Theor Appl Genet (2004) 108:851–863
DOI 10.1007/s00122-003-1510-y

ORIGINAL PAPER

H.-M. Ma · S. Schulze · S. Lee · M. Yang · E. Mirkov · J. Irvine · P. Moore · A. Paterson

An EST survey of the sugarcane transcriptome

Received: 7 February 2003 / Accepted: 30 September 2003 / Published online: 29 November 2003

Abstract Its large genome and high polyploidy makes sugarcane *(Saccharum* spp.) a singularly challenging crop to study and improve using genetic approaches. To provide large numbers of functionally characterized candidate genes that might be tested for direct association (rather than distant linkage) with economically important traits, we sequenced the 5′ ends of 9,216 clones from three cDNA libraries (apex, leaf and mature internode), representing 3,401 non-redundant sequences. About 57% of these sequences could be assigned a tentative function based on statistically significant similarity to previously characterized proteins or DNA sequences. Another 28% corresponded to previously identified, but uncharacterized, sequences. Some of the remaining unidentified sequences were predicted to be genes which could potentially be new to plants or unique to sugarcane. Comparisons of the sugarcane ESTs to a large sorghum EST database revealed similar compositions of expressed genes between some different tissues. Comparison to a detailed *Arabidopsis* protein database showed some highly conserved sequences, which might be useful DNA markers for pan-angiosperm comparative mapping. These EST sequences provide a foundation for many new studies to accelerate isolation of agronomically important genes from the cumbersome sugarcane genome.

Communicated by D.A. Hoisington

H.-M. Ma · S. Schulze · S. Lee · A. Paterson (✉)
Plant Genome Mapping Laboratory,
University of Georgia,
111 Riverbend Rd., Athens, GA 30602, USA
e-mail: paterson@dogwood.botany.uga.edu
Tel.: +1-706-5830162
Fax: +1-706-5830160

M. Yang · E. Mirkov
Department of Plant Pathology and Microbiology,
Texas A&M Agricultural Experiment Station,
2415 East Hwy. 83, Weslaco, TX 78596, USA

J. Irvine
Department of Soil and Crop Sciences,
Texas A&M Agricultural Experiment Station,
241 Hwy. 83, Weslaco, TX 78596, USA

P. Moore
USDA-ARS at Hawaii Agriculture Research Center,
99-193 Aiea Heights Dr, Aiea, HI 9, USA

Present address:
H.-M. Ma, USDA/ARS,
US National Arboretum,
3501 New York Ave. NE, Washington, DC 20002, USA

Introduction

Sugarcane, the sucrose-storing member of the genus *Saccharum,* is a perennial grass with solid stems that have become adapted for high sucrose storage. More than 60% of sucrose, a leading commodity important in the human diet, is produced from sugarcane. Sugarcane also ranks number one worldwide in biomass production among cultivated plants (Food and Agriculture Organization of the United Nations, http://www.fao.org), and may hold much potential as a bio-system to manufacture novel value-added molecules by engineering existing or new metabolic pathways.

Sugarcane is an extreme example of the polyploidy that is common among flowering plant species. Sugarcane, an autopolyploid (hybrid species that incorporates multiple chromosome sets that can pair and recombine freely), has chromosome numbers ranging from about 80 to 140 (Ming et al. 1998). Modern sugarcane cultivars, mainly derived from interspecific hybrids between *S. offcinarum* and *S. spontaneum,* with backcrossing to *S. officinarum* (Berding and Roach 1987), typically have more than eight homologous copies of each basic chromosome of *S. officinarum* and several copies of homeologous chromosomes from *S. spontaneum* (Ming et al. 1998). As such, sugarcane cultivars are highly heterozygous, with several different alleles of each locus. Such genomic redundancy may confer an evolutionary advantage (to buffer mutation load), or encourage the divergence of duplicated genes to adopt new functions. However, its large genome size, complicated genome

organization and high level of molecular diversity present special challenges for sugarcane genetic analysis, and generally slow rates of gain from selection in crop improvement programs. Many agronomically important traits, such as sucrose yield (Ming et al. 2001) and disease resistance (Daugrois et al. 1996), have been identified in sugarcane, but isolation of the corresponding genes has proven to be a daunting task due to the size and complexity of its genome. Even with the efforts of several large research groups, the collective coverage of the world's best available sugarcane maps is estimated at roughly 70% of the genome (Ming et al. 1998; however, the data are largely current). The high chromosome number, large amount of recombination and reduced power to detect linkage due to virtually all marker loci being dominant (Wu et al. 1992) are considerable hindrances to mapping additional populations and traits.

The availability of large numbers of functionally defined sugarcane genes would obviate many of the difficulties associated with genetic analysis of sugarcane. For example, given clones of the key genes in the carbohydrate metabolism pathways, one could directly investigate the extent to which these candidate genes account for genetic variation in sugar content (for example, Ming et al. 2001). Similar strategies could be applied to selecting for disease resistance, plant architecture, flowering time and other traits, given a large collection of sugarcane genes for which functions were known. While candidate genes from other taxa might be employed in some cases, because of the lack of genomic information for sugarcane it is difficult to assess the general degree of similarity of sugarcane genes to those of other taxa, or the extent to which the sets of genes expressed in key sugarcane tissues (such as the mature internodes in which sucrose is stored) parallel those of other related taxa (most of which are seed crops, such as sorghum, and therefore partition photosynthate very differently). Finally, by having large collections of candidate genes that may themselves be the target gene affecting a trait, one avoids many of the problems associated with searching for distantly-linked diagnostic markers in such a large and complex genome.

Partial sequencing of cDNAs to produce expressed sequence tags and comparing the resulting sequences to the ever-increasing public databases to annotate their putative functions, provides researchers with means to gather a large amount of genetic information and potentially identify new genes. A growing number of bio-informatic approaches permit one to explore ESTs for "signatures of selection", providing clues to the identities of important genes (Nielsen et al. 1998; Yang 1998; Schmid et al. 1999; Yang et al. 2000; Swanson et al. 2001a, 2001b) that are especially valuable in a genome as complex as sugarcane. Particularly interesting opportunities are expected to accrue to future investigation of selection that differentiates sugarcane from sorghum, in that these biomass and grain crops (respectively) may have diverged from a common ancestor as little as 5 million years ago and share much similarity in genome structure (Ming et al. 1998). ESTs are also valuable molecular markers, especially useful for comparative mapping projects since the expressed gene sequences are more conserved among different species.

With the goal of expanding our knowledge of sugarcane biology and genomics by providing large numbers of functionally characterized candidate genes that might be tested for direct association (rather than distant linkage) with economically important traits, we present an analysis of sugarcane ESTs sampled from three cDNA libraries (apex, leaf and internode). From each library, 18,432 clones were arrayed onto nylon membranes and hybridized with complex probes derived from RNAs extracted from different tissues (apex, leaf internode 2 and 7) as well as different developmental stages, namely, before and after floral induction. A total of 9,216 clones were selected from these libraries based on hybridization signal intensities. The initial sequence analysis allowed us to identify 3,401 non-redundant sequences. Further comparison using two different search programs against four different databases, resulted in the association of about 85% of our ESTs with homologs in public databases. We have increased the number of GenBank entries for *Saccharum* by about 8-fold, to 8,985 from 992 (as of 8 June 2003, with prior entries largely from Carson and Botha 2000; Carson and Botha 2002; Carson et al. 2002), opening new doors into the study of the biology and productivity of this major crop with a complex genome. We note that a very substantial Brazilian sugarcane EST database (SUCEST) has also been generated and partially described (see entire volume 24 of Genetics and Molecular Biology 2001; http://www.scielo.br/scielo.php?script=sci_issuetoc&pid=1415-475720010001&lng=en&nrm=iso), which we were able to access to a limited degree for comparative purposes; however, as of this writing, access to SUCEST remains restricted.

