Sorghum Expressed Sequence Tags Identify Signature Genes for Drought, Pathogenesis, and Skotomorphogenesis from a Milestone Set of 16,801 Unique Transcripts[1][w]

Lee H. Pratt*, Chun Liang, Manish Shah, Feng Sun, Haiming Wang, St. Patrick Reid[2], Alan R. Gingle, Andrew H. Paterson, Rod Wing[3], Ralph Dean[4], Robert Klein, Henry T. Nguyen[5], Hong-mei Ma, Xin Zhao, Daryl T. Morishige, John E. Mullet, and Marie-Michèle Cordonnier-Pratt

Department of Plant Biology (L.H.P., C.L., M.S., F.S., H.W., S.P.R., M.-M.C.-P.), Center for Applied Genetic Technologies (A.R.G.), Plant Genome Mapping Laboratory (A.H.P., H.-m.M.), and Department of Statistics (X.Z.), University of Georgia, Athens, Georgia 30602; Clemson University Genomics Institute (R.W.) and Department of Plant Pathology and Physiology (R.D.), Clemson University, Clemson, South Carolina 29634; United States Department of Agriculture Agricultural Research Service, Southern Plains Agricultural Research Center, College Station, Texas 77845 (R.K.); Department of Plant and Soil Sciences, Texas Tech University, Lubbock, Texas 79409 (H.T.N.); and Department of Biochemistry and Biophysics, Texas A&M University, College Station, Texas 77843 (D.T.M., J.E.M.)

Improved knowledge of the sorghum transcriptome will enhance basic understanding of how plants respond to stresses and serve as a source of genes of value to agriculture. Toward this goal, *Sorghum bicolor* L. Moench cDNA libraries were prepared from light- and dark-grown seedlings, drought-stressed plants, Colletotrichum-infected seedlings and plants, ovaries, embryos, and immature panicles. Other libraries were prepared with meristems from *Sorghum propinquum* (Kunth) Hitchc. that had been photoperiodically induced to flower, and with rhizomes from *S. propinquum* and johnsongrass (*Sorghum halepense* L. Pers.). A total of 117,682 expressed sequence tags (ESTs) were obtained representing both 3′ and 5′ sequences from about half that number of cDNA clones. A total of 16,801 unique transcripts, representing tentative UniScripts (TUs), were identified from 55,783 3′ ESTs. Of these TUs, 9,032 are represented by two or more ESTs. Collectively, these libraries were predicted to contain a total of approximately 31,000 TUs. Individual libraries, however, were predicted to contain no more than about 6,000 to 9,000, with the exception of light-grown seedlings, which yielded an estimate of close to 13,000. In addition, each library exhibits about the same level of complexity with respect to both the number of TUs preferentially expressed in that library and the frequency with which two or more ESTs is found in only that library. These results indicate that the sorghum genome is expressed in highly selective fashion in the individual organs and in response to the environmental conditions surveyed here. Close to 2,000 differentially expressed TUs were identified among the cDNA libraries examined, of which 775 were differentially expressed at a confidence level of 98%. From these 775 TUs, signature genes were identified defining drought, Colletotrichum infection, skotomorphogenesis (etiolation), ovary, immature panicle, and embryo.

The Poaceae contains numerous species of importance to human nutrition. A thorough exploration of the transcriptome of this important plant family is an important step in understanding its fundamental biology, as well as in identifying genes that will continue to improve its agricultural productivity. Defining the transcriptome of a complex, multicellular eukaryote is, however, a daunting challenge. The two most widely used and comprehensive approaches are whole-genome sequencing coupled with application of gene prediction algorithms (Mathé et al., 2002) and single-pass sequencing of cDNAs to obtain expressed sequence tags (ESTs; Adams et al., 1991). Among newer approaches that have not yet been used as widely are targeted sequencing of gene-rich regions, identified either as being hypomethylated (Rabinowicz et al., 1999; Bedell et al., 2005) or enriched in single-copy sequences (Peterson et al., 2002), and serial analysis

[1] This work was supported by the National Science Foundation Plant Genome Research Program (grant nos. DBI–9872649 to A.H.P. and DBI–0110140 to L.H.P.), by a gift from the National Grain Sorghum Producers, and by grants to H.T.N. from the Texas Advanced Technology Research Program and the U.S. Department of Agriculture National Research Initiative.

[2] Present address: Department of Microbiology, Mount Sinai School of Medicine, 1 Gustave L. Levy Place, New York, NY 10029.

[3] Present address: Arizona Genomics Institute and Department of Plant Sciences, University of Arizona, Tucson, AZ 85721.

[4] Present address: Fungal Genomics Laboratory, North Carolina State University, Raleigh, NC 27606.

[5] Present address: Department of Agronomy, Plant Sciences Unit, University of Missouri, Columbia, MO 65211.

* Corresponding author; e-mail lpratt@plantbio.uga.edu; fax 706–583–0210.

[w] The online version of this article contains Web-only data.

Article, publication date, and citation information can be found at www.plantphysiol.org/cgi/doi/10.1104/pp.105.066134.

of gene expression (Velculescu et al., 1995). No one methodological approach, however, is sufficient. Gene prediction algorithms are as yet imperfect (Mathé et al., 2002), while other methods are in a practical sense incapable of identifying every potentially expressed gene. Ultimately, a combination of strategies employed in parallel will be required to provide a near-complete description of any complex transcriptome.

Among available approaches, an appropriately designed EST project offers a number of substantial advantages: (1) It most often is a much less expensive route to gene discovery than is whole-genome sequencing; (2) it offers unambiguous identification of transcribed genomic sequences; (3) it results in a cDNA resource that can serve a broad scientific community; (4) it provides at no additional cost templates suitable for cDNA-based microarray applications as well as (5) information about gene expression as a function of developmental stage, organ, and/or environmental parameters at the time plant material is harvested for RNA isolation; and (6) it can reveal information about several transcript properties, including untranslated region (UTR) structures, polyadenylation signals, and alternative splicing. Because of these and other advantages, several EST projects in commercially important plant species have been initiated (Michalek et al., 2001; Miller et al., 2001; Fedorova et al., 2002; Fernandes et al., 2002; Shoemaker et al., 2002; Van der Hoeven et al., 2002; Kikuchi et al., 2003; Ogihara et al., 2003; Ronning et al., 2003; Vettore et al., 2003; Fei et al., 2004; Ramírez et al., 2005).

The cereals are among the agriculturally most important members of the Poaceae. The extensive synteny among their genomes (Hulbert et al., 1990; Ahn et al., 1993; Paterson et al., 1995; Bennetzen and Freeling, 1997; Gale and Devos, 1998; Draye et al., 2001; Mullet et al., 2002; Bowers et al., 2003) means that what is learned about any one of them increases our knowledge about all. We have chosen to utilize sorghum as a representative of the cereals for several reasons. Not only is it an important cereal crop in its own right (Doggett, 1988), but it has a relatively small diploid genome of approximately 750 Mb and is closely related to maize (*Zea mays*) and sugarcane (*Saccharum officinarum*), both of which have much larger polyploid genomes (Arumuganathan and Earle, 1991). Well-developed physical and genetic maps and large bacterial artificial chromosome libraries are available (Woo et al., 1994; Lin et al., 1999; Klein et al., 2000; Childs et al., 2001; Draye et al., 2001; Menz et al., 2002; Bowers et al., 2003). Moreover, sorghum is unusually well adapted to harsh environments, including high temperature, drought, and low nutrient availability, and it has been investigated extensively with respect to many important parameters such as C_4 photosynthesis, drought resistance, variation in flowering time, and acid-soil resistance. Consequently, sorghum is an excellent model system for advancing our understanding of what is almost certainly the single most important group of plants with respect to human nutrition.

We characterize and explore here 117,682 sorghum ESTs derived from approximately half that number of independent cDNAs, most of which were sequenced at both 5′ and 3′ ends. A Milestone set (freeze) of 16,801 unique transcripts, or tentative UniScripts (TUs), has been identified from 55,783 3′ ESTs and is in use for microarray applications (Buchanan et al., 2005; Salzman et al., 2005). Of the 16,801 TUs, 7,769 are singletons. These data provide for sorghum an estimate of about 31,000 for the total number of TUs in the 13 cDNA libraries investigated here. Because the overwhelming majority of cDNAs were obtained from unamplified libraries that were neither normalized nor subtracted, these data also provide quantitative information about the expression pattern of each TU among the cDNA libraries examined. This information has been used to identify 775 genes that were differentially expressed at a confidence level of 98%, as well as signature genes for drought, pathogenesis, skotomorphogenesis (etiolation), ovaries, immature panicles, and embryos.

All ESTs are available for examination and download at http://fungen.org/Sorghum.htm and http://cggc.agtec.uga.edu/cggc. Additional information concerning data access is provided in ''Materials and Methods.''

RESULTS

EST Characteristics

The 13 cDNA libraries from which ESTs were obtained are summarized in Table I. Three major considerations went into the choice of libraries. Some were selected to provide linkage to other EST projects (e.g. pathogen-infected plants, incompatible [PI1] and pathogen-infected plants, compatible [PIC1]; Ronning et al., 2003), all were selected to provide a high rate of gene discovery, and several were selected to satisfy specific biological targets. PI1 and PIC1 were selected to provide genes responding to biotic stress. Water-stressed plants (WS1) and leaves from plants stressed after flowering (DSAF1) or before flowering (DSBF1) were selected to identify drought-induced genes. Dark-grown seedlings (DG1) were included to provide insight into skotomorphogenesis. Floral-induced meristems (FM1) are intended to help reveal transcriptome changes occurring in response to photoperiodic induction of flowering. Rhizomes (RHIZ1 and RHIZ2) were selected to identify genes expressed solely or preferentially in rhizomes, with a view toward identifying candidate genes that might help in the control of johnsongrass (*Sorghum halepense* L. Pers.), which requires control in several parts of the world largely because of its rhizomatous growth habit. Time points for harvesting tissues were intended to offer a high probability of identifying genes of interest. For

Table I. *cDNA libraries from which ESTs described here were obtained*

Library Designation	Species	RNA Source
DG1	*S. bicolor*	Seedlings, including roots, germinated and grown for 5 d in total darkness at 25°C and near-saturating humidity
DSAF1	*S. bicolor*	Leaves from greenhouse-grown plants stressed by drought after flowering (library amplified and subtracted)
DSBF1	*S. bicolor*	Leaves from greenhouse-grown plants stressed by drought before flowering (library amplified and subtracted)
EM1	*S. bicolor*	Embryos 24 hr after the onset of imbibition at 25°C on white filter paper in Petri dishes
FM1	*S. propinquum*	Meristems from mature plants treated with 15 short, 8-hr photoperiods in a growth chamber, followed by 16 d in a greenhouse during the natural long days of late April and early May in Athens, GA
IP1	*S. bicolor*	Immature panicles from field-grown plants near College Station, TX
LG1	*S. bicolor*	Greenhouse-grown seedlings, 10–14 d old, including roots, 9–10 cm in height
OV1/OV2	*S. bicolor*	Immature ovaries from 8-week-old, greenhouse-grown plants
PI1	*S. bicolor*	Leaves of 2-week-old seedlings harvested 48 hr after inoculation with isolate FRM421 of *Colletotrichum graminicola* (incompatible challenge)
PIC1	*S. bicolor*	Leaves of 4-week-old plants sprayed with a spore suspension of isolate FRM421 of *C. graminicola* (compatible challenge)
RHIZ1	*S. halepense*	Rhizome apices harvested from field-grown plants near College Station, TX (library amplified)
RHIZ2	*S. propinquum*	Rhizome apices (approximately 1 cm) harvested from mature vegetative greenhouse-grown plants in Athens, GA
WS1	*S. bicolor*	Greenhouse-grown plants, 5 weeks old, including roots, on days 7 and 8 after cessation of watering

example, embryos were harvested 24 h following the onset of germination because prior work demonstrated substantial new transcript accumulation at this early time point (Hauser et al., 1998). RHIZ1 is included because it was immediately available for methodological development. RHIZ2 was produced as a replacement for RHIZ1 because it was from *Sorghum propinquum* (Kunth) Hitchc. rather than johnsongrass because the cDNA inserts in RHIZ1 were short and because RHIZ1 had been amplified. *S. propinquum* was preferred over johnsongrass because the former is the species used for one of the two comprehensive sorghum genetic maps (Bowers et al., 2003).

