Theor Appl Genet (2006) 113:186–195
DOI 10.1007/s00122-006-0285-3

ORIGINAL PAPER

EST sequencing and phylogenetic analysis of the model grass *Brachypodium distachyon*

John P. Vogel · Yong Qiang Gu · Paul Twigg · Gerard R. Lazo · Debbie Laudencia-Chingcuanco · Daniel M. Hayden · Teresa J. Donze · Lindsay A. Vivian · Boryana Stamova · Devin Coleman-Derr

Received: 3 November 2005 / Accepted: 31 March 2006 / Published online: 18 May 2006

Abstract *Brachypodium distachyon* (*Brachypodium*) is a temperate grass with the physical and genomic attributes necessary for a model system (small size, rapid generation time, self-fertile, small genome size, diploidy in some accessions). To increase the utility of *Brachypodium* as a model grass, we sequenced 20,440 expressed sequence tags (ESTs) from five cDNA libraries made from leaves, stems plus leaf sheaths, roots, callus and developing seed heads. The ESTs had an average trimmed length of 650 bp. Blast nucleotide alignments against SwissProt and GenBank non-redundant databases were performed and a total of 99.9% of the ESTs were found to have some similarity to existing protein or nucleotide sequences. Tentative functional classification of 77% of the sequences was possible by association with gene ontology or clusters of orthologous group's index descriptors. To demonstrate the utility of this EST collection for studying cell wall composition, we identified homologs for the genes involved in the biosynthesis of lignin subunits. A subset of the ESTs was used for phylogenetic analysis that reinforced the close relationship of *Brachypodium* to wheat and barley.

Communicated by T. Lübberstedt

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00122-006-0285-3 and is accessible for authorized users.

J. P. Vogel (✉) · Y. Q. Gu · G. R. Lazo · D. Laudencia-Chingcuanco · D. M. Hayden · B. Stamova · D. Coleman-Derr
USDA Western Regional Research Center, 800 Buchanan St., Albany, CA 94710, USA
e-mail: jvogel@pw.usda.gov

P. Twigg · T. J. Donze · L. A. Vivian
University of Nebraska at Kearney, 905 W. 25th St., Kearney NE 68849, USA

Introduction

The application of model systems toward the study of both basic and applied problems in plant biology has become routine. Researchers have employed the model dicot *Arabidopsis thaliana* to study topics ranging from nutrient uptake and metabolism to plant–pathogen interactions. Unfortunately, due to its distant relationship to monocots, *Arabidopsis* is not suitable to study biological features unique to the grasses (e.g. cell wall composition). With its sequenced genome and established research community, rice can serve as a model grass for some applications. Unfortunately, as a specialized semi-aquatic tropical grass that diverged from the most important forage grasses and temperate grains approximately 50 million years ago (Gaut 2002), rice is not a good model for temperate grasses. Rice also has physical and physiological attributes that limit its utility as a model system. Its large size and long generation time make experiments involving growing large numbers of plants under controlled conditions very expensive. It is also challenging to grow rice under the conditions typically present in greenhouses in northern climates.

Brachypodium distachyon (*Brachypodium*) is a small temperate grass with all the attributes needed to be a modern model organism including simple growth requirements, fast generation time, small genome and self-fertility (Draper et al. 2001). *Brachypodium* is also

readily transformed by *Agrobacterium* or biolistics facilitating many biotechnological applications (Christiansen et al. 2005; Vogel et al. 2006). The haploid genome size of diploid *Brachypodium* is approximately 0.36 pg, slightly over twice the size of *Arabidopsis* (Bennett and Leitch 2005; Vogel et al. 2006). Thus, *Brachypodium* possesses one of the smallest genomes of any grass and is suitable for both functional and structural genomic research.

Brachypodium belongs to the subfamily Pooideae and diverged just prior to the radiation of the small grain crops and forage and turf grasses, making it ''sister'' to the temperate grasses of greatest economic significance (Kellogg 2001). This placement is based on limited sequence data from internal transcribed spacers (ITS), the 5.8S subunit of nuclear ribosomal DNA and the chloroplast *ndf*H gene (Catalán and Olmstead 2000; Hsaio et al. 1994). RFLP and RAPD markers also indicate a similar placement of *Brachypodium* (Catalan et al. 1995; Shi et al. 1993). However, as was seen with rice, phylogenies based on different genes can produce phylogenetic trees with different topologies (reviewed in Kellogg 1998). Thus, it would be desirable to establish the relationship of *Brachypodium* to the temperate grasses using additional genes.

Brachypodium can also serve as a model to study polyploidy because polyploid and diploid accessions are available. An interesting observation is that accessions that are hexaploid by chromosome counts only have approximately twice as much DNA as diploids (Vogel et al. 2006). This may have arisen through the loss of DNA after the event leading to polyploidy through the process called diploidization (Wolfe 2001). Alternatively, the hexaploid accessions may have arisen from the combination of ancestors with chromosomes of different sizes.

