DNA Sequence-Based "Bar Codes" for Tracking the Origins of Expressed Sequence Tags from a Maize cDNA Library Constructed Using Multiple mRNA Sources[1]

Fang Qiu[2,3], Ling Guo[2], Tsui-Jung Wen, Feng Liu[4], Daniel A. Ashlock, and Patrick S. Schnable*

Department of Agronomy (F.Q., T.-J.W., P.S.S.) and Mathematics (D.A.A.), Interdepartmental Graduate Program in Bioinformatics and Computational Biology (L.G.), Interdepartmental Genetics Graduate Programs (F.L.), Center for Plant Genomics (P.S.S.), Iowa State University, Ames, Iowa 50011

To enhance gene discovery, expressed sequence tag (EST) projects often make use of cDNA libraries produced using diverse mixtures of mRNAs. As such, expression data are lost because the origins of the resulting ESTs cannot be determined. Alternatively, multiple libraries can be prepared, each from a more restricted source of mRNAs. Although this approach allows the origins of ESTs to be determined, it requires the production of multiple libraries. A hybrid approach is reported here. A cDNA library was prepared using 21 different pools of maize (*Zea mays*) mRNAs. DNA sequence "bar codes" were added during first-strand cDNA synthesis to uniquely identify the mRNA source pool from which individual cDNAs were derived. Using a decoding algorithm that included error correction, it was possible to identify the source mRNA pool of more than 97% of the ESTs. The frequency at which a bar code is represented in an EST contig should be proportional to the abundance of the corresponding mRNA in the source pool. Consistent with this, all ESTs derived from several genes (zein and *adh1*) that are known to be exclusively expressed in kernels or preferentially expressed under anaerobic conditions, respectively, were exclusively tagged with bar codes associated with mRNA pools prepared from kernel and anaerobically treated seedlings, respectively. Hence, by allowing for the retention of expression data, the bar coding of cDNA libraries can enhance the value of EST projects.

To exploit the power of functional genomic technologies (e.g. microarrays, proteomics, and reverse genetics) in a particular species, it is desirable to have available a large collection of genes from that species. The sequencing of random cDNAs, i.e. an expressed sequence tag (EST) approach, is an attractive method for the high-throughput discovery of genes in organisms with complex genomes. However, to fully explore the gene space of an organism, EST-based gene discovery projects must overcome the challenge that genes are differentially expressed. Specifically, many genes are expressed only in specific tissues, organs, developmental states, genotypes, or under particular environmental conditions. Because of this, EST projects often make use of cDNA libraries produced using mixtures of mRNAs isolated from multiple sources. This approach, however, suffers from the disadvantage that expression data are lost because the origins of the resulting ESTs cannot be determined. Another approach is to prepare multiple libraries, each from a fairly uniform source of mRNA (e.g. one organ at a particular developmental stage). Although this approach allows the origins of ESTs to be determined, the production of multiple libraries requires a great deal of labor and time.

We report here an alternative approach for conducting EST-based gene discovery. To maximize the representation and complexity of cDNAs and thereby facilitate gene discovery, multiple sources of maize (*Zea mays*) mRNAs were pooled to construct a single cDNA library. Distinct 6-bp "bar codes" were added to the 3′ ends of each mRNA source during first-strand cDNA synthesis. It was possible to identify the source mRNA pool of more than 97% of the ESTs from this library.

RESULTS

Library Construction and EST Sequencing

A cDNA library was prepared using a complex mixture of mRNAs from the maize inbred line B73. To maximize the gene representation in this library and to thereby facilitate gene discovery, mRNA samples were extracted from 60 different plant samples

[1] This research was supported by the National Science Foundation Plant Genome Program (grant no. DBI–9975868 to P.S.S., D.A.A., and others) and by the Iowa Corn Promotion Board (to P.S.S.). This is journal paper number J–19730 of the Iowa Agriculture and Home Economics Experiment Station (Ames; project no. 3409, supported by Hatch Act and State of Iowa funds).

[2] These authors contributed substantially to this report. F.Q. constructed the bar-coded cDNA library. L.G. conducted the bioinformatic analysis of the EST data and extracted expression data from the bar-coded ESTs.

[3] Present address: Department of Statistics, Iowa State University, Ames, IA 50011.

[4] Present address: Department of Biological Chemistry, University of California at Irvine, Irvine, CA 92697.

