A New Set of Arabidopsis Expressed Sequence Tags from Developing Seeds. The Metabolic Pathway from Carbohydrates to Seed Oil[1][w]

Joseph A. White[2], Jim Todd, Tom Newman, Nicole Focks, Thomas Girke, Oscar Martínez de Ilárduya, Jan G. Jaworski, John B. Ohlrogge, and Christoph Benning*

Departments of Biochemistry and Molecular Biology (J.A.W., N.F., C.B.), Botany and Plant Pathology (J.T., T.G., O.M.d.I., J.B.O.), and United States Department of Energy-Plant Research Laboratory (T.N.), Michigan State University, East Lansing, Michigan 48824; and Department of Chemistry and Biochemistry, Miami University, Oxford, Ohio 45056 (J.G.J.)

Large-scale single-pass sequencing of cDNAs from different plants has provided an extensive reservoir for the cloning of genes, the evaluation of tissue-specific gene expression, markers for map-based cloning, and the annotation of genomic sequences. Although as of January 2000 GenBank contained over 220,000 entries of expressed sequence tags (ESTs) from plants, most publicly available plant ESTs are derived from vegetative tissues and relatively few ESTs are specifically derived from developing seeds. However, important morphogenetic processes are exclusively associated with seed and embryo development and the metabolism of seeds is tailored toward the accumulation of economically valuable storage compounds such as oil. Here we describe a new set of ESTs from Arabidopsis, which has been derived from 5- to 13-d-old immature seeds. Close to 28,000 cDNAs have been screened by DNA/DNA hybridization and approximately 10,500 new Arabidopsis ESTs have been generated and analyzed using different bioinformatics tools. Approximately 40% of the ESTs currently have no match in dbEST, suggesting many represent mRNAs derived from genes that are specifically expressed in seeds. Although these data can be mined with many different biological questions in mind, this study emphasizes the import of photosynthate into developing embryos, its conversion into seed oil, and the regulation of this pathway.

To understand the regulatory networks governing metabolism in developing oil seeds, we initiated a genome-wide analysis of gene expression in seeds of Arabidopsis, taking advantage of recently developed genomic tools (Hieter and Boguski, 1997; Bouchez and Hofte, 1998). Although the Arabidopsis genomic sequence is now fully available (www.arabidopsis.org; Meinke et al., 1998), expressed sequence tags (ESTs) derived from single-pass sequencing of cDNAs in Arabidopsis provide an invaluable resource for the annotation of genomic sequences and the analysis of gene expression associated with specific plant tissues or growth conditions (Newman et al., 1994; Cooke et al., 1996; Rounsley et al., 1996). Cloning of genes encoding enzymes of specific biochemical pathways by single-pass sequencing of cDNAs has been a very successful strategy, particularly when the cDNA libraries have been prepared from tissues with high activity for the respective enzymes. For example, sequencing of cDNAs derived from endosperm of developing castor bean seeds led to the identification of the enzyme involved in ricinoleic acid biosynthesis (Van de Loo et al., 1995a, 1995b). In a similar manner, genes essential for the biosynthesis of conjugated double bond-containing fatty acids were recently identified among ESTs from oleogenic tissues of *Momordica charantia* and *Impatiens balsamina* (Cahoon et al., 1999) and ESTs from wood-forming tissues of trees have proven to be an ideal source for the isolation of cDNAs encoding enzymes of cell wall biosynthesis (Allona et al., 1998; Sterky et al., 1998). ESTs and their accompanying cDNAs also provide the means to construct inexpensive microarrays on glass slides, which can be used to study the expression of genes on a genome-wide scale (DeRisi et al., 1997; Ruan et al., 1998). A careful bioinformatic analysis to identify tissue-specific ESTs is a prerequisite to obtain a comprehensive and representative set of cDNAs for gene expression studies by microarrays (Loftus et al., 1999). Thus, given that only a small number of plant ESTs in the public databases have been derived from seeds, it was essential in the context of the genome-wide analysis of seed metabolism to obtain and analyze a large number of these ESTs first.

Even without subsequent microarray analysis, a sufficiently large number of ESTs derived from a specific tissue can provide a clue toward the expres-

[1] This work was supported in parts by the National Science Foundation (grant nos. MCB–94–06466 and IBN–97–23778), the Midwestern Consortium for Plant Biotechnology Research, Dow Agroscience, and the Michigan Agricultural Experiment Station.

[2] Present address: The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850.

[w] The online version of this article contains Web-only data. The supplemental material is available at www.plantphysiol.org.

* Corresponding author; e-mail benning@pilot.msu.edu; fax 517–353–9334.

sion of specific genes in the tissue (Rafalski et al., 1998; Ewing et al., 1999; Mekhedov et al., 2000). In most cases and within statistical limitations (Audic and Claverie, 1997) the abundance of a specific cDNA in the EST collection is a measure for gene expression. Here we apply this technique also referred to as "electronic or digital northern" to address the questions about the primary metabolic route for the conversion of photosynthate into oil in developing seeds of Arabidopsis. The described analysis of 10,500 cDNAs by single-pass sequencing provides a rich data set, which we can only begin to explore here. For this reason the data set will be available at our web page for further studies.

RESULTS AND DISCUSSION

Single-Pass Sequencing of 10,522 cDNAs from Developing Seeds

Despite the fact that over 45,000 Arabidopsis ESTs have already been deposited in dbEST (release 030300; Boguski et al., 1993), these are not necessarily representative with regard to genes specifically expressed in developing seeds, because siliques, but not isolated developing seeds were used as source of seed cDNAs. To initiate a "functional genomic" analysis of seed metabolism, we sequenced cDNAs derived exclusively from developing Arabidopsis seeds in a single pass from the 5′ end. Because seeds contain highly abundant mRNAs, e.g. those derived from genes encoding storage proteins, we probed nylon filters with 9,136 (data set I) and 18,432 arrayed clones (data set II), respectively, employing cDNA probes as summarized in Table I. From data set I, 4,641 clones (51%) were sequenced and analyzed with BLASTX. Additional clones were selected from data set I (Table I) for probing of the second filter set to further reduce the redundancy in data set II. In this case, 5,922 clones (32%) were sequenced and analyzed. The average read lengths after trimming were 350 bp for clones from data set I and 259 bp for clones from data set II. Taken together, 10,522 clones were analyzed at the level of BLASTX searches equivalent to 38% of the clones on the filters. A total of 11,873 sequences were generated and kept in a FASTA file (complete raw data set), which includes 1,141 sequence runs from the 3′ ends of selected clones, a small number of repeats, and clones for which only poor sequence is available. The sequences have been deposited at GenBank and will be available along with annotations at our web site. The longest clones from each contig as well as singletons (see below) have been deposited at the Arabidopsis Biological Resource Center.

Classification of ESTs According to Predicted Function

To obtain qualitative information about the ESTs, each sequence was searched (BLASTX) against the non-redundant protein database of GenBank. The top scoring hits were automatically extracted and manually annotated according to the description of the sequence(s) returned by BLASTX. The number of clones falling into each class are shown in Table II. It must be emphasized that this procedure provides only tentative clues toward the function of the en-

Table I. *Clones corresponding to highly abundant messages used for prescreening*

Plasmid inserts of pools of up to five clones were used as probes for hybridization to denatured colonies arrayed on filters. Data set I refers to a filter set with 9,136 colonies and data set II to a filter set with 18,432 colonies.

