Plant Physiol. (1993) 103: 359–370

Generation of Expressed Sequence Tags of Random Root cDNA Clones of *Brassica napus* by Single-Run Partial Sequencing[1]

Yu Shin Park, June Myoung Kwak, O-Yu Kwon, Yong Sung Kim, Dae Sil Lee, Moo Je Cho, Hyung Hoan Lee, and Hong Gil Nam*

Department of Life Sciences, Pohang Institute of Science and Technology, P.O. Box 125, Pohang, Kyungbuk, 790–600, South Korea (Y.S.P., J.M.K., O-Y.K., H.G.N.); Genetic Engineering Research Institute, KIST, P.O. Box 17, Taedok Science Town, Taejon, 305–606, South Korea (Y.S.K., D.S.L.); Plant Molecular Biology and Biotechnology Research Center, Kajoa-Dong 900, Chinju, Kyungnam, 660–701, South Korea (M.J.C., H.G.N.); and Department of Biology, Konkuk University, 93–1 Mojin-Dong, Seongdong-Gu, 133–701, Seoul, South Korea (H.H.L.)

Two hundred thirty-seven expressed sequence tags (ESTs) of *Brassica napus* were generated by single-run partial sequencing of 197 random root cDNA clones. A computer search of these root ESTs revealed that 21 ESTs show significant similarity to the protein-coding sequences in the existing data bases, including five stress- or defense-related genes and four clones related to the genes from other kingdoms. Northern blot analysis of the 10 data base-matched cDNA clones revealed that many of the clones are expressed most abundantly in root but less abundantly in other organs. However, two clones were highly root specific. The results show that generation of the root ESTs by partial sequencing of random cDNA clones along with the expression analysis is an efficient approach to isolate genes that are functional in plant root in a large scale. We also discuss the results of the examination of cDNA libraries and sequencing methods suitable for this approach.

Plants, in addition to being important targets for improvement by genetic engineering, provide opportunities to study distinct biological aspects, including plant-specific developmental patterns and the intricate interactions with many environmental factors such as plant-microbe interactions and unique defense processes (Goldberg, 1988). To investigate these phenomena at a detailed molecular level and to genetically engineer plants to produce better characteristics, isolation of necessary molecular clones is a prerequisite. Despite recent technological improvements and a large effort to isolate plant genes, plant gene resources are still very limited. For example, the number of plant genes currently reported in the GenBank data base is only 6195 (GenBank, release 72).

Current approaches for isolating plant genes, such as chromosome walking or gene tagging, are, in most cases, fairly expensive and time consuming and do not provide enough plant gene resources, but recent efforts in plant genome research may provide the necessary gene resources (Magnien et al., 1992). One approach to obtaining a vast amount of gene resources at a substantially lower cost compared to genome sequencing is to sequence cDNA clones. Adams et al. (1991, 1992) were the first to show that, in humans, even partial sequencing of randomly chosen cDNA clones can provide a large amount of information concerning the genetic makeup of an organism and can generate sequence-tagged markers for the genome mapping. More recently, random cDNA sequencing has been performed with cDNA clones of cultured cells of rice and maize leaf and has proven to be an efficient approach to revealing a variety of expressed genes in plants (Uchimiya et al., 1992; Keith et al., 1993). To test the feasibility of this cDNA approach to obtaining ESTs of plant genes that are functional in the plant root organ, we have performed partial sequencing of randomly chosen cDNA clones of the root organ of *Brassica napus*.

B. napus is an agronomically important crop plant for production of vegetable oils and fodder for livestock. Many of the plants in the genus *Brassica*, including Chinese cabbage, cauliflower, cabbage, and broccoli, are also economically important crop plants as sources of foods. Because there has been intensive breeding research using *Brassica* plants, a large amount of genetic resources is available (Olsson and Ellerström, 1980). In addition, the *Brassica* plants have been used as favorite model plants for plant molecular biological studies, including studies of some plant-specific processes such as self-incompatibility (Chen and Nasrallah, 1990) and pollen-specific gene expression (Albani et al., 1990, 1991). Generating a gene resource for *B. napus* will contribute significantly to genetically improving this and related crop plants by genetic engineering with the combined effort of classical breeding, as well as to utilization of the advantages

[1] Supported by grants from the Plant Molecular Biology and Biotechnology Research Center and Pohang Institute of Science and Technology.

* Corresponding author; fax 82–562–79–2199.

Abbreviations: EST, expressed sequence tag; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; PIR, Protein Identification Resource; STS, sequence-tagged site.

of these plants to solve many of the plant-specific processes at the molecular level.

Here we report the result of partial sequencing and data base comparison of cDNA clones of the root organ of *B. napus* and show that with reasonable effort this approach can provide a large number of valuable molecular clones for the genes by which the functions or the structures of the root organ are encoded in this important crop plant. We also compare the characteristics of cDNA clones and sequencing methods appropriate for this cDNA approach. We further report the expression patterns of the 10 root cDNA clones that were identified by data base search.

MATERIALS AND METHODS

Plant Material

Seeds of *Brassica napus* L. cv Naehan were originally obtained from Youngnam Agricultural Station, South Korea. Plants were grown in a compound soil mixture of vermiculite:peat moss:perlite (1:1:1) under temperature-controlled greenhouse conditions with supplementary lighting from 400-W high-pressure sodium lamps. The plants were watered with Hoagland solution (Asher and Edwards, 1983).

Total and Poly(A)$^+$ RNA Isolation

Root tissue was homogenized in the presence of 4 M guanidium isothiocyanate and total RNA was extracted according to the procedure of Cox and Goldberg (1988). Poly(A)$^+$ RNA was purified by three rounds of oligo(dT)-cellulose chromatography with the poly(A)$^+$ Quik mRNA Purification Kit (Stratagene, La Jolla, CA), according to the manufacturer's instructions.

Construction of cDNA Library

cDNA was synthesized from 2 μg of root poly(A)$^+$ RNA with a cDNA synthesis kit (Pharmacia P-L Biochemicals, Piscataway, NJ) according to the manufacturer's instructions, using oligo(dT) as a primer. The *Eco*RI/*Not*I adaptor (Pharmacia P-L Biochemicals) was then ligated to the double-stranded cDNA. A fraction of the adaptor-ligated cDNA was ligated to the *Eco*RI-digested and phosphatase-treated plasmid vector pUC19. The ligated cDNA was then introduced into *Escherichia coli* DH5α cells by electroporation, using the Gene-Pulser electroporater (Bio-Rad). A plasmid library with approximately 1.5×10^5 recombinants was obtained. A fraction of the synthesized cDNA was subjected to partial deletion with the Erase-a-Base kit (Promega, Madison, WI) under conditions designed to remove approximately 200 nucleotides from 5′ and 3′ ends of the cDNA, following the manufacturer's instructions. The cDNA was then ligated to the *Sma*I-digested and dephosphorylated plasmid vector pUC19. The recombinant plasmid cDNAs were electroporated into *E. coli* DH5α cells as described above. The insert size of the cDNA clones was examined by 0.8% agarose gel electrophoresis after digestion of cDNA clones with *Eco*RI (the normal cDNA clones) or *Eco*RI and *Xba*I (the deletion cDNA clones).

Nucleotide Sequencing

Manual nucleotide sequencing was performed by the standard dideoxy chain termination method on double-stranded plasmid DNA using the Sequenase, version 2.0, DNA-sequencing kit (United States Biochemical) and [α-^{35}S]-dATP (Amersham, Buckinghamshire, UK) as label. The reaction mixture was separated by conventional 6% PAGE. For the automated nucleotide sequencing, the sequencing reaction was performed with the Tag Dye Primer Cycle Sequencing kit (Applied Biosystems, Foster City, CA) using the fluorescent dye-labeled M13 universal or reverse primer (Applied Biosystems). The cycle sequencing reaction was carried out in an Ericomp thermal cycler (Ericomp, Inc., San Diego, CA), and the nucleotide sequences were obtained by electrophoresis on an automated DNA sequencer (model 373A; Applied Biosystems). Ambiguous base callings were manually assigned by examination of the printouts of the fluorescence pattern. Plasmid DNA for sequencing reactions was prepared by the alkaline lysis method (Sambrook et al., 1989) with minor modifications.