Materials and methods

cDNA libraries

Apices, leaves and 7th internodes (the internode belonging to the sixth leaf below the uppermost visible collar (dewlap) leaf, according to the Kuijper system as described by Benda 1969) of sugarcane cultivar CP72-2086 (a high sucrose-accumulating variety) were harvested from the field after floral induction. Three cDNA libraries were constructed from poly(A)RNA extracted from these tissues, using first-strand cDNA directionally cloned (5′ *Eco*RI–*Xho*I 3′) into the Uni-ZAP XR vector (Stratagene, La Jolla, Calif.), according to the manufacturer's instructions. The apex library was based on tissue from the innermost portion of the internode that included the floral meristem after removal of as much leaf roll as possible. The leaf library was based on tissue from several leaves immediately above the internode that contained the floral apex.

In vivo excision

Mass-excision of the pBluescript phagemid (containing the cloned cDNA insert) from the Uni-ZAP XR vector was performed using ExAssist helper phage with strain SOLR, according to the

manufacturer's instructions (Stratagene). The resulting titer of excised phagemids of apex, leaf and internode cDNA libraries was 1.22×10^7, 1.6×10^5 and 4.5×10^5 colonies/ml, respectively.

Preparation of high density filters

A total of 18,432 clones from each cDNA library were picked, stored in 48 384-well micro titer plates, replicated and gridded onto four 500 cm^2 nitrocellulose membranes (Hybond N+; Amersham, Piscataway, N.J.) using a Q-BOT (Genetix, New Milton, Hampshire, UK). Clones were double-spotted using a 2×2 pattern. Each filter contained 4,608 colonies (i.e., clones from 12 plates). Membranes were placed on Q-Trays (Genetix) containing Luria Broth (LB) with 50 μg/ml of ampicillin, and grown for 16–20 h at 37°C, followed by alkaline lysis fixation (Nizetic et al. 1991).

Preparation, labeling, and hybridization of complex probes

Apex (after removal of as much leaf roll as possible), leaf (middle one third to one half of the top visible dewlap leaf), internode 2 (the middle portion of that internode) and internode 7 (the middle portion of that internode) tissue from sugarcane variety H32-8560 were harvested both before and after floral induction. RNAs were extracted from all eight tissues using the KOAc method described by Cashmore et al. (1978). mRNAs were purified from total RNA using oligonucleotide (dT) magnetic beads (ATtract Systems III, IV; Promega, Madison, Wis.).

Complex probes were generated by direct labeling in the presence of [^{32}P]dCTP during first-strand cDNA synthesis primed with oligo(dT). Before labeling, 100 ng mRNA or 10 μg total RNA was added to 1 μg of oligo$(dT)_{12-18}$ and heated at 70°C for 10 min, then immediately cooled on ice to denature. This (m)RNA/oligo(dT) mixture (7 μl) was then mixed with 4 μl 5×first-strand buffer, 1 µl 10 mM (dATP, dGTP and dTTP), 2 µl 0.1 M DTT, and 40 U RNasout (Invitrogen, Carlsbad, Calif.), 40 μCi [α-^{32}P]dCTP (Amersham Pharmacia Biotech, Piscataway, N.J.) and 200 U SuperScript RTII (Invitrogen). The reaction was incubated at 37°C for 1.5–2.0 h, and then 21 μl of 10× stop buffer (2 M NaOH, 2 mM EDTA) was added to degrade the template RNA. This labeled probe was usually sufficient for the hybridization of two membranes.

The membranes were prehybridized (100 ml) and hybridized (20 ml) in 0.5 M NaP_i (pH 7.2), 7.0% SDS, 1 mM EDTA, 1% (w/v) BSA (Church and Gilbert 1984) at 65°C. For reused membranes, prehybridization was carried out for 4–5 h. New membranes required an additional overnight prehybridization. After hybridization for 20–22 h, filters were washed twice in 0.25×SSPE and 0.25% SDS, all at 65°C for 30 min at 12 rpm. The filters were then rinsed briefly in 2× SSC and exposed to X-ray film.

Scoring

EST clones were scored manually based on their hybridization intensity (qualitatively) as high, medium, low and rare categories, as illustrated in Fig. 1. The selected colonies were rearrayed into 96-well microtiter plate by a Q-BOT (Genetix). The sequence names reflect the position of the clones in the 96-well plate.

Automated DNA sequencing

Clones selected for sequencing were grown at 37°C overnight in 96-deep-well culture plates with 1.5 ml LB with 50°μg/ml ampicillin per well. Plasmid DNAs were prepared by alkaline lysis with modifications for the 96-well plate format. Sequencing reactions were carried out using ABI PRISM BigDye Terminator Cycle Sequencing Ready Reaction Kits Version 2 (Applied Biosystems, Foster City, Calif.). Reactions were set up in 96-well PCR plates according to the manufacturer's instructions except for using only one twelfth BigDye strength for each reaction. An M13 reverse primer (5′-CAGGAAACAGCTATGACC-3′) was used to generate 5′ EST sequences. Cycle sequencing was performed in a PTC-100 thermocycler (MJ Research, Waltham, Mass.) with a program as follows: preheat at 94°C for 5 min followed by 75 cycles of 94°C denature for 20 s, 50°C annealing for 5 s, 60°C extension for 4 min, and then held at 4°C. Reactions were filtered through Sephadex filter plates (Krakowski et al. 1995), then transferred directly into MicroAmp 96-well reaction plates (Applied Biosystems). High-throughput sequencing was carried out in an ABI PRISM 3700 (Applied Biosystems). Polymer POP-5 was used as the separation matrix to control electro-osmotic flow.

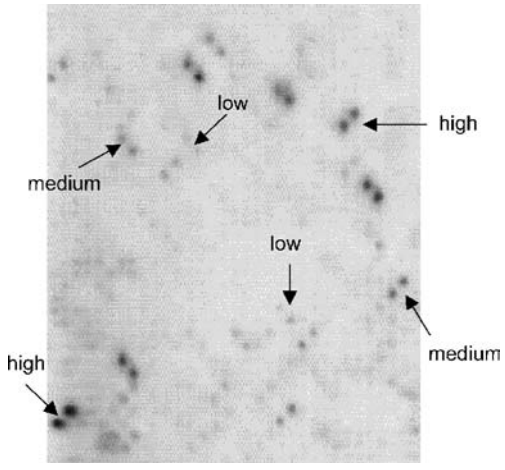

Fig. 1 Example of categories of hybridization intensity (reflecting gene expression levels)

Sequence data processing

Trace files were initially processed using PHRED/PHRAP/CONSED software (Ewing and Green 1998; Ewing et al. 1998). Sequences containing at least 100 continuous nucleotides with a phred score greater than 16 were clustered by Phrap with a minscore of 80. Assembled contigs were viewed/edited using Consed (Gordon et al. 1998). A total of 7,993 high-quality sequences have been deposited in GenBank as accession numbers BQ529595–BQ537587.

Two local alignment search programs of the BLAST suite, blastx and blastn were used for EST sequence similarity searches, with the default matrix BLOSUM 62, and cut-off E value of 10^{-10} for blastx (corresponding to a blast score of about 60) and 10^{-25} for blastn (corresponding to a blast score of about 120). Blast searches against public databases were performed in the following order: SwissProt (SP), GenBank non-redundant protein database (NR), GenBank non-redundant nucleotide database (nr), and GenBank plant database (pl). The plant database was downloaded in a manner that included all plant genes [nr, est, gss (genomic survey sequences), htgs (high throughput genomic sequences), sts (sequence tagged sites)] available from the National Center for Biotechnology Information (NCBI), as of 28 January 2002.