* Corresponding author; e-mail schnable@iastate.edu; fax 515–294–5256.

www.plantphysiol.org/cgi/doi/10.1104/pp.103.025015.

that included various organs, at various stages of development, and that had been subjected to various treatments. These mRNA samples were grouped into 21 pools (Table I). First-strand cDNA synthesis was conducted on each pool using unique *Not*I/oligo(dT) primers that differed by the inclusion of unique 6-bp DNA sequence bar codes embedded between the *Not*I cloning site and $(dT)_{18}$. First-strand cDNAs were pooled and used to construct a single cDNA library (ISUM6) that contained approximately 1.15×10^6 clones. On the basis of double restriction enzyme digestion analysis of 96 random clones, the average length of cDNA inserts is 850 to 900 bp, and the frequency of empty vectors is 2% (data not shown).

Sequencing reactions were performed on 5,184 cDNA clones from library ISUM6 using a primer that provides data from the 3′ end of the cDNAs. Of these attempts, 3,684 (71%), resulted in EST sequences that included a poly(T) tail and more than 200 bp of high-quality, non-vector sequences. These ESTs have been deposited in GenBank as GenInfo Identifier nos. 18177912 to 18181595.

Extracting Expression Data from Bar-Coded ESTs

A method to decipher bar codes was developed so that the mRNA source pool from which an individual EST was derived could be ascertained (see "Materials and Methods"). Of the 3,684 sequences that were passed to this decoding algorithm, 3,531 (95.8%) had exact bar code matches, 70 (1.9%) had errors in their bar codes that were decodable, and 83 (2.3%) were not decodable (see "Materials and Methods"). Hence, the origins of more than 97% of the ESTs from this cDNA library could be determined. The distribution of the bar codes among the ESTs is provided in Table I. Even though efforts were made to use equal amounts of first-strand cDNA from each of the 21 mRNA source pools for the construction of the cDNA library, there are approximately 3-fold differences in the representation of pools within this collection of ESTs. This could be a consequence of differences in the quality of the pools of first-strand cDNAs and/or errors in measuring the concentration of first-strand cDNAs in these pools.

Unlike EST projects that are composed of 5′ sequences from a variety of genetic backgrounds, it is possible for 3′ ESTs, all of which are from the same genetic background (the inbred B73), to be assembled into a set of unique sequence clusters (i.e. genes) with a high degree of confidence. Using CAP3 (Huang and Madan, 1999), the 3,684 ESTs were clustered into 2,250 genes consisting of 483 contigs and 1,767 singletons (Table II).

Because library ISUM6 was not normalized, the frequencies at which particular bar codes appear within a contig should correspond to the relative expression levels of the corresponding gene in the 21 pools of mRNA. The numbers of bar codes detected within each of the 483 EST contigs are shown in Table

Table I. *Bar codes, mRNA sources, and distribution of bar codes among 3,684 ESTs*

Bar Code Identification No.	Bar Code	mRNA Sources[a]	No. ESTs (% of Total)	Redundancy[b]
				%
0	No tag	Unknown	85 (2.3)	4
1	ATACGC	Germinating seeds and seedlings (1, 2, 8, and 11 DAPl[c])	173 (4.7)	9
2	ACTGGC	Mixed tissues (17, 21, 38, 69, and 77 DAPl)	184 (5.0)	7
3	CACAGC	Kernels (3, 5, 10, 15, 20, 25, and 30 DAPo[d])	250 (6.8)	24
4	TAACCC	Adventitious roots (65 DAPl)	252 (6.8)	15
5	CAGGCG	Tassels (3–39 cm, 53 and 56 DAPl)	65 (1.8)	0
6	AGGTAC	Immature ears (0.2–3.0 cm, 53, 56, 59 DAPl)	187 (5.1)	10
7	TGAGCG	Husks (73 DAPl)	89 (2.4)	4
8	GACCAC	Silks (73 DAPl)	189 (5.1)	15
9	AATCGG	First ears (73 DAPl, unpollinated)	209 (5.7)	6
10	CTAAGG	Ear shanks (73 DAPl)	227 (6.2)	9
11	GAAGAG	Etiolated seedlings (8 DAPl)	141 (3.8)	11
12	AGTGAG	Callus	184 (5.0)	22
13	GTGGAC	Cycloheximide-treated callus	206 (5.6)	15
14	GTCACC	Anaerobically-treated seedlings	214 (5.8)	19
15	CGTCCA	α-Naphthalene acetic acid-treated seedlings	142 (3.9)	15
16	GATGCC	Kinetin-treated seedlings	82 (2.2)	17
17	AAGACC	1-Aminocyclopropane-1-carboxylix acid-treated seedlings	154 (4.2)	18
18	GCCTCA	Brassinolide-treated seedlings	177 (4.8)	20
19	CTAGCC	Abscisic acid (ABA)-treated seedlings	128 (3.5)	9
20	TACGGA	GA_3-treated seedlings	247 (6.7)	21
21	GCAGGA	Jasmonic acid-treated seedlings	99 (2.7)	21

