and the Sanger-ESTs to dbEST (accession numbers in Table 2). In addition, all data can be retrieved and explored at http://switchgrassgenomics.noble.org, including PviESTs, PviUTs, PviGEA, and GBrowser view of alignments of ESTs to PviUTs.

AUTHORS' CONTRIBUTIONS

MKU, YT, RAD, MS, and ECB planned the research. MKU, YT, and JYZ designed experiments and the data process-

ing strategy. JYZ, JH, and YS selected data processing methods. JYZ, YCL, IT-J, HS, and ACS carried out the laboratory experiments. CP, EL, JG, and JS made the libraries, prepared samples for sequencing, processed, and submitted the sequence data to NCBI. MW, YY, WCC, JH, and YS carried out sequence data analysis. MW, JH, and JYZ developed the switchgrass genomics database. LEB and PCR provided bacterial artificial chromosome sequencing and gene model prediction. JYZ, YT, and MKU interpreted the results and drafted the manuscript. All authors proofread and approved the manuscript.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Length distribution of VS16 454 expressed sequence tags.

Figure S2. Switchgrass genotype AP13 (a, b) and VS16 (c, d) node cultures on media with antibiotics and fungicide.

Figure S3. Homology of switchgrass AP13 unitranscripts compared with the reference genome.

Figure S4. Assembly of subsets of expressed sequence tags and pooling strategy for integrated switchgrass unitranscripts.

Figure S5. BLASTN *E*-values of integrated switchgrass unitranscripts versus annotated foxtail millet transcript sequences

Figure S6. Functional annotation of integrated switchgrass unitranscripts using the Conserved Domain Database.

Figure S7. Switchgrass seed germination and time points for sampling.

Figure S8. Stages of switchgrass inflorescence development.

Figure S9. Dissection of switchgrass vascular fiber from an internode (a) and a magnified view (b).

Figure S10. Signal intensity distribution of Affymetrix probe sets.

Figure S11. Number of probe sets detecting transcripts from various switchgrass organs.

Table S1. Microbial sequence contamination and estimation of 454 expressed sequence tag-derived consensus sequences.

Table S2. 454 expressed sequence tag assembly of VS16 with different programs.

Table S3. NEWBLER assembly metrics of 454 expressed sequence tags of VS16 with different parameter settings.

Table S4. Expressed sequence tag assembly and sequence resources used for the switchgrass unique transcript curation.

Table S5. Estimated transcript and transcriptome coverage of switchgrass unitranscripts.

Table S6. Alamo-AP13 × Alamo (non-AP13) tissues included in the gene expression atlas.

Table S7. Summary of transcript sequence sources of probe sets and present/absent in the switchgrass gene atlas.

Table S8. Most stably expressed switchgrass genes.

Table S9. Organ-enhanced genes.

Table S10. Putative phenylpropanoid–monolignol biosynthesis genes and normalized average transcript levels.

Table S11. Candidate phenylpropanoid–monolignol biosynthesis pathway genes in switchgrass and expression in 24 organs or tissues.

REFERENCES

Abrouk, M., Murat, F., Pont, C. *et al.* (2010) Palaeogenomics of plants: synteny-based modelling of extinct ancestors. *Trends Plant Sci.* **15**, 479–487.

Alexandrova, K.S., Denchev, P.D. and Conger, B.V. (1996) In vitro development of inflorescences from switchgrass nodal segments. *Crop Sci.* **36**, 175–178.

Benedito, V.A., Torres-Jerez, I., Murray, J.D. *et al.* (2008) A gene expression atlas of the model legume *Medicago truncatula*. *Plant J.* **55**, 504–513.

Bennett, M.D., Bhandol, P. and Leitch, I.J. (2000) Nuclear DNA amounts in angiosperms and their modern uses – 807 new estimates. *Ann. Bot.*, **86**, 859–909.

Boerjan, W., Ralph, J. and Baucher, M. (2003) Lignin biosynthesis. *Annu. Rev. Plant Biol.* **54**, 519–546.

Bouton, J.H. (2007) Molecular breeding of switchgrass for use as a biofuel crop. *Curr. Opin. Genet. & Devel.* **17**, 553–558.