Library correlation

The chi-square contigency test was used to estimate whether there was any relationship between three of our cDNA libraries and nine different sorghum cDNA libraries or 17 different sugarcane cDNA libraries (see Results section for details on these libraries). The blast search results of our three cDNA libraries (the number of best hits derived from individual sorghum or sugarcane libraries) were tabulated. To further refine the test and isolate the groups of

libraries that accounted for this relationship, the contingency table was subdivided into two tables based on the difference between observed and expected frequencies in individual cells. If the resulting χ^2 value was less than the critical χ^2 value associated with the corresponding degree of freedom at $P \leq 0.01$, libraries in that contingency table were excluded from the test.

Results

General expression levels of different ESTs

To gain some insight into the gene expression level in the three sugarcane tissues, the arrayed cDNA libraries were pre-screened by probing with RNAs from the corresponding tissues. Specifically, the leaf library was probed with first-strand cDNA from leaf mRNA extracted before flowering (v-leaf), the apex library with mRNA from apex tissue harvested before and after floral induction (apex/panicle), and the internode library with mRNA from young (2nd) and old (7th) internodes collected before floral induction (v-Int2/-Int7). Figure 1 illustrates the categories of expression level, and Table 1 summarizes the hybridization intensity (assessed qualitatively) associated with 9,216 clones picked from these three cDNA libraries.

Consistent with our qualitative classification of expression levels, a higher percentage of sequences of clones from the highly/moderately expressed group were in those contigs (see the following section) containing no less than five EST reads, and a lower percentage in the singletons than that from the low/rarely expressed group (Table 1).

Initial sequence analysis

A total of 9,216 clones (4,032 from apex, 2,688 from leaf, and 2,496 from internode) from the three libraries were sequenced from the 5′ end to obtain maximally informative coding sequence. The three resulting sequences were first processed for quality by the base-calling program 'Phred' (Ewing and Green 1998; Ewing et al. 1998), giving 7,993 EST sequences (86.7%) that met satisfactory quality standards (100 nt of Q16, corresponding to an approximate error probability of 0.01 nt; Mullikin and McMurray 1999), all of which have been deposited in the GenBank EST database. The average length of cleaned sequences (after removing vector sequence, trimming polyA tails, etc.) was 461 bases, long enough to ensure a sensitive sequence homology comparison with protein databases in addition to the comparison at the nucleotide level.

Sequences were organized according to the source tissue of the library. Sequence assembly based on Phrap resulted in 353 apex contigs, 394 leaf contigs and 419 internode contigs. Including singletons, we generated 532, 1,629 and 1,240 unique sequences from apex, leaf and internode cDNA libraries, respectively. This may represent a somewhat smaller number of genes, since 5′ sequences from the same transcript may not overlap. Sequence similarity searches as seen in the following section would necessitate some adjustment to the number of unique transcripts.

Contigs assembled by Phrap included so-called singleton contigs that consist of only one read with a match to other contigs, but which could not be merged consistently with another contig. Six apex contigs, 122 leaf contigs and 135 intemode contigs are singleton contigs. Therefore, sequence redundancy (number of unique ESTs sequenced more than once/total number of unique ESTs) for the apex, leaf and intemode libraries, based on sequence clustering, was 65.2%, 16.7% and 22.9%, respectively.

Sequence homology search

The contigs/singletons were compared with public databases by using two of the blast search programs (blastx, blastn). Blast searches against public databases were performed in the following order: blastx_SP, blastx_NR, blastn_nr, and blastn_pl. If there were significant matches, the best one was recorded by an in-house program to convert the blast output text file to a spreadsheet file with information about the query ID, best match description, blast score, E value and percentage identity.

Combining the results from the four types of blast comparisons, 459/532 apex ESTs (86.3%), 1,340/1,629 leaf ESTs (82.3%) and 1,060/1,240 internode ESTs (85.5%) found matches in one or more of the public databases, leaving 542 sequences without a close counterpart. Among those having matches, about one third were not characterized or only annotated with descriptions such as "putative" and "-like protein", "unknown protein" or "hypothetical protein".

ESTs from apex, leaf and internode found sequence homology to 431, 1,206 and 883 unique genes, respectively, of which 145, 956 and 652 genes were represented

Table 1 Number of colonies selected based on hybridization intensity (combined into two groups, high/medium and low/rare) and their distribution between contigs (with no less than 5 reads) and singletons

Library	High/medium			Low/rare		
	No. of colonies	Contig with 5 ESTs	Singleton	No. of colonies	Contig with 5 ESTs	Singleton
Apex	290/566	90.8%	2.0%	3176/0	75.4%	6.4%
Leaf	113/176	72.3%	10.2%	775/1624	14.3%	56.5%
Internode	122/1.85	74.7%	7.0%	2093/96	28.3%	40.4%

Table 2 Identified highly abundant expressed genes, with tag count 50, from each sugarcane cDNA library

Library/ accession no.	Database[a]	Organism: gene description	Count
Apex			
Q40680	SP	Orysa: elongation factor 1-beta (EF-1-beta)	117
P14641	SP	Maize: tubulin alpha-2 chain (alpha-2 tubulin)	114
BE593753.1	pl	Sorbi: water-stressed 1 (WS1) sorghum bicolor cDNA	105
P50156	SP	Orysa: tonoplast intrinsic protein gamma	102
BAB02414.1	NR	Arath: chloroplast nucleoid DNA binding protein-like	102
P10979	SP	Maize: glycine-rich RNA-binding abscisic acid-inducible protein	92
AA577646.1	pl	Sacof: sugarcane leaf roll *Saccharum* sp. cDNA clone F77-rev	77
AAK76714.1	NR	Arath: unknown protein	69
P12857	SP	Maize: ADP ATP carrier protein 2 precursor (ADP/ATP translocase 2)	62
O24575	SP	Maize: S-adenosylmethionine decarboxylase proenzyme (adometdc)	61
P12653	SP	Maize: glutathione S-transferase I (GST-I) (GST-29) (GST class-phi)	53
AAF86307.1	NR	Wheat: EF-Hand Ca^{2+}-binding protein Ccd1	50
Leaf			
BE364814.2	pl	Sorbi: pathogen-induced 1 (P11) sorghum bicolor cDNA	57
Internode			
NP_195744.1	NR	Arath: LAX1/AUX1-like permease	94
P94029	SP	Orysa: metallothionein-like protein type 2	63
P50156	SP	Orysa: tonoplast intrinsic protein gamma	63

[a] *SP* SwissProt, *NR* GenBank non-redundant protein database, *nr* GenBank non-redundant nucleotide database, *pl* GenBank plant database (downloaded in a manner that included all plant genes [*nr* (non-redundant), *est* (expressed sequence tags), *gss* (genomic survey sequences), *htgs* (high throughput genomic sequences), *sts* (sequence tagged sites)] available in the National Center for Biotechnology Information (NCBI), as of 28 January 2002)

by ESTs sequenced only once in our experiments. Here, those predicted by Phrap to be non-redundant sequences but sharing the same best hit to a protein/cDNA, were represented by the sequence with highest score. Therefore, the redundancy among those sequences having matches was 66.8% (apex), 20.8% (leaf) and 26.3% (internode).

Table 2 lists the putative genes that are most abundant among our ESTs, occurring 50 times or more. The most highly represented gene is elongation factor 1-beta (Q40680, SwissProt) from the apex library. One tonoplast intrinsic protein gamma gene (P50156, SwissProt) was highly represented in both apex and internode libraries, with EST counts of 102 and 63, respectively. It is noteworthy that the redundant copies of each gene are not all identical to one another. The constituent reads of each contig displayed different degrees of SNP (single nucleotide polymorphism) or INDEL (insertion/deletion) variation, some organized into patterns that suggested a smaller number of alleles.

Functional categorization

Our ESTs showed significant deduced amino acid sequence homology to many known proteins, including enzymes from ubiquitous metabolic pathways, structural proteins, components of the transcriptional or translational apparatus, members of signal transduction pathways and proteins involved in plant growth and development.