[a] Treatments and hormones are defined in "Materials and Methods." [b] (1 − no. of unigenes/no. of ESTs)*100 with indicated bar codes. No. of unigenes equals no. of contigs plus no. of singletons obtained following CAP3 analysis of all ESTs having an indicated bar code. [c] DAPl, Days after planting. [d] DAPo, Days after pollination

Table II. *Nos. of genes defined by contigs of between 1 and 135 clustered ESTs*

No. of ESTs per Contig	No. of Genes
1	1,767
2	234
3	106
4	48
5	26
6	23
7	13
8	9
9	4
10	2
12	4
13	1
14	2
15	1
16	1
17	1
19	3
23	2
27	1
50	1
135	1

III. Approximately 12% of the EST contigs are derived from a single mRNA source pool. This is almost certainly an overestimate of the number of maize genes that are expressed in a single mRNA source, because many of the EST contigs in this study are not large enough to adequately sample the expression space.

The utility of using bar codes to extract expression data can, however, be confirmed by the analysis of several of the larger EST contigs. The distribution of bar codes among ESTs that comprise the 20 contigs with 10 or more members is shown in Table IV. All of the ESTs in two contigs (297 and 305) that have decipherable bar codes have the bar code corresponding to the kernel mRNA pool (bar code 3). On the basis of BLASTX analysis both of these contigs are derived from genes that encode proteins (i.e. zeins) that accumulate predominately, if not exclusively, in kernel endosperms (Woo et al., 2001). On the basis of BLASTX analysis, the eight ESTs in contig 448 are derived from the *adh1* (*alcohol dehydrogenase 1*) gene (data not shown). All of these ESTs were isolated from mRNA pools 12 and 14 that had been subjected to anaerobic stress (data not shown), a treatment that is known to induce the expression of *adh1* (Bailey-Serres and Dawe, 1996). These results suggest that if a larger number of bar-coded ESTs were to be examined, it should be possible to identify genes that are differentially expressed among mRNA pools.

Table III. *The nos. of distinct bar codes detected among each of 483 EST contigs*

No. of mRNA Sources per Contig	No. of Contigs (%)
No tag	1 (0.21)
1	58 (12.0)
2	226 (47.0)
3	103 (21.0)
4	40 (8.3)
5	24 (5.0)
6	8 (1.7)
7	8 (1.7)
8	3 (0.62)
9	4 (0.83)
10	1 (0.21)
12	4 (0.83)
16	2 (0.41)
18	1 (0.21)

The largest contig (196) consists of 135 ESTs and is derived from a gene that encodes a metallothionein-like protein (Table IV). The distribution of bar codes associated with ESTs in contig 196 was used to examine the expression pattern of this metallothionein-like gene. The numbers of ESTs in this contig that were isolated from the various mRNA pools differs from that expected based on the distribution of bar codes in the entire EST collection (Fig. 1; Table I). This gene is overexpressed in mRNA pools 18 and 20 to 21 and is underexpressed in pools 2 to 3, 6, 8 to 10, and 13. Hence, this gene is apparently up-regulated in seedlings treated with the plant hormones brassinolide, GA_3, and jasmonic acid; down-regulated in the presence of cycloheximide; and not well expressed in mature tissues, kernels, and female reproductive structures.

Five of the contigs encode proteins that are novel or are most similar to proteins that lack even a predicted functional assignment (contigs 180, 257, 264, 329, and 439). Any expression data that can be extracted from the bar-coded ESTs will provide clues as to the functions of these genes. Nine of the 20 ESTs in contig 264 carry bar code 20; this suggests that GA_3 induces this gene. Five of the 19 ESTs in contig 439 carry bar code 14, suggesting that this gene is induced under anaerobic conditions.

Contigs 55 and 339 encode proteins that are similar to ABA-induced proteins. Interestingly, only one of 14 and zero of 14 of the ESTs from these contigs carry the bar code (19) that is associated with the mRNA pool from ABA-treated seedlings. Because these rates are less than the rate of ESTs with this bar code in the entire library (3.5%, Table I), it appears that the genes defined by these contigs are not induced by ABA, at least at the level of mRNA accumulation under the induction conditions used in this study. In contrast, the gene associated with contig 111, which encodes a protein similar to one induced by drought, does appear to be overexpressed in adventitious roots, in that five of 23 ESTs from this contig are derived from this mRNA pool.