Chang, S., Puryear, J. and Cairney, J. (1993) A simple and efficient method for isolating RNA from pine trees. *Plant Mol. Biol. Rep.* **11**, 113–116.

Chaudhary, B., Flagel, L., Stupar, R.M., Udall, J.A., Verma, N., Springer, N.M. and Wendel, J.F. (2009) Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (*Gossypium*). *Genetics*, **182**, 503–517.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Muller, W.E., Wetter, T. and Suhai, S. (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* **14**, 1147–1159.

Chuck, G.S., Tobias, C., Sun, L. *et al.* (2011) Overexpression of the maize Corngrass1 microRNA prevents flowering, improves digestibility, and increases starch content of switchgrass. *Proc. Natl Acad. Sci. USA*, **108**, 17550–17555.

Czechowski, T., Stitt, M., Altmann, T., Udvardi, M. and Scheible, W. (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. *Palnt Physiol.* **139**, 5–17.

Fu, C., Mielenz, J.R., Xiao, X. *et al.* (2011) Genetic manipulation of lignin reduces recalcitrance and improves ethanol production from switchgrass. *Proc. Natl Acad. Sci. USA*, **108**, 3803–3808.

Guo, D., Chen, F., Inoue, K., Blount, J.W. and Dixon, R.A. (2001) Downregulation of caffeic acid 3-O-methyltransferase and caffeoyl CoA 3-O-methyltransferase in transgenic alfalfa: impacts on lignin structure and implications for the biosynthesis of G and S lignin. *Plant Cell*, **13**, 73–88.

Hoagland, D. and Arnon, D. (1950) The water-culture method for growing plants without soil. In *Circular*. Univ. of Calif. Agric. Exp. Station, Berkley, CA, pp. 4–32.

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

Hu, Z.-L., Bao, J. and M., R.J. (2008) CateGOrizer: a web-based program to batch analyze gene ontology classification categories. *Online J. Bioinformatics*, **9**, 108–112.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Keshwani, D.R. and Cheng, J.J. (2009) Switchgrass for bioethanol and other value-added applications: a review. *Bioresour. Technol.* **100**, 1515–1523.

Kumar, S. and Blaxter, M. (2010) Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics*, **11**, 571.

Lemus, R., Brummer, E.C., Moore, K.J., Molstad, N.E., Burras, C.L. and Barker, M.F. (2002) Biomass yield and quality of 20 switchgrass populations in southern Iowa, USA. *Biomass Bioenergy*, **23**, 433–442.

Marchler-Bauer, A., Anderson, J.B., DeWeese-Scott, C. *et al.* (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* **31**, 383–387.

Marchler-Bauer, A., Lu, S., Anderson, J.B. et al. (2011) CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.* **39**(D22), 5–D229.

Martinez-Reyna, J.M., Vogel, K.P., Caha, C. and Lee, D.J. (2001) Meiotic stability, chloroplast DNA polymorphisms, and morphological traits of upland x lowland switchgrass reciprocal hybrids. *Crop Sci.* **41**, 1579–1583.

McLaughlin, S.B. and Kszos, L.A. (2005) Development of switchgrass (*Panicum virgatum*) as a bioenergy feedstock in the United States. *Biomass Bioenergy*, **28**, 515–535.

Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S. and Thomas, P.D. (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.* **38**, D204–D210.

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

Moore, K.J., Moser, L.E., Vogel, K.P., Waller, S.S., Johnson, B.E. and Pedersen, J.F. (1991) Describing and quantifying growth stages of perennial forage grasses. *Agron. J.* **83**, 1073–1077.

Murashige, T. and Skoog, F. (1962) A revised medium for rapid growth and bio assays with tobacco tissue cultures. *Physiol. Plant.* **15**, 473–497.

Okada, M., Lanzatella, C., Saha, M.C., Bouton, J., Wu, R. and Tobias, C.M. (2010) Complete switchgrass genetic maps reveal subgenome collinearity, preferential pairing and multilocus interactions. *Genetics*, **185**, 745–760.

Ouyang, S., Zhu, W., Hamilton, J. *et al.* (2007) The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Res.* **35**, D883–D887.

Paterson, A.H., Bowers, J.E., Bruggmann, R. *et al.* (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, **457**, 551–556.