Partially sequencing large numbers of random cDNA clones to generate a collection of expressed sequence tags (ESTs) is a fast and efficient way to provide a wealth of genomic information about a particular species (Adams et al. 1991). Such ESTs can be used for functional genomic experiments including the construction of cDNA microarrays and reverse genetics through gene silencing. ESTs can also serve as the raw material from which molecular markers suitable for mapping experiments can be made. Currently, wheat, corn, rice, barley, sorghum and sugarcane all have > 200,000 ESTs deposited in the dbEST division of GenBank. These ESTs have been used for many projects including: microarrays, analysis of gene expression patterns, generation of molecular markers, and physical mapping (including the assignment of genes to delineated regions in the wheat genome). Here we report the construction of five *Brachypodium* cDNA libraries, the sequencing of 20,440 ESTs, the identification of transcripts for the genes involved in lignin monomer biosynthesis and the construction of a phylogenetic tree for *Brachypodium*, rice, wheat, barley, corn, sorghum, sugarcane, *Arabidopsis*, soybean, tomato, poplar and pine based on 11,118 bp from 20 highly expressed nuclear genes.

Materials and methods

Plant material and growth conditions

Two lines of diploid *B. distachyon* derived from a single wild collection were used for this study. One line, Bd21, underwent five generations of single seed-descent and was used for the stem plus sheath, leaf, callus and root cDNA libraries (Vogel et al. 2006). The developing seed head cDNA library was constructed using another line, Bd21-0, that had not undergone single seed-descent. Plants for the stem plus sheath, leaf and developing seed head libraries were grown in a greenhouse as described (Vogel et al. 2006). Stems plus leaf sheaths, and leaves were harvested from the same plants shortly after anthesis. Entire developing seed heads containing seeds that ranged in maturity from anthesis to almost mature as well as the subtending bracts and stems were harvested and mixed together. Callus was initiated and grown as described (Vogel et al. 2006). A mixture of embryogenic and non-embryogenic callus was used.

To isolate clean roots, plants were grown hydroponically using a raft prepared from aluminum foil and a ring of styrofoam. The raft was placed on the surface of water containing 1 g/l of Peter's 20-20-20 fertilizer (Scotts, Marysville, OH, USA). The fertilizer mix was contained in a 14 cm deep reservoir. Exposed water was covered with aluminum foil to prevent algal growth. Seeds were placed on paper towels and incubated at 4°C for 7 days to synchronize germination. The seeds were then placed at 22°C for 2 days to promote germination. Germinated seeds were inserted into small holes in the aluminum foil such that the emerging root was submerged in the fertilizer solution. The hydroponic apparatus was placed in a growth chamber under continuous illumination at 24°C. Roots were harvested after 3 weeks. All plant material used for library construction was flash frozen in liquid nitrogen prior to RNA extraction.

Molecular techniques

Total RNA was extracted from 4 g aliquots of stems plus sheaths, leaves, roots, developing seed heads, and callus tissue by first grinding in liquid nitrogen. The resulting powder was scraped into an RNase-free 50 ml tube and processed using Plant RNA Reagent (Invitrogen, Carlsbad, CA, USA). Poly A+ RNA was extracted from the total RNA samples using the FastTrackMAG Maxi mRNA isolation kit (Invitrogen, Carlsbad, CA, USA). The yield and integrity of total and poly A+ RNA were assessed by agarose gel electrophoresis and UV spectrophotometry at 260 nm (Beckman DU-640, Fullerton, CA, USA).

Library construction was performed using the Superscript II system for cDNA synthesis and cloning (Invitrogen, Carlsbad, CA, USA) with 5 μg of poly A+ RNA as starting material. First strand synthesis was performed using the *Not*I adapter-primer to add a *Not*I site on the downstream portion of the cDNA. Following second strand synthesis, the resulting cDNA was ligated to *Sal*I adapters, cut with *Not*I, size selected, and ligated into the pSPORT1 plasmid vector prior to transformation into OmniMax 2 T1 phage-resistant chemically competent cells (Invitrogen, Carlsbad, CA, USA). The resulting transformants were selected on LB agar plates containing 100 μg/ml ampicillin. Plasmid minipreps and sequencing using an M13 reverse sequencing primer was carried out as described (Tobias et al. 2005).

Bioinformatics

Raw sequence files were processed using the Phred base-calling program (Ewing and Green 1998; Ewing et al. 1998). Phred also trimmed the sequences based on data quality using a probability cutoff value of 0.05 (Phred score ≥ 20) to retain only the high quality segment of the sequence. The trimmed sequences were further processed to mask the ends of reads that contained vector and adapter sequence using the program Crossmatch (http://www.phrap.org). Masked sequences were then removed from the sequence and quality files using an in-house Perl script (Lazo et al. 2004). Sequences less than 100 bp in length after processing were excluded from analysis. Sequence tracefiles and quality scores are available at the project website (http://wheat.pw.usda.gov/bEST).