Computer Analysis of Nucleotide and Protein Sequences

All of the computer analyses of nucleotide or protein sequences were performed with the software package IG Suite (IntelliGenetics Co., Mountain View, CA), installed on a Sun SPARCstation 2 computer (Sun Microsystems, Inc., Mountain View, CA). The data bases used for homology search were the PIR protein sequence data base (PIR release 31, December 1991), Prosite protein motif data base (Prosite release 8.1, March 1992), and GenBank nucleic acid data base (GenBank release 71, March 1992). PIR and GenBank searches were performed with the Fast Pairwise Comparison of Sequences (FASTDB) program (Brutlag et al., 1990). Protein motif search was conducted with the Quick User-directed Expression Search Tool (QUEST) program (Boyer and Moore, 1977; Knuth et al., 1977; Abarbanel et al., 1984; Cohen et al., 1986), with a minor modification to allow for sequential comparison of the translated sequences. Comparison of the EST sequences to one another was performed with the FASTDB program.

Northern Blot Analysis

Total RNA (30 μg) was denatured in formamide-containing buffer and subjected to denaturing agarose gel electrophoresis as described by Selden (1987). After samples were blotted onto the Biotrans nylon membrane (ICN Biomedicals, Inc., Irvine, CA), the blots were prehybridized in 0.5 M $NaPO_4$ (pH 7.4), 1 mM EDTA, 1% BSA, and 7% SDS for 1 h at 65°C (Church and Gilbert, 1984) and hybridized in the same solution containing 2×10^6 cpm mL^{-1} of probes labeled with [α-^{32}P]dCTP (Amersham) by nick translation. The membranes were then washed twice in 0.1× standard saline citrate, 0.1% SDS at room temperature for 10 min and twice at 42°C for 10 min.

RESULTS

cDNA Libraries

The strategy we used to generate the root ESTs of *B. napus* was the same as that developed for the human brain ESTs (Adams et al., 1991). This strategy involves (a) partial sequencing of cDNA clones by only single-run sequencing reactions to generate ESTs, rather than by full sequencing of cDNA clones by multiple sequencing reactions, and (b) searching for possible homologies either in nucleic acid data bases or in protein data bases after translation of EST sequences into peptide sequences.

To generate the root ESTs of *B. napus* by the single-run partial sequencing of the cDNA clones, we constructed cDNA libraries from mRNA of the root organ of *B. napus* using oligo(dT) as a primer for the synthesis of the first-strand cDNA. We expected that the oligo(dT)-primed cDNA library would contain a smaller proportion of cDNA clones with organelle or repetitive sequences like ribosomal clones than a random hexamer-primed cDNA library would, because only the mRNAs with poly(A)$^+$ sequences would be represented in the cDNA library. Two kinds of cDNA libraries were constructed to obtain an insight into desirable characteristics for a large-scale generation of plant ESTs. One library (the normal cDNA library) was a standard oligo(dT)-primed cDNA library, and the average size of the cDNA inserts in this library was approximately 1 kb. The other library (the deletion cDNA library) was constructed with cDNAs that have been partially deleted in their 5′ and 3′ regions to remove some of the noncoding regions, because we wanted to examine the possibility of identifying genes by searching the protein data base as well as nucleic acid data base after generation of ESTs. Finding similarities between genes is far more sensitive when the peptide sequences are compared (Adams et al., 1991), and we hoped this deletion treatment would generate sequences in the protein-coding regions even after single-run partial sequencing reaction of only the 5′ or 3′ ends of the cDNA inserts.

For generation of a cDNA library with partial deletions in the noncoding regions, the double-stranded cDNAs synthesized by the standard oligo(dT) priming were subjected to exonuclease III treatment, followed by the S1 nuclease treatment to delete approximately 200 nucleotides from both 5′ and 3′ ends of the cDNAs. This deletion treatment generated a cDNA library with an average insert size of 650 bp.

cDNA Sequencing

For generation of the root ESTs, the clones with cDNA insert sizes longer than 300 bp were selected after agarose gel electrophoresis of restriction enzyme-digested cDNA clones. Most of the clones from the normal cDNA library contained cDNA inserts longer than 300 bp, but approximately 65% of the cDNA clones from the deletion cDNA library contained cDNA inserts longer than 300 bp. One hundred six clones from the normal cDNA library and 91 clones from the deletion library were finally subjected to the single-run partial sequencing of the inserted cDNA fragments. The clones with relatively short cDNA inserts were sequenced from only one end of the inserts, whereas the other clones were sequenced from both ends.

Table I. *Characterization of Brassica EST sequences*

The figures in parentheses indicate the percentage of ESTs in each class.

	cDNAs with Partial Deletion (Manual)	cDNAs with No Deletion		Total
		Automated	Manual	
No. of sequenced clones	91	35	74	197
No. of ESTs	101	32	104	237
Data base match	13	5	3	21
	(12.9)	(15.6)	(2.9)	(8.9)
No data base match	77	85		162
	(76.2)	(62.5)		(68.4)
Poly(A)$^+$ insert	10	42		52
	(9.9)	(30.9)		(21.9)
rRNA genes	1	1		2
	(1.0)	(0.7)		(0.8)

As shown in Table I, sequencing of these clones resulted in 136 informative sequence tags from the normal cDNA library and 101 informative sequence tags from the deletion cDNA library. The total number of nucleotide residues of these ESTs corresponds to 54.4 kb. Of the 136 sequence tags from the normal cDNA library, 104 ESTs were generated by conventional manual sequencing, and 32 ESTs were generated by an automated DNA sequencer to compare the efficiency of generating ESTs by these two sequencing approaches (see below). All of the 101 sequence tags from the deletion cDNA library were generated by manual sequencing, and the average length (± SD) of the EST sequences was 213 (±47) nucleotides by the manual sequencing and 336 (±80) nucleotides by the automated DNA sequencing.

The error rate of the sequences of these ESTs was estimated by a pilot experiment with a few known genes. The estimated error rate was approximately 2% in the manual sequencing and approximately 2.5% in the automated sequencing on the ABI 373A automated DNA sequencer. The sequences generated by manual sequencing contained 30 base miscallings, including 8 ambiguous callings, 18 deletions, and 2 insertions of 2479 bp sequences. In contrast, the sequences generated by automated sequencing contained 70 base miscallings, including 67 ambiguous callings and 4 deletions of 2951 bp sequences.

Characterization of the EST Sequences

To characterize the root ESTs generated in this experiment, the EST sequences were first compared to the sequences in the existing data bases. Comparison of EST sequences with the PIR protein data base and GenBank nucleic acid data base was performed with the FASTDB similarity search program (Brutlag et al., 1990) with the Point Acceptable Mutation 150 matrix and the unitary matrix, respectively. The FASTDB program is similar to the FASTA program (Lipman and Pearson, 1985; Gribskov et al., 1987; Pearson and Lipman, 1988) in algorithms but is more sensitive than FASTA in detecting distantly related sequences, because FASTDB

uses similarity matrices in the first step of comparison (Brutlag et al., 1990). In addition, FASTDB can detect local homology regions such as the helix-turn-helix motif.