When categorizing the putative functional roles of the sugarcane ESTs, if matches were found in more than one database, the general order of databases used to record the gene annotation was SP (blastx), NR (blastx), nr (blastn) and pl (blastn). This was based on the notions that (1) protein sequence/translated DNA sequence comparison is much more sensitive than DNA sequence comparison, due to both the noise from the "wobble" third base in each DNA codon, and the capacity for recognition based on chemical similarity of different amino acids, and (2) better-curated databases such as SwissProt tend to be less redundant, which improves the statistical significance of a match. However, it is also less comprehensive and up-to-date. As of 23 January 2002 (the date on which the dataset was "frozen" for analysis), there were only 104,559 entries in SP. The use of multiple databases was intended to exhaust the search for homologous counterparts among the existing large number of sequences.

ESTs with a match in the databases (both characterized and uncharacterized) were categorized into 12 groups based primarily on the catalogs established for *Arabidopsis thaliana* (http://www.tigr.org/docs/tigr-scripts/edb2_scripts/neuk_gene_table.spl?db=ath1) (Fig. 2). Those grouped under the HPU category (Hypothetical/Putative/Unknown Protein) also included sequences with matches to genomic DNAs. In all three libraries, this category had the highest number of sequences (26%–37% of the total sequences). Among the known proteins, the protein expression group included the highest percentage (12–18%) of sequences, followed by metabolism (9%–11%).

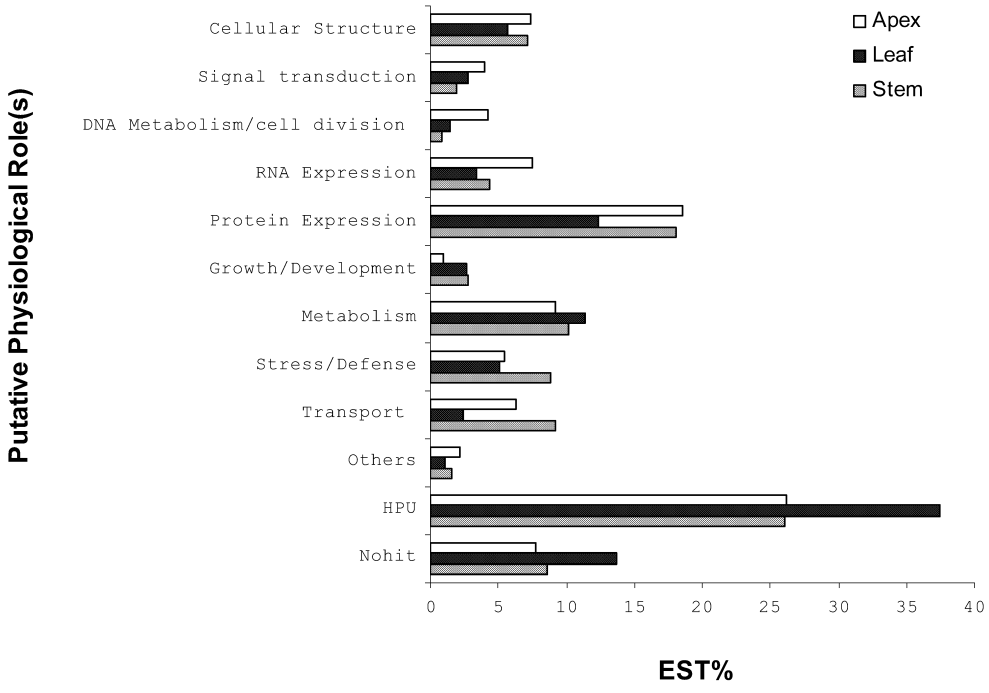

Fig. 2 Functional category of sugarcane ESTs based on their putative biochemical and physiological role(s) and their relative frequencies (as calculated by the number of EST counts/total number of non-redundant ESTs from each library). EST counts and the number of total non-redundant ESTs are estimated such that contigs or singletons showing homology with the same gene (accession number) are considered to represent one gene, even if sequence assembly put them into different contigs or singletons

ESTs shared among and between libraries

The accession number of the database entry for each match was used as a guide to find out how many sequences were represented in more than one cDNA library. If two sequences shared the same accession number as the best match, they were considered to be derived from the same gene. We identified a total of 238 genes that were expressed in more than one tissue. A total of 34 genes were expressed in all three cDNA libraries (Table 3). Among the rest, 29 were found in both apex and leaf libraries, 56 in both apex and internode libraries, and 119 in both leaf and internode libraries (Fig. 3). Many of the 34 shared genes were in the protein expression category (5 ribosomal proteins, 3 elongation factors, 1 initiation factor, 1 DnaJ protein, 1 ubiquitin conjugating enzyme). Besides the three uncharacterized genes, the others were distributed over the functional categories of abiotic stress response, cellular structure/organization, DNA metabolism, sugar metabolism, amino acid metabolism, transport and signaling.

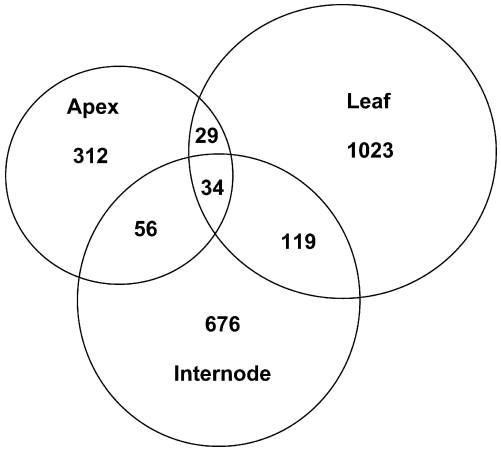

Fig. 3 Number of sugarcane ESTs represented by more than one library

Sequences similar to that of plant origin at the nucleotide level

At the time of the sequence comparisons, there were only 755 sugarcane sequences deposited in GenBank, therefore most of the best matches were to sequences from other plants. Blast results against the plant database (blastn_pl) were sorted and supplemented with the results against non-redundant protein databases (blastx_NR). Although substantially more sequences are available from barley, maize and rice, nearly 60% of the best matches were from sorghum, which is thought to be the closest relative of sugarcane among grasses (Al-Janabi et al. 1994). The runner-up was maize, with about 19% of the best hits, followed by rice (3%). The 542 sequences with no match in the public databases were compared to a sugarcane EST proprietary database (SUCEST; http://sucest.lad.ic.unicamp.br/en/), which ultimately reduced the number of unidentifiable sequences to 168. The open reading frames of these 168 sequences were predicted by EMBOSS (European Molecular Biology Open Software Suite; http://www.hgmp.mrc.ac.uk/Software/EMBOSS/) "getorf" program, and 37 of them appeared to be genes, based on minimum lengths of the translated DNA sequence exceeding 50 amino acids, and ratios of open