Analyses of the bar-coded ESTs generated by this project support the prior observation that the protein inhibitor cycloheximide deregulates gene expression (Koshiba et al., 1995). If cycloheximide did not affect

Table IV. *Distributions of bar codes within the 20 contigs consisting of ten or more ESTs*

Contig Identification No.	Sequence Similarity			Bar Codes[a]																						
	Accession No.	Protein	E Value[b]	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Total
10	P12653	Glutathione *S*-transferase I (maize)	6e-76		1		1	2				3				2	1	2			1	1		1		15
55	P10979	ABA-inducible protein (maize)	8e-21			1						1		1	2	2				1		3	1	2		14
68	P21569	Peptidylprolyl isomerase (maize)	3e-57			2	3			1		3		3	2	4		3	1		2	1		2		27
111	BAB68268	Drought-inducible protein (sugarcane [*Saccharum officinarum*])	3e-20	2	1	3		5	1	1		3		1	1				1		2		1		1	23
180	BAC16424	Unknown protein (rice [*Oryza sativa*])	8e-22	1	1	2	1	1			1	1	1	1	2	1		3	1		3	1	1	1		23
196	P30571	Metallothionein-like protein (maize)	1e-19	2	7	1	1	5		1	4				6	5		3	11	10	14	19	1	31	14	135
257	N/A	No hit	N/A		1			2		2	1					1		3					2			12
264	N/A	No hit	N/A		1										2			2	1					9	2	17
269	S60453	Glc starvation-induced protein (maize)	2e-59									3				1	1		1	2		2	6		3	19
276	Q42443	Thioredoxin H-type	7e-39			2				1	1	2	1			3	1					1			1	13
290	BAA25394	Light-harvesting chlorophyll *a*/*b*-binding protein (*Nicotiana sylvestris*)	3e-88		1	1						1			2				1		1	2	1			10
297	AAL16980	15-kD β-zein (maize)	2e-49	1			18																			19
305	P08031	16-kD β-zein precursor (maize)	2e-45				16																			16
329	N/A	No hit	N/A				1	1	1	1	1	1	1	1				1	1			1	1			12
339	T02663	ABA- and stress-inducible protein (rice)	5e-15		1		1	3				1	1	1				2	2					2		14
353	AAL56401	Cystatin (wheat [Triticum aestivum])	3e-14				1	1					1	2	1	5						1				12
385	P28814	Barwin (barley [*Hordeum vulgare*]	7e-49							1		1			2		2		1	1		2	1		1	12
387	T01354	Herbicide safener-binding protein (maize)	2e-62				1			1		2			1			1					1	1	2	10
439	N/A	No hit	N/A		1		1	1					1	2	2			5	1		1	1	2	1		19
470	S20846	Gly-rich protein (maize)	1e-26		1	3	9	7	1	6	1	2	1	4	2	1		1	3		1	4	2	1		50
Total				6	16	15	54	28	3	15	9	24	7	16	25	25	5	26	25	14	25	39	20	51	24	472
Expected no.[c]				12	23	25	29	33	9	25	12	24	30	31	17	23	30	28	17	10	18	20	16	29	11	472
Chi-square value[d]							*						**	*			**					*		*	*	

[a] The mRNA sources associated with bar code identification nos. are provided in Table I. [b] E value of BLASTX. [c] No. expected if the relative frequencies of these bar codes are the same in this set of highly expressed ESTs as in the entire EST collection. Calculated as 472 (the no. of highly expressed ESTs in this table) times the frequency of a particular bar code in the entire EST collection as shown in Table I. [d] Deviations from 1:1 are indicated by * (0.05 confidence level) or ** (0.01 confidence level).

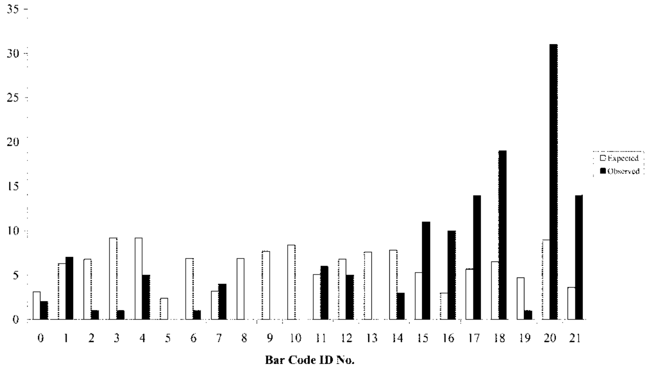

Figure 1. The distribution of bar codes associated with EST contig 196. The predicted distribution was calculated by multiplying the number of ESTs in contig 196 by the frequency at which each bar code appears in the library (Table I). The observed rates at which bar codes 2, 3, 6, 8, 9, 10, 13, 18, 20, and 21 were recovered are significantly different from predicted (0.05 or 0.01 level of significance).