The BlastX algorithm was used to compare ESTs to the EBI UniProt database (Release 4.2). Gene ontology (GO) terms were assigned to individual ESTs by cross referencing the UniProt hits to GO terms using the association tables found on the Gene Ontology Consortium site (http://www.geneontology.org). ESTs without GO associations were compared to the NCBI non-redundant database (Release 144) using BlastX and BlastN. The matches to the non-redundant database were then matched to descriptions from the NCBI clusters of orthologous groups (COG) Index (http://www.ncbi.nlm.nih.gov/COG/). ESTs without GO terms or non-redundant matches were compared to ESTs contained in the dbEST database using BlastN. The Phrap assembly program was used to assemble the ESTs into contigs. The Phrap parameters used (penalty –5; minmatch 50; minscore 100) resulted in EST clusters of > 90% identity over a 100 bp window.

Identification of genes involved in lignin biosynthesis

Homologs for ten genes involved in the synthesis of monolignols (PAL, phenylalanine ammonia-lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-(hydroxy)cinnamoyl CoA ligase; CST, hydroxycinnamoyl CoA:shikimate hydroxycinnamoyltransferase; C3H, *p*-coumarate 3-hydroxylase; CCoAOMT, caffeoyl CoA *O*-methyltransferase; CCR, cinnamoyl CoA reductase; F5H, ferulate 5-hydroxylase; COMT, caffeic acid/5-hydroxyferulic acid *O*-methyltransferase; CAD, cinnamyl alcohol dehydrogenase) (reviewed in Humphreys and Chapple 2002) were identified through BLAST searches of the *Brachypodium* ESTs using known sequences for each gene. *Brachypodium* ESTs with highly significant BLAST scores were then used to query the GenBank non-redundant database. ESTs that hit the correct lignin biosynthetic gene were then compared to a list of ESTs contained in *Brachypodium* contigs to determine the number of tentatively unique genes.

Phylogenetic analysis

Brachypodium contigs containing the most ESTs were selected as candidates for highly expressed genes to be used for phylogenetic analysis. Several steps were then used to select 20 suitable genes from which to construct phylogenetic trees. First, candidate contigs were compared to the NCBI non-redundant database using the BlastN algorithm to select contigs with highly significant hits to known genes. In cases where multiple contigs matched the same gene only one contig was selected for further analysis. Contigs corresponding to known genes were then used to retrieve ESTs from barley (*Hordeum vulgare*), rice (*Oryza sativa*), sugarcane (*Saccharum officinarum*), sorghum (*Sorghum bicolor*), wheat (*Triticum aestivum*) and corn (*Zea mays*)

using BlastN with organism limits against dbEST. Four dicots, *A. thaliana*, soybean (*Glycine max*), tomato (*Lycopersicon esculentum*), poplar (*Populus trichocarpa* or *P. trichocarpa* × *deltoides*) and the gymnosperm pine (*Pinus taeda*) were included in the analysis to provide a buffer against artifacts in the grass lineage stemming from systematic bias. Pine also serves as an outgroup. The tBlastX algorithm was used to identify ESTs from soybean, tomato, poplar and pine that were the most similar to the *Brachypodium* genes. For three genes we could not identify suitable *P. trichocarpa* ESTs and used ESTs derived from the hybrid *P. trichocarpa* × *deltoides*. To identify the corresponding genes from *Arabidopsis*, the contigs were compared to the TAIR database of *Arabidopsis* proteins (http://www.arabidopsis.org/Blast/) using BlastX. For each individual contig, DNA sequences from the top four scoring ESTs for each of the six grasses, the top three scoring ESTs from soybean, tomato, poplar and pine and the coding sequence from *Arabidopsis* were aligned by the ClustalW algorithm using MegAlign software (DNAstar, Madison, WI, USA). A portion of the alignment that contained sequence information for all ESTs was selected to generate a phylogenetic tree using MegAlign. If ESTs from the same species were adjacent to one another in the cladogram, then the gene was selected for further study.

The partial coding sequences (average 556 bp/gene) from the 20 highly expressed genes selected above were then used for phylogenetic analysis. The aligned sequences were combined to produce one sequence for each species that was used for phylogenetic analysis (supplemental data S1). The consensus sequence was 11,118 bp. Phylogenetic trees based on this alignment were constructed using seven programs: MegAlign, ClustalX (ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW/; Thompson et al. 1997), and five programs (DNApars, DNAcomp, DNAml, DNApenny, DNAmlk) contained in the PHYLIP software package version 3.6 (Felsenstein 1989). Bootstrap analysis was conducted using the Seqboot (to create bootstrapped datasets) and Consense (to create a consensus tree) programs contained in the PHYLIP package along with the phylogeny analysis programs. Bootstrap analysis for ClustalX used the internal bootstrap feature.

Results

Library construction and sequencing

The cDNA libraries constructed were of high quality with approximately 400,000 clones per library. The average insert size based on restriction digests of 96 clones per library was 1.5 kb for all the libraries with the exception of the root library that had an average insert size of 1 kb (not shown).

In total, 20,587 high quality sequences were obtained from 27,648 sequence reads and deposited in the dbEST division of GenBank where they were assigned individual accessions numbers DV468879–DV489465. Subsequent to depositing the sequences in genbank, we identified 147 ESTs that matched a database of *E. coli* and grass plastid sequences and requested that Genbank delete these sequences leaving a total of 20,440 EST sequences. Summary statistics by library are presented in Table 1. The average length of high quality non-vector sequence was 650 bp (Table 1). Sequences from the stem plus sheath, root, callus, developing seed head and leaf libraries comprised 19, 19, 20, 23, and 19% of the collection, respectively.