GenBank Annotation	Data Set	Clone ID
12S Seed storage protein	I	M8B2STM, M8D10STM, M8I11STM (Cru3), M8H8STM (Cru1), M8G5STM (CRB)
2S Seed storage protein	I	M7A7STM, M8F5STM
12S Seed storage protein	II	M24E9STM, M15E9STM, M20G5STM, M18H9STM, M20D11STM, M16A1STM, M20C11STM, M21B1STM, M20E6STM, M9C1STM, M22G1STM, M13H5STM, M22A10STM
Vicilin precursor	II	M20C12STM, M31B2STM
Cruciferin BNC1 (11S globulin)	II	M19H3STM
Translation elongation factor eEF-1α	II	M16D2STM
ω-6 Fatty acid desaturase (endoplasmic reticulum [ER])	II	M20G10STM
S-Adenosylmethionine decarboxylase	II	M13F12STM
Chlorophyll *a/b*-binding protein	II	M19H6STM, M25G5STM, M22A8STM
Eukaryotic initiation factor 4A-1	II	M15A10STM
Tubulin α	II	M19E9STM
Enolase	II	M16E4STM
Peroxidase ATP4α	II	M16G3STM
RuBP carboxylase small subunit	II	M17T6STM
Oleosin type 4	II	M19C9STM
Orf gene product	II	M20H4STM
F1N21.11	II	M24E4STM

Table II. *Distribution of cDNAs in classes of putative function*

The class assignment presented in alphabetical order is based on the description of the best match from BLASTX similarity searches to the non-redundant GenBank protein databases. The number of EST clones and the percentage of the total from each category in the seed EST database are listed. (Note: cDNAs that were sequenced from both 5′ and 3′ ends were counted only once. Forty-one clones were not counted, which repeatedly returned incomplete BLAST results.

Code	Description	Count	Percentage
AA	Amino acid metabolism	278	2.6
C	Carbohydrate metabolism	701	6.7
CDC	Cell division cycle	31	0.3
CHP	Chaperonin, heat shock, folding	83	0.8
CSK	Cytoskeleton	143	1.4
CW	Cell wall	262	2.5
D	Defense, disease	65	0.6
DEV	Development	315	3.0
DNA	DNA-modifying enzymes	122	1.2
ER	Endoplasmic reticulum proteins	2	<0.1
HOR	Hormone biosynthetic enzymes	45	0.4
L	Lipid metabolism	490	4.7
LEA	Lea (late embryogenesis abundant)	2	<0.1
MT	Membrane, transporters, receptors	216	2.1
NM	Nitrogen metabolism	25	0.2
NSH	Non-significant homology	2,560	24.3
NUC	Nucleus	28	0.3
NUM	Nucleotide metabolism	50	0.5
OX	Oxygen-detoxifying enzymes, peroxidases	179	1.7
PA	Proteinases, ubiquitin	327	3.1
PS	Photosynthesis	159	1.5
RB	Ribosome, protein translation	330	3.1
RNA	Acting on RNA	162	1.5
RR	Ribosomal RNA	78	0.7
RS	Respiration	114	1.1
SEM	Secondary metabolism	152	1.4
SM	Sulfur metabolism	15	0.1
SP	Storage protein	1,518	14.4
STD	Signal transduction (kinases, calmodulin, etc.)	418	4.0
T	Transcription factors	169	1.6
TON	Tonoplast	38	0.4
TRF	Subcellular trafficking	21	0.2
UF	Unidentified function	1,424	13.5
	Total	10,522	100.0

coded proteins, due to the fact that relatively few of the descriptions associated with GenBank entries have been verified by wet-lab experiments (Boguski, 1999).

Two classes, “non-significant homology” (NSH) and “unidentified function” (UF) represent approximately 40% of the clones and warrant further explanation. Sequences that returned BLASTX scores (high scoring segment pairs) of less than 100 were grouped under NSH (24.3%), indicating that no protein similar to the translation product was present in the public databases at the time of the analysis. This group of sequences was repeatedly resubmitted for analysis. To rule out that the NSH class is enriched in low quality sequences as the primary cause for low BLASTX scores in this class, we compared the average quality values assigned by PHRED to each chromatogram and found similar average quality values for the NSH class and the total EST set. Based on this analysis one can assume that approximately 24% of the clones in the seed database encode novel proteins. The UF class (13.5%) contains ESTs that show significant similarity (BLASTX scores >100) at the level of predicted amino acid sequence to proteins from different organisms for which no function is known.

Despite the prescreening there is still a considerable number of storage protein entries (14.4%) present in the database (Table II) representing the largest class of clones for which a putative function can be assigned. A similar observation was made for ESTs from castor bean and was explained by the presence of short incomplete cDNAs encoding storage proteins that would not hybridize efficiently to the probe during prescreening (Van de Loo et al., 1995b). Considering the number of storage protein clones and other abundant clones identified by hybridization (62%), a minimum of 75% of mRNAs are derived from less than 50 genes in developing seeds. Three classes of particular importance to the analysis of carbon flow in developing oil seeds include 701 entries classified as carbohydrate metabolism, 490 lipid metabolism entries, and 216 entries for putative membrane transporters.

How Many Novel ESTs and How Many Genes Are Represented in the Seed EST Set?

To evaluate whether novel, seed-specific ESTs were present we compared our entire 5′-sequence data set against the Arabidopsis set in "The Arabidopsis Information Resource" available at http://www.Arabidopsis.org/seqtools.html. Of the 10,552 BLASTN results returned, 4,173 (39.5%) showed BLASTN scores (high scoring segment pairs) of less than or equal to 50. Based on these scores it can be estimated that approximately 40% of the ESTs described here are not represented in the public Arabidopsis EST set and many of these therefore may correspond to genes specifically expressed in developing seeds of Arabidopsis.

Because multiple ESTs can be derived from a single gene, sequences were assembled into contigs to estimate the number of genes giving rise to the ESTs. Of the 11,850 sequences used for contig analysis, 7,567 (64%) assembled into 1,569 contigs and 4,283 (36%) remained as singletons. Thus the maximal number of unique cDNAs represented in the entire data set is 5,852. To estimate how many genes are represented in our data set that may be specifically expressed in developing seeds, we determined the number of contigs and singletons represented by the 4,173 ESTs not represented in the public data set. These were 743 contigs and 2,306 singletons representing a maximal number of 3,049 genes. Thus based on this analysis up to one-half of all genes represented by our data set may be specifically expressed in seeds. However, there are three caveats concerning this estimation. First, although in most cases each contig represents one gene, sometimes more than one contig of non-overlapping sequences exist per gene resulting in an overestimation. Second, in some cases due to the limited quality of single-pass sequences, closely related gene families cannot be resolved into individual contigs resulting in an underestimation. Third, because silique-derived cDNA sequences are present in the public database, some of the ESTs in dbEST already represent genes specifically expressed in seeds, e.g. storage protein genes. These have not been taken into account above and will lead to an underestimation of seed-specific genes represented by the seed EST data set.

Mapping ESTs onto the Arabidopsis Genome

One step toward the determination of the exact number of genes represented by ESTs would be to map all ESTs and contig consensus sequences onto the Arabidopsis genome. For this purpose we searched (BLASTN) all sequences in the raw sequence file, as well as all contig consensus sequences against an Arabidopsis genomic sequence subset of all sequences longer than 10 kb. This set should primarily contain sequenced bacteria artificial chromosomes (BACs), phage artificial chromaosomes (PACs), and P1 clones from the Arabidopsis Genome Initiative. The individual results of this analysis can be found in the database and provide a location for most ESTs on the physical map of Arabidopsis by linking these results to the map locations of sequenced clones available at http://www.Arabidopsis.org/seqtools.html. In the past this information could only be obtained by direct PCR mapping approaches (Agyare et al., 1997) due to the absence of large scale genomic sequence information. Because BACs contain on the average 20 to 30 genes each, further analysis on an individual basis is required to ultimately determine whether two contigs are derived from one or several genes on a particular BAC.