the distribution of mRNAs, the ratios of ESTs with bar code 12 (tissue culture without cycloheximide treatment) or bar code 13 (tissue culture with cycloheximide treatment) to the predicted numbers of these bar codes should be the same. Although the observed number of ESTs that carry bar code 12 is close to the predicted number, the observed number that carry bar code 13 is substantially less than predicted. Hence, treatment with cycloheximide is likely to increase the gene representation in cDNA libraries. The efficiency of gene discovery (Contigs/EST) was higher in cycloheximide-treated calli than in untreated calli (Table I). In contrast, treatment with jasmonic acid increased the representation of the highly expressed genes.

The bar codes associated with unpollinated ears (bar code 9) and ear shanks (bar code 10) were observed at substantially lower than predicted rates (Table IV; Fig. 1). These results suggest that these structures have reduced expression of the 20 genes that were most prevalent in this collection of ESTs. Libraries prepared from these structures might therefore be ideal subjects for further gene discovery.

DISCUSSION

Application of Bar Codes to EST Projects

Data mining tools have been developed to extract gene expression data from EST databases (Zhang et al., 1997; Lal et al., 1999; Schmitt et al., 1999; Scheurle et al., 2000; Wheeler et al., 2000). All such methods require databases that include information regarding the mRNA sources from which ESTs were derived. In typical EST projects, this is accomplished by preparing multiple cDNA libraries from a limited number of mRNA pools, each of which was isolated perhaps from a particular organ or after a particular treatment. As such, the mRNA source for all ESTs from a given library is known. This procedure, however, requires the production of multiple libraries.

Here, we describe a method to tag cDNAs from different mRNA pools with unique DNA sequence bar codes before the preparation of a library and the means to interpret the resulting data. In a non-normalized cDNA library, the frequency at which a bar code is represented in an EST contig should be proportional to the abundance of the corresponding mRNA species in the source mRNA pool. The utility of this approach was established by demonstrating that all of the ESTs derived from zein and *adh1* genes, which are preferentially expressed in kernels and under anaerobic stress were exclusively tagged with bar codes associated with mRNA pools prepared from kernels and anaerobically treated seedlings, respectively. Although only a few thousand ESTs were sequenced in this study, analysis of the largest EST contig provided quantitative data concerning the expression of a metallothionein gene. Data were also obtained that suggest that several novel genes are up-regulated by the plant hormone GA_3 or anaerobic stress. In addition, several EST contigs that exhibit similarity to putative ABA-responsive genes do not exhibit strong evidence of ABA induction at the level of mRNA accumulation.

The utility of bar-coded EST data increases in proportion to the size of the data set. Hence, EST data would be substantially more useful if the bar coding of cDNA libraries were widely adopted. Significantly, this could be achieved for little additional cost.

Enhanced Bar Codes

Although bar codes have been used previously in the construction of cDNA libraries (Bonaldo et al., 1996) the purpose was to prevent mix-ups between libraries. Hence, all of the cDNAs in a given library contained the same bar code. Therefore, these bar codes did not provide "within library" expression data. In addition, the bar codes used by Bonaldo et al. (1996) were random 2- to 6-bp oligonucleotides. Hence, it was not possible to correct errors caused by mutation or sequencing errors in the bar codes. In contrast, in the current study, bar codes were designed to allow for error correction. This is important because mutations can be created during oligonucleotide synthesis or the in vivo propagation of cDNAs, or errors can occur during the sequencing of ESTs that would interfere with the decoding of bar code data from EST sequences. Decoding consists of locating the bar code within the EST sequence by identifying the vector and the poly(T) sequences and then determining whether the bases at the approximate location of the bar code match any of the bar codes used in the construction of the library. If bar codes are designed to allow for error correction, it is possible to identify the mRNA pool from which an EST is derived, even if one or more mutations have occurred in the associated bar code.

The natural metric for the design of DNA bar codes is the edit metric (Gusfield, 1997) where the distance between two strings of DNA (the bar code sequences) is the minimal number of (one base) insertions, deletions, or substitutions required to transform one string into the other. The edit metric is thus the natural method of enumerating errors. Once the error metric has been identified, an error-correcting code consists of strings (i.e. bar codes) that are well separated in that notion of distance. The bar codes described in this study (Table I) have a minimum pair wise edit distance of three, which permits the correction of one error. In general, a minimum distance of $2n + 1$ errors permits the correction of n errors, because only concatenations of n errors are unambiguously closer to one particular bar code.

