Plant Physiol. (1995) 108: 1141–1150

Expressed Sequence Tags from Developing Castor Seeds[1]

Frank J. van de Loo[2], Simon Turner, and Chris Somerville*

Carnegie Institution of Washington, Department of Plant Biology, 290 Panama Street, Stanford, California 94305

To expand the availability of genes encoding enzymes and structural proteins associated with storage lipid synthesis and deposition, partial nucleotide sequences, or expressed sequence tags (ESTs), were obtained for 743 cDNA clones derived from developing seeds of castor (*Ricinus communis* L.). Enrichment for seed-specific cDNA clones was obtained by selecting clones that did not detectably hybridize to first-strand cDNA from leaf mRNA. Similarly, clones that hybridized to storage proteins or other highly abundant mRNA species from developing seeds were selected against. To enrich for endomembrane-associated proteins, some clones were selected for sequencing by immunological screening with antibodies prepared against partially purified endoplasmic reticulum membranes. Comparison of the deduced amino acid sequences of the ESTs with the public data bases resulted in the assignment of putative identities of 49% of the clones selected by differential hybridization and 71% of the clones selected by immunological screening. Open reading frames in 100 of the ESTs exhibited higher homology to 78 different nonplant gene products than to any previously known plant gene product.

Many of the enzymes involved in lipid synthesis in animals, fungi, and bacteria are integral membrane proteins. Similarly, the corresponding plant enzymes are generally thought to be integral membrane proteins, although many have proven to be difficult or impossible to purify by conventional biochemical criteria (Browse and Somerville, 1991). Because of these and related difficulties, there is broad interest in obtaining genes for most or all of the enzymes associated with plant lipid synthesis, deposition, and metabolism. Because many plant gene products exhibit significant amino acid sequence homology to the corresponding gene products from microorganisms or animals, the proliferation of amino acid sequence information for many enzymes from nonplant sources has provided novel technical approaches to the isolation of the corresponding plant genes. In particular, several groups have produced large numbers of partial cDNA sequences, or ESTs, and have shown that these EST collections can be used to identify putative clones for a wide range of gene products (Adams et al., 1991, 1992). ESTs have been reported in the published literature for 3089 rice cDNAs (Uchimiya et al., 1992; Sasaki et al., 1994), 200 maize cDNAs (Keith et al., 1993), 197 *Brassica napus* cDNAs (Park et al., 1993), and 2652 Arabidopsis cDNAs (Höfte et al., 1993; Newman et al., 1994), and many thousands of additional rice and Arabidopsis ESTs are available in the public data bases (for a recent review of methods to access these sequences, see Newman et al., 1994).

We have examined the utility of the EST approach for the isolation of cDNA clones for enzymes associated with storage lipid synthesis in developing seeds of castor (*Ricinus communis* L.). Castor seeds have a very high lipid content, most of which is triacylglycerol-containing ricinoleic acid (D-12-hydroxyoctadec-*cis*-9-enoic acid). By contrast, leaves contain very low levels of triacylglycerol and do not contain significant amounts of ricinoleate (van de Loo, 1993). We have previously found that a castor gene encoding a fatty acid desaturase was much more highly expressed in developing seeds than in leaves (Shanklin and Somerville, 1991). Therefore, we enriched for cDNA clones associated with storage lipid synthesis by selecting for cDNA clones that were not highly expressed in leaves. In addition, because storage lipid synthesis takes place in the ER, we attempted to enrich for clones encoding ER proteins by immunologically screening an expression cDNA library with antibodies raised against total microsomal proteins or purified ER. These clones were subsequently used to obtain partial nucleotide sequences, which were translated in all six reading frames, and the deduced sequences were compared with the amino acid sequences in public data bases. Here, we present the results of this analysis. Although our goal was to enrich for a particular subset of genes, the results are of general utility.

MATERIALS AND METHODS

Purification of RNA

Total RNA for differential screening was purified from developing stage III through V (Greenwood and Bewley, 1982) castor (*Ricinus communis* L. cv Baker 296) cellular endosperm plus embryo, or from castor leaves, as described by Puissant and Houdebine (1990). Poly(A)$^+$ RNA was purified by chromatography on oligo(dT)-cellulose (Sambrook et al., 1989).

Total RNA for cDNA library construction was prepared from stage III through V endosperm tissue. Poly(A)$^+$ RNA was purified using oligo(dT) spin columns according to the

[1] This work was supported in part by a grant from the U.S. Department of Energy (DE-FG02-94ER20133). The first two authors contributed equally to this work.

[2] Present address: U.S. Department of Agriculture-Agricultural Research Service, N-322D Ag Science North, University of Kentucky, Lexington, KY 40546–0091.

* Corresponding author; e-mail crs@andrew.stanford.edu; fax 1–415–325–6857.

Abbreviations: ACP, acyl carrier protein; dbEST, data base for expressed sequence tags; EST, expressed sequence tag; FAS, fatty acid synthase; NCBI, National Center for Biotechnology Information; ORF, open reading frame.

manufacturer's instructions (5 Prime-3 Prime, Inc., Boulder, CO).

Construction of cDNA Library

An oriented λZAPII cDNA library was prepared using a ZAP-cDNA synthesis kit (Stratagene) according to the manufacturer's instructions. The cDNA (200 ng) was ligated into λZAPII digested with *Xho*I and *Eco*RI, such that 5′ ends of the inserts should be found at the T3 side of the polylinker. This yielded 5×10^5 primary plaques, which were eluted in buffer (100 mM NaCl, 8 mM $MgSO_4$, 50 mM Tris-HCl, pH 7.5, 0.1% gelatin) and stored at 4°C. *Escherichia coli* strain XL1-blue22 MRF′ (Δ[*mcrA*]183, Δ[*mcrCB-hsdSMR-mrr*]173, *endA1*, *supE44*, *thi-1*, *recA1*, *gyrA96*, *relA1*, *lac*, [F′ *proAB*, *lacI*^{q}ZΔM15, Tn10 (*tet*^{r}), *Amy*, *camr*]) was used for propagating the library and derived plasmids.