Abundance of ESTs Derived from Specific Genes

The number of sequences assembled in the contigs gives an indication of the degree of expression of the respective gene in developing seeds. Table III lists contigs containing more than eight ESTs. The accession numbers provide direct access to the sequence in GenBank (whenever possible, a cDNA sequence) that shows the best match to the contig consensus sequence. As predicted by the initial classification of individual ESTs (Table II), the most abundant ESTs form contigs that encode seed storage proteins. In agreement with the high demands for protein synthesis in developing seeds, ESTs for translational elongation factors were abundant in contigs (Table III, RB). ESTs for proteins possibly involved in storage protein body formation such as vacuolar processing enzyme (Kinoshita et al., 1995; Table III, TON) or proteases in general (Table III, PA) are highly abundant. In a similar manner, genes encoding enzymes involved in protein folding (Table III, CHP) such as protein disulfide isomerase genes are highly expressed in seeds (Boston et al., 1996). Developing embryos of Arabidopsis are green. Thus it is not surprising that ESTs encoding chlorophyll-binding proteins are present in high numbers (Table III, PS). The most highly abundant enzyme-encoding ESTs are those for *S*-adenosyl-Met decarboxylase (Table III, AA). This is a key enzyme of polyamine biosynthesis (Walden et al., 1997). However, ESTs encoding other enzymes of this pathway are not very abundant or are absent. Thus *S*-adenosyl-Met decarboxylase may be involved in addition in a pathway unrelated to polyamine biosynthesis. Among the contigs of abundant ESTs are 20 for which the consensus sequence did not have a match in GenBank or which are similar to proteins of unknown function (Table III, NSH and UF). These provide an interesting pool

Table III. *Most abundant contigs in the Seed EST database*

The closest cDNA match in GenBank is provided (accession no.). When no cDNA entry is available, BAC sequences marked with an asterisk are provided.

Name	Code	Count	Accession No.	Description
1967[a]	AA	87	Y07765	*S*-Adenosylmethionine decarboxylase
1949	AA	25	U97200	Cobalamin-independent methionine synthase
1931	AA	17	M55077	*S*-Adenosylmethionine synthetase 1
1911	AA	13	*Z97335	Hydroxymethyltransferase
1904	AA	12	D83025	Pro oxidase precursor
1902	AA	12	*Z97335	Adenosylhomocysteinase
1953	C	27	M64736	Pyruvate kinase, chloroplast isozyme
1948	C	25	*Z99708	β-Galactosidase-like protein
1943[a]	C	22	X58107	Enolase
1939	C	19	AC006200	Putative aldolase
1936	C	18	M83534	Isocitrate lyase
1933	C	17	Z28374	Pyruvate kinase, chloroplast isozyme
1979[a]	C	16	X13611	Ribulose bisphosphate carboxylase small subunit
1926	C	16	AL031986	Cytoplasmatic aconitase
1923	C	16	U80185	Pyruvate dehydrogenase E1 α-subunit
1908	C	13	*AC002292	ATP-citrate lyase
1900	C	12	S72926	Glc and ribitol dehydrogenase
1897	C	11	L12042	Aldehyde reductase (EC 1.1.1.21), NADPH-dependent
1888	C	11	Y10380	Plastidic aldolase
1982	C	9	U43283	Insect-type alcohol dehydrogenase
1859	C	9	*AC006919	Pyruvate kinase, cytosolic
1942	CHP	21	M82973	Protein disulfide-isomerase
1958[a]	CSK	39	M84696	Tubulin α-2/α-4 chain
1927	CSK	16	M84700	Tubulin β-2/β-3 chain
1913	CSK	14	U37281	Actin 2/7
1950	CW	26	U12757	Diphenol oxidase
1903	CW	12	L34685	Cell wall protein
1873	CW	10	L41869	β-Glucosidase, barley
1857	CW	9	X61280	Hydroxyproline-rich glycoprotein, rice
2000	DEV	59	M62991	Dessication-related protein
1938	DEV	19	AF067857	Embryo-specific protein 1
1893	DEV	11	U75192	Germin-like protein
1889	DEV	11	AF067858	Embryo-specific protein 3
1881	DEV	10	D29803	Prepro MP27-MP32
1891	HOR	11	X83381	Gibberellin 20-oxidase
1961[a]	L	49	X91918	Oleosin type 4
1977	L	38	X91956	Oleosin, 21-kDa isoform
1946	L	24	M64116	Glyceraldehyde 3-phosphate dehydrogenase, cytosolic
1945[a]	L	22	L26296	ω-6 Fatty acid desaturase, ER (Δ-12 desaturase)
1934	L	17	U29142	Fatty acid elongase 1
2001	L	18	AC002396	β-Oxoacyl-(acyl carrier protein) reductase
1929	L	16	L40954	Oleosin, 21-kDa isoform
1914	L	14	*AC002334	3-Ketoacyl-CoA thiolase
1907	L	13	X68150	Ketol acid reductoisomerase
1874	L	10	*AC002333	Stearoyl-ACP desaturase
1868	L	9	S63400	Corticosteroid 11-β-dehydrogenase, isozyme 1
1865	L	9	X62353	Oleosin
1875	MT	10	X65549	ADP/ATP carrier protein 1
1956	NSH	33	Y08726	MtN3
1947	NSH	24	Z29649	Periaxin
1922	NSH	15	*Z00044	Photosystem II 10-kDa phosphoprotein
1918	NSH	15	M60590	α-Agglutinin attachment subunit precursor
1910	NSH	13	AF057357	Lipid transfer protein 2
1909	NSH	13	U83865	NADH dehydrogenase subunit 4
1896	NSH	11	*AC004261	Putative bzip protein
1892	NSH	11	AF015608	SR protein
1885	NSH	10	X97197	Spliceosomal protein

Table III continues on next page.