Table 3 The 34 sugarcane ESTs represented by all three cDNA libraries

Accession no.	Database[a]	Organism: gene description
P49397	SP	Orysa: 40S ribosomal protein S3a (cyc07 protein)
O22424	SP	Maize: 40S ribosomal protein 84
O24573	SP	Maize: 60S acidic ribosomal protein P0
P42791	SP	Arath: 60S ribosomal protein L18
Q07761	SP	Tobac: 60S ribosomal protein L23a (L25)
Q9M352	SP	Arath: 60S ribosomal protein L36-2
P49625	SP	Orysa: 60S ribosomal protein L5
P30171	SP	Soltu: actin 97
P12857	SP	Maize: ADP ATP carrier protein 2 precursor (ADP/ATP translocase 2)
P34991	SP	Human: cyclin A/cdk2-associated protein P19 (RNA poly II elongation factor-like protein)
P46524	SP	Wheat: dehydrin Cor410 (cold-induced Cor410 protein)
P27347	SP	Maize: DNA-binding protein Mnblb (hmgi-like protein)
Q04960	SP	Cucsa: DnaJ protein homolog (DnaJ-1)
Q41803	SP	Maize: elongation factor I-alpha (EF-1-alpha)
Q9ZR17	SP	Orysa: elongation factor 1-gamma (EF-1-gamma) (EF-1b-gamma)
Q23755	SP	Betvu: elongation factor 2 (EF-2)
P42895	SP	Maize: enolase 2 (2-phosphoglycerate dehydratase 2)
P12653	SP	Maize: glutathione S-transferase I (Gst-I) (Gst-29)
P08735	SP	Maize: glyceraldehyde 3-phosphate dehydrogenase cytosolic 1
P02308	SP	Wheat: histone H4
P80639	SP	Maize: initiation factor 5a (EIF-5a) (EIF-4d)
O24575	SP	Maize: S-adenosylmethionine decarboxylase proenzyme (Adometdc)
P46611	SP	Orysa: S-adenosylmethionine synthetase 1 (methionine adenosyl-transferase 1)
P50156	SP	Orysa: tonoplast intrinsic protein gamma (gamma TIP) (aquaporin-TIP)
P35681	SP	Orysa: translationally controlled tumor protein homolog (TCTP)
P35134	SP	Arath: ubiquitin-conjugating enzyme E2-17 Kda 11 (ubiquitin-protein ligase 11)
Q96558	SP	Soybn: UDP-glucose 6-dehydrogenase (UDP-Glc dehydrogenase)
AAK91441.1	NR	Arath: Atlg55500/T5A14_10
AAF86307.1	NR	Triticum: EF-hand Ca^{2+}-binding protein CCD 1
NP_190104.1	NR	Arath: putative protein
NP_181523.1	NR	Arath: unknown protein
AB071694.1	nr	Saccharum: SoDip22 mRNA for drought inducible 22 kD protein
AF168884.1	nr	Zea mays: 18S small subunit ribosomal RNA gene complete sequence
AJ130830.1	nr	Zea mays: gene encoding putative cell wall protein

[a] *SP* SwissProt, *NR* GenBank non-redundant protein database, *nr* GenBank non-redundant nucleotide database, *pl* GenBank plant database [downloaded in a way that included all plant genes (nr, est, gss, htgs, sts) available in the National Center for Biotechnology Information (NCBI), as of 28 January 2002]

reading frame length/translated sequence length of no less than 70%.

Correspondence to sorghum ESTs

As of 28 January 2002, there were 109,847 sorghum sequences in the public database. At an E value of 10^{-25}, blastn searches exclusively against the sorghum sequence database found matches to about 72% of our sugarcane ESTs. Among the remaining sugarcane EST sequences that bad or no matches in the sorghum database, 44% had hits from other plants, mostly maize.

The vast majority of the sorghum sequences matching our ESTs were from one of the following ten different types of cDNA libraries (total of 12 libraries, for more details, see http://cggc.agtec.uga.edu/cggc/data/sequenceView/sequence.asp): nine of them were made from *Sorghum bicolor,* including dark grown (DG1), embryo (EM1), immature panicle (IP1), light grown (LG1), ovary 1 (OV1), ovary 2 (OV2), pathogen-induced (P11), compatible pathogen-infected (PIC1) and water-stressed (WS1); two were made from *Sorghum propinquum,* which included floral-induced meristem (FM1) and rhizome2 (RHIZ2); only one was from *Sorghum halepense,* rhizome1 (RHIZ1).

The distribution of matches from different cDNA libraries could be used to assess the degree of similarity between different tissues, under the assumption that similar tissues or tissues under similar growth conditions would share a similar gene expression patterns (Ewing et al. 1999). Since the number of available sequences from each sorghum cDNA library greatly influences the outcomes of the matches to be found with our ESTs, sorghum libraries (OV1, OV2 and RHIZ1) with substantially fewer than 10,000 sequences were not considered in this test. The average number of ESTs from the remaining libraries was 10,522±742. As described in the Materials and methods, the number of best hits derived from these nine sorghum libraries was tabulated (Table 4) and found to be correlated (χ^2=40.95; $P \leq 0.01$). To identify the libraries that account for this correlation, the contingency

Table 4 Number of top hits found among nine sorghum cDNA libraries and 17 other sugarcane libraries

Library[a,b]	DG1	EM1	FM1	IO1	LG1	PT1	PIC1	RHIZ2	WS1	AD1	AM1	AM2	FL1	FL3	FL4	FL5	HR1	LB2	LR1	RT2	RT3	RZ3	SB1	SD1	SD2	ST3
Apex	50	26	41	32	60	25	16	42	52	25	20	24	18	30	30	19	10	26	18	18	25	29	19	26	14	19
Leaf	121	75	91	60	163	108	99	84	166	103	42	55	76	70	76	44	50	62	87	60	49	91	54	100	108	30
Internode	143	71	84	52	112	80	49	88	141	59	39	58	47	53	56	38	50	43	63	63	42	76	57	55	46	44

[a] *Sorghum bicolor* cDNA library: *DG1* dark grown, *EM1* embryo, *PP1* immature panicle, *LG1* light grown, *OV1* ovary 1, *OV2* ovary 2, *PI1* pathogen induced, *PIC1* compatible pathogen-infected, *WS1* water-stressed *Sorghum propinquum* cDNA library: *FM1* floral-induced meristem, *RHIZ2* rhizome2

[b] *AD1 Acetobacter diazotroficans*-infected plantlet, *AM1* mature plant apical meristem, *AM2* immature plant apical meristem, *FL1* sugarcane flower-Icm, *FL3* sugarcane flower-base of 5 cm flowers, *FL4* sugarcane flower-50 cm flower stem, *FLS* sugarcane flower 20 cm, *HR1 Herbaspirillum rubrisubalbicans*-infected plantlet without developed leaves and roots, *LB2* lateral bud, *LR1* leaf roll-large insert, *RT2* root-large insert, *RT3* root, *RZ3* leaf-root transition zone, *SB1* stem bark-large insert, *SD 1* seeds-large insert, *SD2* seeds-small insert, *ST3* fourth internode

table was subdivided based on the value of $(O-E)^2/E$. One was a 6×3 table consisting of best hit counts from EM1, FM1, LG1, PI1, RHIZ2 and WS1 libraries, the other was a 3×3 table with DG1, IP1, and PIC1 libraries. The results indicated that the sorghum DG1, IP1 and PIC1 libraries shared a significant relationship with at least one of our libraries(χ^2=26.65; $P\leq$0.01), but EM1, FM1, LG1, PI1, RHIZ2 and WS1 libraries did not (χ^2=14.14; $P\leq$0.01). The sign of (O-E) reflected whether the relationship was positive or negative. Our tests suggested that (1) the sorghum IP1 library had a positive correlation with our apex library, but no correlation with leaf or internode libraries; (2) the sorghum PIC 1 library had a very strong positive relationship with our leaf library and a strong negative relationship with both our apex and internode libraries; (3) DGl did not show a significant relationship with the apex library, but displayed a negative correlation with the leaf library and a positive correlation with the internode library.

Correspondence to sugarcane ESTs from SUCEST

Our entire EST collection was directly compared with the SUCEST database (http://sucest.lad.ic.unicamp.br/en/). A total of 302 of our sequences (38 from apex, 172 from leaf and 92 from internode) are absent from SUCEST based on blastn searches at an E value of 10^{-25}. Among those 302 sequences, 134 were found to have a close counterpart in other plants.