Differential Screening

Randomly chosen phage from widely separated plaques were picked into 200 μL of buffer (100 mM NaCl, 8 mM $MgSO_4$, 50 mM Tris-HCl, pH 7.5, 0.1% gelatin) in 96-well microtiter plates. To prepare filter replicas, phage were replicated by transferring a small drop of lysate with a 96-prong device onto bacterial lawns prepared by mixing 0.2 mL of a saturated overnight culture of bacteria with 7.5 mL of molten top agar (Luria broth, 0.5% agar at 45°C) and pouring onto a 132-mm Petri dish containing agar-solidified Luria broth. The blunt, approximately 1-mm-diameter metal prongs carried sufficient phage to give plaques of consistent size, without significant encroachment between neighboring plaques. Following plaque development, duplicate or triplicate nylon filter replicas of the plaques were prepared. Phage DNA was fixed to the filters (Hybond N^+, Amersham) by alkaline denaturation (Sambrook et al., 1989). The filters were prehybridized at 65°C for approximately 1 h in the hybridization solution (0.6 M NaCl, 0.12 M Tris-HCl, pH 7.4, 8 mM EDTA, 0.1% sodium PPi, 0.2% SDS, 5% dextran sulfate, 0.1% heparin, and 1 μg mL^{-1} polyadenylic acid) before addition of the probe and hybridization overnight at 65°C. The filters were washed three times in 2× SSC, 0.1% SDS at room temperature. Hybridization was visualized by phosphor-imaging (Molecular Dynamics, Sunnyvale, CA).

The first set of clones (dbEST 39704–39883) was selected for sequencing by probing plaque lifts with ^{32}P-labeled first-strand cDNA from 1 μg of poly(A) mRNA from leaf or developing seed mRNA. Reverse transcription of mRNA was primed in 50-μL reactions with oligo$(dT)_{12-18}$ in the presence of 4.8 μM [α-^{32}P]dCTP (400 Ci/mmol) as described by Sambrook et al. (1989). Clones were selected that had no detectable signal (above background) with the leaf probe and had a weak to moderately strong signal with the developing-seed probe.

The second set of clones (dbEST 39884–40169) was selected by probing plaque lifts as described above, and, in addition, a third filter was probed with a randomly primed (Sambrook et al., 1989) probe representing the pooled inserts of clones that had been sequenced most frequently in the first batch. These included 6 clones with homology to 2S storage proteins, 10 clones with homology to 12S storage proteins, 4 enolase clones, 2 clones for HSP70, one clone for a low molecular weight heat-shock protein (GenBank accession No. J50710), and 9 clones for ribosomal proteins (P0, L27, S8, S9, S13, S14, S17, S28, YL27). Clones that hybridized to the pooled probes were not selected for sequencing.

Production of Antibody against Crude Endomembranes

Microsomal membranes were prepared from endosperm of developing castor seeds, stages III to V (Greenwood and Bewley, 1982), as follows. Endosperm (10 g) was ground using a mortar and pestle in homogenization buffer (0.15 M Tricine-KOH, pH 7.5, 10 mM KCl, 1 mM $MgCl_2$, 2 mM DTT, 0.6 M Suc). The preparation was centrifuged successively at 10,000*g* for 15 min and 27,000*g* for 20 min, and in each case the pellet was discarded. Crude microsomes were pelleted by centrifugation at 100,000*g* for 100 min. The pellet was resuspended in 100 mM $NaCO_3$, pH 11.5, and layered over 7 mL of 0.5 M Suc, 100 mM $NaCO_3$, pH 11.5, and centrifuged at 100,000*g* for 100 min. The final pellet was resuspended in 1 mL of sterile 150 mM NaCl, 10 mM $NaPO_4$, pH 7.4.

Prior to injection into rabbits, the microsomal membrane preparation was boiled in 1% SDS and then dialyzed exhaustively against several changes of 150 mM NaCl, 10 mM $NaHPO_4$, pH 7.4. Injections of protein (approximately 200 μg) were made subcutaneously at 5-week intervals, using Freund's complete adjuvant for the first injection and incomplete Freund's adjuvant for the second two injections. IgG was purified from serum, collected 10 d after the final injection, by the use of a protein A-Sepharose column (Pharmacia, Uppsala, Sweden) according to the manufacturer's instructions. This IgG was then further treated by passing it (three times) over a column containing soluble protein from castor endosperm, which was bound to a mixture of Affigel 10 and Affigel 15 (Bio-Rad) according to the manufacturer's instructions. After each passage any IgG that bound to the columns was discarded.

Selection of cDNA Clones using Antibodies

The cDNA library was plated at a density of approximately 2000 plaques per 132-mm Petri plate. Preparation of nitrocellulose replicas, incubations with the anti-ER IgG, described above, and selection of plaques were carried out according to instructions for the λZAPII vector (Stratagene). No further plaque purification was performed; 500 immunopositive plaques were selected by this method. One hundred of these clones were sequenced without further selection. The remaining 400 phage were placed in 96-well microtiter dishes and hybridized, as described above, with cDNA clones of commonly occurring sequences. Positively hybridizing plaques were excluded from those sequenced. Clones selected for sequencing using the anti-ER IgG have dbEST accession Nos. 52290 to 52341, 60978 to 61068, and 61073 to 61180.

Clones were also selected using an antiserum raised against crude microsomes isolated solely by differential centrifugation (a gift of Dr. Clint Chapple). These clones were isolated after two rounds of plaque purification. The 41 sequences from clones selected using these antibodies have dbEST accession Nos. 60941 to 60977 and 61069 to 61072.