Initial choices for species and genotype were made for four reasons. (1) Genotype BTx623 was selected for most libraries because it is one of the most widely used *Sorghum bicolor* L. Moench accessions in breeding programs and (2) has been used as one of the parents for the construction of both of the most detailed genetic maps for sorghum (Menz et al., 2002; Bowers et al., 2003). (3) *S. propinquum* was selected in part because it was used for one of the two sorghum genetic maps and in part because it possesses rhizomes, which *S. bicolor* lack. (4) Similarly, to understand better changes that might occur during the transition of a meristem from vegetative to reproductive, the photoperiodic behavior of *S. propinquum* was exploited in order to provide appropriate starting material. DSAF1 and DSBF1 were not originally part of this project. They were included subsequently because of the added information they provide concerning drought, the major abiotic focus of this work.

From a total of 151,870 sequence attempts, 117,682 high-quality 3′ and 5′ ESTs were obtained (Table II). With the exceptions of RHIZ1, DSAF1, and DSBF1, about one- half of the clones contain full-coding-length cDNAs. Estimates for cDNAs cloned backwards from expectations range from 0.5% to 3.5%. Most libraries were sequenced to a depth of about 5,000 cDNAs. After trimming for vector, adapter, and quality, ESTs as submitted to GenBank averaged 516 and 529 nucleotides (nt) for 3′ and 5′ reads, respectively. The greatest number of trimmed sequences had lengths between 500 and 599 nt, with 89% exceeding 300 nt (Fig. 1). These sequences can be explored and downloaded as fasta files as described in ''Materials and Methods.''

Milestone TUs

Only 3′ ESTs were clustered for two reasons. First, ESTs deriving from the same gene would be expected to have substantial sequence overlap. Conversely, 5′ ESTs would be expected to start at different places

Table II. *EST characteristics and cluster information*

Library	ESTs (No.)		Average Trimmed Length		3′ ESTs Included in TUs (No.)[a]	TUs (No.)[b]	Singletons (No.)[c]	Full Coding Length	Inverted Clones
	3′	5′	3′	5′					
			nt					%	%
DG1	5,642	6,627	501	528	5,149	2,858	1,938	56	3.2
DSAF1	3,057	3,497	516[d]	516[d]	2,960	1,995	1,401	14	2.3
DSBF1	2,961	785	543[d]	543[d]	2,882	1,904	1,399	18	2.0
EM1	5,126	5,405	508	521	4,793	2,690	1,779	52	1.6
FM1	4,976	5,336	507	504	4,861	2,716	1,850	50	1.2
IP1	4,936	5,067	500	538	4,731	2,582	1,700	49	0.5
LG1	5,015	5,316	485	543	4,822	2,966	2,210	52	2.6
OV1	2,578	2,810	516	546	2,458	1,575	1,152	53	2.9
OV2	2,615	2,787	571	504	2,476	1,647	1,245	52	2.3
PI1	5,077	5,203	523	527	4,774	2,539	1,711	53	1.4
PIC1	5,042	4,522	551	592	4,823	2,646	1,699	28	0.7
RHIZ1	1,179	0	426	n.a.[e]	1,087	789	598	n.d.[f]	n.d.[f]
RHIZ2	5,308	5,978	510	513	5,175	2,610	1,665	64	1.2
WS1	5,437	5,400	537	515	4,792	2,580	1,856	50	3.5
Total	58,949	58,733	516	529	55,783	16,801	7,769[g]		

[a]Some 3′ ESTs were excluded from the Milestone 3′ EST assembly as explained in "Materials and Methods." [b]The total number of TUs represented within the indicated library; because a single TU can have ESTs from many libraries, the sum of the values in this column exceed 16,801. [c]The number of TUs in which an EST from this library appears only once. [d]Average is for both 3′ and 5′ sequences together. [e]Not applicable. [f]Not determined. [g]Total number of singletons in the Milestone assembly.

depending upon where reverse transcription (RT) terminated. Thus, by using only 3′ ESTs the much greater frequency of error associated with clustering 5′ ESTs is avoided. Wang et al. (2004) have subsequently documented with Arabidopsis (*Arabidopsis thaliana*) that the rate of erroneously separating ESTs into two or more clusters is 30% with 5′ ESTs, but only 3% with 3′ ESTs. Since we have sequenced both ends of the vast majority of cDNAs, little or nothing is lost by taking the more rigorous approach used here. Moreover, because the 3′ and 5′ ESTs from the same cDNA are linked, once the 3′ ESTs have been clustered the useful length of consensus sequences can be extended by directed incorporation of the 5′ sequences. The second reason derives from the expectation that UTRs of 3′ ESTs should discriminate better among members of a multigene family than would 5′ ESTs, thereby providing a potentially better measure of gene discovery.

The 55,783 3′ ESTs clustered here identify 6,114 singletons, 1,655 contigs-of-one, and 9,032 clusters of two or more members (Table II). When a sequence is sufficiently similar to one or more other sequences, phrap attempts to assemble it with them. If phrap ultimately fails to do so, however, the sequence is designated by phrap as a contig-of-one. The identifier of a TU in this category begins with 1. A sequence that bears so little resemblance to any other sequence that no attempt is ever made to assemble it with other sequences is designated by phrap as a singleton. The identifier for this category begins with 0. While both categories contain only one EST, it can be important to be aware that those originally identified as a contig-of-one do have a strong resemblance to one or more other TUs. The identifier of a TU with two or more members begins with a 2. For simplicity, the term singleton in the following will also refer to contigs-of-one. Collectively, singletons and assemblies with two or more members will be referred to here as TUs, as already defined. The distribution of TU consensus sequence lengths is presented in Figure 1. With few exceptions, they are little more than about 100 nt longer than individual sequences (Fig. 1). The number of TUs as a function of the number of ESTs per TU indicates that very few genes are observed to be expressed at high frequency (Fig. 2). Only 42 TUs are detected at a frequency exceeding one transcript per 1,000, while only 2,158 exceed a frequency of one per 10,000.

The relative coverage of this EST data set has been evaluated by BLASTn to 255,964 sugarcane, 416,090

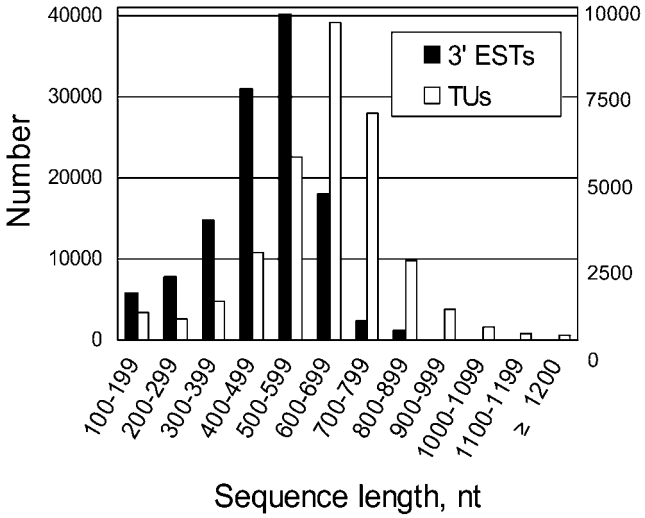


Figure 1. The number of 3′ ESTs (left-hand scale) or TUs (right-hand scale) binned by sequence or contig length, respectively.

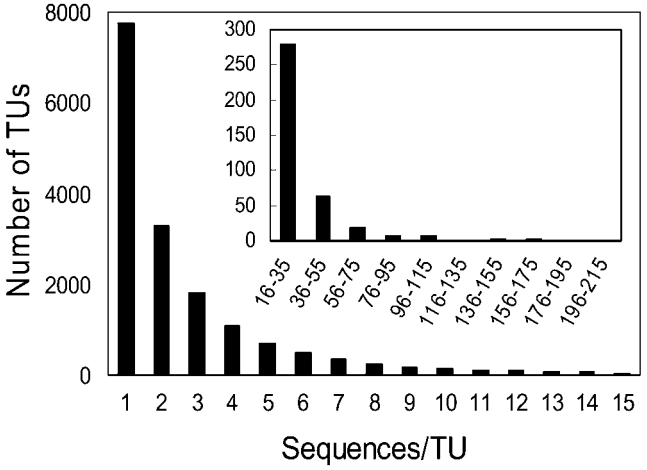


Figure 2. The number of TUs as a function of the number of 3′ ESTs per TU.

maize, and 284,234 rice (*Oryza sativa*) ESTs downloaded from GenBank on September 13, 2004. The best return for each TU from each database was binned, revealing the expected inverse relationship between frequency of high-quality hits and evolutionary distance. The percentage of TUs returning an Expect value ≤E-100 was 54.9%, 43.1%, and 11.6% for sugarcane, maize, and rice, respectively. Conversely, these percentages for Expect values >E-5 were 19.6%, 23.4%, and 35.5%, respectively. A bar chart that includes these data is presented in Supplemental Figure 1.

Discovery Rate and Distribution of TUs

Because the overwhelming majority of cDNAs were randomly selected from unamplified and nonnormalized libraries, results of clustering 3′ ESTs can be used to estimate the rate of discovery of new TUs as a function of the number of 3′ ESTs accumulated. Because the required information has been entered into the same Oracle database that also contains the results of EST clustering, it is possible to calculate and display the number of TUs as a function of the number of 3′ ESTs included in the data set (Fig. 3), to do the same for each cDNA library separately, and to do the same cumulatively, as additional libraries are added (Fig. 4). From the theoretical curve in Figure 3, obtained as described in ''Materials and Methods,'' it is then possible to define the rate of TU discovery at any number of 3′ ESTs and to extrapolate in order to obtain an estimate of the total number expected if one or more libraries were sequenced to infinite depth (Fig. 4). The rate of gene discovery remains substantial, even after sampling 55,783 cDNAs. At this point the rate of discovery of new TUs by sequencing new cDNA clones picked at random from these same libraries is predicted from the slope of the theoretical curve to be 13.6% (Fig. 3). At infinite sequencing depth, the result predicts that these libraries contain representatives of approximately 30,600 TUs (Fig. 4). Each library individually is predicted to contain representatives of no more than about 13,000 TUs, with most containing only about 6,000 to 9,000 (Fig. 4).

The richest library in terms of the maximum number of TUs predicted is that prepared from young, light-grown seedlings (LG1 in Fig. 4). TUs enhanced in their expression, however, were no more frequent in LG1 than in other libraries. This is the case whether fold induction relative to the average expression across all libraries is measured (Fig. 5), or the frequency with which TUs consisting of two or more 3′ ESTs is observed in only one library is determined (Fig. 6). For each library or subgroup, fold induction is the frequency with which that library or subgroup was represented in a TU (the number of 3′ ESTs in the TU from that library or subgroup divided by the total number of 3′ ESTs in the library or subgroup) divided by the ratio of the total number of 3′ ESTs in that TU to the total number of 3′ ESTs (55,783).