Table III. *Continued from previous page.*

Name	Code	Count	Accession No.	Description
1884	NSH	10	M86720	Ribulose bisphosphate carboxylase/oxygenase activase
1879	NSH	10	M73714	Fatty aldehyde dehydrogenase, microsomal
1925	OX	16	AJ004810	Cytochrome P450 monooxygenase
1920	OX	15	*AC004683	Peroxidase
1895	OX	11	X98313	Peroxidase
1877	OX	10	AF021937	Catalase 1
1957	PA	36	AC000375	Aspartic protease
1955	PA	31	AF065639	Cucumisin-like serine protease
1894	PA	11	*AC002131	Aspartic proteinase
1870	PA	9	D13042	Cys proteinase
1855	PA	9	*AC004786	Putative Ser carboxypeptidase I
1952[a]	PS	27	X03909	Chlorophyll *a/b*-binding protein of LHCII type I
1928	PS	16	X03907	Chlorophyll *a/b*-binding protein of LHCII type I
1906[a]	PS	13	X64460	Chlorophyll *a/b*-binding protein type I precursor Lhb1B2
1890	PS	11	X98108	23-kDa Polypeptide of oxygen-evolving complex
1866	PS	9	AF139470	Chlorophyll *a/b*-binding protein CP24
1858[a]	PS	9	X03909	Chlorophyll *a/b*-binding protein of LHCII Type I
1963	RB	60	AJ223969	Elongation factor 1 α-subunit
1951[a]	RB	26	X16430	Elongation factor 1-α
1915[a]	RB	14	X16430	Elongation factor 1-α
1899	RB	12	*AC000132	Elongation factor 1-γ
1887	RB	10	Z97178	Elongation factor 2
1878[a]	RB	10	X65052	Eukaryotic initiation factor 4α-1
1869	RB	9	*AC002339	40S Ribosomal protein S2
1861	RNA	9	*AC006260	Putative RNA-binding protein
1966	RR	59	gi|1785673	26S RNA
1941	RS	20	U23082	ATP synthase β-chain
1880	SEM	10	U80668	Homogentisate 1,2-dioxygenase
1972[a]	SP	727	U66916	12S Cruciferin seed storage protein
1971[a]	SP	409	M37247	12S Seed storage protein precursor (CRA1)
1970	SP	211	M22033	2S Seed storage protein 3 precursor (2S albumin)
1969[a]	SP	92	M37248	12S Seed storage protein precursor (CRB)
1968[a]	SP	88	Y00722	Vicilin precursor (α-globulin)
1965[a]	SP	71	M22033	2S seed storage protein 2 precursor (2S albumin)
1962[a]	SP	50	Z99708	Globulin-like protein
1959[a]	SP	41	AC003027	12S Seed storage protein
1954[a]	SP	28	M22033	2S Seed storage protein 1 (2S albumin)
1924[a]	SP	16	M22033	2S Seed storage protein 4
1871[a]	SP	9	X65039	2S Storage protein (black mustard)
1921	STD	15	U77381	Guanine nucleotide-binding protein β-subunit-like
1886	STD	10	D83531	GDP dissociation inhibitor
1867	STD	9	AB015138	Vacuolar proton pyrophosphatase
1898	TON	12	AF026275	β-Tonoplast intrinsic protein
1872	TON	9	D61394	Vacuolar processing enzyme
1940[a]	UF	19	AC002130	F1N21.11
1937[a]	UF	18	X91954	Putative protein
1935	UF	18	X91953	F19K23.1, F19K23.15
1919	UF	15	X03496	Chloroplast 30S ribosomal protein S11
1917	UF	15	P43349	Translationally controlled tumor protein homolog
1912	UF	13	*AC005698	T3P18.6 (putative Pro-rich cell wall protein)
1901	UF	12	*AC004557	F17L21.27
1905	UF	13	AF035385	Putative protein
1882	UF	10	*AL021636	Putative protein
1864	UF	9	*AL021811	Putative protein
1863	UF	9	AF013628	Reversibly glycosylated polypeptide-2
1860	UF	9	*AC000375	F19K23.3,F19K23.15. ESTs

[a] The number of clones in these contigs is biased due to prescreening with cDNAs given in Table I.

of novel proteins with a function that may be of special relevance for developing seeds and further functional analysis may lead to the discovery of molecular processes crucial to developing seeds. An obvious class missing in the contig list of most abundant ESTs (Table III) is that containing ESTs with similarity to transcription factor genes, even though the entire data set contains a considerable number of such ESTs (169, 1.6%; Table II, T). It is clear that regulatory genes are not as highly expressed as storage protein genes or genes essential for the biosynthesis of other storage compounds. Although this notion may be trivial, it nevertheless confirms that the observed abundance of ESTs in each contig or class is in agreement with common knowledge about the biology of plant cells and of developing seeds in particular.

Different Representation of Genes in the Seed EST Set and the Public Arabidopsis EST Set

The public EST data set for Arabidopsis available March 2000 consists of over 45,000 sequences derived from cDNA libraries produced from a range of tissues. The largest group of sequences (approximately 31,000) originated from sequencing a mixed population of cDNAs from etiolated seedlings, tissue culture-grown roots, and aerial tissue from flowering plants (Newman et al., 1994). The 10,522 sequences from a developing seed cDNA library described in this study represent the largest set of public Arabidopsis ESTs currently available from a narrowly defined developmental stage of the plant. How different is this new set from those sequences already deposited? To answer this question we compared the percentage of ESTs in the seed database for several genes with their abundance among the non-seed Arabidopsis ESTs previously deposited in dbEST. For example, for glyceraldehyde-3-P dehydrogenase, a gene that might be considered constitutive, or "housekeeping," the relative abundance in the two data sets is identical (0.3%). In contrast and as expected, genes that are known to be highly expressed in seeds were found to be abundant in the seed EST data set. For example, storage proteins represent at least 50% of the clones in the seed library, which is at least 500-fold more abundant than in the non-seed set. Likewise, oleosins are approximately 100-fold more prevalent in the seed library than in the non-seed data.

In mature Arabidopsis seeds, lipid in the form of triacylglycerol is the major form of carbon storage, representing 30% to 40% of the seed dry weight. It might be expected that higher flux of carbon into lipid synthesis in seeds would be reflected in a higher proportion of clones for fatty acid synthesis within the seed data set than in dbEST. This is in fact the case: approximately 0.5% of the seed ESTs encode proteins of the plastidic fatty acid synthase compared with approximately 0.15% of Arabidopsis ESTs found in dbEST for the same reactions. Furthermore, we detected ESTs for seed-specific genes that are completely missing from the public data set. For example, clones corresponding to *FAE1* encoding a protein that controls seed-specific fatty acid elongation occurred 20 times in our database, but not at all in dbEST. In general, the vast majority of these comparisons validate that this new EST set provides the expected tissue-specific representation of gene expression in seeds and contains a very different population of ESTs than previously available.

The Conversion of Photosynthate into Fatty Acids

Figure 1 depicts the major pathways involved in the conversion of Suc into fatty acids. These include the conversion of imported Suc by a cytosolic glycolytic pathway (reactions 1–16), transfer of intermediates across the plastid envelopes (reactions 17–20), intermittent starch biosynthesis and degradation in the plastid (reactions 21–26), a plastidic glycolytic pathway (reaction 27–36), the oxidative pentose phosphate cycle (reactions 37–42), the plastidic pyruvate dehydrogenase complex (reaction 44), as well as reactions involved in fatty acid biosynthesis and modification (reactions 45–52). In Figure 1 the thickness of arrows represents the number of ESTs in data sets I and II, which encode the respective enzyme. Because different enzymes have different turnover numbers and other kinetic factors, this number cannot be used to compare the magnitude of flux through the different reactions. However, EST numbers in many cases can provide useful comparisons between the same reaction in different compartments, or between similar biochemical reactions. The assignment of the plastidic and cytosolic isoforms was generally based on BLASTX results showing sequence similarity of the respective ESTs or contigs to genes encoding proteins of known function and subcellular location. In ambiguous cases, e.g. for Glc-6-P dehydrogenase (Fig. 1, reaction 37) we used multiple alignment of the respective ESTs from the seed database with all known Glc-6-P dehydrogenase-encoding plant genes in conjunction with cluster analysis. Further refinement could be achieved by predicting the presence of chloroplast transit peptides from genomic DNA sequences that correspond to the ESTs. However, in the absence of biochemical data these assignments must be considered preliminary. A list of each enzyme, the number of ESTs, and the clone and contig identifiers are given in Table IV. Reactions for which no corresponding EST is present are drawn with a dashed line in Figure 1.