DNA Sequencing and Analysis

Following excision of the phagemid from the λZAPII vector, a single colony was selected and grown overnight in 5 mL of Luria broth medium containing 100 μg/mL ampicillin. Plasmid DNA was purified using Magic Minipreps (Promega), quantitated, and then sequenced using the T3 primer by *Taq* cycle sequencing on an ABI Catalyst 8000 Molecular Workstation (Applied Biosystems) and ABI373A sequenators (Applied Biosystems). Sequence data were edited manually to remove vector and ambiguous sequences from the ends and compared against the nonredundant combined data bases using the program BLASTX (Altschul et al., 1990) provided by the NCBI e-mail server (blast@ncbi.nlm.nih.gov). Alignments returning scores of at least 80 were considered significant (for discussion, see Newman et al., 1994). Sequences have been submitted to the NCBI data base of ESTs (dbEST), and all sequence comparisons described in this paper were done on the date of submission to the data base (January or August 1994). For dbEST accession retrieval instructions, send the following two-line electronic mail message to retrieve@ncbi.nlm.nih.gov

DATALIB DBEST
HELP

RESULTS

DNA Sequence Analysis

An oriented cDNA library was constructed in the expression vector λZAPII using RNA purified from endosperm tissue of developing castor seeds. The quality of the library was assessed by hybridizing a castor stearoyl-ACP desaturase cDNA clone (Shanklin and Somerville, 1991) to 20,000 plaques. Sixteen clones were recovered, three of which were full length. Therefore, it was concluded that the library contained adequate representation of moderately abundant clones. Clones were selected from this library by differential screening and immunological screening procedures described below. An average of more than 400 bp of nucleotide sequence from 743 selected clones was obtained by single sequencing runs from the putative 5′ end of the insert DNA. These sequences have been deposited in the dbEST.

After manual editing to remove vector sequences, each EST sequence was compared with the protein sequences in the public data bases by using the program BLASTX (Altschul et al., 1990) provided by the NCBI e-mail server. BLASTX scores of greater than 80 were considered indicative of potentially significant sequence homology (Newman et al., 1994). A summary of the results of these analyses are presented in Tables I and II. Three clones with significant homology to a phosphate translocator, an acyltransferase, and a bacterial regulatory protein have been omitted from the tables but will be described in detail elsewhere. The 321 clones for which the highest scoring match was a sequence previously obtained from castor or from another higher plant are listed in Table I. The 100 clones representing 78 different genes for which the highest homology was to an organism other than a higher plant are listed in Table II. In some of these cases, in which the highest homology was to a nonplant sequence, weaker homology to plant sequences was also apparent. This suggests that the EST is related to a sequence previously determined from higher plants but appears to represent a new isozyme. The rapidity with which sequence data are currently being added to public data bases makes a more detailed analysis of these categories rapidly obsolete (Newman et al., 1994). A current analysis of the castor sequences can be conveniently obtained by using the NCBI World Wide Web server at http://www.ncbi.nlm.nih.gov as described by Newman et al. (1994).

Clones Selected by Differential Screening

The first 222 clones were selected for sequencing by probing 839 randomly chosen clones with labeled first-strand cDNA from leaves and developing seeds. Of the 839, 280 gave no detectable signal when probed with the leaf-derived cDNA and 162 clones produced a strong signal when hybridized to the ^{32}P-labeled probe made from developing seed poly(A)$^+$ RNA. Of these 280, 222 exhibited weak hybridization to the seed-derived probe and were selected for sequencing. Thus, approximately 26.5% of clones were in the category "seed-specific and not highly abundant." Of the 162 clones having a strong signal when probed with seed-derived cDNA, only 58 appeared to be seed specific.

An additional 348 clones were selected for sequencing by probing 851 randomly chosen clones with leaf-derived cDNA, seed-derived cDNA, and a pool of the most frequently sequenced clones from the first batch of sequences. In this experiment, 370 of 851 plaques gave a strong signal when probed with seed-derived cDNA, 512 gave no detectable signal with leaf-derived cDNA, and 141 gave a signal with a probe made from the pool of abundant sequences. This resulted in the selection of 348 clones to be sequenced (i.e. 40.9% of all clones). The difference in the number of clones selected in the two experiments is attributed to differences in the specific activity of the various probes, which led to varying degrees of discrimination. Nevertheless, these results are in general agreement with determinations of organ-specific RNA complexity in tobacco derived from RNA excess/single-copy DNA hybridization experiments (Kamalay and Goldberg 1980, 1984).

Of the 570 clones selected for sequencing, informative sequence data were obtained for 468 clones. Among these were three clones that had sequence similarity to the Fad2 fatty acid Δ12 desaturase of *Arabidopsis thaliana* (Table I). These three clones were all derived from the same gene, which has subsequently been shown to encode an oleate

Table I. *Inventory of castor ESTs encoding ORFs with significant sequence homology to gene products from higher plants*

For each EST No. (the accession No. assigned by dbEST) the screen by which the clone was selected for sequencing (d, differential; i, immuno), the BLASTX score of the highest scoring alignment, the protein description, and the source organism are indicated. Where more than one EST had the same putative identification, the number of "hits" is indicated in parentheses.