In the current study, the rates of single mutations and multiple mutations in bar codes were 1.9% and 2.3%. Because it was possible to correct single mutations, mRNA source data were unavailable for only 2.3% of the ESTs. The rate of uncorrectable errors might, however, be higher in other bar-coded libraries. This is because the error rate is likely to depend on a number of factors, e.g. the quality of the oligonucleotides used for first-strand cDNA synthesis, the *Escherichia coli* host strain in which the library is propagated, and the DNA sequencing protocol. It might therefore be desirable to use in the future a set of bar codes that would allow for the correction of two errors, i.e. that are at least five edits apart. If the length of bar codes is increased by just 2 bp (to 8 bp), it is possible to design 34 unique bar codes that meet this criterion (Ashlock et al., 2002).

MATERIALS AND METHODS

Sources of mRNAs

Sixty tissue samples that included different stages of development, organs, and various treatments were collected from the maize (*Zea mays*) inbred line B73. Before mRNA extraction, RNA samples were grouped into 21 pools (Table I). Pool 1 consisted of RNAs from germinated seeds and seedlings grown in paper rolls and collected 1, 2, 8, and 11 d after planting. Pool 2 contained RNAs from a mixture of tissues from field-grown plants 17, 21, 38, 69, and 77 d after planting. Pool 3 consisted of RNAs from kernels collected 3, 5, 10, 15, 20, 25, and 30 d after pollination. Pool 4 consisted of RNAs from adventitious roots collected from field-grown plants 65 d after planting. Pool 5 consisted of RNAs from tassels with lengths between 3 and 39 cm collected from plants 53 and 56 d after planting. Pool 6 consisted of RNAs from immature ears with lengths of between 0.2 and 3.0 cm collected from plants 53, 56, and 59 d after planting. Pools 7 to 10 consisted of RNAs from husks, silks, unpollinated first ears, and ear shanks collected from three plants 73 d after planting, respectively. Pool 11 consisted of RNAs from etiolated seedlings grown in paper rolls in the dark and collected 8 d after planting. Pool 12 consisted of RNAs from calli derived from immature zygotic embryos that had been tissue-cultured in medium based on N6 salts for 72 d (Songstad et al., 1991). Pool 13 consisted of RNAs from similar callus tissue but cultured in the presence of 5, 10, and 20 μM cycloheximide and collected 15, 30, and 60 min after treatment. Cycloheximide is an inhibitor of protein synthesis that can derepress gene expression (Koshiba et al., 1995). Pool 14 consisted of RNAs from seedlings having coleoptiles between 0.5 and 4 cm in length that were subjected to anoxia via immersion for 6, 8, and 12 h in a pH 7.0 solution of 5 mM Tris-HCl, 100 mg L^{-1} ampicillin (Lemke-Keyes and Sachs, 1989). Pools 15 to 20 consisted of RNAs from 11-d-old seedlings grown in paper rolls partially immersed in water containing 10^{-5}, 10^{-6}, or 10^{-7} M of the plant hormones α-naphthalene acetic acid, kinetin, 1-aminocyclopropane-1-carboxylix acid, brassinolide, ABA, or GA_3, respectively. Pool 21 consisted of RNAs from 11-d-old seedling grown in paper rolls partially immersed in water containing 10^{-7} or 10^{-8} M jasmonic acid.

Construction of the Bar-Coded cDNA Library

RNAs were extracted from the 60 samples using Trizol Reagent (Invitrogen, Carlsbad, CA) and examined for RNase activity using the RNase Alert Kit (Ambion, Austin, TX). Equal amounts of the RNA samples were combined to form the 21 pools as shown in Table I. Pooled RNA samples were digested with DNase I (Invitrogen) to remove DNA contamination and precipitated using LiCl. Extraction of mRNA from these 21 RNA pools was performed using Oligotex mRNA kits (Qiagen USA, Valencia, CA). First-strand cDNAs were prepared from the 21 mRNA pools by priming with 21 distinct *Not*I/oligo(dT) primers that contained distinguishable bar code tags, $(N)_6$, 5′-AAC TGG AAG AAT TCG CGG CCG CNN NNN NTT TTT TTT TTT TTT TTT T-3′. The bar code tags associated with specific pools are shown in Table I and can be used to identify the mRNA pool from which a particular cDNA clone was derived.