Hierarchical Clustering, Differentially Expressed TUs, and Signature Genes

Hierarchical clustering of 3′ ESTs representing the 258 TUs with 20 or more members revealed that few of these highly expressed genes were expressed uniformly among all libraries (data not shown). Similarly, an evaluation of the 10 most abundantly expressed genes indicated that most were expressed preferentially in only a few libraries, with the three drought-related libraries (WS1, DSAF1, DSBF1) accounting for the majority of expression in half of these 10 (data not shown). To explore in greater detail the ability of this EST data set to discriminate among the different environmental conditions or plant organs from which the individual libraries were obtained, the *R* statistic of Stekel et al. (2000) was calculated for the 3′ ESTs in every TU. Stekel et al. documented that the relationship between *R* and the probability that expression differs from the null hypothesis that expression is uniform among libraries must be determined independently

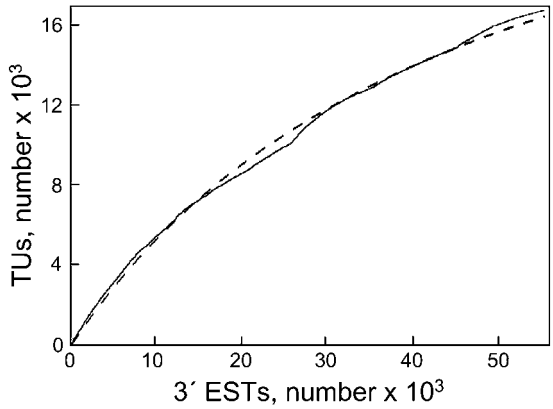


Figure 3. The number of TUs as a function of the number of 3′ ESTs accumulated. The solid line represents experimental data; the dashed line represents a best fit to those data. Inflection points occur where new libraries were introduced.

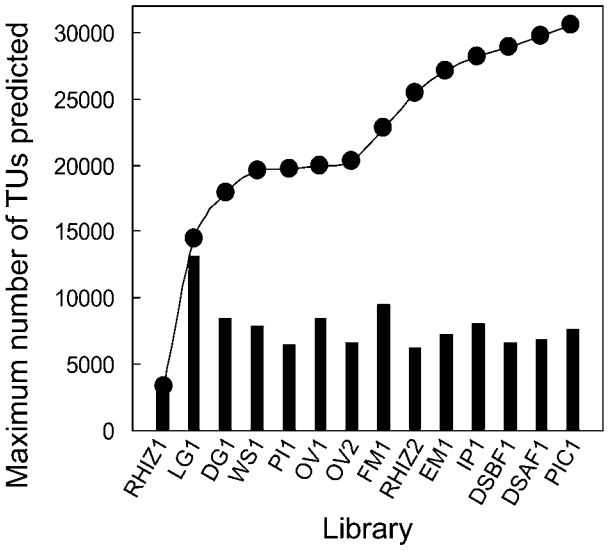


Figure 4. The maximum number of TUs predicted for each library if sampled to infinite depth (bars) and of TUs predicted as the number of libraries increases in cumulative fashion from left to right (black circles). In most cases, as an additional library is introduced the total number of TUs predicted at infinity increases.

for each data set examined. Consequently, this relationship was determined for the data set examined here (Table III). The results indicate that for $R \geq 6$, there are 3,174 TUs observed to be differentially expressed, of which 1,272 are expected to be false positives. Consequently, this analysis indicates that close to 2,000 TUs in this dataset are differentially expressed, although the believability value associated with any one TU is only 59.9%. In order to focus on TUs with greater certainty of being differentially expressed, a subset of 775 with an *R* statistic equal to or greater than 11.55, equivalent to a believability of 98% or greater, was selected for further analysis. Because DSAF1 and DSBF1 were normalized libraries, they were excluded from the analysis presented here.

This subset of 775 TUs was evaluated by hierarchical clustering, yielding the result illustrated in Figure 7. The number of members in these TUs ranged from 4 to 215. The two pathogen libraries clustered as a group, as did the three drought libraries when the analysis was repeated with the inclusion of DSAF1 and DSBF1 (data not shown). FM1 and RHIZ2, both from *S. propinquum*, also clustered together. Individual examination of several of the TUs that identify these latter two libraries (green bar to right of heat map in Fig. 7) reveal that they most often represent genes whose orthologs in *S. propinquum* and *S. bicolor* differ enough that the ESTs derived from them were separated into different TUs.

With the further exclusion of RHIZ1, RHIZ2, and FM1, 70 TUs were selected from Figure 7 and resubmitted to hierarchical clustering. These 70 TUs consisted of 10 representing each of seven subgroups or libraries. The results identify well-defined signature TUs for each of the environmental conditions (drought, pathogenesis, skotomorphogenesis, photomorphogenesis) or tissues (embryo, immature panicle, ovary) examined (Fig. 8). Two size fractions of the ovary library were picked and sequenced for a practical reason described in "Materials and Methods." Comparison of the ESTs derived from these two library fractions in Figures 7 and 8 indicates that variable size distribution in these two library fractions does lead to minor differences in the TUs identified (Fig. 7), even though ovary 1 (OV1) and OV2 nonetheless cluster well with one another (Figs. 7 and 8).

Eighteen signature genes identified by hierarchical clustering were evaluated by quantitative RT-PCR. To evaluate the utility of these signature genes, seven comparisons were made between abscisic acid (ABA)-treated and light-grown seedlings and 11 between dark- and light-grown seedlings. The former comparisons were designed to assess the utility of the signature genes with respect to ABA response, which is a subcomponent of dehydration stress, and to connect these signature genes to an in-depth microarray evaluation of ABA and dehydration responsive genes in sorghum (Buchanan et al., 2005). The latter comparisons focused on five or six TUs expressed preferentially in dark- or light-grown seedlings, respectively. Fold induction for each of these comparisons is reported in the right-most column of Figure 8. With only two exceptions, TUs 2_8855 and 2_7723, the results are consistent with those obtained by hierarchical clustering.

The entire Milestone 1.0 data set is available in comma-delimited format as Supplemental Table I. It is also available for download, together with all consensus sequences, using MAGIC Gene Discovery at http://fungen.org/Sorghum.htm. Supplemental Table I contains TU identification (ID), number of 3′ ESTs in the TU, number of 3′ ESTs in each library for

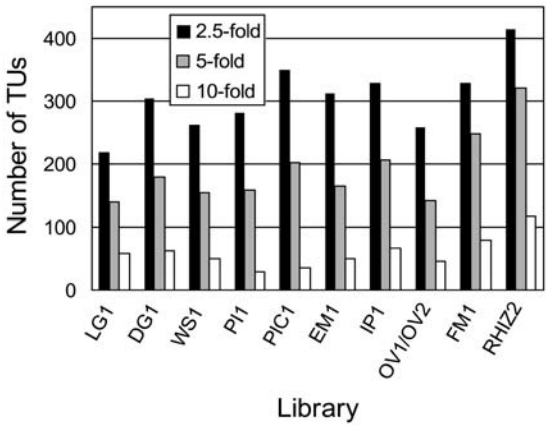


Figure 5. The number of TUs whose expression is 2.5-, 5-, or 10-fold greater in the indicated library as compared to the average expression for all libraries. Fold induction was calculated only for TUs in which at least three 3′ ESTs were detected in the indicated library and for those libraries randomly sampled to approximately the same depth of approximately 5,000 3′ ESTs.

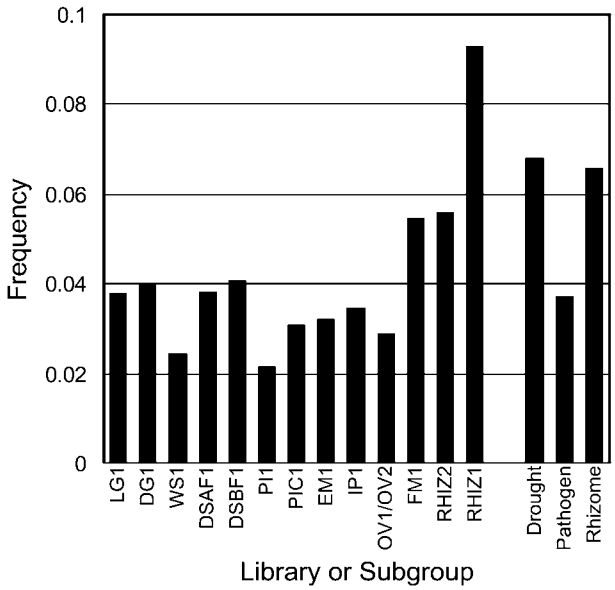


Figure 6. The frequency of TUs consisting of two or more 3′ ESTs observed exclusively in only one library or library subgroup. Drought consists of WS1, DSAF1, and DSBF1; pathogen of PI1 and PIC1; and rhizome of RHIZ1 and RHIZ2.

that TU, BLASTx target description, Expect value, Protein Information Resource Non-redundant Reference Protein (PIR-NREF) ID, and the 3′ EST that represents the TU (TU anchor sequence). Supplemental Table II provides the same information for the data in Figure 7, the *R* statistic, and the order in which TUs appear in the heat map.

DISCUSSION

The analysis of sorghum ESTs presented here is an early step in taking advantage of sorghum as a model organism for genome-scale investigations of stress-related genes among the Poaceae. It complements the more extensive effort that has already been put into mapping the sorghum genome (Whitkus et al., 1992; Chittenden et al., 1994; Paterson et al., 1995; Klein et al., 2000; Menz et al., 2002; Bowers et al., 2003) and will facilitate annotation of an eventual sorghum genome sequence. Direct association of some ESTs to an emerging physical map of the sorghum genome (Childs et al., 2001; Draye et al., 2001), together with mapping of ESTs and TU consensus sequences to the rice genome at Gramene (Ware et al., 2002), provides added value to both the sorghum ESTs described here and the physical and genetic maps.

By random sampling of a relatively large number of mostly nonnormalized, unamplified, and diverse cDNA libraries to a uniform depth of about 5,000 cDNAs, and by sequencing both 3′ and 5′ ends of each cDNA (Tables I and II), the advantages enumerated in the introduction have been realized. The random sampling permits more rigorous interpretation of the results of hierarchical clustering. Sequencing both ends of each cDNA permitted more rigorous clustering, as compared to the large majority of other plant EST projects, which focused almost exclusively on 5′ ESTs (e.g. Shoemaker et al., 2002; Ronning et al., 2003; Vettore et al., 2003; Fei et al., 2004; Ramírez et al., 2005). As Wang et al. (2004) have documented recently for Arabidopsis, clustering of 5′ ESTs resulted in a 30% overestimation of the number of unique clusters, as opposed to only 3% for 3′ ESTs. This difference results largely from a far greater frequency of insufficient overlap among 5′ as compared to 3′ ESTs. Simultaneously, however, the additional 5′ ESTs obtained here provide additional coding information and substantial amounts of 5′ UTR sequence for the many cDNA clones that are full coding length (Table II). The average high-quality trimmed read length of over 500 nt (Table II, Fig. 1), coupled with both 3′ and 5′ sequences for most cDNAs, yields more than 1 kb of sequence for the majority of TUs. Consequently, the sequences reported here, together with the cDNA clones from which they were obtained, provide not only both qualitative and quantitative information about the sorghum transcriptome, but also a rich resource for downstream applications. This resource is already in use for microarray applications (Buchanan et al., 2005; Salzman et al., 2005).

The PIR-NREF database was selected for default provisional electronic annotation for several reasons (Wu et al., 2002). It is comprehensive, incorporating sequences from six other databases, and current, with biweekly updates. It is nonredundant and well curated, with extensive source attribution. The best hit for each sequence is provided irrespective of Expect value, permitting independent judgments concerning the significance of a hit. MAGIC Gene Discovery at http://fungen.org/genediscovery (Cordonnier-Pratt et al., 2004) displays the alignment for each high-scoring pair as illustrated in Supplemental Figure 2, as well as extensive information about each BLAST return as enumerated in the legend for this figure.