There are about 300,000 ESTs in the SUCEST database, mainly derived from 17 diverse cDNA libraries, each with more than 10,000 sequences. They are *Acetobacter diazotroficans* infected plantlet (AD1), mature plant apical meristem (AM1), immature plant apical meristem (AM2), sugarcane flower-1 cm (FL1), sugarcane flower-base of 5 cm flowers (FL3), sugarcane flower-50 cm flower stem (FL4), sugarcane flower-20 cm (FL5), *Herbaspirillum rubrisubalbicans* infected plantlet without developed leaves and roots (HR1), lateral bud (LB2), leaf roll-large insert (LR1), root -large insert (RT2), root (RT3), leaf root transition zone (RZ3), stem bark-large insert (SB1), seeds-large insert (SD1), seeds-small insert (SD2) and fourth internode (ST3). As with the different sorghum cDNA libraries, the number of best hits derived from these 17 sugarcane libraries was tabulated (Table 4) and found to be correlated χ^2=69.55; $P\leq$0.01).

To find out which libraries contributed to the pattern of EST abundance between different sugarcane libraries, two sub-contingency tables of 4×3 (composed of HR1, RT3, SD2 and ST3 sugarcane libraries) and 13×3 were formed. The results indicated that HR1, RT3, SD2 and ST3 sugarcane libraries largely accounted for the significant relationship with our libraries χ^2=36.95; $P\leq$0.01). In addition, it also revealed that (1) HR1 displayed a negative correlation with the apex library and a positive one with the internode library, but at a confidence level of only 0.25<P<0.1; (2) RT3 showed a positive correlation

with the apex library, but no correlation with the leaf or internode libraries; (3) SD2 exhibited a very strong positive correlation with the leaf library but very strong negative correlation with both apex and internode libraries; and (4) ST3 showed a very strong negative correlation with the leaf library, and a marginally significant positive correlation with both apex and internode libraries.

Sequences similar to those of non-plant origin

Among sequences that showed no match to the plant database, but shared matches to the non-redundant protein database, ten sequences from the leaf library had homology to sequences of non-plant origin. Nine of these sequences were from fungi (four from *Schizosaccharomycetacea,* three from *Saccharomycetaceae,* and the other two from the *Amanitacease* and *Erysiphaceae* families). None of these sequences showed homology at the DNA sequence level, suggesting that the homology shared at the protein level was due to evolutionary conservation rather than sample contamination. The tenth sequence appeared to be chimeric, with its 5′ region (138–225 bp) homologous to a *Sorghum bicolor* glycine-rich RNA-binding protein mRNA, and its 3′ region (239–551) to the *Escherichia coli* K12 putative ATP-binding component of a transport system. The chimeric sequence is presumed to be due to cloning error.

Matches with *Arabidopsis* proteins

Our 3,401 sugarcane ESTs were searched against 25,545 *Arabidopsis* proteins (obtained from NCBI) using blastx with an E value of 10^{-10}. A total of 53.66% (1,825/3,401) of our ESTs found a close relative in the *Arabidopsis* protein set. Among the 1,825 homologs, 270 sequences were represented more than once, accounting for 495 sequences. Despite the evolutionary divergence of these two species, 39 non-redundant sequences showed almost perfect match ($E \leq 10^{-100}$). Of these 39 sequences, more than half (19) were involved in amino acid or protein metabolism. In addition to two unknown proteins, the rest participate in sugar metabolism (e.g., sucrose-synthase like protein), signal transduction pathways (e.g., 14-3-3 protein; Shaw 2000), or function as transporters (e.g., plasma membrane intrinsic protein 2a), components of cell wall structure (e.g., endoxyloglucan transferase), cytoskeleton (e.g., actin, tubulin) or photosynthetic apparatus (e.g., photosystem II type I chlorophyll a/b-binding protein). Notably, one gene among them, *pinhead,* is a developmental protein required for reliable formation of primary and axillary shoot apical meristems, a component of a hypothetical meristem-forming competence factor.

About 145 non-redundant sugarcane ESTs had no close counterpart in the *Arabidopsis* protein set, but shared homology with protein sequences derived from other organisms. Some of those, such as C4 plant-specific genes (bundle sheath specific protein 1) and disease resistance genes (NADPH HC toxin reductase *hm1,* strip rust resistance protein, subtilisiralchymotrypsin inhibitor), are probably not expressed in *Arabidopsis.* Others appear to be genes that are relatively rapidly evolving and no longer correspond closely to the *Arabidopsis* sequence. For instance, our ESTs included relatively good matches with jacalin from barley and caffeic acid 3-methyltransferase from sugarcane, but were unable to draw a match from the *Arabidopsis* protein set, even though it carries homologs of these genes.

Discussion

EST sequencing from three sugarcane libraries has expanded public resources by 8-fold for this important crop, providing a substantial resource of candidate genes that might be tested genetically for direct association with traits important to quality, productivity and development in the complex polyploid genome of this major crop. Early insights can also be gained into the similarities and differences in gene expression that are associated with its differentiation (apex library), photosynthate production (leaf library) and photosynthate deposition/storage (internode library).

While virtually all of the functional categories of ESTs include members that have been implicated in key aspects of growth and development of other plants, a few categories stand out in their potential importance. For example, variation in source strength (photoassimilate production) might be investigated based on genes encoding components of the photosynthetic apparatus, of which 23 were found (corresponding to 99 reads, exclusively in the leaf library) or genes encoding key steps in carbohydrate metabolism (of which, 67, 33 and 18 were found in the apex, leaf and internode libraries, respectively). Key aspects of plant growth and development such as the timing and regulation of flowering, and plant architecture (particularly sink size) might be investigated based on 64 genes (14, 24 and 26 from apex, leaf, and internode libraries, respectively) implicated in plant growth and development and 97 genes (20, 48 and 29 from apex, leaf and internode libraries, respectively) implicated in signal transduction. A total of 132 genes are associated with plant defense against stress, including detoxification, abiotic stress and biotic stress (21, 47 and 64 from apex, leaf, and internode libraries, respectively). Given the very slow rate of traditional sugarcane improvement, estimated at 1% yield improvement per year, significant association of even a small fraction of these candidates with economically important traits could substantially accelerate improvement of this singularly complex crop.

ESTs from non-normalized cDNA libraries provide a snapshot of the processes and pathways that are active in plant growth and development. Comparison to a qualitative assessment of array hybridization signal verifies that EST abundance is a meaningful representation of transcript expression levels, and striking differences among

tissues (libraries) are consistent with tentative functional assignments of many ESTs. Comparisons of EST abundance in these libraries to other EST libraries from sugarcane or closely related sorghum were largely congruent, but with interesting exceptions. In particular, the lack of correspondence of sorghum libraries to the sugarcane internode library, representing the primary economically important sink of the sugarcane plant, suggests that much additional investigation of this particularly important tissue may be warranted.

Sugarcane ESTs generally show phylogenetically-congruent similarities to genes from other taxa, with sorghum being by far the best botanical model for sugarcane. The relatively poor correspondence of sugarcane to rice genes raises questions about the extent to which the rice genomic sequence will translate into the meeting key needs of sugarcane genomics. Application of a "skimming" approach such as Cot-based cloning and sequencing (Peterson et al. 2002a) to sugarcane, or perhaps better to sorghum (to take advantage of the smaller genome and lower redundancy, while benefiting from the close relationship to sugarcane), may be important to ongoing progress in sugarcane genomic biology. A small but significant population of ESTs showed no matches to other taxa, pointing to the existence of populations of rapidly-evolving genes that may contribute disproportionally to taxonomic differentiation. Further, some of these showed no matches in a large private-sector sugarcane database (SUCEST), suggesting possible impacts of environmental factors on gene expression.