Pertea, G., Huang, X., Liang, F. *et al.* (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.

Saathoff, A.J., Sarath, G., Chow, E.K., Dien, B.S. and Tobias, C.M. (2011) Downregulation of cinnamyl-alcohol dehydrogenase in switchgrass by RNA silencing results in enhanced glucose release after cellulase treatment. *PLoS One*, **6**, e16416.

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

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D. and Lohmann, J.U. (2005) A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet.* **37**, 501–506.

Sekhon, R.S., Lin, H., Childs, K.L., Hansey, C.N., Buell, C.R., de Leon, N. and Kaeppler, S.M. (2011) Genome-wide atlas of transcription during maize development. *Plant J.* **66**, 553–563.

Shen, H., Fu, C., Xiao, X., Ray, T., Tang, Y., Wang, Z. and Chen, F. (2009) Developmental control of lignification in stems of lowland switchgrass variety Alamo and the effects on saccharification efficiency. *Bioenergy Res.* **2**, 233–245.

Simmons, B.A., Loqué, D. and Ralph, J. (2010) Advances in modifying lignin for enhanced biofuel production. *Curr. Opin. Plant Biol.* **13**, 312–319.

Soderlund, C., Johnson, E., Bomhoff, M. and Descour, A. (2009) PAVE: program for assembling and viewing ESTs. *BMC Genomics*, **10**, 400.

Srivastava, A., Palanichelvam, K., Ma, J., Steele, J., Blancaflor, E. and Tang, Y. (2010) Collection and analysis of expressed sequence tags derived from laser capture microdissected switchgrass (*Panicum virgatum* L. Alamo) vascular tissues. *Bioenergy Res.* **3**, 278–294.

Tanaka, T., Antonio, B.A., Kikuchi, S. *et al.* (2008) The rice annotation project database (RAP-DB): 2008 update. *Nucleic Acids Res.*, **36**, D1028–D1033.

The International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, **463**, 763–768.

Tobias, C., Twigg, P., Hayden, D., Vogel, K., Mitchell, R., Lazo, G., Chow, E. and Sarath, G. (2005) Analysis of expressed sequence tags and the identification of associated short tandem repeats in switchgrass. *Theor. Appl. Genet.* **111**, 956–964.

Tobias, C.M., Sarath, G., Twigg, P., Lindquist, E., Pangilinan, J., Penning, B.W., Barry, K., McCann, M.C., Carpita, N.C. and Lazo, G.R. (2008) Comparative genomics in switchgrass using 61,585 high-quality expressed sequence tags. *Plant Gen.* **1**, 111–124.

Triplett, J.K., Wang, Y., Zhong, J. and Kellogg, E.A. (2012) Five nuclear loci resolve the polyploid history of switchgrass (*Panicum virgatum* L.) and relatives. *PLoS One*, **7**, e38702.

Vanholme, R., Demedts, B., Morreel, K., Ralph, J. and Boerjan, W. (2010) Lignin biosynthesis and structure. *Plant Physiol.* **153**, 895–905.

Wang, L., Xie, W., Chen, Y. *et al.* (2010) A dynamic gene expression atlas covering the entire life cycle of rice. *Plant J.* **61**, 752–766.

Wang, Y., Samuels, T. and Wu, Y. (2011) Development of 1,030 genomic SSR markers in switchgrass. *Theor. Appl. Genet.* **122**, 677–686.

Wu, T.D. and Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

Xu, B., Escamilla-Treviño, L.L., Sathitsuksanoh, N., Shen, Z., Shen, H., Percival Zhang, Y.H., Dixon, R.A. and Zhao, B. (2011) Silencing of 4-coumarate:coenzyme A ligase in switchgrass leads to reduced lignin content and improved fermentable sugar yields for biofuel production. *New Phytol.* **192**, 611–625.

Yuan, J.S., Tiller, K.H., Al-Ahmad, H., Stewart, N.R. and Stewart, C.N. Jr (2008) Plants to power: bioenergy to fuel the future. *Trends Plant Sci.* **13**, 421–429.

Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R. and Siebert, P.D. (2001) Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques*, **30**, 892–897.

Zhulidov, P.A., Bogdanova, E.A., Shcheglov, A.S. *et al.* (2004) Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res.* **32**, e37.