Table III. *Relationship between R statistic and believability*

R	TUs Observed (No.)[a]	TUs from Randomized Data (No.)[b]	Believability
			%
4	6,168	4,522	26.7
6	3,174	1,272	59.9
8	1,731	280	83.9
10	1,051	54.5	94.8
12	680	10.1	98.5
14	499	1.8	99.6
16	378	0.31	99.9
18	305	0.05	100.0

[a]The number of TUs with an *R* statistic equal to or greater than the indicated *R* value. [b]The mean number of TUs with an *R* statistic equal to or greater than the indicated *R* value calculated following 1,000 randomizations of data as described by Stekel et al. (2000).

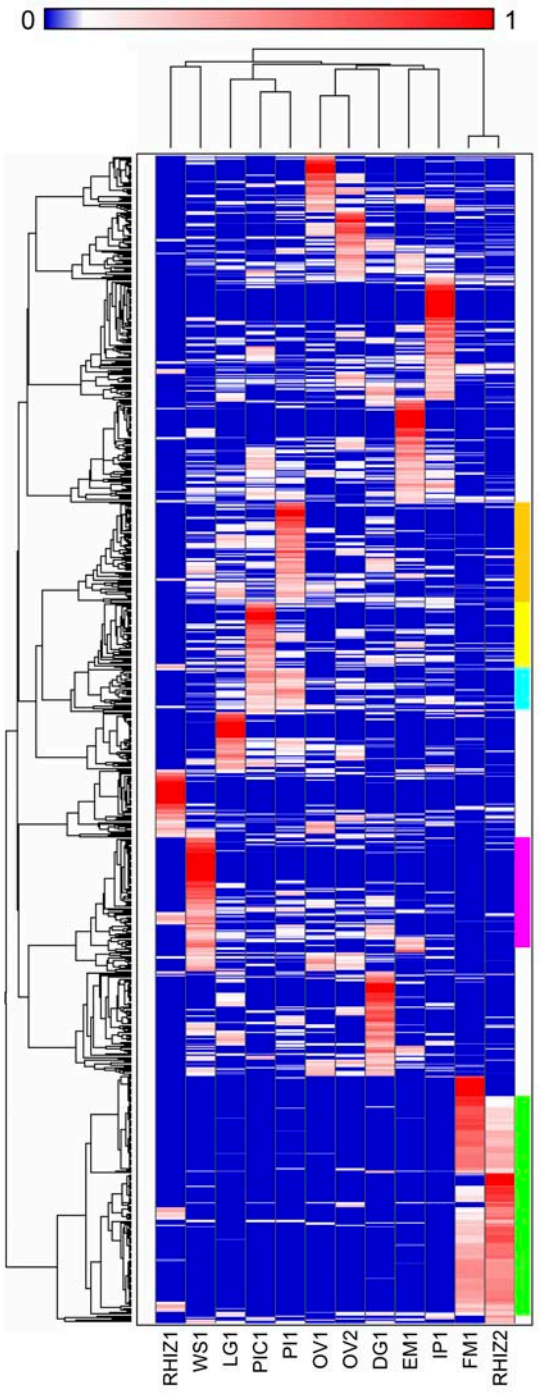


Figure 7. Hierarchical clustering of 3′ ESTs for 775 TUs with an *R* statistic equal to or greater than 11.55, which is equivalent to a believability of 98%. White represents expression at the average value observed for all libraries, while blue and red represent reduced and enhanced expression, respectively. Along the right margin, the green bar identifies signature TUs for *S. propinquum*, while the gold, yellow, blue, and violet bars identify TUs expressed preferentially in PI1, PIC1, PI1 + PIC1 together, and WS1, respectively.

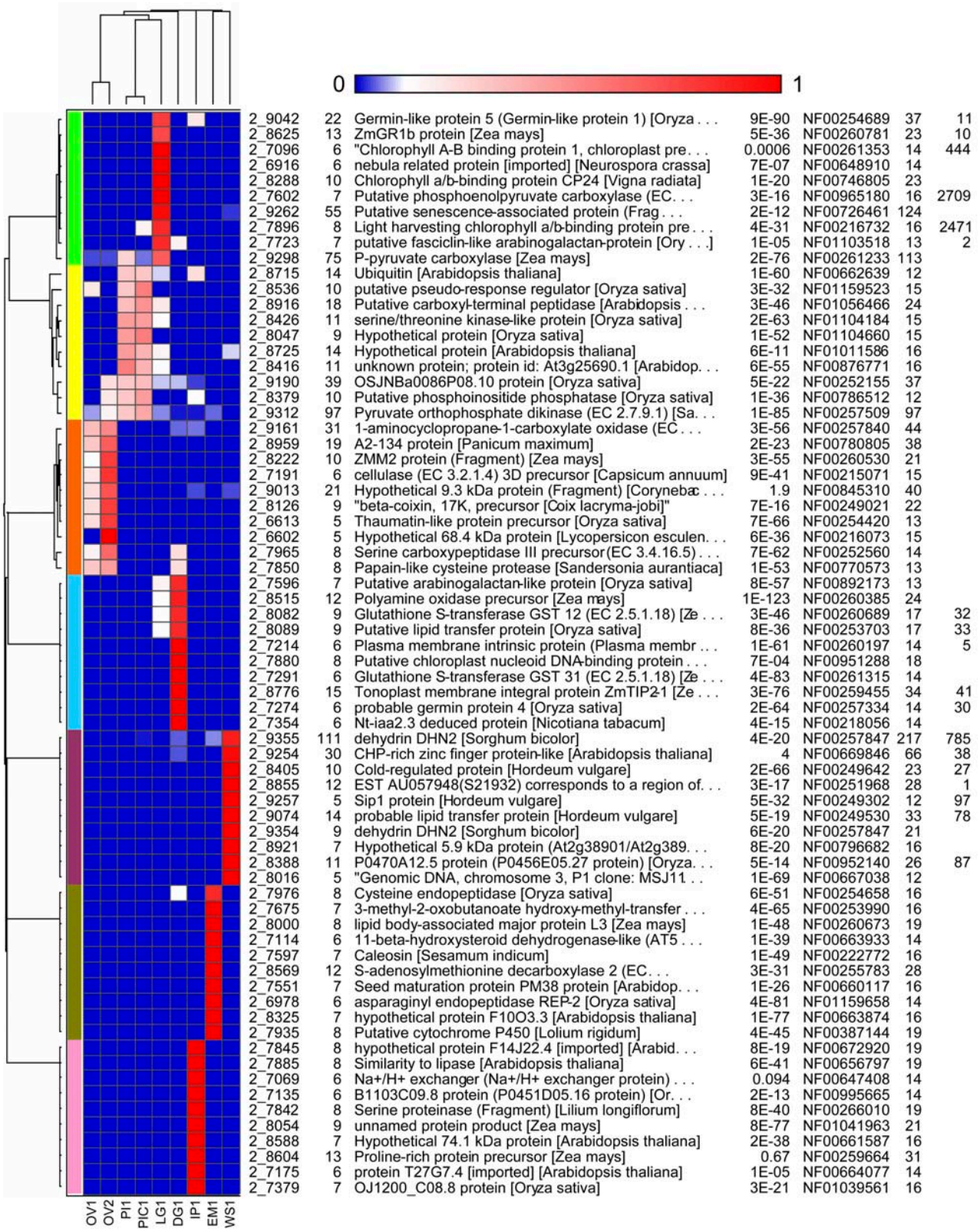
This Java graphical user interface also permits query and visualization of the results of BLAST returns from other databases, including full-length rice cDNA, both nt and protein, rice ESTs downloaded from dbEST, and the rice genome.

EST Clustering, Gene Discovery, and Transcriptome Utilization

The observation that TU consensus sequences are only slightly longer than individual 3′ ESTs (Fig. 1) is one indication that the TUs identified here are of good quality. Because all 3′ ESTs should start at or near the same position, depending upon differential polyadenylation sites, consensus sequences should never be much longer than individual sequences (Fig. 1). MAGIC Gene Discovery at http://fungen.org permits visual inspection of individual TUs, with discrepancies from the consensus sequence highlighted, and identifies ESTs that have been assembled as their reverse complement, thereby permitting independent judgments of quality (Supplemental Fig. 3). Moreover, as an additional consequence of curating the assembly in a relational database and of assigning to each TU an anchor cDNA clone, as the size of this assembly grows TU identifiers will inasmuch as possible be retained, and, when necessary, a means for tracking necessary ID changes will be provided (C. Liang, F. Sun, H. Wang, D. Kolychev, L.H. Pratt, and M.-M. Cordonnier-Pratt, unpublished data). Consequently, the value of this EST assembly will have more permanence than is usually the case, which is an important consideration when used for microarray and other downstream applications.

Given the relatively low cost, sequencing both ends of cDNAs randomly selected from predominantly unamplified and nonnormalized libraries and to a relatively shallow depth provides an excellent compromise between cost and benefit. This approach maintains a substantial rate of gene discovery (Fig. 3) without unnecessarily reducing the information content of the cDNA libraries (Figs. 4–8). The rate of gene discovery remains acceptable in part because the overwhelming majority of TUs have few members (Fig. 2) and in part because each time a new library is introduced transcripts deriving from new genes become available (Fig. 4).

Although collectively these libraries appear to contain in excess of 30,000 TUs, with the exception of LG1 no one library is predicted to contribute more than about 6,000 to 9,000 (Fig. 4). Each library also exhibits about the same level of complexity. This same observation holds when considering either the number of TUs preferentially expressed in individual libraries (Fig. 5) or the frequency with which TUs consisting of two or more 3′ ESTs coming from only one library or library subgroup are observed (Fig. 6). While the data suggest that LG1 is the richest library in terms of total number of genes being expressed (Fig. 4), it appears to provide, if anything, fewer preferentially expressed