EST No.	Screen	Score	Putative Identification	Organism
40046	d	155	3-Oxoacyl-[ACP] reductase	*Cuphea lanceolata*
61137	i	476	Actin	Carrot
39913	d	320	Actin depolymerizing factor (3)	*Lilium longiflorum*
61090	d + i	349	Acyl carrier protein (4)	*Cuphea lanceolata*
61021	i	373	Acyl-CoA-binding protein (2)	Oilseed rape
39728	d	267	Acyl-[ACP] desaturase (3)	Safflower
39912	d	484	ADP,ATP carrier protein (3)	Rice
40135	d + i	570	Agglutinin (3)	Castor
40149	d	267	Ala transaminase	*Panicum miliaceum*
61033	i	160	α-Amylase	Rice
60958	i	255	Annexin (3)	Alfalfa
39736	d	228	Anthranilate synthase component I-1	*A. thaliana*
40128	d	134	Ascorbate peroxidase	Spinach
39733	d	411	Asparaginyl-tRNA synthetase	*E. coli*
40159	d	369	Aspartate aminotransferase	Soybean
40150	d	133	Auxin-induced protein aux28	Soybean
39898	d	554	β-ketoacyl-ACP synthase	Castor
40110	d	191	Bispecific *O*-methyltransferase	*Populus tremuloides*
40169	d	204	Bowman-Birk trypsin inhibitor (4)	Potato
40127	d	308	Calcium-dependent protein kinase	Soybean
52328	i	630	Calnexin (5)	*A. thaliana*
60992	i	402	Calreticulin	Barley
39963	d + i	407	Chaperonin, 20 kD (2)	Spinach
52323	i	186	Cysteine proteinase inhibitor	*A. thaliana*
40112	d	143	Cyt b_5	Rice
39910	d	137	Dehydrin	Pea
40142	d	498	Δ 12 oleate desaturase (3)	*A. thaliana*
39932	d	245	Dessication-related protein	*Craterostigma plantagineum*
40152	d	153	Dihydrolipoamide dehydrogenase (2)	Pea
39887	d	131	DNA-binding protein GT-2	Rice
40030	d + i	305	DNAJ protein (2)	Cucumber
39883	d	93	Ec protein	*A. thaliana*
61082	i	128	Elongation factor eEF-1a	Soybean
39939	d	120	Embryonic abundant protein	*Vicia faba*
39804	d	628	Enolase (5)	Castor
39928	d	112	Fructokinase (2)	Potato
40132	d	541	Fru-bisP aldolase (2)	Spinach
60943	i	189	Gast1 gene product	Tomato
61115	i	135	GSH peroxidase	Tobacco
61037	d + i	439	Glyceraldehyde-3-P-dehydrogenase (3)	*Atriplex nummularia*
61145	i	153	Gly-rich RNA-binding protein	*A. thaliana*
60999	i	512	GRP94 homolog (2)	Barley
60946	i	234	Guanine nucleotide-binding protein	Tobacco
39865	d	462	Guanine nucleotide regulatory protein	Fava bean
40016	d	325	Heat-shock protein, class I (3)	Rice
39706	d	247	Heat-shock protein (2)	Spinach
61051	d + i	302	Heat-shock protein, 70 kD (5)	Kidney bean
40119	d	312	Histone H2A	*Picea abies*
61102	i	365	Histone H2B	Maize
39807	d	188	Initiation factor 5A	*Nicotiana plumbaginifolia*
39962	d	117	LEA protein	*A. thaliana*
39900	d	435	Legumenin (bean endopeptidase)	*Vigna mungo*
61012	i	142	Low molecular weight Cys-rich protein	Soybean
39805	d	480	Luminal binding protein BLP-4	Tobacco
39894	d	354	Malate dehydrogenase	Watermelon
39860	d + i	150	Metallothionein I (4)	Castor
61070	i	339	NADH-ubiquinone reductase	*A. thaliana*
39998	d	140	Naringenin 3-dioxygenase	*Callistephus chinensis*
39890	d	190	Nonspecific lipid-transfer protein A	Castor

Table I. *(Continued)*

EST No.	Screen	Score	Putative Identification	Organism
39806	d	120	Oleosin (2)	Oilseed rape
39786	d	93	Opaque2 heterodimerizing protein 1	Maize
40002	d	150	ORF (pir|S22499|S22499)	*Bromus secalinus*
39976	d	109	Pathogenesis-related protein 5	*A. thaliana*
61111	i	327	Peptidyl-prolyl *cis-trans*-isomerase	*Allium cepa*
39956	d	328	Pollen-specific protein	*A. thaliana*
40012	d	96	Poly(A)-binding protein	*A. thaliana*
60959	i	612	Polyubiquitin	Sunflower
39826	d	142	PPLZ02-like protein (2)	*A. thaliana*
39863	d	150	Pro-rich protein	*Phaseolus vulgaris*
61001	d + i	538	Protein disulfide isomerase (12)	Alfalfa
39993	d	549	Protein kinase (5)	Soybean
39948	d	137	Protein P10	*N. plumbaginifolia*
39906	d	459	PPi-F,6P-1-phosphotransferase	Potato
39755	d	444	RAS-related GTP-binding protein (2)	Tobacco
39975	d	130	Ribosomal protein L2	Tomato
61006	i	336	Ribosomal protein L26	*A. thaliana*
60961	d + i	439	Ribosomal protein L27 (6)	Pea
39961	d	498	Ribosomal protein L29 (2)	*A. thaliana*
40020	d	498	Ribosomal protein L3 (2)	*A. thaliana*
39750	d	83	Ribosomal protein L34	Tobacco
61028	i	305	Ribosomal protein L5	*A. thaliana*
39921	d	466	Ribosomal protein L8	Tomato
52294	d + i	615	Ribosomal protein PO (27)	Rice
39945	d	88	Ribosomal protein S11	*Epifagus virginiana*
39836	d	243	Ribosomal protein S13	Maize
39848	d	236	Ribosomal protein S14	Maize
60964	d + i	521	Ribosomal protein S15 (3)	*A. thaliana*
61013	i	445	Ribosomal protein S18	*A. thaliana*
40134	d	227	Ribosomal protein S25	Tomato
39965	d	522	Ribosomal protein S3AE	*Catharanthus roseus*
39827	d	110	Ribosomal protein S4	Tobacco
39888	d	275	Ribosomal protein S6	Tobacco
61151	d + i	628	Ricin (2)	Castor
40116	d	557	RNA polymerase II subunit	Soybean
39930	d	117	*S*-Adenosylmethionine decarboxylase	Potato
39929	d	372	*S*-Adenosylmethionine synthetase	*A. thaliana*
39951	d + i	394	Seed storage protein, 12S (52)	*Cucurbita maxima*
60956	i	85	Seed storage protein, 7S (2)	Pea
61113	d + i	523	Seed storage protein, 2S (45)	Castor
40027	d	99	Ser proteinase	*L. longiflorum*
39715	d	259	St12p protein	*A. thaliana*
39811	d	108	Stellacyanin	*Rhus vernicifera*
61144	i	233	Suc synthase 1	*V. faba*
40031	d	256	Superoxide dismutase-1 (Cu-Zn)	Rice
40168	d	506	T complex polypeptide 1	Oat
61138	i	487	Thioredoxin H	*A. thaliana*
40082	d + i	517	Tonoplast intrinsic protein α (5)	*A. thaliana*
39716	d	84	Transmembrane protein TMP-B (2)	*A. thaliana*
40033	d	626	Tubulin α chain (2)	*Anemia phyllitidis*
60942	i	592	Ubiquitin (4)	*A. thaliana*
39771	d	324	Ubiquitin carrier protein (2)	Alfalfa
39896	d + i	568	Ubiquitin/ribosomal protein (3)	Turnip
39983	d	179	UTP-Glc-1-P uridylyltransferase	Potato
39901	d	264	Vacuolar ATP synthase catalytic subunit A	Cotton