It is interesting that those reactions are often found in clusters, e.g. reactions 25 through 29 (plastidic glycolysis) or 38 through 41 (oxidative pentosephosphate cycle). It is tempting to speculate that the observed clustering reflects the coordinated regulation of gene expression according to metabolic pathways and may provide a first glimpse at the regulatory network governing seed metabolism. However, it must be emphasized that even though this new data set is large, it still is incomplete and the resolution for

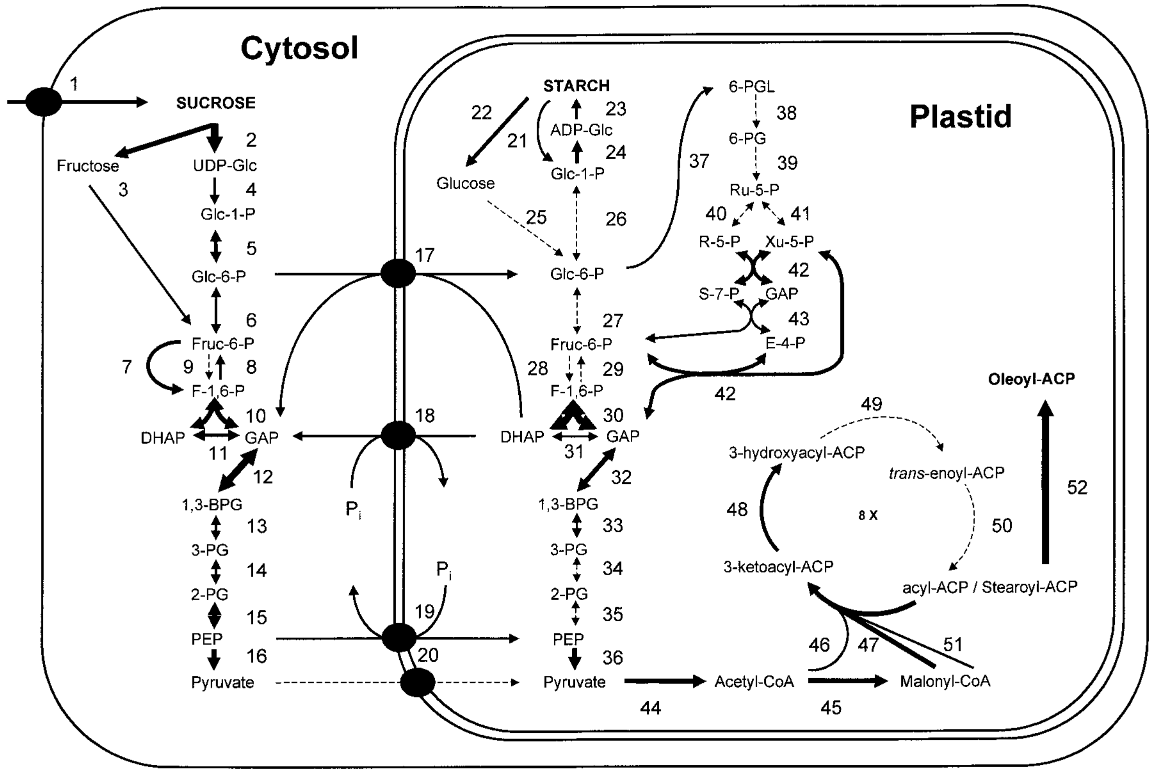

Figure 1. Schematic representation of metabolic pathways in a typical oil storing cell of a developing Arabidopsis embryo. The selective focus presented here is on carbohydrate metabolism and fatty acid biosynthesis. Only cytosolic and plastidic isoforms are considered. Double-headed arrows indicate readily reversible reactions, single headed arrows indicate typically irreversible reactions. Cosubstrates such as water or nucleotides have been omitted. Abbreviations are conventional, but can also be deduced from the enzyme descriptions given in Table IV. Numbers correspond to individual reactions and serve to identify the respective enzyme in Table IV. The thickness of arrows provides a coarse indication of the number of ESTs present in the seed EST data set for the respective reaction. The exact numbers for each reaction can be found in Table IV.

differential expression is lost for reactions that are not represented by ESTs.

Membrane Transporters

Suc is the transport form of CO_2 fixed by photosynthesis and must be imported into the developing embryo. Studies with developing bean seeds suggest that Suc and hexose transporters located in the epidermis of the embryo are involved (Weber et al., 1997). Two Suc transporter genes are known for Arabidopsis, *SUC1* and *SUC2* (Sauer and Stolz, 1994) and corresponding ESTs are present in the seed database (Table IV; Fig. 1, reaction 1). Most ESTs correspond to *SUC2*, but there is also a contig of ESTs that are more similar to the Suc transporter from bean (Tab IV). Whether this class of ESTs represents a third Suc transporter gene from Arabidopsis specific for developing seeds needs to be further investigated. Furthermore, several ESTs with similarity to hexose transporters are present, which may be involved in the import of hexoses derived from Suc cleavage by apoplastic invertase.

Hexose metabolites enter the plastid to provide precursors for starch and fatty acid biosynthesis. Using isolated plastids of developing embryos of oilseed rape, it has been shown that labeled Glc-6-P and pyruvate are the most efficient of all the different possible substrates tested in labeling starch and triacylglycerols, respectively (Kang and Rawsthorne, 1994). Furthermore, fatty acid biosynthesis was stimulated if Glc-6-P and pyruvate were present (Kang and Rawsthorne, 1996). ESTs with similarity to a plastid Glc-6-P/phosphate (or triosephosphate) antiporter (Kammerer et al., 1998) are abundant in the seed EST database (Table IV; Fig. 1, reaction 17). However, we were unable to identify a set of ESTs with similarities to any known pyruvate or monocarboxylic acid transporter (Table IV; Fig. 1, reaction 20). Either pyruvate does not require a specific translocator, the respective protein cannot be identified without further biochemical or molecular information, or pyruvate is not the metabolite imported into plastids in vivo. It has been previously suggested that a plastid phosphoenolpyruvate/phosphate antiporter may be providing the plastid with pyruvate following metabolism of the imported phosphoenolpyruvate (Fischer et al., 1997). There are several ESTs present encoding proteins with similarity to a phosphoenolpyruvate translocator (Table IV; Fig. 1, reaction 19). A

Table IV. *Enzymes involved in carbohydrate and lipid metabolism*

The enzymes are organized according to the reaction scheme shown in Figure 1. For contigs the contig number is provided, for singletons the seed database accession no. Hits refers to the total number of 5′-sequences in the seed ESTs data base.