2_9042	22	Germin-like protein 5 (Germin-like protein 1) [Oryza . . .	9E-90	NF00254689	37	11
2_8625	13	ZmGR1b protein [Zea mays]	5E-36	NF00260781	23	10
2_7096	6	"Chlorophyll A-B binding protein 1, chloroplast pre. . .	0.0006	NF00261353	14	444
2_6916	6	nebula related protein [imported] [Neurospora crassa]	7E-07	NF00648910	14	
2_8288	10	Chlorophyll a/b-binding protein CP24 [Vigna radiata]	1E-20	NF00746805	23	
2_7602	7	Putative phosphoenolpyruvate carboxylase (EC. . .	3E-16	NF00965180	16	2709
2_9262	55	Putative senescence-associated protein (Frag . . .	2E-12	NF00726461	124	
2_7896	8	Light harvesting chlorophyll a/b-binding protein pre . . .	4E-31	NF00216732	16	2471
2_7723	7	putative fasciclin-like arabinogalactan-protein [Ory . . .]	1E-05	NF01103518	13	2
2_9298	75	P-pyruvate carboxylase [Zea mays]	2E-76	NF00261233	113	
2_8715	14	Ubiquitin [Arabidopsis thaliana]	1E-60	NF00662639	12	
2_8536	10	putative pseudo-response regulator [Oryza sativa]	3E-32	NF01159523	15	
2_8916	18	Putative carboxyl-terminal peptidase [Arabidopsis . . .	3E-46	NF01056466	24	
2_8426	11	serine/threonine kinase-like protein [Oryza sativa]	2E-63	NF01104184	15	
2_8047	9	Hypothetical protein [Oryza sativa]	1E-52	NF01104660	15	
2_8725	14	Hypothetical protein [Arabidopsis thaliana]	6E-11	NF01011586	16	
2_8416	11	unknown protein; protein id: At3g25690.1 [Arabidop. . .	6E-55	NF00876771	16	
2_9190	39	OSJNBa0086P08.10 protein [Oryza sativa]	5E-22	NF00252155	37	
2_8379	10	Putative phosphoinositide phosphatase [Oryza sativa]	1E-36	NF00786512	12	
2_9312	97	Pyruvate orthophosphate dikinase (EC 2.7.9.1) [Sa. . .	1E-85	NF00257509	97	
2_9161	31	1-aminocyclopropane-1-carboxylate oxidase (EC . . .	3E-56	NF00257840	44	
2_8959	19	A2-134 protein [Panicum maximum]	2E-23	NF00780805	38	
2_8222	10	ZMM2 protein (Fragment) [Zea mays]	3E-55	NF00260530	21	
2_7191	6	cellulase (EC 3.2.1.4) 3D precursor [Capsicum annuum]	9E-41	NF00215071	15	
2_9013	21	Hypothetical 9.3 kDa protein (Fragment) [Corynebac . . .	1.9	NF00845310	40	
2_8126	9	"beta-coixin, 17K, precursor [Coix lacryma-jobi]"	7E-16	NF00249021	22	
2_6613	5	Thaumatin-like protein precursor [Oryza sativa]	7E-66	NF00254420	13	
2_6602	5	Hypothetical 68.4 kDa protein [Lycopersicon esculen. . .	6E-36	NF00216073	15	
2_7965	8	Serine carboxypeptidase III precursor (EC 3.4.16.5) . . .	7E-62	NF00252560	14	
2_7850	8	Papain-like cysteine protease [Sandersonia aurantiaca]	1E-53	NF00770573	13	
2_7596	7	Putative arabinogalactan-like protein [Oryza sativa]	8E-57	NF00892173	13	
2_8515	12	Polyamine oxidase precursor [Zea mays]	1E-123	NF00260385	24	
2_8082	9	Glutathione S-transferase GST 12 (EC 2.5.1.18) [Ze . . .	3E-46	NF00260689	17	32
2_8089	9	Putative lipid transfer protein [Oryza sativa]	8E-36	NF00253703	17	33
2_7214	6	Plasma membrane intrinsic protein (Plasma membr . . .	1E-61	NF00260197	14	5
2_7880	8	Putative chloroplast nucleoid DNA-binding protein . . .	7E-04	NF00951288	18	
2_7291	6	Glutathione S-transferase GST 31 (EC 2.5.1.18) [Ze . . .	4E-83	NF00261315	14	
2_8776	15	Tonoplast membrane integral protein ZmTIP2-1 [Ze . . .	3E-76	NF00259455	34	41
2_7274	6	probable germin protein 4 [Oryza sativa]	2E-64	NF00257334	14	30
2_7354	6	Nt-iaa2.3 deduced protein [Nicotiana tabacum]	4E-15	NF00218056	14	
2_9355	111	dehydrin DHN2 [Sorghum bicolor]	4E-20	NF00257847	217	785
2_9254	30	CHP-rich zinc finger protein-like [Arabidopsis thaliana]	4	NF00669846	66	38
2_8405	10	Cold-regulated protein [Hordeum vulgare]	2E-66	NF00249642	23	27
2_8855	12	EST AU057948(S21932) corresponds to a region of. . .	3E-17	NF00251968	28	1
2_9257	5	Sip1 protein [Hordeum vulgare]	5E-32	NF00249302	12	97
2_9074	14	probable lipid transfer protein [Hordeum vulgare]	5E-19	NF00249530	33	78
2_9354	9	dehydrin DHN2 [Sorghum bicolor]	6E-20	NF00257847	21	
2_8921	7	Hypothetical 5.9 kDa protein (At2g38901/At2g389. . .	8E-20	NF00796682	16	
2_8388	11	P0470A12.5 protein (P0456E05.27 protein) [Oryza. . .	5E-14	NF00952140	26	87
2_8016	5	"Genomic DNA, chromosome 3, P1 clone: MSJ11. . .	1E-69	NF00667038	12	
2_7976	8	Cysteine endopeptidase [Oryza sativa]	6E-51	NF00254658	16	
2_7675	7	3-methyl-2-oxobutanoate hydroxy-methyl-transfer . . .	4E-65	NF00253990	16	
2_8000	8	lipid body-associated major protein L3 [Zea mays]	1E-48	NF00260673	19	
2_7114	6	11-beta-hydroxysteroid dehydrogenase-like (AT5 . . .	1E-39	NF00663933	14	
2_7597	7	Caleosin [Sesamum indicum]	1E-49	NF00222772	16	
2_8569	12	S-adenosylmethionine decarboxylase 2 (EC. . .	3E-31	NF00255783	28	
2_7551	7	Seed maturation protein PM38 protein [Arabidop. . .	1E-26	NF00660117	16	
2_6978	6	asparaginyl endopeptidase REP-2 [Oryza sativa]	4E-81	NF01159658	14	
2_8325	7	hypothetical protein F10O3.3 [Arabidopsis thaliana]	1E-77	NF00663874	16	
2_7935	8	Putative cytochrome P450 [Lolium rigidum]	4E-45	NF00387144	19	
2_7845	8	hypothetical protein F14J22.4 [imported] [Arabid. . .	8E-19	NF00672920	19	
2_7885	8	Similarity to lipase [Arabidopsis thaliana]	6E-41	NF00656797	19	
2_7069	6	Na+/H+ exchanger (Na+/H+ exchanger protein) . . .	0.094	NF00647408	14	
2_7135	6	B1103C09.8 protein (P0451D05.16 protein) [Or. . .	2E-13	NF00995665	14	
2_7842	8	Serine proteinase (Fragment) [Lilium longiflorum]	8E-40	NF00266010	19	
2_8054	9	unnamed protein product [Zea mays]	8E-77	NF01041963	21	
2_8588	7	Hypothetical 74.1 kDa protein [Arabidopsis thaliana]	2E-38	NF00661587	16	
2_8604	13	Proline-rich protein precursor [Zea mays]	0.67	NF00259664	31	
2_7175	6	protein T27G7.4 [imported] [Arabidopsis thaliana]	1E-05	NF00664077	14	
2_7379	7	OJ1200_C08.8 protein [Oryza sativa]	3E-21	NF01039561	16	

Figure 8. Hierarchical clustering of 70 TUs selected from Figure 7, 10 for each of seven subgroups or libraries: pathogen (PI1, PIC1), ovary (OV1, OV2), WS1, DG1, IP1, EM1, and LG1. From left to right, annotations are TU ID, number of 3′ ESTs in the TU, PIR-NREF target description, Expect value, PIR-NREF ID, *R* statistic, and where available fold induction as measured by quantitative RT-PCR. Colored bars along the left-hand margin of the heat map identify sets of signature genes. The color scale is comparable to that in Figure 7.

genes (Fig. 5) and, as compared to other libraries, about the same frequency of TUs with two or more members coming from only one library (Fig. 6). Combined with the need for redundancy in order to explore events such as alternative polyadenylation and differential splicing (Burke et al., 1998; Gautheret et al., 1998; Beaudoing and Gautheret, 2001), as well as to identify polymorphisms (Buetow et al., 1999; Batley et al., 2003), the quantitative analysis presented here (Figs. 3–6) strongly supports for resource development sequencing a greater number of libraries picked randomly and to a shallow depth, as opposed to sequencing a smaller number of subtracted or normalized libraries more deeply, as has fortunately most often been the case (e.g. Fedorova et al., 2002; Shoemaker et al., 2002; Vettore et al., 2003; Fei et al., 2004).

The results just discussed, especially those in Figure 4, also document that sorghum expresses only a small fraction of its genome either in any one organ, at any one developmental stage, or in response to any specific environmental influence. We are unaware of a quantitative analysis similar to that presented here for any other plant, but see no reason why this observation should not have general validity.

The apparent enhancement in the frequency of TUs with two or more members coming from only one library that is observed for FM1, RHIZ1, and RHIZ2 (Fig. 6) results from the observation that some of the genes in *S. propinquum* (FM1, RHIZ2) and johnsongrass (RHIZ1) differ enough from their orthologs in *S. bicolor* to be grouped into different TUs when the genes are highly expressed. This outcome is not surprising given that the clustering performed here was intended to discriminate among different members of a gene family and, as a consequence, was sensitive to relatively small differences in sequence, especially in the 3′ UTR. Hierarchical clustering of the 258 TUs with 20 or more members (data not shown) provides an outcome much like that seen in Figure 7, which is to say that FM1 and RHIZ2 form a distinct cluster separate from all other libraries. Manual inspection of these *S. propinquum* signature genes (Fig. 7, green bar) reveals that they appear to be orthologs of *S. bicolor* genes that were separated into individual TUs.

Comparison to Other Plant EST Projects

Comparison by BLASTn of all 16,801 sorghum TUs to ESTs from sugarcane, maize, and rice reveal, as anticipated, decreasing similarity with increasing phylogenetic distance. Even in the case of sugarcane, however, close to 20% of sorghum TUs returned an Expect value >E-5 and almost 5% returned a value >1 (>E0; Supplemental Fig. 1). For maize, the equivalent values are just over 23% and 10%, while for rice they are just over 35% and 10%. It is evident that even when compared to these 956,288 Poaceae ESTs, the results documented here for sorghum indicate, at least superficially, that there remains a large pool of genes to be discovered by this approach. It should be noted, however, that since the bulk of sugarcane, maize, and rice ESTs are 5′ while the TUs are defined by 3′ ESTs, then one might expect insufficient overlap in at least some instances. Nonetheless, we have observed that 5′ sorghum ESTs derived from TU anchor clones often find fewer and/or poorer matches than do the 3′ TU sequences (data not shown; see also below). This outcome indicates that average cDNA lengths for other EST projects have often been relatively short such that even 5′ ESTs are near the 3′ terminus.

The 16,801 TUs identified here from 55,783 cDNAs is consistent with observations from other plant EST projects. A comparable estimate for potato (*Solanum tuberosum*; Ronning et al., 2003) from 61,940 ESTs yielded 19,892 tentative consensus sequences (TCs) and singletons. Larger data sets of >120,000 and 152,635 ESTs for soybean (*Glycine max*; Shoemaker et al., 2002) and tomato (*Lycopersicon esculentum*; Fei et al., 2004) yielded estimates of 34,264 and 31,012, respectively. In none of these three cases, however, was any correction made for the relatively high level of redundancy to be expected when 5′ ESTs are clustered (Wang et al., 2004). In contrast, while Vettore et al. (2003) estimated 43,141 putative unique transcripts from 237,954 predominantly 5′ ESTs, they also estimated redundancy to be 22%. After correction, 33,620 remained. Consequently, the 16,801 identified from 55,783 sorghum cDNAs, together with the estimate of just under 31,000 if all libraries were sequenced to infinite depth (Fig. 4), appears consistent with results of other EST projects, at least after correction for the redundancy to be expected when clustering 5′ ESTs.

Hierarchical Clustering, the *R* statistic, and Differential Expression

It is important to note that like Fei et al. (2004), but unlike most other plant EST projects, the clustering performed here was done with a data set normalized not only for the depth of sequencing of each cDNA library, but also for the strength of expression of each TU. Consequently, relative changes in expression level are more readily apparent in a heat map (e.g. Figs. 7 and 8), thereby permitting more rigorous interpretation of the results.

Evaluation as described by Stekel et al. (2000) of the potential differential expression of TUs compiled from this data set establishes that an *R* statistic of 11.55 or greater is equivalent to a believability of 98% or greater (Table III). Thus, of the 775 TUs submitted here to hierarchical clustering (Fig. 7), only 15 or 16 are expected to be false positives, leaving about 760 whose expression differs significantly from the null hypothesis that expression is uniform across all libraries. Many more than these 760 TUs are expressed differentially, however, as indicated by the excess of differentially expressed TUs observed experimentally over the number of false positives (Table III). For example, for $R \geq 6$, 3,174 differentially expressed TUs are observed. Because 1,272 are expected to be false

positives, however, only 1,902 TUs are expected to be truly expressed differentially. Since $R \geq 6$ corresponds to a believability of only 59.9% for any one TU, however, further investigation would be required to evaluate more rigorously each candidate for differential expression. Nonetheless, the point to be made is that many more genes are differentially expressed than the approximately 760 identified here with a high degree of statistical confidence.