For PIR data base search of the root EST sequences, the sequences were first translated into all six possible translational reading frames, and the translated peptide sequences were compared with the PIR data base. The similarity was considered significant when the percentage of the amino acid identity between two sequences was more than 35% or the significance value of the similarity was more than 8.0 and when further manual examination confirmed that the computer comparison was meaningful. Although the program indicates that the similarity between two sequences is significant when the significance value is greater than 4.0, we found that the criteria we used identified more definitive similarity between two sequences. The ESTs without significant similarity to the sequences in the PIR data base were further compared to the sequences in the GenBank nucleic acid data base. The similarity of two sequences was considered highly significant when the nucleotide sequence identity was more than 65%.

The characteristics of the root ESTs generated in this experiment are summarized in Table I. Of the 237 sequence tags generated, 21 sequences (8.9%) showed significant homology with the sequences in the data base, excluding two rRNA clones—one 18S clone and one 16S rRNA clone—and the ESTs containing poly(A)$^+$ tracks. The ESTs with significant similarities to the protein-coding sequences in the PIR and GenBank data bases are listed in Table II. One hundred sixty-two sequence tags (68.4%) did not show significant homology with the sequences in the data bases and thus represent previously unidentified plant genes. We also found few ESTs with significant homologies with organelle sequences or other repetitive sequences, in contrast to the case of human brain cDNA library. The lower proportion of the ribosomal clones and organelle sequences in the cDNA library we used indicates that most of the root ESTs in our experiment represent nuclear-encoded protein-coding genes.

As expected, when we constructed the cDNA libraries for EST generation, the ESTs with poly(A)$^+$ sequences were found in only 10 sequence tags (9.9%) of the 101 sequence tags generated from the deletion library (Table I). However,

Table II. *Brassica ESTs putatively identified by the PIR and GenBank data base search*

The parameters used in the search were mostly default values of the 1G Suite software package (see "Materials and Methods").

Clone	Putative Identification	Species[a]	DB[b]	No.[c]	Length[d]	Percentage ID[e]	Sig.[f]
R50	Met adenosyltransferase	A*	P	JN0131	64	92	23.02
R80	Heat-shock protein 84	M*	P	HHMS84	85	68	24.83
R98	GSH transferase III	Z	P	XUZM32	26	41	12.53
R103	pPLZ02 protein	L	P	S08665	48	35	8.06
R121	Histone H3 protein	D*	P	S09655	115	77	21.05
R167	H^+-transporting ATP synthetase β-chain	H*	P	S07041	106	82	37.88
R198	Tublin β-chain	Z*	P	S14702	46	69	20.91
R216	Cyt b_5	R*	P	CBRT5M	38	54	16.56
R10D	Calmodulin	Al*	P	S10533	49	87	5.70
R11D	Calreticulin	M*	P	S06763	56	42	9.32
DR9	Glucanase	B*	P	A25455	48	31	10.42
DR34	Pathogenesis-related protein	T	P	S00513	60	37	6.62
DR66	GSH transferase III	Z*	P	XUZM31	41	39	6.69
DR69	*Fix*R protein	Bj	P	S01065	39	45	8.46
DR91	RNA polymerase II subunit RPB10	Y	P	S13348	41	47	4.83
DR117	GAPDH (cytosolic)	S*	P	DEIS3C	54	98	24.81
DR261	Elongation factor-1α	A*	G	X16432	154	69	14.54
DR297	Met adenosyltransferase	R*	P	S06114	52	72	12.78
DR298	Actin	Sb*	P	ATSY3	51	58	17.62
DR358	Thioredoxin	C*	P	S16090	50	30	11.30
DR407	Hyp-rich protein	P*	P	B29356	17	61	8.17

[a] Species abbreviations: A, *Arabidopsis*; Al, *Alfalfa*; B, barley; Bj, *Bradyrhizobium japonicum*; C, *Chlamydomonas reinhardtii*; D, *Drosophila hydei*; H, human; L, large-leaved lupine; M, mouse; P, *Phaseolus vulgaris*; R, rat; S, *Sinapsis alba*; Sb, soybean; T, tobacco; Y, yeast; Z, maize. * These ESTs match sequences from other organisms as well as the indicated species. [b] Data base (DB) abbreviations: P, PIR; G, GenBank. [c] Number indicates accession numbers or locus names of the matched sequences. [d] Length indicates lengths of identical sequences for nucleotide matches and of identical of similar amino acid residues for peptide matches. [e] Percentage ID, Percentage identity. [f] Sig., Significance value of similarity. This value is the probability that the optimized score of a sequence alignment can be reached in the comparison of a query with the data base. The value is calculated from the formula: (optimized score − mean score)/SD. For example, a significance value of 4 means an alignment score is 4× SD above the mean of the optimized score, and a sequence alignment with the significance value more than 4 is considered to be very significant (Reference Manual, IntelliGentics Suite, Release 5.4; Brutlag et al., 1990).

the poly(A)$^+$ sequences were found in 42 sequence tags (30.9%) of the 136 sequence tags generated from the normal cDNA library. In addition, the ESTs from the deletion library showed a much higher probability of matching the protein-coding sequences in the data bases than those from the normal cDNA library (see below).

Among the 21 clones that matched the protein-coding genes in the data bases, 13 clones (12.9%) were from the deletion cDNA library and 3 clones (2.9%) were from the normal cDNA library, when the EST sequences were obtained by the conventional manual sequencing. This result indicates that the deletion treatment to remove a part of the noncoding regions in the cDNA was very effective in obtaining ESTs in the protein-coding region. In contrast, the EST sequences obtained from the automated sequencer showed a slightly higher probability (15.6%) of matching the sequences in the data bases, even when the normal cDNA clones without the deletion treatment were sequenced (see "Discussion"). In more than 80% of the ESTs with the poly(A)$^+$ sequences, the length of the poly(A)$^+$ track was very short. The length of the non-poly(A)$^+$ sequences in these poly(A)$^+$-containing ESTs was mostly between 100 and 200 nucleotides. Even though these EST sequences are mostly noncoding regions and have no data base matches, they can be used as extremely valuable ESTs for genome mapping, because the sequences in the noncoding regions are highly divergent even among the genes in a multigene family (Ko, 1990).

The ESTs with Similarities to the Sequences in the Data Base

As shown in Table II, among the ESTs with significant similarities to the PIR and GenBank data bases, 17 sequences (7.2%) matched previously reported plant genes and 4 sequences (R216, R11D, DR69, and DR91 in Table II; 1.7%) had similarities to no plant sequences but only to sequences of nonplants. Examples of peptide sequence alignment of some of these data base-matched ESTs are shown in Figure 1.

Many of the data base-matched ESTs are similar to known housekeeping genes, including histone H3 protein (R121, 77% identity), actin (DR298, 58% identity), *Arabidopsis* Met adenosyltransferase (R50, 92% identity), rat Met adenosyltransferase (DR297, 72% identity), tubulin β-chain (R198, 69% identity), Cyt b_5 (R216, 54% identity), and thioredoxin genes (DR358, 30% identity). In addition, the clone DR261 is similar to *Arabidopsis* (69% nucleotide identity), tomato (61% nucleotide identity), and yeast (57% nucleotide identity) elongation factor-1α, an abundant soluble protein (Axelos et al., 1989). The DR91 clone is similar to the yeast RNA polymerase II subunit RPB10 (47% identity). The DR117 clone is nearly identical with the cytosolic form of the *Arabidopsis* GAPDH gene (98% identity) and also shows strong homology (75% identity) with yeast GAPDH genes. GAPDH was reported to have higher basal level expression in root (Russell and Sachs, 1989), although its expression level also increases by hypoxic treatment and heat stress. The clone R167 is very similar to maize (92% identity), human (82% identity), and bovine (82% identity) mitochondrial H$^+$-transporting ATP synthase β-chain.