12-hydroxylase by expression in transgenic plants (van de Loo et al., 1995).

The largest class of clones showed homology to storage proteins or components of the protein biosynthetic apparatus. These included enzymes of amino acid metabolism such as Thr synthase, anthranilate synthase, carbamoylphosphate synthase, Orn carbamoyltransferase, and three aminotransferases. Other components of the translation apparatus were asparaginyl-tRNA synthetase, initiation factor 5, and elongation factor 2. A putative regulatory protein of storage protein synthesis, an opaque2 heterodimerizing protein homolog, was also identified. Vari-

Table II. *Inventory of castor ESTs encoding ORFs with significant similarity to nonplant gene products*

For each EST the column headings indicate the accession No. assigned by dbEST, the screen by which the clone was selected for sequencing (d, differential; i, immuno), the BLASTX score of the highest scoring alignment, the protein description, and the source organism. Where more than one EST had the same putative identification, the number of "hits" is indicated in parentheses.

EST No.	Screen	Score	Putative Identification	Organism
40083	d	442	14–3-3 protein, ε isoform	Mouse
39766	d	153	4-Aminobutyrate aminotransferase	*E. coli*
39969	d	91	Adenylate cyclase	*Bordetella pertussis*
40061	d	131	Adult intestinal protein AdRab-F	Rabbit
39854	d	124	Aldehyde dehydrogenase	Chicken
52333	i	90	ATP-dependent protease	*Streptococcus salivrius*
39971	d	230	Biotin carboxylase	*Anabaena* sp.
60973	i	156	Btf3 transcription factor (2)	Human
39908	d	118	Carbamoyl-phosphate synthase	*E. coli*
40009	d	150	Coatomer	Bovine
39774	d	295	Deoxyribonuclease	*Drosophila melanogaster*
40097	d	502	Elongation factor 2	*Chlorella kessleri*
52338	i	152	EMP70 endosomal protein	Bakers' yeast
40086	d	171	Ferrochelatase	Mouse
39801	d	114	Gly 1 gene product	Bakers' yeast
39777	d	152	Gst1-hs GTP-binding protein (2)	Human
39990	d	162	Histone H4	*Tetrahymena pyriformis*
52313	i	88	Indole acetamide hydrolase	*Agrobacterium rhizogenes*
40048	d	82	Integral membrane protein	Mouse
40107	d	97	Internal membrane protein	Bakers' yeast
39886	d	154	Lamin b receptor	Chicken
40090	d	264	Lipid A biosynthetic protein, 17 kD	*Rickettsia rickettsii*
61005	i	115	Met synthase (2)	*E. coli*
39775	d	189	MO25, embryo-specific gene	Mouse
60947	i	84	Myosin (2)	Chicken
39871	d	114	NAM8, suppressor of splicing defect	Bakers' yeast
39878	d	117	NIFM, processing of nitrogenase	*Klebsiella pneumoniae*
39742	d	263	ORF (gpID14662IHUMORF06 1)	Human
39779	d	158	ORF, 17.4 kD, glpf-hslu region	*E. coli*
39934	d	134	ORF, 31.0 kD, rnpb 3′ region	*E. coli*
39851	d	106	ORF, 42.5 kD, carb-kefc region	*E. coli*
39762	d	89	ORF, photosynthetic cluster	*Rhodobacter capsulatus*
40059	d	219	ORF, PEP3 5′ region	Bakers' yeast
40025	d	189	Orn carbamoyltransferase	*Mycobacterium bovis*
39923	d	408	P23 transplantation antigen	Bovine
39708	d	169	Peroxisomal membrane protein A	*Candida boidinii*
52331	i	203	Potassium channel	Rat
61010	d + i	184	Pyruvate dehydrogenase E1 component (2)	*Bacillus subtilis*
39741	d	165	Pyruvate kinase	Chicken
40055	d	162	Riboflavin synthase	*B. subtilis*
40013	d	144	Ribokinase	*E. coli*
40003	d	319	Ribosomal protein L11	Rat
52306	i	395	Ribosomal protein L12 (5)	Human
61064	i	90	Ribosomal protein L19E	Slime mold
40124	d	80	Ribosomal protein L21.e	Bakers' yeast
40105	d	217	Ribosomal protein L31	*Chlamydomonas reinhardtii*
39914	d	212	Ribosomal protein L36	Rat
40074	d	440	Ribosomal protein L5	Bakers' yeast
40064	d	159	Ribosomal protein L7	Human
39911	d	137	Ribosomal protein P1	*Artemia salina*
61154	d + i	193	Ribosomal protein P2 (10)	*Cladosporium herberum*
39704	d	211	Ribosomal protein S17 (2)	Chinese hamster
61059	i	171	Ribosomal protein S19	Rat
39746	d	480	Ribosomal protein S28 (2)	Human
52327	i	195	Ribosomal protein S29	Rat
40163	d	108	Ribosomal protein S5	*Podocoryne carnea*
61107	i	280	Ribosomal protein S6	Fission yeast
39829	d	139	Ribosomal protein S7	Human