Drought

One of the two foci of this effort was to investigate the influence of drought on the sorghum transcriptome. The 9,656 ESTs derived from libraries WS1, DSAF1, and DSBF1 define 717 TUs with two or more members and 1,517 singleton TUs containing ESTs from only these three libraries. As a group, they exhibit the highest frequency of context-specific gene expression as estimated in Figure 6. Because DSAF1 and DSBF1 when constructed were not originally designed to be included in this project, however, they were both subtracted libraries. Consequently, they were not included in the hierarchical clustering presented here. Of the 775 TUs in Figure 7, 72 are preferentially expressed in WS1 (Fig. 7, violet bar; TUs 454–525 in Supplemental Table II). Of the 1,591 ESTs in these 72 TUs, 1,042 or 65% derive from WS1. DSAF1 and DSBF1 contributed another 225 ESTs to these 72 TUs. They include three dehydrins, four heat-shock proteins, a late embryogenesis-abundant protein, a drought-inducible protein, a dehydration-responsive protein, a tonoplast-intrinsic protein, and a pore-protein homolog. About half have at least a putative or hypothetical function assigned based upon BLASTx returns from PIR-NREF.

Comparison of this entire Milestone EST data set to ABA-induced sorghum TUs identified by microarray and confirmed by quantitative RT-PCR (Buchanan et al., 2005) reveals close correspondence. A total of 55 TUs for this comparison were derived from genes up-regulated by ABA in hydroponically grown sorghum seedlings. Of the 55, 28 were observed here to be expressed only in WS1, DSAF1, and/or DSBF1, while an additional 11 were expressed predominantly in only these three libraries. Four more are included if embryo (EM1) is also considered to be a drought-related library, reflecting the desiccation that occurs during seed maturation. Of the 12 TUs remaining, one contains two out of four ESTs from WS1 and DSBF1, while three are singletons. Of the remaining eight, none had a differential digital expression profile with believability ≥98%. Given that a perfect correlation between enhanced gene expression induced by drought and ABA should not be expected in sorghum since the ABA response is only a subcomponent of the drought response (Buchanan et al., 2005; see below), the agreement between these two data sets is substantial. These data confirm, at least with respect to drought, that the signature genes identified here are useful as starting points for other drought-related investigations.

Pathogenesis

Similar to the characterization here of sorghum genes expressed preferentially in response to both compatible and incompatible infections, an earlier potato EST project obtained about 5,000 ESTs from each of two cDNA libraries, prepared from plants challenged with either a compatible or incompatible pathogen (Ronning et al., 2003). As in the case of potato singletons and TCs, a substantial number of sorghum TUs contained ESTs deriving solely from the incompatible interaction. Of those TUs with two or more members, 102 were specific to PI1 as compared to 100 specific for the incompatible challenge in potato. Of TUs with only one member, 492 were from PI1 as compared to 1,100 singletons for potato. Most of these pathogen-specific sorghum TUs or potato TCs and singletons, however, are not expressed differentially with statistical significance. Unfortunately, comparison of TUs and TCs expressed differentially with the same level of statistical significance is not possible because the relationship between the R statistic and believability appears not to have been determined for the potato data set. As Stekel et al. (2000) pointed out, this relationship is an empirical function of the data set examined. Consequently, the apparent assumption of Ronning et al. (2003) that the relationship for potato was the same as that for the four human cDNA prostate libraries investigated by Stekel et al. (2000) was invalid. Nonetheless, a comparison between sorghum TUs differentially expressed with a believability of 98% and potato TCs with an R statistic >12, identifies in both cases a subset of TUs enriched in representatives from both incompatible and compatible challenges (Fig. 7, cyan bar). For sorghum, this subset of 28 TUs is identified in Supplemental Table II as numbers 342 through 369. Also, as for potato, a second subset of sorghum sequences is relatively specific to the incompatible challenge (Fig. 7, gold bar). This subset consists of 66 TUs, numbers 233 through 297 in Supplemental Table II. Annotations for these PI1-specific TUs include pathogenesis-related proteins, chalcone synthase, chitinases, peroxidase, oxidase, and glycotransferases, as well as numerous TUs that effectively returned no meaningful annotation. Although not described for potato, a third subset of 44 sorghum TUs was preferentially expressed in PIC1 (Fig. 7, yellow bar). These TUs are numbers 298 through 341 in Supplemental Table II. Only two of these TUs have annotations comparable to those just enumerated for the PI1-specific subset, consisting of an oxidase and a wound-inductive mRNA.

A more recent tomato EST data set (Fei et al., 2004) also includes ESTs from both a compatible and an incompatible interaction, again with about 5,000 ESTs from each, as well as another 9,135 prepared from plants treated with a mix of elicitors. The tomato data set contains 169 TCs differentially expressed in these three libraries with a P value <0.05 (http://ted.bti.cornell.edu/digital/supplement/diff/disease.xls).

After excluding the data from the mixed-elicitor library and all data from TCs represented by fewer than four ESTs in the remaining two libraries, 114 remain. These 114 were then divided into 46 susceptible-specific TCs, 40 compatible-specific TCs, and 28 TCs induced about equally by both. Those expressed at least twice as frequently in one library as compared to the other were identified as specific, with the remainder being attributed to both equally. Viewed in this manner, the two data sets provide quantitatively comparable outcomes.

Signature Genes

While coexpressed TUs are sometimes expected to identify genes encoding proteins that interact with one another, the data set here is too small to provide statistically meaningful information within this context (Price and Rieffel, 2004). These same results can, however, also be used to discriminate among the different expression patterns associated with the cDNA libraries by identifying signature TUs and, by extrapolation, signature genes. This objective was accomplished by hierarchical clustering of a limited number of TUs, representing each of seven different subgroups or libraries. Hierarchical clustering of this subset provides a clear set of signature genes for each situation, as highlighted by the colored bars along the left-hand side of the heat map in Figure 8. From top to bottom these are light-grown control, pathogenesis, immature panicles, ovary, etiolation, embryo, and drought.

Annotations of the signature genes are often informative and consistent with expectations (Fig. 8). Obvious examples include a seed maturation protein in the embryo library, a dehydrin in the drought libraries, and a chlorophyll *a/b*-binding protein in the light-grown library. In addition, it will be of interest to follow up several of the annotated signature genes in order to obtain further insight into their biological function in sorghum. For example, differential representation of a pseudo-response regulator gene in the pathogen libraries may indicate that modified clock gating is important in mobilizing responses to pathogens. Similarly, the pathogen signature gene encoding a putative phosphoinositide phosphatase suggests that down-regulation of phospholipid signaling may play a role in the response of sorghum to pathogens (Laxalt and Munnik, 2002). In ovaries, increased expression of genes encoding 1-aminocyclopropane-1-carboxylate oxidase, a thaumatin-like protein and a Cys protease, is consistent with elevated levels of ethylene and jasmonate in this tissue. Elevated expression of a chloroplast nucleoid DNA protein in shoots of dark-grown as compared to light-grown sorghum seedlings is noteworthy because leaf growth, and presumably chloroplast development and gene expression is significantly inhibited in dark-grown sorghum. High expression of this protein may indicate that plastid DNA synthesis occurs in dark-grown plants and that nucleoid compaction may in some way regulate gene expression.

Other signature genes, however, are annotated as hypothetical, putative, similarity to, or some other designation indicating that annotation is at best highly speculative. Yet other genes are effectively not annotated at all, returning in two cases Expect values greater than 1. Thus, functions of the products of many or most of these signature TUs are effectively unknown. Nonetheless, their differential digital expression patterns can provide assistance in elucidating their functions, although such investigations are beyond the scope of this analysis.

Of the 70 signature TUs, eight returned from PIR-NREF an Expect value ≥E-5 (Fig. 8). Six of these eight failed to align with a region in the rice genome at Gramene (http://www.gramene.org/), thereby becoming candidates for sorghum-specific genes. Two, 2_8604 and 2_7723, returned Expect values of 5E-10 and 4E-14, respectively, following BLASTn to the rice genome. Of the six remaining, 2_7175 and 2_7723 returned values of 7E-12 and 2E-8, respectively, following BLASTx to a rice full-length mRNA database. Additionally, BLASTn of both 3′ and 5′ sequences for these six remaining TUs to dbEST at the National Center for Biotechnology Information on May 11, 2005, revealed that all but one returned multiple, significant hits from sugarcane and maize. The only exception was 2_7880. Although many additional hits were returned from sorghum ESTs produced subsequent to this analysis, only one hit with an Expect value <0.11 was returned. This hit was to sugarcane inoculated with *Gluconacetobacter diazotroficans*, returning a value of 1E-19. Thus, from these 70 TUs examined individually, only one appears to be either sorghum specific or at least expressed preferentially in this species as compared to other plants, including other grasses.

The probability is quite high that each of the 775 TUs characterized in Figure 7 is expressed differentially (Table III) and that the signature genes identified in Figure 8 are truly diagnostic. For example, the 10 TUs representing the 10 drought signature genes contain not only 203 ESTs deriving from WS1, but an additional 88 ESTs deriving from DSAF1 and DSBF1, which as noted previously were not included in the hierarchical clustering. Moreover, evaluation of 18 TUs by quantitative RT-PCR is consistent with the results of hierarchical clustering with only two exceptions (Fig. 8). In the case of TU 2_8855 no induction by ABA was detected by RT-PCR. Similarly, however, none of the 2,342 3′ ESTs from a subsequently produced sorghum ABA-induced cDNA library (http://fungen.org) can be associated with this TU. Thus, it represents a gene induced by drought, but apparently not by ABA. In the case of TU 2_7723, fold induction as assayed by RT-PCR was very low. Because arabinogalactan-proteins derive from a relatively large gene family (Gaspar et al., 2001), however, it is possible that the RT-PCR assay was responding to one or more different

members of this family than were assessed here by EST clustering. Consequently, it is evident that the analysis presented in Figures 7 and 8, and available as Supplemental Table II, provides a large number of excellent candidates for future investigation. The signature genes will be among the most useful when exploring responses that are specific to one of the treatments or tissues investigated here.

MATERIALS AND METHODS

cDNA Libraries

A total of 13 libraries were prepared from *Sorghum bicolor* L. Moench, *Sorghum propinquum* (Kunth) Hitchc., or johnsongrass (*Sorghum halepense* L. Pers.) as summarized in Table I. With the exception of libraries DSAF1 and DSBF1, *S. bicolor* libraries were prepared from genotype BTx623. DSAF1 and DSBF1, which were initially prepared for a different purpose, were from genotypes B35 and Tx7000, respectively. B35 is an inbred line with stay-green, post-flowering drought tolerance, while Tx7000 is an elite, high-yielding accession with nonstay-green, preflowering drought tolerance. For DSBF1, water was withheld after 4 weeks to impose gradual water deficit and to simulate natural preflowering drought stress. For DSAF1, final irrigation was administered 3 d after anthesis (about 2 months after sowing) to impose gradual water deficit and to simulate natural postflowering drought stress. Harvested tissue was frozen by immersion in liquid nitrogen and stored at −80°C.

Embryos were isolated by milling imbibed grain with a Quaker model 4-E plate mill (Clinton Separators), immersing the ground grain in liquid nitrogen, and filtering it through a sieve with pores of 0.84 mm. Endosperm, which was ground to a powder by the mill, passed through while embryos were retained.

With the exception of RHIZ1, DSAF1, and DSBF1, libraries were constructed by Stratagene, beginning with total RNA extracted from plant material finely ground under liquid nitrogen. cDNAs were cloned into the *Eco*RI (5′ end) and *Xho*I (3′ end) sites of lambda ZAPII. Average insert sizes, as reported by Stratagene for 12 randomly picked clones from each library, were between 1.25 and 2.0 kb. RHIZ1, DSAF1, and DSBF1 were similarly prepared in the same vector, but in the laboratories of Andrew Paterson (RHIZ1) or Henry Nguyen (DSAF1, DSBF1).