Gene ontology

Of the 20,587 *B. distachyon* ESTs sequenced, 15,595 (76%) were assigned GO terms in any category (biological, cellular, molecular), 14,617 (71%) were assigned GO molecular terms, 13,770 (69%) were assigned GO biological terms, and 11,898 (58%) were assigned GO cellular terms (the complete list of GO

Table 1 *Brachypodium* EST summary

Library	Sequencing reads	High quality ESTs	Contigs	Singletons	Average length (bp)	Library-specific sequences[a]
Developing seed head	5,952	4,688	857	2,015	610	1,439
Leaf	4,608	3,780	600	1,721	661	1,095
Stem plus sheath	4,800	3,907	667	1,801	665	1,024
Root	4,800	3,869	638	1,932	654	1,461
Callus	7,488	4,196	712	1,749	665	1,222
Pooled ESTs	27,648	20,440	3,832	4,945	650	NA

[a]The library-specific sequences represents the number of singletons plus the number of contigs unique to that library after Phrap assembly of the pooled sequences

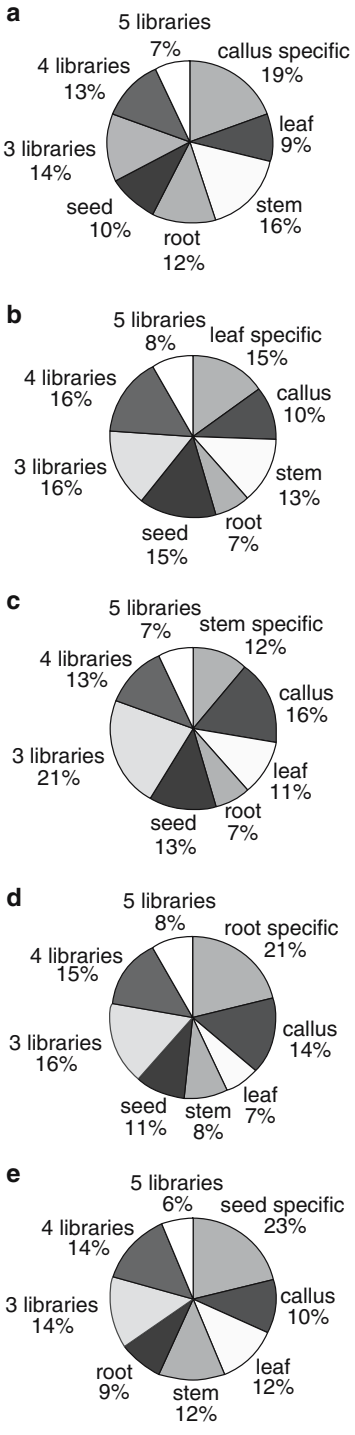

Fig. 1 Library contig comparisons. The percentage of contigs from individual libraries that were library specific, or contained ESTs from one, two, three or four other libraries. For contigs containing ESTs from one other library the other library is indicated. For contigs containing ESTs from more than two libraries the numbers for all combinations were pooled. The library and total number of contigs per library were **a** callus, 1,490 **b** leaf, 1,286 **c** stem plus sheath, 1,499 **d** root, 1,275 and **e** developing seed head, 1,545

matches is available at: http://wheat.pw.usda.gov/pubs/2006/Vogel/). Of the remaining 4,992 ESTs, 4,982 had matches to the NCBI non-redundant database; 6% of these were placed in functional categories as classified by the NCBI COG index and 94% were placed in unknown or genome categories (not shown). Of the ten ESTs that did not match the NCBI non-redundant database three of these matched sequences in the dbEST database and the rest were highly repetitive.

EST contig assembly and library comparisons

Phrap assembly of the 20,440 individual sequences resulted in 3,832 contigs and 4,945 singletons (Table 1). Thus, 8,777 tentatively unique genes (contigs + singletons) are contained in this collection. The average contig contained four ESTs and 273 contigs contained ten or more ESTs. The developing seed head and root libraries had the highest number of library-specific sequences (Table 1). To further assess the gene diversity in each library, the number of libraries contributing ESTs to each contig was determined (Fig. 1). The developing seed head and root libraries had the highest percentage of library unique contigs at 23 and 21%, respectively (Fig. 1d, e). The root library was the least similar to the stem plus sheath and leaf libraries with only 8 and 7% shared contigs, respectively.

Lignin biosynthetic genes

Since we plan to use *Brachypodium* as a model for the development of biofuels and lignin content has a negative impact on the conversion of biomass into ethanol, we wanted to know if the genes required for the biosynthesis of lignin monomers were present in the EST collection. We found *Brachypodium* homologs for all ten genes thought to be required for the biosynthesis of monolignols (Humphreys and Chapple 2002) (Table 2). The root and stem plus sheath libraries contained the greatest number of ESTs from lignin biosynthetic genes, 53 and 32, respectively. By contrast, the leaf library only contained eight ESTs from lignin biosynthetic genes. This is consistent with the higher degree of lignification and vascularization found in roots and stems as compared to leaves. When examining individual genes, PAL had the highest number of ESTs from all libraries (32 ESTs) which is not surprising given that PAL feeds into several biosynthetic pathways, not just lignin. We only found one EST for F5H suggesting that this enzyme is found in much lower abundance.