Bar codes are potentially subject to mutations such as insertions, deletions, and substitutions during primer synthesis and subsequent in vitro and in vivo manipulations that can confuse the origins of clones that have nearly identical bar codes. An edit-distance lexicode algorithm (Ashlock et al., 2002) was used to generate a large set of bar codes with a specified minimal pair wise edit distance. For this library, the bar-code tags produced by the algorithm were 6 bp in length with a minimum edit distance of three. This permits the unambiguous correction of one error, which could be a base pair insertion, deletion, or substitution.

The lexicode algorithm for locating error correcting codes, which in this case are collections of DNA bar codes, is as follows. The members of a set of potential bar codes, e.g. all length 6 DNA words, are placed in alphabetical order. The algorithm is given a minimum pair wise edit distance *d* that must exist between any two barcodes. An empty set of barcodes B is initialized. Traversing the list of potential barcodes in order, a word is added to B if it is at least *d* edits from every word already in B. Because it takes the next possible word, the lexicode algorithm is an example of a greedy algorithm.

The unmodified lexicode algorithm does not locate maximal size barcode sets for a given length and minimum distance. The algorithm can be modified by handing it a nonempty set B of initial words that are in the barcode set by fiat. Such initial sets of included barcodes are called seeds. Most seeds yield smaller codes than found by the unmodified algorithm; some yield larger codes. An evolutionary algorithm is used to search for three member seeds that yield larger codes. This algorithm acts on a population of seeds in a manner analogous to biological evolution (for details, see Ashlock et al., 2002). The fitness of a seed within the evolutionary algorithm is the size of the code “grown” from the seed by the lexicode algorithm. Use of the size of a greedy closure of a partial structure is a new type of evolutionary algorithm called a greedy closure evolutionary algorithm.

The algorithm requires one additional modification to be used to produce embeddable barcodes. Barcodes located with the modified lexicode algorithm as given do not respect restriction sites for enzymes or other biological constraints. At any point where a potential barcode is checked for minimum distance from words already in the code, or when words in seeds are chosen, the words are also checked for compliance and biological constraints.

Various biological restrictions were considered in the design of the bar codes. Bar codes were not accepted that ended in T, contained the strings TT or AAA, or contained *Eco*RI (GAATTC) or *Not*I (GCGGCCGC) restriction enzyme sites. Following these rules, 21 unique bar codes were generated to label the 21 mRNA pools (Table I).

Approximately equal amounts of first-strand cDNA from each pool were combined and used as templates for DNA PolI-catalyzed second-strand synthesis. After the addition of *Eco*RI adapters, double strand-cDNAs were digested with *Not*I. Molecules between 0.5 and 2.0 kb were directionally cloned into the *Eco*RI and *Not*I sites of the pSlip7 expression vector (F. Liu and P.S. Schnable, unpublished data; GenBank accession no. AY217101). Plasmid DNA isolated from the resulting library was digested with *Not*I to remove empty vector clones. Linear DNA molecules of between 5.4 to 7 kb were gel purified and self-ligated at low concentration to promote recircularization. Ligation products were precipitated and transformed into DH10B host cells.

Sequencing Methods

Plasmid DNAs of cDNA clones from the ISUM6 library were isolated in a 96-well format using a modified alkaline lysis method adapted from one provided by the Clemson University Genomic Institute (http://www.genome.clemson.edu). Sequencing reactions were conducted using 4 μL (0.5 μg) of plasmid DNA, 2 μL of BigDye Version2 mix (Applied Biosystems, Foster City, CA), 2 μL of 5× sequencing reaction buffer (10 mM $MgCl_2$ and 0.4 M Tris, pH 9.0) and 1 μL of 3.2 μM universal primer (GTAAAACGACGGCCAGT) and the following PCR program using a PTC-225 Tetrad thermal cycler (MJ Research, Waltham, MA): 25 cycles of 96°C for 30 s, 50°C for 15 s, and 60°C for 4 min. Unincorporated dye terminators were removed from the sequencing reactions using Sephadex G-50 columns. Sequence reactions were subjected to electrophoresis on an ABI PRISM 3700 DNA analyzer at the Iowa State University DNA Sequence and Synthesis Facility.

Bioinformatic Analysis of ESTs

Base calling was performed by using Phred (Ewing et al., 1998). Overall sequence quality assessment and vector trimming were conducted using the Lucy software (v1.16s; Chou and Holmes, 2001). Lucy parameters were set to ensure an overall trimmed quality of 97.5% or better without any vector fragments in the high-quality region of each sequence. Low-quality bases between the poly(T) and the high-quality region were replaced with N to serve as spacers. EST sequences were assembled using CAP3 (Huang and Madan, 1999). The following CAP3 parameters were used: (a) a minimum 50-bp overlap; (b) greater than 95% similarity in the overlap; and (c) a clipping range of 60 bp. Sequence similarity analyses were performed using BLAST software (Altschul et al., 1990). Because 3′ ESTs contain only limited coding information, EST contigs were first compared with all public-sector maize ESTs in ZmDB (http://www.zmdb.iastate.edu/) using BLASTN to identify the associated tentative unique genes. Tentative unique genes were then compared with GenBank using BLASTX. The cutoff E value for BLASTX was less than 10^{-11}. The cutoff E value for BLASTN was less than 10^{-11}.