ESTs

Library phage were received from Stratagene in two or three fractions per library, with each fraction representing a different insert size range. With one exception, plasmids derived from these libraries were obtained from the fraction with the longest insert size range. In the case of the ovary library, the fraction with the longest inserts (OV2) yielded too few clones. Hence, the second of three fractions, which had the next-longest insert size range, was also used (OV1). RHIZ1, DSAF1, and DSBF1 plasmids were obtained from libraries that were amplified, but not size fractionated. DSAF1 and DSBF1 were also subtracted using driver cDNA prepared from poly(A)$^+$-RNA obtained from nonstressed sorghum leaves essentially as described by Soares and Bonaldo (1998).

Following transformation by electroporation, bacteria were plated, clones randomly picked into freezing medium in 96- or 384-well plates, and frozen at −80°C after overnight growth at 37°C in a HiGro (Genomic Solutions). All colonies used for sequencing were grown in triplicate: two sets of shallow 96-well plates for subsequent clone distribution, and one set of deep-well blocks for preparation of template DNA. The latter was prepared in the same deep-well blocks in which the bacteria were cultured, using an alkaline lysis procedure essentially as described by Roe et al. (1996; http://www.genome.ou.edu/proto.html).

ABI BigDye Terminator Cycle Sequence Ready Reaction version 2 or 3 was used at 12-fold dilution, as described by Roe et al. (1996; http://www.genome.ou.edu/big_dyes_plasmid.html). For 384 reactions, a master mix contained 268 μL BigDye, 56 μL primer, 532 μL of 400 mM Tris-Cl (pH 9.0 at 22°C), 130 μL dimethyl sulfoxide, and 214 μL water. The reverse primer for 5′ sequences was 5′-CAGGAAACAGCTATGACC-3′ (300 pmol μL^{-1}). Most 3′ sequences were primed with 5′-TAATACGACTCACTATAGGG-3′ (T7 primer, 150 pmol μL^{-1}). When poly(A) tails were sufficiently long to significantly reduce the yield of high-quality sequences, 3′ sequences were obtained with an anchored poly(T) primer (5′-T_{21}[C/G/A]-3′, 450 pmol μL^{-1}; Roe et al., 1996). Thermal cycling was done in a GeneAmp 9700 (Applied Biosystems) in either 96- or 384-well format. With a Hydra96 (Matrix Technologies), 2 μL of water was added to each well followed by 2 μL of plasmid DNA (approximately 100–200 ng μL^{-1}) dissolved in water. With a stepper pipet, 3 μL of master mix was added to each well. Thermal cycling continued to saturation (99 cycles of 96°C for 10 s, 50°C for 5 s, 60°C for 4 min) followed by a hold at 4°C. Sequencing products were cleaned by centrifugal filtration through water-equilibrated Sephadex G-50 in either 96- or 384-well filter plates.

Data Analysis

Data-processing pipelines and an Oracle database were created for this project (Cordonnier-Pratt et al., 2004). Information about each 96-well plate of plasmid DNA was entered into the database at the same time electropherograms were uploaded into the server where bases were called with phred (Ewing et al., 1998; Ewing and Green, 1998). Base calls and associated phred quality scores were parsed into the database. Vector, linker, and low-quality ends were identified using an in-house processing script (C. Liang, F. Sun, H. Wang, J. Qu, R.M. Freeman Jr., L.H. Pratt, and M.-M. Cordonnier-Pratt, unpublished data).

Vector/adaptor- and quality-trimmed 3′ ESTs were clustered and assembled with phrap (http://www.phrap.org). To reduce the frequency of poorly assembled TUs, members of each TU were resubmitted to phrap one TU at a time. Because phrap discriminates among sequences far better when assembling them in smaller groups, this resubmission eliminated most of the poorly assembled TUs by subdividing them. Extensive data for each TU was entered into database tables designed for this purpose. These data included the first and last base positions of a sequence relative to the consensus, the offset of each sequence relative to the consensus, whether a sequence had been reverse complemented to match the consensus, the length of each sequence including pads required for alignment, and all discrepancies from the consensus. From this information a normalized percentage of alignment of each trimmed and padded sequence to the consensus for its TU was determined in order to identify poorly assembled TUs. The latter were eliminated from the Milestone 1.0 assembly presented here. While these poorly assembled TUs are not included in the 16,801 TUs reported and characterized here, they have not been disregarded. Instead, they have been flagged in the database as poorly assembled and added to the Milestone TUs for use in microarray applications (Buchanan et al., 2005; Salzman et al., 2005). It is for this reason that the Milestone assembly was created from only 55,783 of the available 58,949 3′ ESTs (Table II).

The following relationship, obtained from Dr. Bruce Roe and James White of the University of Oklahoma, was used to estimate the number of TUs in a library or group of libraries as a function of sequencing depth (Figs. 3 and 4):

$$\hat{y} = G/[1 + (G \times S/n)],$$

where $\hat{y}$ is the estimated number of TUs when n number of ESTs has been obtained, G is the maximum number of TUs expected as n approaches infinity, and S is an empirically derived parameter that when multiplied by G corresponds to the number of ESTs required to obtain one-half of the maximum number of TUs. S therefore effectively determines the slope of the curve. An iterative process is used to determine the G and S values that yield a curve that best fits the experimental data as shown in Figure 3. Note that as n goes to infinity, the function is simplified to $\hat{y} = G$. This relationship is a special case of a widely used pharmacological drug responsiveness model:

$$y = b_0 - b_0/\left[1 + (x/b_2)^{b_1}\right],$$

where x is the dose level, usually in coded form such that $x \geq 1$, y is the response as a percentage of the maximum, b_0 is the expected response at saturating dose, b_2 is the concentration that produces half-maximal response, and b_1 determines the slope of the function. With the substitutions described below, the drug responsiveness model can be transformed to that used here to model the rate of gene discovery while simultaneously retaining its usefulness for modeling a saturation curve of the sort to be expected as a randomly picked cDNA library is sequenced to increasing depth. It is of general applicability and thus useful with other EST datasets. The substitutions and the rationales behind them are as follows. (1) The number of ESTs that have been obtained (n) is substituted for dose level (x), which in both cases is the

independent variable. (2) The estimated number of TUs when n ESTs have been sequenced as a proportion of the maximum number as n approaches infinity ($\hat{y}$) is substituted for drug response (y), which in both cases is the dependent variable. (3) The number of TUs expected as n approaches infinity (G) is substituted for the expected response at saturating dose (b_0), which in both cases is the maximum to be anticipated. (4) The number of ESTs required to identify one-half of the maximum number of TUs ($G \times S$) is substituted for the dose required to give the half-maximal drug response (b_2), which in both cases determines the slope of the function. (5) In addition, b_1 is set to 1, which reflects the assumption that as a randomly picked cDNA library is sequenced, the rate of gene discovery as a function of the number of ESTs sequenced remains relatively unchanged. In our experience, this has proven to be the case when one examines curves like that in Figure 3 for each individual cDNA library (data not shown).

Provisional electronic annotation of all ESTs and TU consensus sequences was obtained by BLASTx (Altschul et al., 1990, 1997) against full-coding-length entries from the PIR-NREF database (Wu et al., 2002). Output from the best hit for each sequence was parsed into the database. Only when the Expect value was greater than 10 was no entry made. Provisional electronic annotations are therefore always provided together with the Expect value to permit an independent judgment concerning significance.

BLASTx returns from this curated PIR-NREF database were also used to estimate the percentage of clones containing full-coding-length inserts and of inserts cloned inversely from expectations. For each library, BLASTx returns with an Expect value less than E-13 and with three or fewer high-scoring pairs were identified. From this subset, the percentage of query 5′ ESTs that either matched the initiating Met or contained sufficient 5′ sequence upstream of the match to encode the initiating Met was determined. This calculation assumes that a target protein is the same length as that encoded by the query sequence. While not always correct, it is nonetheless a reasonable assumption that targets are as likely to be shorter than the query as they are to be longer such that on average the assumption is reasonable. The percentage of inverted clones was estimated from the same subset of 5′ ESTs. If the reading frame was negative, that was taken as evidence that a presumed 5′ EST was in fact a 3′ EST. These calculations can be redone with different parameters using MAGIC Gene Discovery described in ''Data Access'' below.

The R statistic of Stekel et al. (2000) was determined for all Milestone TUs. Those TUs providing a value equal to or greater than 11.55 (98% believability determined as described below) were evaluated by hierarchical clustering using Spotfire Decision Site ver. 7.2. Results shown here were obtained by UPGMA clustering using Pearson's correlation as the similarity measure and average value as ordering function. Data were normalized both with respect to the number of 3′ ESTs sampled in a given library and the number of total 3′ ESTs within a TU. For each library and each TU, an initially normalized value, $I_{i,j}$, was calculated as:

$$I_{i,j} = 55,723C_{i,j}/L_iC_j,$$

where 55,723 is the total number of 3′ ESTs, L_i is the number of 3′ ESTs in the ith library (Table II), $C_{i,j}$ is the number of 3′ ESTs from the ith library in the jth TU, and C_j is the total number of 3′ ESTs in the jth TU. Fully normalized values, $N_{i,j}$, were calculated as:

$$N_{i,j} = I_{i,j}/\sum_{i=1}^{m} I_{i,j},$$

where m is the number of cDNA libraries. Hence, the expression level, $N_{i,j}$, among libraries for every TU is expressed on a scale of 0 to 1. Expression identical to that for the average across all libraries is either $1/12 = 0.0833$ (Fig. 7) or $1/9 = 0.111$ (Fig. 8).

The relationship between the R statistic and the likelihood that expression of a given TU differs significantly from the null hypothesis of uniform expression across all libraries was determined following the suggestion of Stekel et al. (2000). This determination was done by creating 1,000 randomized data sets from the experimental data set evaluated here. The R statistic for every TU was then determined for each of the 1,000 randomized data sets in order to identify the number of false positives as a function of R. A believability index was calculated as:

$$(E-F)/E,$$

where E is the number of TUs in the experimental data set with an R value equal to or greater than a specified value, and F is the mean number of TUs from 1,000 randomized data sets identified as false positives (Table III).

Quantitative RT-PCR

Quantitative RT-PCR was performed as described by Salzman et al. (2005). RNA was isolated from aerial portions of 5-d-old dark-grown seedlings, 8-d-old light-grown seedlings, and 8-d-old light-grown seedlings treated with 125 μM ABA for 27 h.

Data Access

ESTs have been deposited in GenBank. Accession numbers and associated laboratory sequence names are available in comma-delimited format in Supplemental Tables III (DG1, DSAF1, DSBF1, EM1, FM1), IV (IP1, LG1, OV1, OV2, PI1), and V (PIC1, RHIZ1, RHIZ2, WS1). Sequences can also be viewed at and downloaded from http://fungen.org/Sorghum.htm and http://cggc.agtec.uga.edu/cggc. The EST clustering analysis can be explored at http://fungen.org/Sorghum.htm. At this URL, JavaServer Pages and a pair of Java graphical user interfaces delivered by Java Web Start provide query access to an Oracle database containing all of the sorghum data reported here (Cordonnier-Pratt et al., 2004). JavaServer Pages provide drill-down access to sequences and include direct links to corresponding GenBank accessions and, where available, to locations on the rice (*Oryza sativa*) genome at Gramene. MAGIC Sequence Viewer provides graphical access to all sequences, including display of phred quality scores for individual base calls, and the ability to download sequences as fasta files, trimmed and reverse complemented as desired. MAGIC Gene Discovery displays color-coded expression profiles of TUs, provides a variety of query and filter functions for retrieving TUs meeting predefined criteria, displays contig alignments, revealing discrepancies between individual sequences and the consensus (Supplemental Fig. 3), lists all sequences in a TU, displays a wide range of information returned from BLAST to PIR-NREF and other databases, including the alignments themselves (Supplemental Fig. 2), and permits downloading in fasta format the complete set of TU consensus sequences. Clones are distributed as described at http://fungen.org/Projects/Sorghum/Clonerequests.htm.