Table 2 Representation of lignin biosynthetic genes in *Brachypodium* ESTs

Genes	Stem plus sheath	Leaf	Root	Developing seed head	Callus	Total ESTs	Tentatively unique genes
PAL	12	0	10	6	4	32	6
C4H	0	0	4	1	1	6	4
4CL	0	1	4	0	1	6	4
CST or CQT	0	2	1	2	6	11	3
C3H	1	0	8	0	1	10	7
CCoAOMT	8	0	7	3	5	23	4
CCR	3	2	3	2	2	12	4
F5H	0	0	1	0	0	1	1
COMT	7	2	5	6	1	21	1
CAD	2	1	9	2	0	14	3
Library total	32	8	53	22	21	136	37

PAL phenylalanine ammonia-lyase, *C4H* cinnamate 4-hydroxylase, *4CL* 4-(hydroxy)cinnamoyl CoA ligase, *CST* hydroxycinnamoyl CoA:shikimate hydroxycinnamoyltransferase, *C3H p*-coumarate 3-hydroxylase, *CCoAOMT* caffeoyl CoA *O*-methyltransferase, *CCR* cinnamoyl CoA reductase, *F5H* ferulate 5-hydroxylase, *COMT* caffeic acid/5-hydroxyferulic acid *O*-methyltransferase, *CAD* cinnamyl alcohol dehydrogenase

Phylogenetic analysis

To better establish the relationship of *Brachypodium* to other grasses, we created a phylogenetic tree based on partial sequences from 20 highly expressed genes similar to: SAM decarboxylase, glyceraldehyde-3-phosphate dehydrogenase, 1-aminocyclopropane-1-carboxylate oxidase, 1,6-bisphosphate aldolase, chlorophyll a/b-binding protein CP26, catalase, α-tubulin, heat shock protein 70, cyclophilin, reversibly glycosylated polypeptide, cytosolic heat shock protein 90, aquaporin PIP1, putative chlorophyll a/b-binding protein type III, an unknown protein, dnaK-type molecular chaperone hsp70, alanine aminotransferase, 23 kDa oxygen evolving protein of photosystem II, endotransglucosylase/hydrolase XTH1, ATP/ADP translocator, and xylose isomerase. We chose to use highly expressed genes to maximize the chance of finding the corresponding gene in the species included in the analysis. Pine was included as an outgroup. Several steps were taken to select the 20 genes for phylogenetic analysis. First, 55 candidate *Brachypodium* contigs were compared to the NCBI nonredundant database using the BlastN algorithm. Of these, 43 contigs produced highly significant hits to the 36 different genes, 11 had no highly significant hits and one appeared to be chimeric. In cases where multiple contigs matched the same gene only one contig was selected for further analysis. Contigs corresponding to the 36 different genes were then used to retrieve the four most similar ESTs from each of the six grasses, the three most similar ESTs from soybean, tomato, poplar and pine and the most similar *Arabidopsis* gene. Nine of these genes were dropped from consideration because we could not reliably identify homologs outside of the grasses. For the remaining 27 genes the ESTs for each individual gene were aligned and used to create separate phylogenetic trees for each gene. ESTs from the same species fell onto different clades of the tree for seven of the genes examined indicating the presence of gene family members that would interfere with the analysis. These seven genes were dropped from the analysis leaving 20 genes for the phylogenetic analysis (Table 3).

The phylogenetic trees created using MegAlign for individual genes were different (not shown) underscoring the need to look at multiple genes to arrive at a correct phylogeny. To perform a more thorough phylogenetic analysis, the aligned sequences of all 20 genes for each species were combined (supplementary data S1). The consensus sequence of this alignment was 11,118 bp with an average of 556 bp per gene. Using this alignment, phylogenetic trees were created using seven different programs. With the exception of the tree produced by MegAlign, all trees were bootstrapped 1,000 times. The topology of the grass clade was identical in all the trees with the exception of rice which was placed basal to the split of the barley–wheat–*Brachypodium* clade and the corn–sorghum–sugarcane clade in the tree produced by DNA penny. The bootstrap values for the grass portion of the trees were extremely high. Thus, we are confident in the placement of the grasses in the tree presented in Fig. 2.