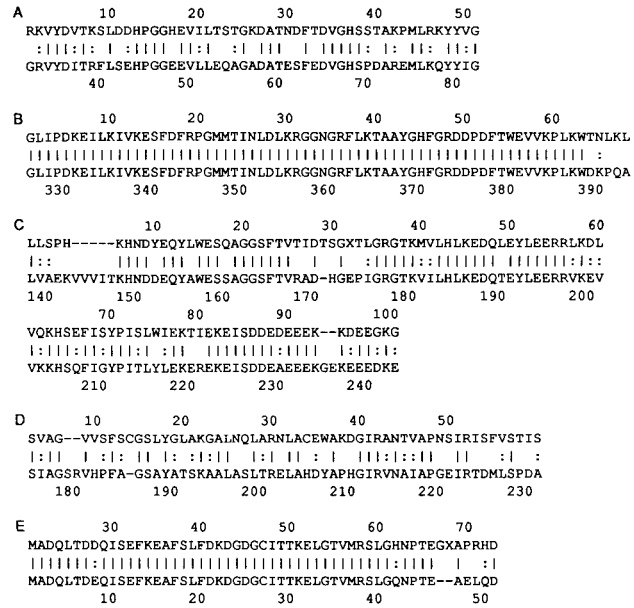

Figure 1. Protein sequence alignments of the data base-matched ESTs. The upper sequences are the *Brassica* ESTs, and the lower sequences are the sequences from the PIR data base. The numbers indicate the positions of the amino acid residues relative to the amino termini of the proteins. Sequence alignment was performed with the FASTDB program. Gaps were introduced to increase identity and similarity. Vertical lines indicate identical residues, and double dots indicate conservative substitutions. A, Alignment of R216 with rat Cyt b_5 (locus name CBRT5M); 54% identity, 75% similarity. B, Alignment of R50 with *Arabidopsis* Met adenosyltransferase (accession number JN0131); 92% identity, 94% similarity. C, Alignment of R80 with mouse heat-shock protein 84 (locus name HHMS84); 68% identity, 85% similarity. D, Alignment of DR69 with *Bradyrhizobium fix*R protein (accession number S01065); 45% identity, 70% similarity. E, Alignment of R10D with alfalfa calmodulin (accession number S10533); 87% identity, 92% similarity.

It is interesting that five defense- or stress-related genes were identified among the 21 data base-matched ESTs (Table II; see "Discussion" for details). The matched sequences include several glucanases (DR9), GSH transferase (DR66), tobacco pathogenesis-related protein genes (DR34), tomato extensin (DR407), and mammalian 90-kD heat-shock protein family (R80).

Two cDNA clones are related to Ca^{2+} ion binding. R11D was found to be similar to rabbit (42% identity), rat (36% identity), and mouse (38% identity) calreticulin, which is a high-affinity calcium-binding protein of animal skeletal muscle sarcoplasmic reticulum (Fliegel et al., 1989; Smith and Koch, 1989; Murthy et al., 1990). The R10D clone shows strong similarity to several (87% identity) calmodulin genes.

The four clones that have not been identified previously in plants but have been identified from this study include the clones R216, R11D, DR69, and DR91, which show homologies with the rat Cyt b_5 gene, the mouse calreticulin gene, the *Fix*R protein of the plant symbiotic bacterium *Bradyrhizobium japonicum*, and the yeast RNA polymerase II subunit RPB10, although the Cyt b_5 gene was isolated from cauliflower shortly after this data base search (Kearns et al., 1992) and

turned out to be highly homologous with the R216 clone. Thus, the origin of the matched sequences appears to be very broad, and this cDNA approach can find related sequences from distantly related organisms.

Expression Analysis of the Data Base-Matched ESTs

In addition to the data base search, one other approach to further characterize the root EST clones is to examine the expression patterns of the clones. We have examined the expression patterns of the 10 clones that matched the sequences in the data base by northern blot analysis. The results shown in Figure 2 reveal that many of these root cDNA clones are expressed at a higher level in root, but they are expressed in other tissues, too. However, two clones (20%), R103 and DR34, show a highly root-specific expression pattern. It is interesting that four clones (40%), DR407, DR261, DR9, and R216, show relatively higher expression levels in root and stem than in other organs. This may be because root and stem share many similar tissues, such as xylem and phloem. Others show a more or less constitutive expression pattern, except the R80 clone, which is not expressed in floral organ.

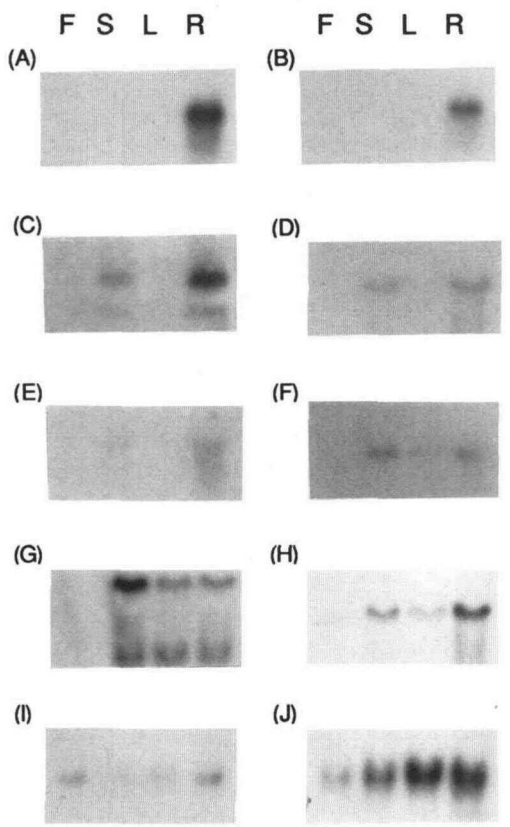

Figure 2. Northern blot analysis of the 10 data base-matched root ESTs. Total RNA (30 μg) from root (R), leaf (L), stem (S), and flower (F) were separated on 1.2% denaturing agarose gels. After blotting, each blot was probed with the EST clone R103 (A), DR34 (B), DR407 (C), DR261 (D), DR9 (E), R216 (F), R80 (G), R11D (H), DR91 (I), or DR69 (J). The sizes of the mRNAs detected by these clones are approximately 2 (A), 1.2 (B), 4.6 and 2.4 (C), 1.7 (D), 1.9 (E), 0.9 (F), 2.5 and 0.8 (G), 1.7 (H), 0.8 (I), and 1.1 kb (J).

DISCUSSION

As a part of our effort to identify genes that function in plant root, the randomly selected 197 cDNA clones of the root organ of *B. napus* were partially sequenced by single-run sequencing reaction and generated 237 ESTs. These EST sequences were characterized mainly by comparison with the sequences in the GenBank and PIR data bases or partly by northern blot analysis. Twenty-one cDNA clones (8.9%) matched the protein-coding sequences from other organisms in the GenBank and PIR with high similarity. These sequences include housekeeping genes, stress- or defense-related genes, calcium-binding genes, and others. After the nature of these ESTs is partially deduced from the sequence comparison, it is possible to further characterize these genes by full sequencing of the clones. We have fully sequenced two putatively identified ESTs through a data base search, and the clone R10D, similar to alfalfa calmodulin, was identified as a calmodulin gene, and the clone DR91, similar to yeast RNA polymerase II subunit RPB10, was found to be a likely candidate for the plant counterpart of the yeast RPB10 gene (data not shown). The results show that the partial sequencing approach, followed by data base search, is an efficient way to isolate functionally or structurally related genes to known sequences from plants, which agrees with previous reports concerning the partial sequencing of random cDNA clones of rice and maize (Uchimiya et al., 1992; Keith et al., 1993).