Use of non-normalized cDNA libraries

One rationale for using non-normalized libraries is to obtain some quantitative measure of gene expression level and preserve different allelic isoforms of the same gene family. Although a normalized cDNA library will give a more or less even chance to collect a transcript expressed at any level in the tissue, tag counts from a non-normalized cDNA library approximately reflect the abundance of each specific transcript present in the tissue used to make the library. Highly expressed genes will be represented by multiple clones in the cDNA library and vice versa. Overlapping multiple ESTs derived from the same gene (family) offer the possibility of being clustered to give a longer DNA region than might be sequenced in a single run. The resulting consensus sequence from independent but overlapping tags would increase confidence in base calling, partially compensating for the error-prone nature of the single-pass sequencing strategy.

The three cDNA libraries used in this study were constructed using a phage (λ) vector system. However, arrayed cDNA library technology was developed for a plasmid system. The three sugarcane cDNA phage libraries were excised into phagemid libraries, a step that may somewhat alter the representation of clones from the original abundance. In addition, the clones selected for sequencing were not entirely random, but instead were the result of the attempt to associate the resulting sequences with a predefined expression level. These two factors could have an effect on the final ratio among unique transcripts.

Tissue-specificity of gene expression patterns

Sequencing from three cDNA libraries can potentially lead to the discovery of tissue-specific and constitutively expressed genes. However, unless a very large number of sequences are available from each tissue, it is not possible to conclude whether a sequence represented by one cDNA library is truly tissue-specific. Our EST project was performed on a relatively small scale, so the presence of an EST in only one cDNA library does not provide sufficient information to predict whether it is tissue-specific. On the other hand, those found in all three libraries (Table 3) are likely to be constitutively expressed and might be useful for the identification of constitutive promoters with different expression levels.

Apex ESTs showed a higher relative abundance in all three functional groups, DNA metabolism/cell division, RNA expression and protein expression, than those from leaf and internode libraries. This is not unexpected since the apex is an actively growing tissue type. However, the relative abundance of ESTs in these categories was largely attributable to a few highly represented proteins, e.g., 117 reads for elongation factor (EF)1-Beta and 46 reads for 40 S ribosomal protein S28. The leaf library had a lower percentage of sequence counts in most functional groups, but a larger number of different proteins present.

What is expressed or not expressed determines the physiological status of a tissue. A detailed analysis of the cell wall, a subgroup of cellular structure, showed that the apex had nearly as many ESTs as that of the internode. However, in contrast to ESTs from the internode, none of the 92 sequences of the apex library-encoded enzymes was specifically involved in the lignin biosynthesis pathway. Apex ESTs were mostly involved in cell expansion (Kohorn 2000), such as expansin (7 reads), arabinogalactan-like protein (4 reads), cellulose synthase (21 reads) and xyloglucan endotransglycosylase-like proteins (40 reads). In the internode, genes for cell wall biosynthesis were mainly enzymes involved either in the xyloglucan network (25 reads) or lignin biosynthesis (40 reads). The majority of leaf genes in the same category were involved either in cellulose (4 reads) or lignin (7 reads) biosynthesis.

Among 10 enzymes in the glycolysis pathway, we found ESTs matching seven in one or more libraries, missing only hexokinase, phosphoglucoisomerase and phosphoglycerate kinase. The tag count ratios of the corresponding enzymes (phosphofructokinase: aldolase: triose phosphate isomerase: glyceraldehyde-3 phosphate dehydrogenase: phosphoglycerate mutase: enolase: pyruvate kinase) were 0:0:1:35:1:11:0 for apex, 2:2:0:5:1:3:5 for leaf and 1:5:1:17:0:5:0 for internode. It seemed that

glyceraldehyde-3-phosphate dehydrogenase and enolase were the most highly expressed glycolytic enzymes in the three tissues. These are the two enzymes that catalyze the two energy-conserving reactions that eventually lead to the formation of ATP during glycolysis.

Among unique transcripts, quite a few encode isoforms of the same functional protein. In several cases, different tissues seemed to express a different isoform. One good example is 4 coumarate-CoA Ligase (4CL), the last enzyme in the general pathway to provide activated CoA-esters of cinnamic acids, used in the biosynthesis of diverse phenolic compounds (e.g., flavonoids, isoflavonoid, and lignin) via specific branch pathways. Curiously, one copy of a different isoform of the 4CL enzyme from different plant origins (4CL1 of soybean, 4CL2 of tobacco and 4CL3 of *Arabidopsis*) was employed by apex, leaf and internode, respectively. All three transcripts are best hits with high scores ($\geq$110) from blastx_SP searches.

There are other instances in which different isoforms were found in different tissues, but with a different expression level (tag counts). Both apex and internode EST sets include homologs to maize sucrose synthase isoforms I (Sh1) and II (Sus1), contrary to Carson and Botha's results (2000). We found *sus1* to be more highly expressed than *sh1* based on EST counts in both apex (42 vs 2) and internode (6 vs 3) libraries. The leaf only had one copy of isoform I. Although it has been characterized to catalyze reversible reactions, sucrose synthase is mostly known to break down rather than synthesize sucrose, and is distributed widely among sink (importing) tissues (Hawker 1985), such as apex and internode. The two maize sucrose synthase isoforms have been found to show contrasting sugar modulation (Koch 1996), *sh1* being expressed at maximum levels under carbohydrate starvation, while *susl* expression is induced in response to abundant sugar supplies.

Similaritiy between our cDNA libraries and other libraries from sorghum and sugarcane

The three cDNA libraries, apex, leaf and internode, represent sugarcane's principle tissues. ESTs derived from the three libraries were quite varied (Fig. 3). The frequencies of sequences shared between different cDNA libraries might be suggestive of similarities in gene expression patterns between them.

When compared to nine sorghum cDNA libraries and 17 other sugarcane cDNA libraries, our results suggest that randomly sampled ESTs do reflect patterns of gene expression which are correlated to the tissue's overall physiology. Each of our three cDNA libraries exhibited a significant correlation with some libraries but not others, based on the χ^2 square test on the number of best hits derived from different libraries. The sorghum IP1 (immature panicles) library showed a positive correlation with our apex library, constructed from the apices after floral induction, but no relationship with our leaf or internode library. Our leaf library was very similar to PIC1, made from sorghum leaf RNA after infection with anthracnose pathogen *Colletotrichum graminicola.* It is interesting to note that our leaf library also displayed a very strong positive correlation with the sugarcane SD2 (seed) library but a negative one with the ST3 (4th internode) library, while the opposite trend existed for both apex and internode libraries, being correlated negatively with SD2, but positively with ST3.

Fishing for rare genes

A perennial criticism of the EST approach is that it may miss many rarely expressed transcripts (Marra et al. 1998). The methodology behind ESTs implicitly favors genes of high/medium expression level. However, pre-screening of arrayed eDNA blots may help to segregate cDNAs into subpopulations with different average degrees of iteration, which can then be sequenced to depths appropriate to represent the respective subpopulations. The finding that our leaf library contributed more unique sequences and had lower redundancy could be due to this colony selection scheme. Complex probes from leaf mRNA were used to interrogate the arrayed leaf cDNA library. The resulting hybridization signal intensity reflected the abundance of the corresponding mRNA molecule in the complex probes. Those that did not develop a detectable hybridization signal were considered rarely expressed transcripts. The number of leaf ESTs sequenced only once in our experiments, which derived from the colonies of high/medium, low and rare expression, respectively, were 8.0% [23/(113+176)], 20.0% (155/775), and 48.0% (778/1624). The internode library showed a comparable pattern with single ESTs accounting for 5.5% [17/(122+185)], 27.9% (584/2093), and 53.1% (51/96) of the respective categories. This result suggests that, with the assumption of good cDNA library coverage and macro-array quality, pre-screening and picking those clones that do not show a hybridization signal may enhance the possibility of obtaining sequences from rarely expressed genes.

It was quite unexpected that the apex library had the highest redundancy. Apex ESTs that were sequenced just once only represented 1.4% [12/(290+566)] and 4.2% (133/3176) from its high/medium and low category, about 4.0–6.0 fold lower than that of the corresponding leaf and internode libraries. This indicated that the high redundancy might well be an artifact attributed to library construction, rather than a true reflection of what was really expressed in the apex.

