

TECHNICAL ADVANCE

An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana*[†]

Herman Höfte*, Thierry Desprez, Joëlle Amselem, Hélène Chiapello and Michel Caboche

Laboratoire de Biologie Cellulaire, INRA, Route de St-Cyr, 78026 Versailles Cedex, France

Annick Moisan, Marie-Françoise Jourjon and Jean-Louis Charpentreau

Laboratoire de Biométrie et d'Intelligence Artificielle, INRA, Auzeville, B.P. 27, 31326 Castanet Tolosan, France

Pierre Berthomieu, Danièle Guerrier and Jérôme Giraudat

Institut des Sciences Végétales, UPR 40 CNRS, 91190 Gif Sur Yvette, France

Françoise Quigley, Frank Thomas, De-Yao Yu and Régis Mache

Laboratoire de Biologie Moléculaire Végétale, Université Joseph Fourier, B.P. 53 X, 38041 Grenoble Cedex, France

Monique Raynal, Richard Cooke, Françoise Grellet and Michel Delseny

Laboratoire de Physiologie et Biologie Moléculaire Végétale, CNRS-Université, Avenue de Villeneuve, 66025 Perpignan Cedex, France

Yves Parmentier, Guy de Marcillac, Claude Gigot, Jacqueline Fleck and Gabriel Philipps

Institut de Biologie Moléculaire des Plantes, UPR 406 CNRS, 12 Rue du Général Zimmer, 67084 Strasbourg Cedex, France

Michèle Axelos, Claude Bardet, Dominique Tremousaygue and Bernard Lescure

Laboratoire de Biologie Moléculaire des Relations Plantes-Micro-organismes. Génétique Moléculaire d'*Arabidopsis thaliana*, UMR 05 CNRS-INRA, B.P. 27, 31326 Castanet-Tolosan, France

Summary

As part of the goal to generate a detailed transcript map for *Arabidopsis thaliana*, 1152 single run sequences (expressed sequence tags or ESTs) have been determined from cDNA clones taken at random

in libraries prepared from different sources of plant material: developing siliques, etiolated seedlings, flower buds, and cultured cells. Eight hundred and ninety-five different genes could be identified, 32% of which showed significant similarity to existing sequences in *Arabidopsis* and an array of other organisms. These sequences in combination with their positioning on the *Arabidopsis* genetic map will not only constitute a new set of molecular markers for genome analysis in *Arabidopsis* but also provide a direct route for the *in vivo* analysis of their gene products. The sequences have been made available to the public databases.

Introduction

Systematic genomic and cDNA sequencing projects are currently under way for several organisms. Despite the development of rapid, automated sequencing methods, the establishment of the complete sequence of even the smallest of these genomes will require several years if not decades. A complementary, but extremely rapid way to identify new genes is the partial sequencing of anonymous cDNAs producing expressed sequence tags or ESTs. Detailed analyses of ESTs have been published for more than 2000 human brain cDNAs (Adams *et al.*, 1992), for *Caenorhabditis elegans* (McCombie *et al.*, 1992; Waterston *et al.*, 1992) and mouse (Höög, 1991).

In plants, accumulation of sequence information has been slow compared with other organisms. In mid-1992, sequences were available for less than 500 different proteins from higher plants, whereas the number of genes expressed during the lifetime of a plant is estimated to be between 15 000 and 60 000 (Flavell, 1980; Goldberg *et al.*, 1978). Recently, the analysis of 830 rice cDNA sequences has been published (Uchimiya *et al.*, 1993) as well as 130 maize ESTs (Keith *et al.*, 1993). These experiments demonstrated that for a high proportion of the anonymous sequences a probable function can be found by comparison with sequence databases.

We chose to generate ESTs from *Arabidopsis thaliana*. This species has been adopted as a model for plant biology for several reasons: it has a short generation time, one of the smallest known plant genomes (100 Mb),

Received 27 May 1993; revised 12 July 1993; accepted 4 August 1993.

* For correspondence (fax + 33 1 30 83 31 11).

[†]A collaboration of Groupement de Recherche 1003, CNRS, France.

is relatively easily transformed and a rapidly increasing collection of visible mutants with their map positions on one of the five chromosome pairs exists (Rédei and Koncz, 1992).

Recently, the first positional cloning experiments in *Arabidopsis* have been reported (Arondel *et al.*, 1992; Giraudat *et al.*, 1992). Despite the small genome size, chromosome walking projects turn out to be laborious and time consuming. Efforts to construct a complete physical map are under way in several laboratories (Hauge and Goodman, 1992). This implies the construction of contigs of overlapping YAC clones and their linkage with the genetic map and requires a large number of polymorphic molecular markers. cDNA clones are not only very useful molecular markers, they also can be used to constitute a detailed genetic map of expressed genes. The availability of such a map in combination with the knowledge of the putative function of the protein product for those sequences that are homologous to known proteins, will facilitate the identification of candidate genes for mutant loci and simplify positional cloning strategies. Even with the limited information currently available, such approaches have already led to the identification of several genes corresponding to previously isolated mutants (Mandel *et al.*, 1992; Reed *et al.*, 1993).