Reaction	Enzyme	Hits	Contig No./Accession No.
1	Suc transporter, plasma membrane	10	1483, 1470, M30D4STM, M64I03STM, M8I14STM, M59G8STM
2	Suc synthase, cytosolic	12	1820, 1365, M76O20STM, M65H04STM, M65F09STM
3	Fructokinase, cytosolic	2	M62E16STM, M72N06STM
4	UDP-glucose pyrophosphorylase, cytosolic	5	889, M68J13STM, M76D11STM, M64P15STM
5	Phosphoglucomutase, cytosolic	7	1538, 1493, M75J18STM
6	Phosphoglucose isomerase, cytosolic	1	M28A5STM
7	Phosphofructokinase (PPi), cytosolic	9	1373, 492, 1420, 1159, M20H10STM, M69E16STM, M77B21STM
8	Fructosebisphosphate phosphatase, cytosolic	1	M28F1STM
9	Phosphofructokinase (ATP),	0	
10	Aldolase, cytosolic	16	1848, 1432, 624, 593, 571, M20F4STM
11	Triosephosphate isomerase, cytosolic	4	1707
12	Glyceraldehyde-3-P DH, cytosolic	26	1946, 1714, M65J17STM
13	Phosphoglyucerate kinase, cytosolic	1	M64A02STM
14	Phosphoglycerate mutase, cytosolic	2	M67H19STM, M75J05STM
15	Enolase, cytosolic	22	1943
16	Pyruvate kinase, cytosolic	11	1859, M75L17STM, M74L17STM, M78D01STM, M68E05STM, M73I18STM, M65E10STM
17	Glc 6-P translocator, plastidic	5	1574, 1043, M76C19STM
18	Triosephosphate translocator, plastidic	1	M18E5STM
19	Phosphoenolpyruvate translocator, plastidic	5	1425, 496
20	Pyruvate translocator, plastidic	0	
21	Starch phosphorylase, plastidic	3	1340, M19B10STM
22	Starch-branching enzyme, plastidic	2	M66D16STM, M15E5STM
23	Starch synthase, plastidic	3	M74E22STM, M75B02STM, M48A10STM
24	ADP-Glc pyrophosphorylase, plastidic	8	1834, M27F2STM
25	Hexokinase, plastidic	0	
26	Phosphoglucose mutase, plastidic	0	
27	Phosphoglucose isomerase, plastidic	0	
28	Phosphofructokinase (ATP), plastidic	0	
29	Fructosebisphosphate phosphatase, plastidic	0	
30	Aldolase, plastidic	34	1939, 1888, 694, M65E15STM, M23A5STM, M72H02STM, M47E9STM
31	Triosephosphate isomerase, plastidic	4	1753, M11D3STM
32	Glyceraldehyde-3-P DH, plastidic	10	1813, M11F9STM, M11F10STM, M69C01STM, M65E08STM
33	Phosphoglycerate kinase, plastidic	4	1463, M26H6STM, M65O05STM
34	Phosphoglycerate mutase, plastidic	0	
35	Enolase, plastidic	0	
36	Pyruvate kinase, plastidic	27	1933, 1953, M12D1STM, M14F8STM, M26H5STM, M31A4STM, M36D1STM, M59H11STM, M63F01STM, M70O13STM, M67C02STM, M77O23STM
37	Glc-6-P dehydrogenase, plastidic	1	M53H3STM
38	Lactonase, plastidic	0	
39	Gluconate-6-P dehydrogenase, plastidic	0	
40	Ribose phosphate isomerase, plastidic	0	
41	Ribulose phosphate epimerase, plastidic	0	
42	Transketolase, plastidic	6	1806
43	Transaldolase, plastidic	5	1767, M56C9TM
44	Pyruvate dehydrogenase E1α, plastidic	13	1923, M78J02STM
	Pyruvate dehydrogenase E1β, plastidic	8	1822, M63B12STM
	Pyruvate dehydrogenase E2, plastidic	6	763, M38G4STM, M76A12STM, M66K17STM, M69I19STM
45	Acetyl-CoA carboxylase, BCCP, plastidic	7	1853
	Acetyl-CoA carboxylase, BC, plastidic	2	M72D24STM, M66H20STM
	Acetyl-CoA carboxylase, α-CT, plastidic	0	
	Acetyl-CoA carboxylase, β-CT, plastidic	0	
46	Ketoacyl-ACP synthase III, plastidic	1	1323
47	Ketoacyl-ACP synthase I, plastidic	13	1774, 1854
48	Ketoacyl-ACP-reductase, plastidic	16	2001, 1849, M40E5STM, M61E6STM, M68L14STM, M72M16STM, M76K16STM, M72D13STM
49	3-Hydroxyacyl-ACP dehydrase, plastidic	0	
50	Enoyl-ACP reductase, plastidic	0	
51	Ketoacyl-ACP synthase II, plastidic	1	M66J06STM
52	Stearoyl-ACP desaturase, plastidic	17	1874, 1699, 1157, M41G8STM

high expression of this antiporter in non-green plant tissues has also been observed using conventional methods (Kammerer et al., 1998). In the same study it was also shown that the gene for the triosephosphate/phosphate translocator is much more highly expressed in green tissues as compared with non-green tissues. Thus, the presence of only one EST for the respective gene in the seed database (Table IV; Fig. 1, reaction 18) is in agreement with the conventional northern analysis.

Glycolysis, Oxidative Pentose Phosphate Cycle, and Starch Metabolism

In general, plants do have a complete glycolytic pathway in the cytosol (Plaxton, 1996) and it has been shown that a complete pathway also exists in the plastids of oil seeds (Dennis and Miernyk, 1982; Kang and Rawsthorne, 1994). The question remains to what extent both pathways are utilized in the conversion of carbohydrates into precursors of fatty acid biosynthesis. All genes encoding glycolytic enzymes of the cytosol are expressed, whereas ESTs encoding plastidic isoforms are absent in many cases (Fig. 1; Table IV). Exceptions are the central reactions 30 through 33 of the plastidic glycolytic pathway, as well as the plastidic isoform of pyruvate kinase (reaction 36). It seems certain that there is differential transcriptional regulation of the two pathways. Assuming that there is no general difference between the specific activities of the cytosolic and plastid enzymes, the data would be consistent with a more active cytosolic pathway. The peculiar high expression of plastidic pyruvate kinase genes (reaction 33) in conjunction with the relatively high abundance of phosphoenolpyruvate transporter ESTs (reaction 19) is consistent with a major route of carbon from Suc into precursors of fatty acid biosynthesis involving the cytosolic glycolytic pathway up to phosphoenolpyruvate, import of this compound into the plastid, and subsequent conversion to pyruvate. It is interesting that ESTs for plastid isoforms of pyruvate dehydrogenase (27 ESTs) are approximately 2-fold more abundant than for mitochondrial isoforms (13 ESTs). This contrasts with the non-seed Arabidopsis EST set in dbEST where ESTs are approximately equal for the two subcellular localizations. These comparisons are clearly consistent with our expectations of the relative flux through fatty acid synthesis and the tricarboxylic acid cycle in seed and non-seed tissues.

Biosynthesis of fatty acids does not only require carbon units, but more than twice as many moles of reduced nicotinamide nucleotides per fatty acid (Ohlrogge et al., 1993). Reductants for fatty acid biosynthesis can be generated in the heterotrophic plastid by the pyruvate dehydrogenase reaction (reaction 44), by the initial reactions (reactions 37 and 39) of the oxidative pentose phosphate cycle, and in green seeds by photosystem I. Although the different subunits of the plastidic pyruvate dehydrogenase complex are highly expressed (Table IV, reaction 44), only one out of seven Glc-6-P dehydrogenase- (reaction 37) encoding ESTs could be clearly identified as plastidic. No ESTs were found for reactions 38 through 41 of the plastidic oxidative pentose phosphate cycle, but ESTs encoding enzymes involved in recycling the carbon moieties were plentiful (reactions 42 and 43). It is known that plastidic Glc-6-P dehydrogenase is allosterically regulated in sophisticated ways in photosynthetic tissues (Wenderoth et al., 1997). Thus it seems possible that this tight regulation of the oxidative pentose phosphate pathway begins already at the level of transcription and is visible in the low abundance of the respective ESTs. Plastids of developing Arabidopsis seeds are transiently green and some of the most abundant ESTs encode proteins of the photosynthetic membrane (Table III), supporting the conclusion (Browse and Slack, 1985; Eastmond et al., 1996; Asokanthan et al., 1997; Bao et al., 1998) that some of the reducing equivalents required for fatty acid biosynthesis are derived from photosynthesis.

Developing seeds of Arabidopsis transiently accumulate starch (Focks and Benning, 1998). In accordance with this, ESTs encoding enzymes involved in starch biosynthesis and degradation are quite abundant (Fig. 1; Table IV, reactions 21–24), similar to those encoding enzymes that catalyze the initial reactions of fatty acid biosynthesis (reactions 45–48). The ESTs of starch metabolism represent an example of the apparent coordinate expression of genes encoding enzymes of the same metabolic pathway and may reveal a regulon.