Is the cDNA approach we describe here really useful in identifying new plant genes, or are the clones identifiable by this approach merely the *Brassica* equivalents of other plant genes that can be isolated by the simple nucleic acid hybridization technique using known plant genes as probes? Although it is true that many of the genes that have been identified by our random sequencing approach are homologous with the genes from other plants (and it is likely that some of these clones can be isolated by a simple cross-hybridization technique), some of the genes cannot be isolated by a simple hybridization technique, especially the genes that have been identified by homologies with the genes of other kingdoms.

The clone R50 shows 92% amino acid identity and 78% nucleotide sequence identity with the Met adenosyltransferase gene of *Arabidopsis thaliana*. The clone DR117 shows 98% amino acid identity and 95% nucleotide sequence identity with the *Sinapsis alba* GAPDH gene. To isolate these clones from *B. napus*, one can simply use the regular or the low-stringency hybridization technique without much difficulty, but this is not always the case. For example, the DR91 showed 48% amino acid identity with the yeast RNA polymerase II subunit RPB10. Full sequencing of this clone revealed that it is likely to encode a plant RNA polymerase subunit as mentioned above; yet cloning of this clone by simple hybridization technique would not have been possible because of a low degree of nucleic acid homology.

Another example is the clone R216. The PIR search showed that the clone has 54% amino acid identity with the rat Cyt b_5 gene, and the GenBank search showed that the clone has 54% nucleotide sequence identity with the chicken Cyt b_5 gene. It is not possible to isolate a plant Cyt b_5 gene by

probing with the chicken gene. The first plant Cyt b_5 gene was isolated from cauliflower soon after our data base search (Kearns et al., 1992), and the R216 clone showed 62% amino acid identity with the cauliflower Cyt b_5 sequences, suggesting that the R216 clone is likely to encode a Cyt b_5 in *B. napus*. The cauliflower clone was isolated by purifying the Cyt b_5 protein from floret microsomes, obtaining partial peptide sequences of the protein, and then performing polymerase chain reaction of the cauliflower cDNA with the oligonucleotides that correspond to the peptide sequence. It would have been much easier if the clone R216 had been available earlier. Thus, it appears that this partial cDNA sequencing is a useful approach to identifying genes that have not been found previously in plants but that have a low degree of protein homology with known plant or nonplant genes, even in the cases in which the simple DNA hybridization assay may not be successful.

The nature of the cDNA library seems to be important in generating ESTs, depending on the purpose of EST generation. It is obvious that a cDNA library that is devoid of noncoding regions and allows sequencing of the coding regions is important for putative identification of the EST sequences by data base search, as demonstrated in our experiment of comparing the level of data base matching between the EST sequences from a normal cDNA library and a cDNA library with some deletions in the 5′ or 3′ untranslated sequences (Table I). This would be particularly true when the manual sequencing that generates relatively short ESTs is to be used to generate ESTs mainly for putative identification of genes by data base search. This type of cDNA library may be obtained by partial deletion of the noncoding regions of cDNA molecules as performed in our experiments or by synthesizing the cDNA molecules with random-hexamer priming, although the latter library would have a higher chance of containing sequences other than cDNA of nuclear mRNA. Our preliminary data with the random-hexamer-primed root cDNA library shows that there are virtually no ESTs with the poly(A)$^+$ sequences (data not shown).

One other type of cDNA library that can be used for EST generation for putative identification of clones by data base search is a directionally cloned cDNA library, because the 5′ untranslated regions of many plant genes are relatively short (Joshi, 1987). Keith et al. (1992) have obtained a relatively high level of identification (20%) by using a directionally cloned cDNA library of maize leaf tissues. Although a direct comparison of the level of data base match between the maize ESTs from the previous study and the *Brassica* ESTs from our study may not be appropriate, because the organisms and the organs used in the two studies are different, the results suggest that a directionally cloned cDNA library is a potential alternative to a deleted cDNA library.

However, even though the maize cDNA clones were directionally cloned, only 20% of the clones were reported to be full-length clones or to include the putative start codon. The result suggests that most (80%) of the cDNA clones used in the maize cDNA sequencing are, in fact, devoid of the 5′ untranslated regions and thus the maize ESTs are likely to contain a higher proportion of protein coding sequences than one would expect from sequencing full-length cDNA clones. However, in most cases, it will be almost impossible to make a cDNA library with proper deletions in the 5′ untranslated sequences by a simple directional cDNA construction method. In addition, it is also true that some plant genes have rather long 5′ untranslated sequences. Thus, the advantage of sequencing the deleted cDNA library over sequencing of the directionally cloned cDNAs is that, by constructing a cDNA library with intended and proper deletions, one can generate sequences with a higher proportion of coding regions upon sequencing, and, thus, the clones will have higher probabilities of data base matching. This deletion treatment would be particularly advantageous for identifying genes with relatively long 5′ untranslated sequences by data base search.

This argument is partly supported by our result. When we assume that half of the EST sequences obtained from the normal library are the sequences from 5′ ends and that all of the three matched sequences from the normal library are from these 5′ sequences, the level of identification is still approximately 6%, which is less than the identification level of the deleted library (13%). In addition, the identification of the EST sequences in the data base can be done with sequences from either N-terminal or C-terminal sequences or perhaps from internal sequences in a deleted library, depending on the extent of the deletion treatment.

The location of related protein sequence motifs between two genes may not always be the amino-terminal region, especially when the distantly related genes are compared. The evolutionarily distant genes may have regions with different degrees of conservation. Thus, when more sequence regions than just the amino-terminal region of a gene can be compared, there are greater chances of identifying related genes or protein motifs in the data base. However, this advantage may not be so great when applied to the comparison of closely related sequences. One disadvantage of sequencing the deleted cDNA clones is that, to further isolate the corresponding full-length clones, one will have to screen another full-length library after identifying a gene by data base matching. However, this is a minor effort when compared to that of initially identifying a gene. Another disadvantage of the deleted cDNA library becomes clear when the purpose of generating the EST is mostly to obtain STSs for genome mapping. The deleted cDNA clones will not produce gene-specific EST sequences when a gene is a member of a multigene family. In this case, sequencing of either the 5′ or 3′ untranslated regions of directionally cloned cDNAs will be much more useful than the deleted cDNA clones, because the untranslated regions are more divergent than the protein-coding regions, as mentioned in "Results."

To examine whether different sequencing methods can affect the level of data base matching of the EST sequences, we performed automated sequencing as well as manual sequencing. The automated sequencing usually produces longer sequences from a single-run sequencing reaction, as was shown in our results, too. In addition, the data base-matching efficiency of the EST sequences obtained from the automated sequencing was much higher (15.6%) than that of the EST sequences obtained from the manual sequencing when the nondeleted cDNA library was examined. A simple explanation for this discrepancy is that the longer sequences obtained by automated sequencing contained more of the protein-

coding sequences, reading the sequences beyond the 5′ or 3′ noncoding regions. Thus, these sequences would have higher chances of data base matching, even when they are obtained from the nondeleted library. In addition, even if the sequences mostly contained the protein-coding regions, longer sequences would have greater chances of data base matching, because two related sequences often consist of domains with different degrees of homologies.

This hypothesis could be easily examined by artificially truncating the longer sequences obtained by automated sequencing to the average length of the sequences obtained by manual sequencing and by comparing the level of data base matching between the longer sequences and the shorter, truncated sequences. Unfortunately, the data we obtained did not provide evidence for our hypothesis. The data base-matched sequences from the automated sequencing were the clones R50, R80, R103, R121, and R167 that match the Met adenosyltransferase, heat-shock protein 84, pPLZ02 protein, histone H3, and H^+-transporting ATP synthetase β-chain, respectively. All of these matched clones mostly contained the protein-coding sequences, two (R167, R50) matching at the carboxy terminus, two (R121, R80) matching at the amino terminus, and one (R103) matching at an internal region, when judged from the alignments of the matched sequences. In addition, the homologies of these sequences were either very high or rather even throughout the sequenced regions. When these sequences were truncated and compared to the sequences in the data base, all of the sequences were still identifiable in the PIR data base. Thus, the high level of the data base matching we observed with the sequences from the automated sequencing appears to be a coincidence, although the differences in the types of sequence errors between the automated sequencing and the manual sequencing might have partially contributed to this difference (see "Results" and discussions below).