Conclusion

While these data considerably improve our ability to conduct targeted evaluations of specific genes in the genetic control of economically important traits in sugarcane, they also raise new questions and point to

new needs. For example, further investigation of gene expression in sugarcane internodes appears especially important, perhaps aided by hybridization-based screening prior to selection of colonies for sequencing to enrich for gene transcripts with low expression levels (thereby increasing the yield of new information). EST homology searches suggest that sorghum data provides far better guidance than rice sequences for genomic study of sugarcane, but the paucity of genomic sequence for sorghum may warrant further investigation of its low-copy DNA using efficient new approaches (Peterson et al. 2002b).

Of special interest are rapidly evolving genes that may contribute to the unique features of sugarcane. We found a substantial population of sugarcane ESTs for which we found no counterpart in the sorghum (or other public) EST databases, and vice versa. The fact that close to 10% of our sequences are absent from SUCEST with blastn searches at an E value of 10^{-25},even though our database is only about 3% of the size of SUCEST, suggests that some apparently "orphan" genes may actually reflect to the impact of environmental factors or genetic backgrounds on gene expression. Investigation of the finished rice sequence for inferred protein products that correspond to these genes may shed light on whether they represent truly novel genes or merely different expression patterns, albeit with the cautions above as to the genetic distance between sugarcane and rice. ESTs can be integrated into genomic maps to correlate genes with agronomically important phenotypes, with the prospect that some of them will be absolutely linked (i.e., at zero recombination) with the trait. Alternatively, ESTs can be used to design overgo primers (Cai et al. 1998) for physical mapping to avoid the necessity of identifying a polymorphism for each EST. The functions of hypothetical or unknown genes may be best defined by using ESTs to create mutations to change the gene expression level to fully study gene function in vivo. (Baulcombe 1999; Wang and Waterhouse 2001).

Acknowledgements This work was funded by the International Consortium for Sugarcane Biotechnology. We thank Meizhu Yang and Dr. Yang Si for making the sugarcane leaf and internode libraries and Dr Veera Padmanabhan for the apex library. We also thank Dr. John Bowers and Dr. Alan Gingle for assistance in data processing.

References

Al-Janabi SM, McClelland M, Peterson C, Sobral BWS (1994) Phylogenetic analysis of organellar DNA sequences in the *Andropogoneae*: *Saccharinae*. Theor Appl Genet 88:933–944

Baulcombe DC (1999) Fast forward genetics based on virus-induced gene silencing. Curr Opin Plant Biol 2:109–113

Benda, GTA (1969) Numbering sugarcane leaves and shoots. Sugarcane Pathol Newsl 3:16–18

Berding N, Roach BT (1987) Germplasm collection, maintenance, and use. In: Heinz DJ (ed) Sugarcane improvement through breeding. Elsevier, Amsterdam, pp 143–210

Cai WW, Reneker J, Chow CW, Vaishnav M, Bradley A (1998) An anchored framework BAC map of mouse chromosome 11 assembled using multiplex oligonucleotide hybridization. Genomics 54:387–397

Carson DL, Botha FC (2000) Preliminary analysis of expressed sequence tags for sugarcane. Crop Sci 40:1769–1779

Carson DL, Botha FC (2002) Genes expressed in sugarcane maturing internodal tissue. Plant Cell Rep 20:1075–1081

Carson DL, Huckett BI, Botha FC (2002) Sugarcane ESTs differentially expressed in immature and maturing internodal tissue. Plant Sci 162:289–300

Cashmore AR, Broadhurst MK, Gray RE (1978) Cell-free synthesis of leaf protein: Identification of an apparent precursor of the small subunit of ribulose-1, 5-bisphosphate carboxylase. Proc Natl Acad Sci USA 75:655–659

Church GM, Gilbert W (1984) Genomic sequencing. Proc Natl Acad Sci USA 81:1991–1995

Daugrois JH, Grivet L, Roques D, Hoarau JY, Lombard H, Glaszmann JC, D'Hont A (1996) A putative major gene for rust resistance linked with a RFLP marker in sugarane cultivar 'R5 70'. Theor Appl Genet 92:1059–1064

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

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred I Accuracy assessment. Genome Res 8:175–185

Ewing RM, Kahla AB, Poirot O, Lopez F, Audic S, Claverie J-M (1999) Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. Genome Res 9:950–959

Gordon D, Abajian C, Green P (1998) Consed: a graphical toll for sequence finishing. Genome Res 8:195–202

Hawker, JS (1985) Sucrose. In: Dey PM, Dixon RA (eds) Biochemistry of storage carbohydrates in green plants. Academic Press, New York, pp 1–51

Koch KE (1996) Carbohydrate-modulated gene expression in plants. Annu Rev Plant Physiol Plant Mol Biol 47:509–540

Kohom BD (2000) Plasma membrane-cell wall contacts. Plant Physiol 124:31–38

Krakowski K, Bunville J, Seto J, Baskin D, Seto, D (1995) Rapid purification of fluorescent dyelabeled products in a 96-well format for high-throughput automated DNA sequencing. Nucleic Acids Res 23:4930–4931

Marra MA, Hillier L, Waterston RH (1998) Expressed sequence tags - ESTablishing bridges between genomes. Trends Genet 14:4–7

Ming R, Liu S-C, Lin Y-R, da Silva J, Wilson W, Braga D, van Deynze A, Wenslaff TE, Wu KK, Moore PH, Burnquist W, Sorrells ME, Irvine JE, Paterson AH (1998) Detailed alignment of the saccharum and sorghum chromosomes: comparative organization of closely related diploid and polyploid genomes. Genetics 150:1663–1682

Ming R, Liu S-C, Moore PH, Irvine JE, Paterson AH (2001) QTL analysis in a complex autopolyploid: genetic control of sucrose content in suarcane. Genome Res 11:2075–2084

Mullikin J, McMurray A (1999) Sequencing the Genome, Fast. Science 283:1867–1868

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936.

Nizetic D, Drmanac R, Lehrach H (1991) An improved bacterial colony lysis procedure enables direct DNA hybridisation using short (10, 11 bases) oligonucleotides to cosmids. Nucleic Acids Res 19:182

Peterson DG, Schulze SR, Sciara EB, Lee SA, Bowers JE, Nagel A, Jiang N, Tibbitts DC, Wessler SR, Paterson AH (2002a) Integration of cot analysis, DNA cloning, and high throughput sequencing facilitates genome characterization and gene discovery. Genome Res 12:795–807

Peterson DG, Wessler SR, Paterson AH (2002b) Efficient capture of sequence complexity using Cot-based cloning and sequencing. Trends Genet 18:547–550

Schmid KJ, Nigro L, Aquadro CF, Tautz D (1999) Large number of replacement polymorphisms in rapidly evolving genes of

Drosophila: implications for genome-wide surveys of DNA polymorphism. Genetics 153:1717–1729

Shaw A (2000) The 14–3-3 proteins. Curr Biol 10:R400

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MT, Aquadro CF (2001a) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc Natl Acad Sci USA 98:7375–7379

Swanson WJ, Yang Z, Wolfner MF, Aquadro CF (2001b) Positive Darwinian selection drives the evolution of several reproductive proteins in mammals. Proc Natl Acad Sci USA, 98:2509–2514

Wang M-B, Waterhouse PM (2001) Application of gene silencing in plants. Curr Opin Plant Biol 5:146–150

Wu KK, Burnquist W, Sorrells ME, Tew TL, Moore PH, Tanksley SD (1992) The detection and estimation of linkage in polyploids using single-dose restriction fragments. Theor Appl Genet 83:294–300

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573

Yang Z, Nielsen R, Goldman N, Krabbe Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449