Table II. (*Continued*)

EST No.	Screen	Score	Putative Identification	Organism
39784	d	279	Ribosomal protein S8	*Xenopus laevis*
39785	d	240	Ribosomal protein YL10	Bakers' yeast
39973	d	267	Ribosomal protein YL37	Bakers' yeast
39958	d	125	Ribosome-binding protein p34	Rat
52309	i	240	*S*-adenosylmethionine:sterol-C-methyltransferase	Bakers' yeast
61073	i	80	Single-stranded binding protein	*Brucella abortus*
40000	d	150	Sm-D autoantigen	Mouse
40017	d	94	Spermidine synthase	Human
61146	i	313	Stress-inducible protein sti35	Imperfect fungus
40137	d	189	Succinate dehydrogenase flavoprotein	*Rickettsia prowazekii*
39832	d	80	Suc α-glucosidase	Rat
39707	d	96	Thr synthase	*B. subtilis*
52290	i	119	Transcription factor SII-related	Mouse
39723	d	97	Tropomyosin-related protein	Rat
40037	d	278	V-ATPase 14 kD subunit (2)	*D. melanogaster*
5232	i	102	Vegetative specific protein H7 (2)	Slime mold
39738	d	89	ORF, YCL311	Bakers' yeast
61117	i	166	ORF, YhR012w	Bakers' yeast
39778	d	119	Zeta-crystallin NADPH-oxidoreductase	*Leishmania amazonensis*
39719	d	293	Zinc-binding protein (Trithorax)	*D. melanogaster*

ous clones of protein disulfide isomerase, heat-shock proteins, and chaperonins were identified. These could be involved in folding and processing of seed storage proteins. Other steps in protein synthesis and storage for which clones may have been identified included a protein involved in Golgi vesicle traffic (St12p) and import of protein into the vacuole (NIFM, which has homology to bacterial protein-processing/export proteins). The identification of a number of tonoplast and vacuolar clones is consistent with expansion of vacuoles for protein storage. Clones for ribosomal proteins and other components of the protein synthesis and protein-processing pathway were also identified by immunoscreening (see below). The recovery of clones for storage proteins appeared to be due to the fact that many small (i.e. incomplete) clones gave weak hybridization signals when probed with seed-derived cDNA or the pool of abundant clones and were, therefore, not excluded.

Approximately 6% of the clones obtained by differential screening were homologous to enzymes involved in carbohydrate or lipid metabolism. This is consistent with the fact that a major metabolic activity of the developing seed is the conversion of imported Suc via glycolysis to pyruvate to fatty acid and lipid synthesis. Particularly interesting examples include pyruvate kinase, three components of the pyruvate dehydrogenase complex, and the biotin carboxylase component of the plastidic acetyl-CoA carboxylase.

In addition to the 570 clones selected for sequencing on the basis of the differential screen, we sequenced five clones that gave a strong signal with the seed probe and would, therefore, normally have been excluded. Four encoded 2S seed storage protein, and the fifth encoded 12S seed storage protein (data not shown). This confirmed that the clones excluded on the basis of strong seed signal were generally not useful for identifying novel genes.

Clones Selected by Immunological Screening

Sequence data were obtained for 290 clones selected by screening with either of two antisera (Tables I and II, clones marked "i"). Sequences obtained from clones isolated with antiserum raised against crude microsomes from developing castor seeds showed little evidence of enrichment for integral ER proteins. The clones selected were primarily homologous to sequences that are commonly found in developing seeds. These sequences included those showing homology to proteins from the ubiquitin pathway, annexin, and seed storage proteins.

Antiserum was also raised against purified microsomal membranes from developing castor seeds and used to select clones. The best evidence that these anti-ER antibodies were at least partially selective for ER integral membrane proteins is deduced from the fact that cDNAs homologous to the calnexin gene were sequenced five times. Calnexin, a calcium-binding protein, is considered to be one of the major integral membrane proteins of the ER (Wada et al., 1991). Another clone homologous to an integral membrane protein of the ER is a putative sterol biosynthetic gene, *S*-adenosylmethionine:Δ24-sterol C-methyltransferase.

Two hundred six of the clones encoded ORFs with significant homology to previously described gene products. The majority of selected clones corresponded to proteins that may be transiently associated with the ER (e.g. ribosomes and storage proteins), although not actually being integral membrane proteins. In view of the fact that the protein composition of developing seeds consists of more that 50% storage proteins, which are synthesized on membrane-bound polysomes (Chrispeels et al., 1982), it is not surprising that many of the ESTs corresponded to proteins involved in protein biosynthesis. A large number of ribosomal genes were isolated. Among these were cDNAs en-

coding the P0 (L10E), P2(L12E1), L5, L12, L19, L26, and L27 proteins of the 60S subunit, as well as the S6, S15, S18, S19, and S29 proteins of the 40S subunit. In addition, a number of sequences appear to correspond to proteins involved in proper folding and assembly of newly synthesized polypeptides. These include sequences showing homology to Hsp70, GRP94, DNAJ, protein disulfide isomerase, and peptidyl-prolyl *cis-trans*-isomerase. Among the diverse array of novel plant sequences that did not show homology to previously characterized plant genes or ESTs from other plant species are those showing homology to the E1 component of pyruvate dehydrogenase, a vitamin B_{12}-dependent Met synthase, and a sequence homologous to the *iaaH* gene of *Agrobacterium*.