However, we still believe that it is valid that the longer sequences obtained by automated sequencing will have a higher level of data base matching, at least in some cases, and performed a simulation test with a few known genes to examine the validity of this hypothesis. One example is the citrate synthase gene of *A. thaliana*. The reported sequence of this clone contains 57 bp of 5′ untranslated sequences. The sequence of this clone was artificially truncated to the length of 213 bp reading (the average reading of the manual sequencing) or to the length of 336 bp reading (the average length of the automated sequencing) either from the 5′ or 3′ end and was compared to the sequences in the PIR data base. Although the shorter sequences from both ends did not find any significantly related sequences from the data base, the longer sequences from the 5′ ends matched the pig citrate synthase gene with 48% homology, and the longer sequences from the 3′ ends also matched the pig citrate synthetase with 32% homology with a local region of rather high homology. Another example is found with the Cyt b_5 gene of cauliflower. When the same procedure was applied to this gene, the longer sequences from the 5′ ends showed sequence homology to the rat Cyt b_5, but the shorter sequences did not show any significant homology with the genes in the data base. Neither the longer nor the shorter sequence from the 3′ ends matches any sequences.

The types of sequence errors found in the EST sequences can also affect the level of data base matching. As mentioned in "Results," most of sequence error types were base miscallings or ambiguous base calls. This type of error did not affect the level of data base matching, because a miscalled base is simply regarded as a change of a single amino acid residue or as a same amino acid residue, depending on the location of the miscalled base and on the frame used to search the data base. However, the deletions or insertions may cause a slight effect on data base matching, because these errors can cause frame shifts in the translated amino acid sequences.

The effect of frame shift errors in data base matching was tested by artificially introducing deletions into some of the EST sequences we had generated. A deletion that is made at the nucleotide position 121 of the clone R50 and DR117 and that shows strong homology with the *Arabidopsis* Met adenosyltransferase (92%) and GAPDH (98%) genes did not affect the data base matching at all. To further examine the effect of the frame shift errors in identifying sequences with weaker homologies, two deletions were made at nucleotide positions 31 and 121 of the DR66 that shows 39% sequence homology with the GSH transferase III gene of maize. In this case, the homology decreased to 28%, but the clone could be still identified as a related clone to the GSH transferase III gene because of the presence of relatively stronger local homology regions. However, when a deletion was made at nucleotide position 121 of DR358 clone that shows 30% homology with the thioredoxin gene of *Chlamydomonas*, the homology declined to 24%, and the clone could not be identified as a related clone. Thus, although the majority of sequence errors that are observed during the single-run partial sequencing of cDNA clones do not affect the level of data base matching, the frame shift errors can affect the data base matching to a certain extent.

The extent of repeated sequencing of the same cDNA clones was very low in our experiment, and only 3 of the 197 clones were sequenced twice. The experiments conducted with the human brain cDNA library, the maize leaf cDNA library, and the rice cultured cell cDNA library also showed that the majority of the ESTs generated without normalization process were represented only once (Adams et al., 1992; Uchimiya et al., 1992; Keith et al., 1993).

However, the question arises about when the repetitive sequencing would be a problem, how feasible it is to pursue further this type of effort, and in which direction this type of approach should head. Estimation by the mRNA/cDNA hybridization kinetic study (Okamura and Goldberg, 1989) of the soybean root mRNAs showed that the approximate percentage of mRNA mass and the number of diverse mRNAs in the superabundant class is 19% and 50, respectively, which is similar in pea and, thus, should apply to *B. napus* without much deviation. The figures mean that the average percentage of a single mRNA species in the superabundant class is 0.38% (19%/50), assuming the average lengths of the root mRNA species are equal, regardless of their abundance. Thus, when we sequence the randomly chosen cDNAs to generate the root ESTs, we would expect the sequencing of an mRNA sequence in the superabundant class 3.8 times out of 1000 clones on average. The approximate percentage of mRNA mass and the number of diverse mRNAs in the abundant

class is 36% and 2000, respectively. Using the same calculation, we would expect sequencing of mRNA sequences in the abundant class 0.18 times out of 1000 clones. The approximate percentage of mRNA mass and the number of diverse mRNAs in the rare class is 45% and 18,000, respectively. In this case, we would expect sequencing of mRNA sequences in this class 0.025 times out of 1000 clones. Thus, we can estimate that 1000 randomly chosen clones would consist of 190 superabundant mRNA sequences (19%) with an average repetitiveness of 3.8, 360 abundant mRNA sequences (36%) with an average repetitiveness of 0.18, and 450 rare mRNA sequences (45%) with an average repetitiveness of 0.02. This estimation shows that the repeated sequencing of the superabundant or abundant mRNAs would not be a problem until a fair number of cDNA clones is sequenced. Perhaps up to sequencing of approximately 5500 clones ($1/0.18 \times 1000$ clones for the abundant class mRNAs), at which point the repetition of the abundant class sequences start to build up, most of sequences except the sequences from the superabundant class ($19\% \times 5500$ clones $= 1045$ clones) may be unique.

However, to reduce the effort of repetitive sequencing and to have access to the rare mRNA sequences, it will be necessary to make a normalized cDNA library in which all of the approximately 20,000 mRNA species expressed in plant root (Okamura and Goldberg, 1989) are present more or less in an equal proportion, for example, by a kinetic approach (Patanjali et al., 1991). Nonetheless, it should be noted that random sequencing of the cDNA clones without a normalization process will provide an approximate but important idea about the types of genes that are expressed in plant root as the superabundant or abundant class.

In the case of the plant root mRNA, the screening procedure to remove the clones in the superabundant class seems to be impractical, because the number of the superabundant class mRNA species are too many to perform hybridization with every clone in this class. Perhaps the best strategy to generate the ESTs for identification of genes that function in plant root, then, might be that the random sequencing approach as has been described here using the deleted library with no normalization process is pursued until the number of the ESTs reaches approximately 1000 to 2000, and at the same time the sequencing of a normalized cDNA library is pursued. One other strategy that has to be considered is to use a subtracted cDNA library that would contain the sequences expressed in root tissues only. The ESTs from this type of library will provide important information about the genes that give the plant root its unique developmental patterns and functions.

Of the 197 root EST clones examined in this experiment, 175 EST clones did not show significant similarity to the sequences in the GenBank or PIR data base and, thus, may represent new plant genes. Given that the identities of the large segment of the EST sequences are unknown, how are we going to move forward in identifying these unidentifiable ESTs by data base search? The recent trend of rapid increase in the plant gene data base will enhance the level of identifying the ESTs that have related structure or related peptide motifs (see below). In addition, among the 21 ESTs identified in our experiment, four clones were identified by homology to nonplant genes only (see "Results") and 11 clones also showed significant homologies with the nonplant genes as well as to the plant genes.

If we consider the rapid increase of the numbers of genes isolated from other organisms, including human, mouse, *Drosophila, Caenorhabditis elegans, Saccharomyces cerevisiae,* and *E. coli* (Maddox, 1991), it is very likely that the identifiable genes by data base search will increase substantially. In addition, development of a better computer algorithm for identifying distantly related genes that show weak amino acid identity but contain structurally or evolutionary related sequences will be necessary to increase the level of identification of the EST sequences. The nature of the ESTs that are not identifiable or are identifiable with only very weak homologies by the PIR and GenBank data base search may be recognized by combining several methods, including full sequencing of the clones with rather weak homology, examination of expression patterns, and expression of the sense or antisense mRNA of these clones in plants. Especially the antisense approach combined with the expression analysis of the EST clone (see below) may be one of the most powerful approaches to investigate the function of the ESTs unidentified by the data base search, because there are numerous reports of reducing expression of a specific clone by expressing the corresponding antisense mRNA (Ecker and Davis, 1986; Smith et al., 1988; Visser et al., 1991).