Fatty Acid Biosynthesis

Given that the major carbon storage in developing oil seeds is associated with triacylglycerol, but not starch, one would expect that ESTs encoding enzymes directly involved in fatty acid biosynthesis are at least as abundant as those encoding starch metabolic enzymes. This seems to be true for the ketoacyl-acyl carrier protein synthases (reactions 46, 47, and 51) as well as for acetyl-coenzyme A (CoA) carboxylase (reaction 45), which provides the malonyl-CoA substrate for fatty acid biosynthesis. In general the relative abundance of the cDNAs encoding different enzymes of fatty acid synthesis is similar in the seed and non-seed EST sets, suggesting that seeds do not alter to a substantial degree the relative expression of genes encoding pathway components to accomplish the increased flux through the pathway in seeds. Rather, the entire pathway is apparently up-regulated, as suggested by the overall higher relative abundance of ESTs noted for fatty acid synthesis ESTs in the seed compared with the non-seed sets (Mekhedov et al., 2000). These data, therefore, confirm tissue mRNA expression data from several studies of genes encoding individual enzymes of fatty acid synthesis (e.g. Fawcett et al., 1994), but furthermore suggest that at least nine genes encoding enzymes or subunits involved in this pathway are coordinately regulated. The broader scale in silico expression analysis presented here has thus uncovered phenomena that were not apparent from the previous studies focusing on single genes.

CONCLUSIONS

We have provided a large data set of ESTs from developing Arabidopsis seeds and have begun to analyze this rich resource. The analysis of this data set is not complete and some of the conclusions may have to be revised as better bioinformatics tools become available. However, based on our preliminary analysis it is clear that this data set is substantially different from the currently available public Arabidopsis EST data set. With few exceptions, there is considerable congruence between conventional biochemical wisdom regarding seed metabolism and the number of ESTs encoding seed metabolic enzymes. Even by examining only 52 reactions (Fig. 1), patterns of expression became obvious. These observed patterns may reflect the existence of metabolic regulons, groups of genes that are coordinately expressed. In many cases the current EST data set provides the first experimental access to these genes and the basis for their in-depth molecular analysis and for the biochemical studies of the encoded proteins.

MATERIALS AND METHODS

Library Preparation and Screening

To construct the Arabidopsis developing seed cDNA library, immature seeds of Arabidopsis ecotype Columbia-2 were collected 5 to 13 d after flowering. RNA was extracted according to Hall et al. (1978) from 1 g of seed tissue and a directional Uni-ZAP XR cDNA library was commercially prepared from poly(A)$^+$ mRNA (Stratagene, La Jolla, CA). The initial titer of the amplified library was 1.9×10^{10} plaque-forming units/mL. Based on 48 randomly selected clones, the average insert size was estimated to be 1.9 kb. Following the excision of phagemids, bacterial colonies were arrayed onto nylon membranes at a density of 36 clones cm^{-2} by Genome Systems (St. Louis). Data were generated in two stages corresponding to a membrane set with 9,136 cDNA clones and a second set containing 18,432 clones. The first set of membranes was hybridized with 12S and 2S seed storage protein cDNA clones. Non-hybridizing clones were selected for sequencing. The second set of membranes was hybridized with six pools of five different probes derived from cDNAs (Table I) that were highly abundant among the EST sequences from the first set. Non-hybridizing clones were sequenced following re-racking.

Sequence Analysis

The first set of cDNAs (data set I) was sequenced at Michigan State University from the 5′ ends using the SK primer for pBluescript II, or from the 3′ ends using the M13 21 primer. The second set of cDNAs (data set II) was sequenced by Incyte Pharmaceuticals (Palo Alto, CA) from the 5′ ends using the Bluescript T3 primer. Chromatograms from the data set I were processed in batches using Sequencher v.3.0 (Gene Codes, Ann Arbor, MI). The 5′- and 3′-ambiguous sequences were trimmed. Vector sequences were removed as part of this process. Sequences that were less than 150 bp long or had >4% ambiguity were not processed. Chromatograms from data set II were processed in bulk using PHRED (Phil Green and Brent Ewing, University of Washington, Seattle). Sequences that were less than 225 bp or >4% ambiguous were not further processed. At this time 95% of the sequences have been deposited at GenBank. The remaining 5% (exclusively derived from data set II) will be available in GenBank by March 2001.

Database Searches

For data set I, sequences were processed with the Genetics Computer Group programs (Wisconsin Package Version 9.1, Madison, WI), and used for similarity searches against GenBank by using shell or PERL scripts that call Genetics Computer Group NETBLAST (BLASTX version 1.4.11; Altschul et al., 1990) for each sequence. Searches were done in batches. For data set II, the FASTA file produced by PHRED/PHD2FASTA was processed by PERL scripts to do BLASTX searches with default parameters. The BLASTX searches were done over a period of 12 months from September 2, 1998 to September 21, 1999 using the most recent releases of GenBank. A subset was periodically retested (see below). The output from BLASTX was processed with PERL scripts to extract the top scoring hit from each result file. The following information for the top scoring entry in each result file was retained: gene identifier, description, BLAST score, probability, percent identity, alignment length, and reading frame. These results were compiled in text files. Each result was manually interpreted and categorized according to predicted biochemical function. BLASTN searches were done against a subset of dbBEST (available at http://www.Arabidopsis.org/seqtools.html) containing only Arabidopsis sequences using a FASTA file with all raw sequences. Stand-alone BLASTN version 2.0.9 running under Linux 5.2 was used for this analysis.

Contig Analysis

Contig analysis was performed with PHRAP (Phil Green, University of Washington, Seattle). Chromatograms from both data sets were processed with PHRED/PHD2FASTA, CROSS_MATCH (to mask vector sequence), and PHRAP. The first 30 bp from each sequence were trimmed during assembly by PHRAP. The .ace output file from PHRAP was processed with a PERL script to obtain the list of ESTs in each contig. Contigs were manually screened and corrected in cases where obviously unrelated sequences were clustered together.

Database

All data were imported into a Microsoft Access 97 relational database. The database was built around unique clone identifiers that refer to clone locations in microtiter plates. In some cases entries for 3′ sequences are available. These can be recognized by the last letter X added to the clone identifier. In a few cases the same clone has been sequenced twice. This has been marked by adding the last letters A and B to the clone identifier. The database and the PERL scripts are available for viewing at our web page at http://benningnt.bch.msu.edu/index.htm.

ACKNOWLEDGMENTS

We thank Sergei Mekhedov for advice and Jay Thelen for analysis of pyruvate dehydrogenase sequences. We would also like to thank Chris Eakin and Chris Beasley for their help with the annotation of data and construction of the web site.

Received April 7, 2000; modified May 24, 2000; accepted July 27, 2000.