In contrast to the robust placement of the grasses within the tree, the placement of the dicots was ambiguous. No two programs gave the same topology and the bootstrap values were very low ranging from 420 to 700 out of 1,000. This lack of resolution is due to

Table 3 Accession numbers of sequences used for phylogenetic analysis

Top BLAST hit	Brachypodium	Barley	Corn	Rice	Sorghum	Sugarcane	Wheat	Arabidopsis	Pine	Poplar	Soybean	Tomato
SAM decarboxylase	DV488148	AL504307	CF009395	CK065613	BE365098	CA243775	CK162989	AT3G02470	DT625809	DT488524	BI785825	BM535147
Glyceraldehyde-3-phosphate dehydrogenase	DV482392	BM816319	CD435701	CX109155	CN151608	CA227717	DR740680	AT3G04120	DR080553	DT472087	CF807304	CN385802
1-Aminocyclopropane-1-carboxylate oxidase	DV485291	BE601938	CO532005	CB617947	DN551771	CA215264	DR738620	AT1G77330	DR102609	DN493561	CA851282	BI933301
1,6-Bisphosphate aldolase	DV483599	BF262521	DR806517	CB625801	CN149917	CA270328	DR740693	AT4G38970	DR024105	DT473175	BG839260	BG643424
Chlorophyll a/b-binding protein CP26	DV482455	BE421616	DR784988	CX109339	CN136506	CA069295	DR740953	AF339718	DR093721	CV243547	CA782594	AI776873
Catalase	DV483509	BF066021	DT643673	CB667927	CX616764	CA272363	CK209694	AT4G35090	DT627191	DT488723	BI945832	AW738726
A-tubulin	DV480590	BE412615	CD438536	CX109371	DN552321	CA106384	CK206375	AT4G14960	DT624400	DT497382	CA784105	BW689166
Heat shock protein 70	DV479856	BF065865	CD440431	CX109183	CN149101	CA226485	CK172239	AT5G02500	DR743409	DT502260	BE661281	DV104301
Cyclophilin	DV489188	CK125207	DT647159	CX109169	CX622874	CA297725	DR741122	AT2G21130	DN458251	DT500949	BI943088	BE462885
Reversibly glycosylated polypeptide	DV489003	CK125023	CO446222	CX109310	CN137898	CA160606	CK167017	AT5G15650	DR743576	BU878466[a]	CA785183	BW692954
Cytosolic heat shock protein 90	DV486551	BI948190	CD445816	CX109067	CF432583	CA223327	BU672348	AT5G56010	CO363395	CV233275	CF806664	BW689531
Aquaporin PIP1	DV478604	BF267876	DR824612	CB618084	CN130637	CA180028	CK162490	AT4G23400	CV134736	DT472648	BE607714	BW688737
Putative chlorophyll a/b-binding protein type III	DV483783	BJ473210	DR829827	CX109019	CN138219	CA183339	CK162712	AT1G61520	DR091581	DT476036	CA783804	AW737962
Unknown	DV488856	DN186964	DY235032	CX109221	CD227406	CA138248	CD864746	AT3G15450	CO362117	DT486808	CD408246	BI422271
dnaK-type molecular chaperone hsp70	DV485940	BM815970	CO459136	CB625798	CX614024	CA125208	CF133977	AT1G56410	DR060357	DT502260[a]	BG838733	BI934097
Alanine aminotransferase	DV487670	BJ472798	DV530870	CK082957	CF433549	CA088190	CV771481	AT1G70580	DT632782	DT502704[a]	BG839268	AW094682
23 kDa oxygen evolving protein of photosystem II	DV479214	DN177880	DV536708	CB621282	CX616974	CA294118	CK209006	AT1G06680	CO361238	DT486794	BG838271	AW093349
Endotransglucosylase/ hydrolase XTH1	DV476675	BF631110	CO453603	CF957628	AW564403	CA269869	CK208716	AT5G57550	DR014592	CV236471	CX548756	BM412485
ATP/ADP translocator	DV488729	CB870567	CO443247	CB650964	CD429771	CA265064	DR740190	AT5G13490	CO164871	DT481613	CX711873	BI932322
Xylose isomerase	DV472423	BQ468505	DV026491	CB661421	CN125028	CA196118	CK162381	AT5G57655	DT638327	BU875211	BU927122	BW692921

[a]Sequence derived from *P. trichocarpa* × *deltoides*. All other poplar sequences were derived from *P. trichocarpa*