Another approach could be the gene disruption experiment by the homologous recombination technique, if this technique becomes more efficient in the future. One other potential approach might be the functional identification of ESTs by complementation in yeast or in *E. coli,* because there are many mutants disrupted in several aspects of cellular processes in these organisms, and there are numerous reports of functional complementation of plant genes in these organisms (for an example, see Nitschke et al., 1992). Although a substantial amount of effort would be involved in identifying the function of the unidentified ESTs, these unidentified novel plant genes could provide a powerful gene resource for molecular genetic dissection of many plant cellular processes.

In addition to searching for the similarity of the root ESTs to the sequences in the PIR or GenBank data base, the EST sequences can be further characterized by comparing to the Prosite protein motif data base to examine the presence of functionally or structurally identifiable peptide motif sequences in EST sequences after translation of the EST sequences into all six possible reading frames. The results of the Prosite search that was conducted with the QUEST program (Boyer and Moore, 1977; Knuth et al., 1977; Abarbanel et al., 1984; Cohen et al., 1986) are shown in Table III. Presently, it is not clear how much valuable insight this motif identification can provide to the investigation of the nature of the ESTs, because most of the motif identification process seems to have a stringency that is too low to specifically identify a sequence with a given function. However, some EST sequences identified in the motif data base, such as R29 and R10D with calcium-binding domain motif or R121 with histone H3 motif, were also identified as data base matching ESTs in the PIR data base, which suggests that in some cases the motif identification can provide information concerning the nature of ESTs. The sequence motifs currently collected

Table III. *Peptide motif matches of Brassica ESTs by Prosite motif data base search*

Motif	Motif-Matched ESTs
ATP/GTP-binding site motif A (P-loop)	R37, DR47, DR360
Cyt c family heme-binding site	R87
EF-hand calcium-binding site	R29, R10D
Farnesyl group-binding site	R53, R108, R128, R173, R206, R184, DR357
Histone H3	R121
Homeobox	R35, R150
Leu zipper	R27, R39, R45, R62, R79, R80, R252, DR293
Mitochondrial energy transfer proteins	R76
Nuclear-targeting sequence	R130, R156, R229, DR299
RNA-binding region RNP-1	R74
Sugar transport protein	DR69
Zinc carboxydase, zinc-binding region 2	R92

in the motif data base are not comprehensive at all, and a more comprehensive collection of structurally, functionally, or evolutionarily related motifs in the data base will enhance the level of identification of the ESTs in the future.

Regardless of whether the EST sequences are identifiable in the data base, they still can be utilized as STSs, a valuable resource for genetic mapping. STSs are genetic markers with nucleotide sequences of 200 to 500 bp, which are unique in the genome of an organism (Olson et al., 1989). STSs can be assayed by the polymerase chain reaction using a pair of oligonucleotide sequences corresponding to two short regions in a STS and can be transferred simply as information but not as biological materials like conventional DNA markers. Although the STSs are not widely used in the genome mapping of plant species yet, they have been used successfully for the physical mapping of human chromosomes (Vollrath et al., 1992). STS mapping of these EST markers in the genome of *B. napus* also may provide genomic information such as distribution of expressed genes in the genome and the relationship of the expression patterns or functions of expressed genes with the location of the genes in the genome.

The results of the northern blot analysis of the data base-matched ESTs show that, in addition to the sequence cataloging of ESTs, expression cataloging of ESTs can provide valuable information concerning the characteristics of the EST clones that are expressed to provide plant organs with specific functions and structures (see "Results"). Although we have performed the expression analysis of the EST clones in several plant organs, the expression cataloging can be extended to determine the regulation patterns of these clones following hormonal treatments or environmental stimuli, such as wound and drought. Perhaps even more valuable expression cataloging can be achieved by in situ hybridization screening of the EST clones. It is also noted that a high proportion (20%) of the randomly chosen root cDNA clones, without prior subtraction of commonly expressed genes in other tissues, was root specific. To obtain root-specific genes, it is sufficient, to a certain extent, to simply examine expression patterns of random EST clones without prior subtraction process of the cDNA library.

The sequencing and the data base search of the random cDNA clones as described in this experiment should be able to provide an idea about what types of genes are functioning in plant organs. One of the immediate interpretations of our results is that many defense- or stress-related genes are expressed in plant root. Plant root, consisting of the underground part of a plant, plays many important roles in the survival of a plant (Schiefelbein and Benfey, 1991): it takes up and transports water and dissolved ions, it is the place of synthesis and storage for plant growth regulators such as cytokinins and GAs, and, in addition, it anchors the whole plant in the soil and shows complex interactions with various environmental factors such as many soil microorganisms, gravity, moisture, nutrients, temperature, and composition of soil.

It is interesting that many (5 of the 21 data base-matched sequences) of the data base-matched sequences from the root ESTs are defense- or stress-related genes (Table II). The clone DR9 matched to several glucanases (31% identity), which are induced by pathogen infection and were reported to have antifungal activity (Boller, 1987). The clone DR66 matched GSH transferase genes (39% identity). In plants, the GSH transferases detoxify a number of herbicides (Mozer et al., 1983). The DR34 clone is similar to the tobacco pathogenesis-related genes (37% identity), which are known to be induced in response to biotic or abiotic elicitors and contain endohydrolytic or peroxidase activity (Van Loon, 1985). The DR407 clone matched to the tomato extensin (61% identity) and several Hyp-rich glycoproteins. The extensin-like Hyp-rich glycoproteins are accumulated in response to pathogens and wounds, potentially influencing wound healing and disease resistance (Showalter and Rumeau, 1990), in addition to being important cell wall components. The clone R80 shows similarity to the mouse (68% identity), human (56% identity), rabbit (55% identity), and *Drosophila* (52% identity) 90-kD heat-shock protein family. This result indicates that the plant root system may express many defense- or stress-related genes to cope with the intensive environmental interactions mentioned above. Although in our experiment we have used a cDNA library from a whole root organ that is composed of several types of tissues and cells, a better EST cataloging procedure to identify genes that function in a specific tissue or a cell type can be achieved from a cDNA library constructed from a tissue or a cell type. In some plants, it is now

quite practical to isolate a certain single cell type such as guard cells of *Vicia faba*.

Plant root also serves as an interesting subject for the study of plant organ development and cellular differentiation, because relatively simple types at several stages of differentiation are present at the same time in a single root organ. Expression cataloging of the root ESTs as performed in our experiment along with the putative identification of the ESTs by data base search may well provide the molecular clones that are needed to dissect the molecular processes of plant root development.

ACKNOWLEDGMENTS

The first two authors have made equal contributions to this work. We thank Jin Yong Jeong for his technical help with the automated sequencing.

Received February 3, 1993; accepted June 1, 1993.