LITERATURE CITED

Agyare FD, Lashkari DA, Lagos A, Namath AF, Lagos G, Davis RW, Lemieux B (1997) Mapping expressed sequence tag sites on yeast artificial chromosome clones of *Arabidopsis thaliana* DNA. Genome Res **7:** 1–9

Allona I, Quinn M, Shoop E, Swope K, Cyr SS, Carlis J, Riedl J, Retzel E, Campbell MM, Sederoff R, Whetten RW (1998) Analysis of xylem formation in pine by cDNA sequencing. Proc Natl Acad Sci USA **95:** 9693–9698

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

Asokanthan PS, Johnson RW, Griffith M, Krol M (1997) The photosynthetic potential of canola embryos. Physiol Plant **101:** 353–360

Audic S, Claverie JM (1997) The significance of digital gene expression profiles. Genome Res **7:** 986–995

Bao XM, Pollard M, Ohlrogge J (1998) The biosynthesis of erucic acid in developing embryos of *Brassica rapa*. Plant Physiol **118:** 183–190

Boguski MS (1999) Biosequence exegesis. Science **286:** 453–455

Boguski MS, Lowe TM, Tolstoshev CM (1993) dbEST: database for "expressed sequence tags." Nat Genet **4:** 332–333

Boston RS, Viitanen PV, Vierling E (1996) Molecular chaperones and protein folding in plants. Plant Mol Biol **32:** 191–222

Bouchez D, Hofte H (1998) Functional genomics in plants. Plant Physiol **118:** 725–732

Browse J, Slack CR (1985) Fatty acid synthesis in plastids from maturing safflower and linseed cotyledons. Planta **166:** 74–80

Cahoon EB, Carlson TJ, Ripp KG, Schweiger BJ, Cook GA, Hall SE, Kinney AJ (1999) Biosynthetic origin of conjugated double bonds: production of fatty acid components of high-value drying oils in transgenic soybean embryos. Proc Natl Acad Sci USA **96:** 12935–12940

Cooke R, Raynal M, Laudie M, Grellet F, Delseny M, Morris PC, Guerrier D, Giraudat J, Quigley F, Clabault G, Li YF, Mache R, Krivitzky M, Gy IJ, Kreis M, Lecharny A, Parmentier Y, Marbach J, Fleck J, Clement B, Philipps G, Herve C, Bardet C, Tremousaygue D, Hofte H (1996) Further progress towards a catalogue of all Arabidopsis genes: analysis of a set of 5,000 non-redundant ESTs. Plant J **9:** 101–124

Dennis D, Miernyk J (1982) Compartmentation of non-photosynthetic carbohydrate metabolism. Annu Rev Plant Physiol **33:** 27–50

DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science **278:** 680–686

Eastmond P, Kolacna L, Rawthorne S (1996) Photosynthesis by developing embryos of oil seed rape (*Brassica napus* L.). J Exp Bot **47:** 1763–1769

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

Fawcett T, Simon WJ, Swinhoe R, Shanklin J, Nishida I, Christie WW, Slabas AR (1994) Expression of mRNA and steady-state levels of protein isoforms of enoyl-ACP reductase from *Brassica napus*. Plant Mol Biol **26:** 155–163

Fischer K, Kammerer B, Gutensohn M, Arbinger B, Weber A, Hausler RE, Flugge UI (1997) A new class of plastidic phosphate translocators: a putative link between primary and secondary metabolism by the phospho*enol*pyruvate/phosphate antiporter. Plant Cell **9:** 453–462

Focks N, Benning C (1998) wrinkled1: a novel, low-seed-oil mutant of Arabidopsis with a deficiency in the seed-specific regulation of carbohydrate metabolism. Plant Physiol **118:** 91–101

Hall TC, Ma Y, Buchbinder BU, Pyne JW, Sun SM, Bliss FA (1978) Messenger RNA for G1 protein of French bean seeds: cell free translation and product characterization. Proc Natl Acad Sci USA **75:** 3196–3200

Hieter P, Boguski M (1997) Functional genomics: it's all how you read it. Science **278:** 601–602

Kammerer B, Fischer K, Hilpert B, Schubert S, Gutensohn M, Weber A, Flugge UI (1998) Molecular characterization of a carbon transporter in plastids from heterotrophic tissues: the glucose 6-phosphate/phosphate antiporter. Plant Cell **10:** 105–117

Kang F, Rawsthorne S (1994) Starch and fatty acid biosynthesis in plastids from developing embryos of oil seed rape. Plant J **6:** 795–805

Kang F, Rawsthorne S (1996) Metabolism of glucose-6-

phosphate and utilization of multiple metabolites for fatty acid synthesis by plastids from developing oilseed rape embryos. Planta **199:** 321–327

Kinoshita T, Nishimura M, Hara-Nishimura I (1995) Homologues of a vacuolar processing enzyme that are expressed in different organs in *Arabidopsis thaliana*. Plant Mol Biol **29:** 81–89

Loftus SK, Chen Y, Gooden G, Ryan JF, Birznieks G, Hilliard M, Baxevanis AD, Bittner M, Meltzer P, Trent J, Pavan W (1999) Informatic selection of a neural crest-melanocyte cDNA set for microarray analysis. Proc Natl Acad Sci USA **96:** 9277–9280

Meinke DW, Cherry JM, Dean C, Rounsley SD, Koornneef M (1998) *Arabidopsis thaliana*: a model plant for genome analysis. Science **282:** 679–682

Mekhedov S, Martinez de Ilarduya O, Ohlrogge J (2000) Towards a functional catalog of the plant genome: a survey of genes for lipid biosynthesis. Plant Physiol **122:** 389–401

Newman T, de Bruijn FJ, Green P, Keegstra K, Kende H, McIntosh L, Ohlrogge J, Raikhel N, Somerville S, Thomashow M (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

Ohlrogge JB, Jaworski JG, Post-Beittenmiller D (1993) De novo fatty acid biosynthesis. *In* TS Moore, ed, Lipid Metabolism in Plants. CRC Press, Boca Raton, FL, pp 3–32

Plaxton WC (1996) Organization and regulation of plant glycolysis. Annu Rev Plant Physiol Plant Mol Biol **47:** 185–214

Rafalski JA, Hanafey M, Miao GH, Ching A, Lee JM, Dolan M, Tingey S (1998) New experimental and computational approaches to the analysis of gene expression. Acta Biochim Pol **45:** 929–934

Rounsley SD, Glodek A, Sutton G, Adams MD, Somerville CR, Venter JC, Kerlavage AR (1996) The construction of Arabidopsis expressed sequence tag assemblies: a new resource to facilitate gene identification. Plant Physiol **112:** 1177–1183

Ruan Y, Gilmore J, Conner T (1998) Towards Arabidopsis genome analysis: monitoring expression profiles of 1,400 genes using cDNA microarrays. Plant J **15:** 821–833

Sauer N, Stolz J (1994) SUC1 and SUC2: two sucrose transporters from *Arabidopsis thaliana*: expression and characterization in baker's yeast and identification of the histidine-tagged protein. Plant J **6:** 67–77

Sterky F, Regan S, Karlsson J, Hertzberg M, Rohde A, Holmberg A, Amini B, Bhalerao R, Larsson M, Villarroel R, Van Montagu M, Sandberg G, Olsson O, Teeri TT, Boerjan W, Gustafsson P, Uhlen M, Sundberg B, Lundeberg J (1998) Gene discovery in the wood-forming tissues of poplar: analysis of 5,692 expressed sequence tags. Proc Natl Acad Sci USA **95:** 13330–13335

Van de Loo FJ, Broun P, Turner S, Somerville C (1995a) An oleate 12-hydroxylase from *Ricinus communis* L. is a fatty acyl desaturase homolog. Proc Natl Acad Sci USA **92:** 6743–6747

Van de Loo FJ, Turner S, Somerville C (1995b) Expressed sequence tags from developing castor seeds. Plant Physiol **108:** 1441–1150

Walden R, Cordeiro A, Tiburcio AF (1997) Polyamines: small molecules triggering pathways in plant growth and development. Plant Physiol **113:** 1009–1013

Weber H, Borisjuk L, Heim U, Sauer N, Wobus U (1997) A role for sugar transporters during seed development: molecular characterization of a hexose and a sucrose carrier in fava bean seeds. Plant Cell **9:** 895–908

Wenderoth I, Scheibe R, von Schaewen A (1997) Identification of the cysteine residues involved in redox modification of plant plastidic glucose-6-phosphate dehydrogenase. J Biol Chem **272:** 26985–26990