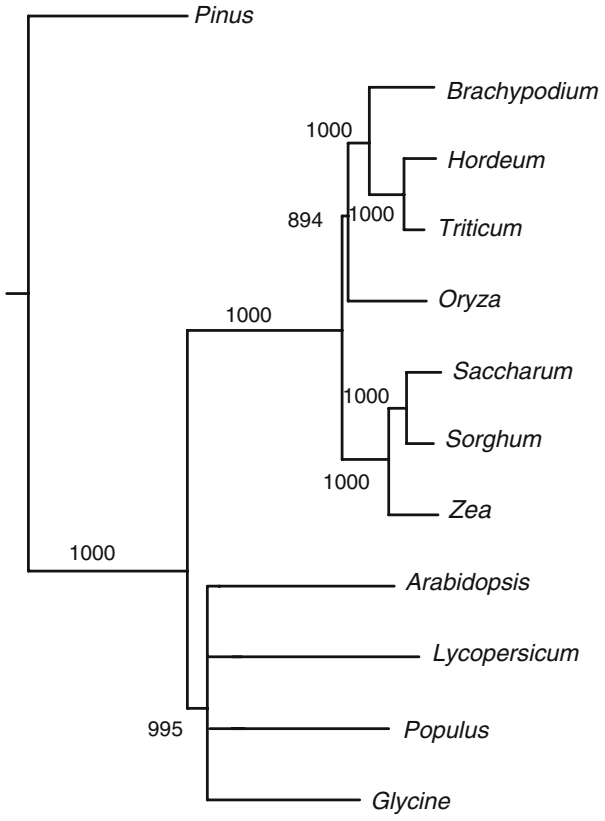

Fig. 2 Relationship of *Brachypodium* to other grasses. Rooted phylogentic tree based on the combined partial nucleotide sequences of 20 highly expressed genes. The tree was produced using clustalX set to ignore positions with gaps or ambiguities and correct for multiple substitutions. DNApars, DNAml, DNAmlk, DNAcomp and MegAlign all produced trees with identical topology of the grasses. The dicot lineage was not resolved because all programs used gave conflicting topologies with low bootstrap values. Bootstrap values out of 1,000 are presented. Branch length is proportional to sequence divergence

the low nucleotide sequence conservation among the dicots. The percent identity among the dicots ranged from 77.4 to 79.6% whereas the percent identity among the grasses ranged from 86.5 to 96.2%. Given this uncertainty, the dicots are presented as unresolved (Fig. 2). Since the role of the dicot sequences was to buffer against artifacts in the grass clade their placement in the tree is not critical.

Discussion

Temperate grasses are extremely important to humans because they supply a large percentage of our food either directly through the consumption of grains or indirectly through the consumption of grass-fed animals. Temperate grasses are also poised to play an increasing role in supplying energy due to the increasing economic, environmental and social expenses associated with petroleum-based fuels. The most important temperate grains and forage grasses are, for the most part, difficult experimental subjects in the laboratory. Taken together, these reasons point to the need for a model temperate grass that can be used to make rapid gains in our knowledge about the unique attributes of the temperate grasses. *Brachypodium* is well suited to serve as such a model grass. The EST sequences generated in this study greatly increase the *Brachypodium* sequence information available. *Brachypodium* now ranks ninth among the grasses in terms of ESTs contained in dbEST. While the modest size of our sequencing effort does not represent a comprehensive analysis of the expressed portion of the *Brachypodium* genome, it does provide enough information to begin functional genomic experiments and also provides the raw material for generating molecular markers.

We chose to sequence ESTs from cDNA libraries constructed from five different plant parts to maximize the number of genes represented in the EST collection. Our comparison of the libraries to one another supported the obvious biological differences between the materials used to make the libraries. For example, the developing seed head library had the highest percentage of library specific contigs which is likely due to the complex nature and specialized tissues contained in a seed head that are not found in the other libraries. As another example, the root library had the least overlap with the leaf and stem plus sheath libraries. This is not surprising given the photosynthetic nature of the leaf and stem.

Due to the increasing economic, environmental and social costs associated with the consumption of fossil fuels, ethanol derived from ligno-cellulosic biomass has become an attractive alternative fuel source. To produce ethanol from biomass, the sugars locked in the cellulose and hemicellulose fraction of the cell wall are degraded to monosaccharides using a combination of physical, chemical and enzymatic treatments and then fermented into ethanol. The lignin fractions of the cell wall is not converted into ethanol and, in fact, interferes with the conversion. Lignin also decreases the digestibility of forage grasses. Thus, manipulation of the lignin biosynthetic pathway is an obvious target for biotechnological enhancement of forage grasses and grasses grown specifically for conversion into ethanol (Chen et al. 2003). We identified homologs to all ten genes currently thought to be required for the biosynthesis of lignin precursors (reviewed in Humphreys and Chapple 2002). Thus, we now have the starting pie-

ces to rapidly evaluate different strategies for altering lignin content or composition in a temperate grass.

Our phylogenetic analysis based on 11,118 bp from 20 genes indicated that *Brachypodium* is much more closely related to wheat and barley than to corn, sugarcane or sorghum. This is consistent with previous reports (Kellogg 2001) based on much smaller data sets. That all seven computer programs used in our study arrived at the same conclusion with extremely high bootstrap values underscores the monophyletic nature of these two clades. In light of the strong support of the relationship among these grasses based upon the combined data set, it is interesting to note that some of the individual genes gave different phylogenetic trees. This underscores the importance of using large data sets from multiple genes to draw conclusions about phylogeny.

In contrast to the placement of the grasses, the position of the dicots within the phylogenetic trees was variable and not well supported by bootstrap analysis. In fact, none of the trees were identical to a phylogeny recently produced for the dicots (Judd and Olmstead 2004). The contrast between the robust clade containing the grasses and the unresolved relationship among the dicots is due to the greater sequence divergence observed among the dicots as compared to the divergence among the grasses. An analysis using protein rather than nucleotide sequence may have helped resolve the dicots, but that would have discarded information critical to resolving the highly related grasses. Since our goal was to define the relationship of *Brachypodium* within the grasses and dicot sequences were only added to act as a buffer against bias within the grasses, the lack of resolution in the dicot clade is acceptable.