One EST showed homology to the *iaaH* gene from the Ti (or Ri) plasmid of *Agrobacterium*. This gene encodes the enzyme indole acetamide hydrolase, one of two enzymes that these bacteria use to synthesize IAA (auxin) (Thomashow et al., 1984). This pathway is not usually considered to be the normal pathway by which plants synthesize auxin (for review, see Hobbie and Estelle, 1994); however, studies of transformed plants (Klee et al., 1987) and recombinant T-DNAs (Binns et al., 1982) suggest that some plant species possess this enzyme activity. Further study of this cDNA may help to resolve this controversy.

DISCUSSION

Putative identifications were made for 49% of clones selected by the differential screen and 71% of clones selected by immunoscreening. These values are substantially higher than those obtained in most similar experiments (Adams et al., 1991, 1992; Hoog, 1991; Uchimiya et al., 1992; Höfte et al., 1993; Keith et al., 1993; Newman et al., 1994). This is primarily because the tissue used for library construction was specialized for protein and lipid accumulation, processes that involve several well-characterized classes of proteins, such as storage proteins, ribosomal proteins, and glycolytic enzymes. Although differential hybridization to various probes was useful for eliminating some highly redundant or ubiquitous clones, the presence of many incomplete cDNAs in the library reduced the effectiveness of the approach because these gave weak hybridization signals that disguised their identity as members of an abundant or ubiquitous class of transcripts. The antibody-based screen was probably also strongly biased toward abundant transcripts, since immunization with whole membranes would be expected to give the highest titer against the most abundant epitopes. Because moderately or abundantly expressed genes tend to have a higher probability than weakly expressed genes of showing homology to a known gene in the sequence data bases (Green et al., 1993), the relatively high frequency of putative identifications is not unexpected.

Although developing castor endosperm was not an ideal choice of tissues for maximizing the identification of novel plant genes by the EST approach, putative plant homologs were identified for 78 new genes (Table II). An examination of the putative identifications of clones selected by the differential screen (Tables I and II, marked "d") indicates that these clones are not all seed specific. This is almost certainly accounted for by the fact that low-abundance messages were not sufficiently represented in the leaf mRNA-derived probe to give a hybridization signal significantly above background and, therefore, these clones were not discriminated against in the preliminary screen.

A large number of the cDNAs selected using antibodies against total microsomal or ER membranes (Tables I and II, marked "i") do not appear to correspond to genes that encode integral membrane proteins of the ER. There are probably two main reasons for this. First, because of the large number of clones that were handled, only one round of plaque purification was performed, and inevitably some of the colonies selected in the subsequent phagemid excision may be from a contaminating phage rather than the phage that gave the initial positive reaction with the anti-ER antibody. Additionally, although many of the soluble proteins and proteins loosely associated with the ER were removed during membrane preparation, some contamination of the microsomal membrane preparation with proteins other than integral membrane proteins probably occurred. Because integral membrane proteins that contain multiple membrane-spanning domains are often very poor antigens, there is an intrinsic bias toward raising antibodies to epitopes from the contaminating soluble proteins. This may explain why no cDNAs were selected for some of the proteins that would be expected in the ER, such as those for phospholipid biosynthesis. This idea is consistent with the observation that calnexin, a 67-kD integral membrane protein of the ER that was selected and sequenced five times, has only a single membrane-spanning domain and a very large luminal N-terminal domain (Wada et al., 1991), which presumably acts as a good antigen.

When one examines EST sequence data, such as that presented here, a number of precautions must be kept in mind (Newman et al., 1994). In particular, we note that, when a number of the castor ESTs are grouped under the same putative identification, they may represent different but related genes. For example, two clones (ESTs 39895 and 40152) had homology to the mitochondrial form of dihydrolipoamide dehydrogenase previously sequenced from pea (Table I). However, these clones share almost no similarity at the nucleotide level. They probably represent the plastid and mitochondrial forms of the enzyme, since 39895 has much lower amino acid identity (45%) with the mitochondrial form than does 40152 (86%).

A further limitation of the data presented in Tables I and II is that the "putative identification" is based simply on the name of the accession with the strongest sequence alignment. In some cases this gives much less information than if the other alignments of a BLASTX search result are also considered. For example, the putative identification of EST 40061 is "adult intestinal protein AdRab-F" (Table II), because the strongest alignment was with a rabbit gene product of unknown function, identified by differential hybridization (Boll et al., 1993). However, an examination of the complete BLASTX output reveals that the castor EST also has sequence similarity, in the same reading frame,

with a range of multifunctional enzymes including bacterial polyketide synthases, mycocerosic acid synthase of *Mycobacterium tuberculosis* (which catalyzes fatty acid elongation), Fix23 of *Rhizobium meliloti* (involved in the synthesis of a secreted lipopolysaccharide), and the type-I FAS of chicken and rat. There are also alignments to alcohol dehydrogenases and quinone oxidoreductases, and the alignment with the type-I FAS enzymes appears to be to the enoyl-reductase domain. Furthermore, similar residues are responsible for the alignment of the castor EST and each of these enzymes, and these appear to be residues that are conserved in the NAD(P)H-binding site of enoyl-reductase. A hypothesis can therefore be made that EST 40061 represents a castor cDNA for the enoyl-reductase component of a type-I FAS. Type-I FAS enzymes are believed to exist in the cytoplasm of the plant cell, performing such reactions as fatty acid elongation, but no clones for components of these enzymes are yet available. The enoyl-ACP reductase component of the plant plastid type-II (prokaryotic) FAS has been cloned from rapeseed but has no detectable homology to this castor EST. A similar investigation of EST 39851 (*E. coli* ORF, Table II) suggests that it may encode an enzyme of fatty acid β-oxidation, acyl-CoA dehydrogenase, that is required by the seed during germination, and that may be synthesized and stored in glyoxysomes during seed development. These examples illustrate that full exploitation of the information content of EST sequence data can require closer examination of individual clones than is feasible to present in Tables I and II. Correct interpretation of sequence homology data remains one of the most challenging aspects of this kind of approach to gene identification.