Bar Code Detection

The position of the vector-*Not*I/cDNA boundary was determined for each EST using the output from the Lucy software (Chou and Holmes, 2001). The 6 bp before and the 12 bp after the vector-*Not*I/cDNA boundary were extracted from each EST. This region should contain the bar code and was subjected to Smith-Waterman alignment with each of the 21 bar codes used to construct the cDNA library. This alignment used the following penalties: gap = −2, mismatch = −1, and match = +1. The local alignment between each EST and the bar code that exhibited the highest Smith-Waterman similarity was checked. If five or more bases in this alignment matched the most similar bar code, then it was assumed that this bar code uniquely defined the source pool from which that EST had been derived. If the alignment with the highest Smith-Waterman similarity contained two or more mismatches, the EST was recorded as having an un-decodable bar code. The requirement for five or more identical bases in an alignment was selected because the bar codes used in this study differed from one another by an edit distance of three, which allows only a single error to be corrected.

Distribution of Materials

Upon request, all novel materials described in this publication will be made available in a timely manner for noncommercial research purposes, subject to the requisite permission from any third-party owners of all or parts of the materials. Obtaining any permissions will be the responsibility of the requestors.

ACKNOWLEDGMENTS

We thank Diane Sickau, Regan Slonecker, Erin Archer, Luke Holst, Ka-Wai Leong, and Shu-Ting Tsao for technical assistance with EST sequencing; Olga Alechina for help isolating RNAs; Dr. Frank Hochholdinger for preparing and collecting hormone-treated seedlings; and the Iowa State University Plant Transformation Facility for providing maize callus tissue.

Received April 16, 2003; returned for revision May 19, 2003; accepted June 29, 2003.

LITERATURE CITED

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

Ashlock D, Guo L, Qiu F (2002) Greedy closure genetic algorithms. *In* Fogel DB, ed, Proceedings of the 2002 Congress on Evolutionary Computation CEC2002. IEEE Press, Piscataway, NJ, pp 1296–1301

Bailey-Serres J, Dawe RK (1996) Both 5′ and 3′ sequences of maize adh1 mRNA are required for enhanced translation under low-oxygen conditions. Plant Physiol **112:** 685–695

Bonaldo MF, Lennon G, Soares MB (1996) Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res **6:** 791–806

Chou HH, Holmes M (2001) DNA sequence quality trimming and vector removal. Bioinformatics **17:** 1093–1104

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

Gusfield D (1997) Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, UK

Huang XQ, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res **9:** 868–877

Koshiba T, Ballas N, Wong LM, Theologis A (1995) Transcriptional regulation of PS-IAA4/5 and PS-IAA6 early gene expression by indoleacetic acid and protein synthesis inhibitors in pea (*Pisum sativum*). J Mol Biol **253:** 396–413

Lal A, Lash AE, Altschul SF, Velculescu V, Zhang L, McLendon RE, Marra MA, Prange C, Morin PJ, Polyak K et al. (1999) A public database for gene expression in human cancers. Cancer Res **59:** 5403–5407

Lemke-Keyes CA, Sachs MM (1989) Genetic variation for seedling tolerance to anaerobic stress in maize germplasm. Maydica **34:** 329–337

Scheurle D, DeYoung MP, Binninger DM, Page H, Jahanzeb M, Narayanan R (2000) Cancer gene discovery using digital differential display. Cancer Res **60:** 4037–4043

Schmitt AO, Specht T, Beckmann G, Dahl E, Pilarsky CP, Hinzmann B, Rosenthal A (1999) Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. Nucleic Acids Res **27:** 4251–4260

Songstad DD, Armstrong CL, Petersen WL (1991) AgN03 increases Type II callus production from immature zygotic embryos of inbred B73 and its derivatives. Plant Cell Rep **9:** 699–702

Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA (2000) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res **28:** 10–14

Woo YM, Hu DW, Larkins BA, Jung R (2001) Genomics analysis of genes expressed in maize endosperm identifies novel seed proteins and clarifies patterns of zein gene expression. Plant Cell **13:** 2297–2317

Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW (1997) Gene expression profiles in normal and cancer cells. Science **276:** 1268–1272