Sequence data from this article can be found in the GenBank/EMBL data libraries under the accession numbers that are provided in Supplemental Tables III to V.

ACKNOWLEDGMENTS

L.H.P. and M.-M.C.-P. thank Drs. Bruce Roe, Doris Kupfer, and Fares Najar (University of Oklahoma) for their thorough and intensive introduction to high-throughput DNA sequencing. A.R.G. thanks Dr. Bruce Roe, Mr. James White, and Mr. Steven Kenton for their helpful introduction to the bioinformatics of high-throughput sequencing. Dr. Bruce Roe and Mr. James White kindly provided the relationship we have used to predict the number of TUs as a function of the number of 3′ ESTs accumulated.

Received May 27, 2005; revised July 20, 2005; accepted July 26, 2005; published September 16, 2005.

LITERATURE CITED

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, et al (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science **252:** 1651–1656

Ahn S, Anderson J, Sorrells M, Tanksley S (1993) Homeologous relationships of rice, wheat, and maize chromosomes. Mol Gen Genet **241:** 483–490

Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol **215:** 403–410

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389–3402

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep **9:** 208–218

Batley FJ, Barker G, O'Sullivan H, Edward KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol **132:** 84–91

Beaudoing E, Gautheret D (2001) Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. Genome Res **11:** 1520–1526

Bedell JA, Budiman MA, Nunberg A, Citek RW, Robbins D, Jones J, Flick E, Rohlfing T, Fries J, Bradford K, et al (2005) Sorghum genome sequencing by methyl filtration. PLoS Biol **3:** 103–115

Bennetzen JL, Freeling M (1997) The unified grass genome: synergy in synteny. Genome Res **7:** 301–306

Bowers JE, Abbey C, Anderson S, Chang C, Draye X, Hoppe AH, Jessup R, Lemke C, Lennington J, Li Z, et al (2003) A high-density genetic recombination map of sequence-tagged sites for *Sorghum*, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. Genetics **165:** 367–386

Buchanan C, Lim S, Salzman RA, Kagiampakis I, Klein RR, Pratt LH, Cordonnier-Pratt M-M, Klein PE, Mullet JE (2005) *Sorghum bicolor*'s transcriptome response to dehydration, high salinity and ABA. Plant Mol Biol (in press)

Buetow KH, Edmonson MN, Cassidy DB (1999) Reliable identification of large numbers of candidate SNPs from public EST data. Nat Genet **21:** 323–325

Burke J, Wang H, Hide W, Davison DB (1998) Alternative gene form discovery and candidate gene selection from gene indexing projects. Genome Res **8:** 276–290

Childs KL, Klein RR, Klein PE, Morishige DT, Mullet JE (2001) Mapping genes on an integrated sorghum genetic and physical map using cDNA selection technology. Plant J **27:** 243–255

Chittenden LM, Schertz KF, Lin YR, Wing RA, Paterson AH (1994) A detailed RFLP map of *Sorghum bicolor* x *S. propinquum*, suitable for high-density mapping, suggests ancestral duplication of sorghum chromosomes or chromosomal segments. Theor Appl Genet **87:** 925–933

Cordonnier-Pratt M-M, Liang C, Wang H, Kolychev D, Sun F, Freeman R, Sullivan R, Pratt LH (2004) MAGIC Database and interfaces: an integrated package for gene discovery and expression. Comp Funct Genom **5:** 268–275

Doggett H (1988) Sorghum, Ed 2. Blackwell Publishing, Ames, IA

Draye X, Lin YR, Qian XY, Bowers JE, Burow GB, Morrell PL, Peterson DG, Presting GG, Ren SX, Wing RA, et al (2001) Toward integration of comparative genetic, physical, diversity, and cytomolecular maps for grasses and grains, using the sorghum genome as a foundation. Plant Physiol **125:** 1325–1341

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

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

Fedorova M, van de Mortel J, Matsumoto PA, Cho J, Town CD, VandenBosch KA, Gantt JS, Vance CP (2002) Genome-wide identification of nodule-specific transcripts in the model legume *Medicago truncatula.* Plant Physiol **130:** 519–537

Fei Z, Tang X, Alba RM, White JA, Ronning CM, Martin GB, Tanksley SD, Giavannoni JJ (2004) Comprehensive EST analysis of tomato and comparative genomics of fruit ripening. Plant J **40:** 47–59

Fernandes J, Brendel V, Gai X, Lal S, Chandler VL, Elumalai RP, Galbraith DW, Pierson EA, Walbot V (2002) Comparison of RNA expression profiles based on maize expressed sequence tag frequency analysis and micro-array hybridization. Plant Physiol **128:** 896–910

Gale MD, Devos KM (1998) Comparative genetics in the grasses. Proc Natl Acad Sci USA **95:** 1971–1974

Gaspar Y, Johnson KL, McKenna JA, Bacic A, Schultz CJ (2001) The complex structures of arabinogalactan-proteins and the journey towards understanding function. Plant Mol Biol **47:** 161–176

Gautheret D, Poirot O, Lopez F, Audic S, Claverie J-M (1998) Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. Genome Res **8:** 524–530

Hauser BA, Cordonnier-Pratt M-M, Pratt LH (1998) Temporal and photoregulated expression of five tomato phytochrome genes. Plant J **14:** 431–439

Hulbert SH, Richter TE, Axtell JD, Bennetzen JL (1990) Genetic-mapping and characterization of sorghum and related crops by means of maize DNA probes. Proc Natl Acad Sci USA **87:** 4251–4255

Kikuchi S, Satoh K, Nagata T, Kawagashira N, Doi K, Kishimoto N, Yazaki J, Ishikawa M, Yamada H, Ooka H, et al (2003) Collection, mapping, and annotation of 28,000 full-length cDNA clones from *Japonica* rice. Science **301:** 376–379

Klein PE, Klein RR, Cartinhour SW, Ulanch PE, Dong J, Obert JA, Morishige DT, Schlueter SD, Childs KL, Ale M, et al (2000) A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. Genome Res **10:** 789–807

Laxalt AM, Munnik T (2002) Phospholipid signaling in plant defense. Curr Opin Plant Biol **5:** 332–338

Lin YR, Zhu LH, Ren SX, Yang JS, Schertz KF, Paterson AH (1999) A *Sorghum propinquum* BAC library, suitable for cloning genes associated with loss-of-function mutations during crop domestication. Mol Breed **5:** 511–520

Mathé C, Sagot M-F, Schiex T, Rouzé P (2002) Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res **30:** 4103–4117

Menz MA, Klein RR, Mullet JE, Obert JA, Unruh NC, Klein PE (2002) A high-density genetic map of *Sorghum bicolor* (L.) Moench based on 2926 AFLP, RFLP and SSR markers. Plant Mol Biol **48:** 483–499

Michalek W, Smilde D, Perovic D, Pleissner KP, Willscher U, Potokina J, Graner A (2001) ESTs as a resource for barley genomics (abstract no. w30). *In* Plant and Animal Genome Conference IX, January 13–17, 2001, San Diego

Miller R, Chao S, Butler E, Kang Y, Rausch C, Seaton C, Wilson C, Hsia C, Tong J, Hummel D, et al (2001) Progress of the Triticeae genome project in the U.S.: EST generation and evaluation (abstract no. P30). *In* Plant and Animal Genome Conference IX, January 13–17, 2001, San Diego

Mullet JE, Klein RR, Klein PE (2002) Sorghum bicolor: an important species for comparative grass genomics and a source of beneficial genes for agriculture. Curr Opin Plant Biol **5:** 118–121

Ogihara Y, Mochida K, Nemoto Y, Murai K, Yamazaki Y, Shin-I T, Kohara Y (2003) Correlated clustering and virtual display of gene expression patterns in the wheat life cycle by large-scale statistical analyses of expressed sequence tags. Plant J **33:** 1001–1011

Paterson A, Lin Y-R, Li Z, Schertz KF, Doebley JF, Pinson SRM, Liu SC, Stansel JW, Irvine JE (1995) Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. Science **269:** 1714–1718

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

Price MN, Rieffel E (2004) Finding coexpressed genes in counts-based data: an improved measure with validation experiments. Bioinformatics **20:** 945–952

Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, Stein L, McCombie WR, Martienssen RA (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nat Genet **23:** 305–308

Ramírez M, Graham MA, Blanco-López L, Silvente S, Medrano-Soto A, Blair MW, Hernández G, Vance CP, Lara M (2005) Sequencing and analysis of common bean ESTs: building a foundation for functional genomics. Plant Physiol **137:** 1211–1227

Roe BA, Crabtree JS, Khan AS, editors (1996) DNA Isolation and Sequencing. John Wiley, New York

Ronning CM, Stegalkina SS, Ascenzi RA, Bougri O, Hart AL, Utterbach TR, Vanaken SE, Riedmuller SB, White JA, Cho J, et al (2003) Comparative analyses of potato expressed sequence tag libraries. Plant Physiol **131:** 419–429

Salzman RA, Brady JA, Finlayson SA, Buchanan CD, Sun F, Klein PE, Klein RR, Pratt LH, Cordonnier-Pratt M-M, Mullet JE (2005) Transcriptional profiling of sorghum induced by methyl jasmonate, salicylic acid, and aminocyclopropane carboxylic acid reveals cooperative regulation and novel gene responses. Plant Physiol **138:** 352–368

Shoemaker R, Keim P, Vodkin L, Retzel E, Clifton SW, Waterston R, Smoller D, Coryell V, Khanna A, Erpelding J, et al (2002) A compilation of soybean ESTs: generation and analysis. Genome **45:** 329–338

Soares MB, Bonaldo MF (1998) Constructing and screening normalized cDNA libraries. *In* B Birren, ED Green, S Klapholz, RM Myers, J

Roskams, eds, Genome Analysis: A Laboratory Manual, Vol 2. Cold Spring Harbor Laboratory Press, New York, pp 49–157

Stekel DJ, Git Y, Falciani F (2000) The comparison of gene expression from multiple cDNA libraries. Genome Res **10:** 2055–2061

Van der Hoeven R, Ronning C, Giovannoni J, Martin G, Tanksley S (2002) Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. Plant Cell **14:** 1441–1456

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science **270:** 484–487

Vettore AL, da Silva FR, Kemper EL, Souza GM, da Silva AM, Ferro MI, Henrique-Silva F, Giglioti EA, Lemos MV, Coutinho LL, et al (2003) Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. Genome Res **13:** 2725–2735

Wang J-P, Lindsay BG, Leebens-Mack J, Cui L, Wall K, Miller WC, dePamphilis CW (2004) EST clustering error evaluation and correction. Bioinformatics **20:** 2973–2984

Ware DH, Jaiswal P, Ni J, Yap IV, Pan X, Clark KY, Teytelman L, Schmidt SC, Zhao W, Chang K, et al (2002) Gramene, a tool for grass genomics. Plant Physiol **130:** 1606–1613

Whitkus R, Doebley J, Lee M (1992) Comparative genetic mapping of sorghum and maize. Genetics **132:** 1119–1130

Woo SS, Jiang J, Gill BS, Paterson AH, Wing RA (1994) Construction and characterization of a bacterial artificial chromosome library of Sorghum bicolor. Nucleic Acids Res **22:** 4922–4931

Wu CH, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z-Z, Ledley RS, Lewis KC, Mewes H-W, Orcutt BC, et al (2002) The protein information resource: an integrated public resource of functional annotation of proteins. Nucleic Acids Res **30:** 35–37