In summary, the 743 sequences obtained in this work have been deposited in the dbEST data base, where they can be conveniently accessed as a new resource for identifying a variety of novel and useful plant genes. As the content of the data bases continues to expand, those EST sequences for which function could not be suggested at present may also yield novel opportunities to connect aspects of plant biology to knowledge gained from the study of other classes of organisms.

ACKNOWLEDGMENTS

We are grateful to Svieta Ndibongo, Linda Danhof, and Jill Hartsig for their expert assistance with the selection of clones and purification of DNA and to Susan Lootens for DNA sequencing.

Received January 10, 1995; accepted March 17, 1995.

LITERATURE CITED

Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, Utterback TR, Nagle JW, Fields C, Venter JC (1992) Sequence identification of 2375 human brain genes. Nature **355:** 632–634

Adams MD, Kelley 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 the human genome project. Science **252:** 1651–1656

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

Binns AN, Sciaky D, Wood NH (1982) Variation in hormone autonomy and regeneration potential of cells transformed by strain A66 of *Agrobacterium tumefaciens*. Cell **31:** 605–612

Boll W, Schmid-Chanda T, Semenza G, Mantei N (1993) Messenger RNAs expressed in intestine of adult but not baby rabbits: isolation of cognate cDNAs and characterization of a novel brush border membrane protein with esterase and phospholipase activity. J Biol Chem **268:** 12901–12911

Browse, Somerville CR (1991) Glycerolipid synthesis: biochemistry and regulation. Annu Rev Plant Physiol Mol Biol **42:** 467–506

Chrispeels MJ, Higgins TJV, Craig S, Spencer D (1982) Role of the endoplasmic reticulum in the synthesis of reserve proteins and the kinetics of their transport to protein bodies in developing pea cotyledons. J Cell Biol **93:** 5–14

Green P, Lipman D, Hillier L, Waterston R, States D, Claverie JM (1993) Ancient conserved regions in new gene sequences and the protein databases. Science **259:** 1711–1716

Greenwood JS, Bewley JD (1982) Seed development in *Ricinus communis* (castor bean). I. Descriptive morphology. Can J Bot **60:** 1751–1760

Hobbie L, Estelle M (1994) Genetic approaches to auxin action. Plant Cell Environ **17:** 525–540

Höfte H, Desprez T, Amselem J, Chiapello H, Caboche M, Moisan A, Jourjon MF, Charpenteau JL, Berthomieu P, Guerrier D, Giraudat J, Quigley F, Thomas F, Yu DY, Mache R, Raynal M, Cooke R, Grellet F, Delseny M, Parmentier Y, Marcillac G, Gigot C, Fleck J, Philipps G, Axelos M, Bardet C, Tremousaygue D, Lescure B (1993) An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana*. Plant J **4:** 1051–1061

Hoog C (1991) Isolation of a large number of novel mammalian genes by a differential cDNA library screening strategy. Nucleic Acids Res **19:** 6123–6127

Kamalay JC, Goldberg RB (1980) Regulation of structural gene expression in tobacco. Cell **19:** 935–946

Kamalay JC, Goldberg RB (1984) Organ-specific nuclear RNAs in tobacco. Proc Natl Acad Sci USA **81:** 2801–2805

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

Klee HJ, Horsch RB, Hinchee MA, Hein MB, Hoffmann NL (1987) The effect of overproduction of two *Agrobacterium tumefaciens* T-DNA auxin biosynthetic gene products in transgenic petunia plants. Gene Dev **1:** 86–96

Newman T, de Bruijn FJ, Green P, Keegstra K, Kende H, McIntosh L, Ohlrogge J, Raikhel N, Somerville S, Thomashow M, Retzel E, Somerville C (1994) Genes galore. A summary of methods for accessing results from large-scale partial sequencing of anonymous *Arabidopsis* cDNA clones. Plant Physiol **106:** 1241–1255

Park YS, Kwak JM, Kim YS, Lee DS, Cho MJ, Lee HH, Nam HG (1993) Generation of expressed sequence tags of random root cDNA clones of *Brassica napus* by single-run partial sequencing. Plant Physiol **103:** 359–370

Puissant C, Houdebine L (1990) An improvement of the single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Biotechniques **8:** 148–149

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

Sasaki T, Song J, Koga-Ban Y, Matsui E, Fang F, Higo H, Nagasaki H, Hori M, Miya M, Murayama-Kayano E, Takiguchi T, Takasuga A, Niki T, Ishimaru K, Ikeda H, Yamamoto Y, Mukai Y, Ohta I, Miyadera N, Havukkala I, Minobe Y (1994) Toward cataloguing all rice genes: large-scale sequencing of randomly

chosen rice cDNAs from a callus cDNA library. Plant J **6:** 615–624

Shanklin J, Somerville CR (1991) Stearoyl-ACP desaturase from higher plants is structurally unrelated to the animal homolog. Proc Natl Acad Sci USA **88:** 2510–2514

Thomashow LS, Reeves S, Thomashow MF (1984) Crown gall oncogenesis: evidence that a T-DNA gene from the Agrobacterium Ti plasmid pTiA6 encodes an enzyme that catalyzes synthesis of indoleacetic acid. Proc Natl Acad Sci USA **81:** 5071–5075

Uchimiya H, Kidou S, Shimazaki T, Aotsuka S, Takamatsu S, Nishi R, Hashimoto H, Matsubayashi Y, Kidou 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 de Loo F (1993) Ricinoleate biosynthesis in *Ricinus communis* L. PhD thesis. Michigan State University, East Lansing

van de Loo F, Broun P, Turner S, Somerville CR (1995) An oleate 12-hydroxylase from *Ricinus communis* (L.) is a fatty acyl desaturase homolog. Proc Natl Acad Sci USA (in press)

Wada I, Rindress ID, Cameron PH, Ou W-J, Doherty JJ, Louvard D, Bell AW, Dignard D, Thomas DY (1991) SSRa and associated calnexin are major calcium binding proteins of the endoplasmic reticulum membrane. J Biol Chem **266:** 19599–19610