We have developed a valuable resource for the emerging *Brachypodium* research community that can be used for functional genomic experiments, as a starting point for the development of molecular markers, and as anchor sequences for BAC-based physical maps. Our initial analysis of these ESTs confirmed the close relationship of *Brachypodium* to wheat and barley and identified homologs for all the genes required for lignin monomer biosynthesis.

Acknowledgements Supported by the United States Department of Agriculture, Agricultural Research Service CRIS 5325-21000-013-00, NP307 Biofuel and Bioenergy Alternatives. This work was also supported in part by NIH Grant P20 RR16569 from the BRIN Program of the National Center for Research Resources, and a University of Nebraska at Kearney Research Services Council University Research & Creative Activity Grant.

References

Adams M, Kelley J, Dubnick M, Polymeropoulos M, Xiao H, Merril C, Wu A, Olde B, Moreno R et al (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656

Bennett MD, Leitch IJ (2005) Nuclear DNA amounts in Angiosperms: progress, problems and prospects. Ann Bot 95:45–90

Catalán P, Olmstead RG (2000) Phylogenetic reconstruction of the genus *Brachypodium* P. Beauv. (Poaceae) from combined sequences of chloroplast *ndh*F gene and nuclear ITS. Plant Syst Evol 220:1–19

Catalan P, Ying S, Armstrong L, Draper J, Stace CA (1995) Molecular phylogeny of the grass genus *Brachypodium* P. Beuav. based on RFLP and RAPD analysis. Bot J Linn Soc 117:263–280

Chen L, Auh C-K, Dowling P, Bell J, Chen F, Hopkins A, Dixon RA, Wang Z-Y (2003) Improved forage digestibility of tall fescue (*Festuca arundinacea*) by transgenic down-regulation of cinnamyl alcohol dehydrogenase. Plant Biotech J 1:437–449

Christiansen P, Didion T, Andersen CH, Folling M, Nielsen KK (2005) A rapid and efficient transformation protocol for the grass *Brachypodium distachyon*. Plant Cell Rep 23:751–758

Draper J, Mur LAJ, Jenkins G, Ghosh-Biswas GC, Bablak P, Hasterok R, Routledge APM (2001) *Brachypodium distachyon*. A new model system for functional genomics in grasses. Plant Phys 127:1539–1555

Ewing B, Green P (1998) Basecalling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194

Ewing B, Hiller L, Wendl M, Green P (1998) Basecalling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175–185

Felsenstein J (1989) Phylogeny inference package (Version 3.2). Cladistics 5:164–166

Gaut BS (2002) Evolutionary dynamics of grass genomes. New Phytol 154:15–28

Hsaio C, Chatterton NJ, Asay KH, Jensen KB (1994) Phylogenetic relationships of 10 grass species: an assessment of phylogenetic utility of the internal transcribed spacer region in nuclear ribosomal DNA in monocots. Genome 37:112–120

Humphreys J, Chapple C (2002) Rewriting the lignin roadmap. Curr Opin Plant Biol 5:224–229

Judd WS, Olmstead RG (2004) A survey of tricolpate (eudicot) phylogenetic relationships. Am J Bot 91:1627–1644

Kellogg EA (1998) Relationships of cereal crops and other grasses. Proc Natl Acad Sci USA 95:2005–2010

Kellogg EA (2001) Evolutionary history of the grasses. Plant Phys 125:1198–1205

Lazo GR, Chao S, Hummel DD, Edwards H, Crossman CC, Lui N, Matthews DE, Carollo VL, Hane D, You FM, Butler E, Miller RE, Close TJ, Peng JH, Lapitan NLV, Gustafson JP, Qi LL, Echalier B, Gill BS, Dilbirligi M, Sandhu D, Gill KS, Greene RA, Sorrells ME, Akhunov ED, Dvorak J, Linkiewicz AM, Dubcovsky J, Hossain KG, Kalavacharla V, Kianian SF, Mahmoud AA, Miftahudin, Ma XF, Conley EJ, Anderson JA, Pathan MS, Nguyen HT, McGuire PE, Qualset CO, Anderson OD (2004) Development of an expressed sequence tag (EST) resource for wheat (*Triticum aestivum* L.): EST generation, unigene analysis, probe selection, and bioinformatics for a 16,000-locus bin-delineated map. Genetics 168:585–593

Shi Y, Draper J, Stace C (1993) Ribosomal DNA variation and its phylogenetic implication in the genus *Brachypodium* (Poaceae). Plant Syst Evol 188:125–138

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 22:4673–4680

Tobias CM, Twigg P, Hayden DM, Vogel KP, Mitchell RM, Lazo GR, Chow EK, Sarath G (2005) Analysis of expressed sequence tags and the identification of associated short tandem repeats in switchgrass. Theor Appl Genet 111:956–964

Vogel JP, Garvin DF, Leong OM, Hayden DM (2006) *Agrobacterium*-mediated transformation and inbred line development in the model grass *Brachypodium distachyon*. Plant Cell Tissue Organ Cult 85:199–211

Wolfe KH (2001) Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet 2:333–341