LITERATURE CITED

Abarbanel RM, Wieneke PR, Mansfield E, Jaffe DA, Brutlag DC (1984) Rapid searches for complex patterns in biological molecules. Nucleic Acids Res **12:** 263–280

Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelly JM, Utterback TR, Nagle JW, Fields C, Venter JC (1992) Sequence identification of 2,375 human brain genes. Nature **355:** 632–634

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

Albani D, Altosaar I, Arnison PG, Fabijanski SF (1991) A gene showing sequence similarity to pectin esterase is specifically expressed in developing pollen of *Brassica napus*. Sequences in its 5′ flanking region are conserved in other pollen-specific promoters. Plant Mol Biol **16:** 501–513

Albani D, Robert LS, Donaldson PA, Altosaar I, Arnison PG, Fabijanski SF (1990) Characterization of a pollen-specific gene family from *Brassica napus* which is activated during early microspore development. Plant Mol Biol **15:** 605–622

Asher CJ, Edwards DG (1983) Modern solution culture techniques. *In* A Läuchli, RL Bieleski, eds, Inorganic Plant Nutrition, Encyclopedia of Plant Physiology, New Series, Vol 15A. Springer-Verlag, Berlin, pp 94–119

Axelos M, Bardet C, Liboz T, Thai ALV, Curie C, Lescure B (1989) The gene family encoding the *Arabidopsis thaliana* translation elongation factor EF-1α: molecular cloning, characterization and expression. Mol Gen Genet **219:** 106–112

Boller T (1987) Molecular and genetic perspectives. *In* T Kosuge, EW Nester, eds, Plant–Microbe Interactions, Vol 2. Macmillan Press, New York, pp 385–413

Boyer RS, Moore JS (1977) A fast string searching algorithm. Comm ACM **20:** 762–772

Brutlag DL, Dautricourt J-P, Maulik S, Relph J (1990) Improved sensitivity of biological sequence database searches. Comput Appl Biosci **6:** 237–245

Chen CH, Nasrallah JB (1990) A new class of S-sequences defined by a pollen recessive self-incompatibility allele of *Brassica oleracea*. Mol Gen Genet **221:** 241–248

Church GM, Gilbert W (1984) Genomic sequencing. Proc Natl Acad Sci USA **81:** 1191–1195

Cohen FE, Abarbanel RM, Kuntz ID, Pletterick R (1986) Turn prediction in proteins using a pattern matching approach. Biochemistry **25:** 266–275

Cox KH, Goldberg RB (1988) Analysis of plant gene expression. *In* CH Shaw, ed, Plant Molecular Biology, A Practical Approach, IRL Press, Oxford, England, pp 1–4

Ecker JR, Davis RW (1986) Inhibition of gene expression in plant cells by expression of antisense RNA. Proc Natl Acad Sci USA **83:** 1359–1370

Fliegel L, Burns K, MacLennan DH, Reithmeier RAF, Michalak M (1989) Molecular cloning of the high affinity calcium-binding protein (Calreticulin) of skeletal muscle sarcoplasmic reticulum. J Biol Chem **264:** 21522–21528

Goldberg RB (1988) Plants: novel developmental processes. Science **240:** 1460–1467

Gribskov M, McLachlan AD, Eisenberg D (1987) Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA **84:** 4355–4358

Joshi CP (1987) An inspection of the domain between putative TATA box and translation start site in 79 plants. Nucleic Acids Res **15:** 6643–6653

Kearns EV, Keck P, Somerville CR (1992) Primary structure of cytochrome b_5 from cauliflower (*Brassica oleracea* L.) deduced from peptide and cDNA sequences. Plant Physiol **99:** 1254–1257

Keith CS, Hoang DO, Barrett BM, Feigelman B, Nelson MC, Thai H, Baysdorfer C (1993) Partial sequencing analysis of 130 randomly selected maize cDNA clones. Plant Physiol **101:** 329–332

Knuth DE, Morris JH, Pratt VR (1977) Fast pattern matching in strings. SIAM J Comput **6:** 323–350

Ko MSH (1990) An 'equalized cDNA library' by the reassociation of short double-stranded cDNAs. Nucleic Acids Res **18:** 5705–5711

Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science **227:** 1435–1441

Maddox JC (1991) The case for the human genome. Nature **352:** 11–14

Magnien E, Bevan M, Planque K (1992) A European 'BRIDGE' project to tackle a model plant genome. Trends Biotechnol **10:** 12–15

Mozer TJ, Tiemeier DC, Jaworski EG (1983) Purification and characterization of corn glutathione S-transferase. Biochemistry **22:** 1068–1072

Murthy KK, Banville D, Srikant CB, Carrier F, Holmes C, Bell A, Patel YC (1990) Structural homology between the rat calreticulin gene product and the *Onchocerca volvulus* antigen Ral-1. Nucleic Acids Res **18:** 4933

Nitschke K, Fleig U, Schell J, Palme K (1992) Complementation of the cs *dis*2–11 cell cycle mutant of *Schizosaccharomyces pombe* by a protein phosphatase from *Arabidopsis thaliana*. EMBO J **11:** 1327–1333

Okamura JK, Goldberg RB (1989) Regulation of plant gene expression: general principles. *In* PK Stumpf, EE Conn, eds, The Biochemistry of Plants, A Comprehensive Treatise. Academic Press, New York, pp 1–82

Olson M, Hood L, Cantor C, Botstein D (1989) A common language for physical mapping of the human genome. Science **245:** 1434–1435

Olsson G, Ellerström S (1980) Polyploid breeding in Europe. *In* S Tsunoda, K Hinata, C Gómez-Campo, eds, *Brassica* Crops and Wild Allies. Japan Scientific Societies Press, Tokyo, pp 223–234

Patanjali SR, Parimoo S, Weissman SM (1991) Construction of a uniform-abundance (normalized) cDNA library. Proc Natl Acad Sci USA **88:** 1943–1947

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad USA **85:** 2444–2448

Russell DA, Sachs MM (1989) Differential expression and sequence analysis of the maize glyceraldehyde-3-phosphate dehydrogenase gene family. Plant Cell **1:** 793–803

Sambrook J, Fritch EF, Maniatis T (1989) Molecular Cloning: A Laboratory Manual, Ed 2. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Schiefelbein JW, Benfey PN (1991) The development of plant roots: new approaches to underground problems. Plant Cell **3:** 1147–1154

Selden RF (1987) Analysis of RNA by northern hybridization. *In* FM Ausubel, R Brent, RE Kingston, DD Moore, JG Seidman, JA Smith, K Struhl, eds, Current Protocols in Molecular Biology, Vol 1. John Wiley and Sons, New York, pp 4.9.1–4.9.8

Showalter AM, Rumeau D (1990) Molecular biology of the plant cell wall hydroxyproline-rich glycoproteins. *In* WS Adair, RP Me-

cham, eds, Organization and Assembly of Plant and Animal Extracellular Matrix. Academic Press, New York, pp 247–281

Smith CJS, Watson CF, Ray J, Bird CR, Morris PC, Schuch W, Grierson D (1988) Antisense RNA inhibition of polygalacturonase gene expression in transgenic tomatoes. Nature **334:** 724–726

Smith MJ, Koch GLE (1989) Multiple zones in the sequence of calreticulin (CRP55, calregulin, HACBP), a major calcium binding ER/SR protein. EMBO J **8:** 3581–3586

Uchimiya H, Kidou S, Shimazaki T, Aotsuka S, Takamatsu S, Nishi R, Hashimoto H, Matsubayashi Y, Kdou N, Umeda M, Kato A (1992) Random sequencing of cDNA libraries reveals a variety of expressed genes in cultured cells of rice (*Oryza sativa* L.). Plant J **2:** 1005–1009

Van Loon LC (1985) Pathogenesis-related proteins. Plant Mol Biol **4:** 111–116

Visser RGF, Somhorst I, Kuipers GJ, Ruys NJ, Feenstra WJ, Jacobsen E (1991) Inhibition of the expression of the gene for granule-bound starch synthase in potato by antisense constructs. Mol Gen Genet **225:** 289–296

Vollrath D, Foote S, Hilton A, Brown LG, Beer-Romero P, Bogan JS, Page DC (1992) The human Y chromosome: a 43-interval map based on naturally occurring deletions. Science **258:** 52–59