Here we present the first results of a concerted effort of several laboratories to make an inventory of expressed genes in *A. thaliana*. Five cDNA libraries were used, prepared from different starting materials chosen as a function of the research interests of the participating laboratories. Thus, the libraries represent mRNAs expressed during floral development, embryogenesis, seed maturation, development of etiolated plants and the cell cycle. The analysis of the first 1152 sequences analysed allowed us to identify 895 non-redundant sequence tags of 250–350 bp, 32% of which match with known genes in *Arabidopsis* or other organisms. Many of these correspond to genes implicated in general cellular function and which had not yet been identified in plants

(e.g. 20% of identified genes encode proteins involved in translation), while others are of particular interest (proteasome alpha subunit) or unexpected in plants (laminin receptor). So far, 903 ESTs have been made available to EMBL and GenBank and also can be obtained via AATDB (an *Arabidopsis thaliana* database, Hauge and Goodman, 1992). The clones are freely available upon request.

Results and discussion

cDNA libraries and selection of clones

cDNA libraries were prepared from five different sources of tissue selected in conjunction with the research interests of the participating laboratories (Table 1): to identify cDNAs corresponding to mRNAs expressed in dividing cells, libraries were constructed from suspension-cultured cells (C) as well as from leaf strips cultured *in vitro* (L) to induce cell division. Libraries from flower buds (F) and developing siliques (S) were used to sequence cDNAs prepared from mRNAs corresponding to flower development and embryogenesis respectively. For the study of genes expressed during hypocotyl elongation, libraries were constructed from 5-day-old etiolated seedlings (E).

Clones were randomly picked for template preparations with the idea to constitute an overall picture of genes expressed in these tissues.

Sequence analysis and database comparisons

One thousand one hundred and fifty-two single run sequences were determined from one or in some cases from both ends. Clones from the oriented libraries were sequenced from the 5' end, to target preferentially the coding region of cDNAs. This allows the detection of more distant sequence homologies using the more sensitive comparison of deduced amino acid sequences. Table 2 shows the results of the similarity searches of the sequences against the public databases. Sequences

Table 1. Characteristics of cDNA libraries used in the sequencing project

Tissue	Research interests	Library	Vector	Oriented	No. ESTs	% redundancy ^a
Flowerbuds (F)	Flower development	GA GB	λZAPII	+	234	34
Developing siliques (S)	Embryogenesis	YA	λZAPII	–	318	29
Suspension cultured cells (C)	Cell cycle	TA	λZAPII	+	353	9
Cultured leaf strips (L)	Cell cycle	SA	λZAPII	+	60	0
Etiolated seedlings (E)	Hypocotyl elongation	VB VD	pHD-1	–	169	2
Total					1152 ^b	31

^a% redundancy refers to the number of sequences overlapping (as defined in Table 3) with one or more other ESTs/total number of ESTs.

^bIncludes 18 ESTs from a leaf cDNA library (Kreis, Orsay, unpublished data).

Table 2. Sequence similarities to *A.thaliana* cDNAs

EST-identifier	Putative identification	Organism	% Id (aa)	Overlap (aa)	Dbase	Acc. No.	No. cDNAs / library
(a) cDNAs 'identical' to <i>Arabidopsis</i> genes*							
ATTS0001	Adenylate translocator		96	98	G	X65549	C
ATTS0886	ATP synthase, 57 kDa subunit, vacuolar		96.4	83	S	P11574	C
ATTS0662	Calmodulin1		98	106	S	P25854	E; S
ATTS0224	Chalcone synthase		100	124	S	P13114	S
ATTS0657	Chlorophyll <i>a/b</i> binding protein, CAB1		100	123	S	P04777	E; F
ATTS0788	Chlorophyll <i>a/b</i> binding protein, CAB140 Lhcb1		98	86	P	A29280	7F; S; E
ATTS0029	Citrate synthase, mitochondrial precursor		97	62	S	P20115	C
ATTS0474	Elongation factor 1 alpha-A1,-A3		100	121	G	X16431	3C; F
ATTS0837	Elongation factor, 1-alpha		100	103	S	P13905	2C
ATTS0449	Enolase		97.6	205b	E	X58107	L, E
ATTS0002	Glyceraldehyde-3-phosphate dehydrogenase		96	52	S	P25858	2C
ATTS0022	H ⁺ -pyrophosphatase, vacuolar		100	69	G	M81892	C; S
ATTS0141	Histone H4		100	47	G	M17132	2C; L; F
ATTS0371	KIN1 protein		100	66	S	P18612	S
ATTS0408	Nitrilase		98.8	329b	E	X63445	L; S
ATTS0109	Phenylalanine ammonia lyase		97	101	P	PQ0140	C
ATTS0348	Polyubiquitin		100	36	E	X12853	S
ATTS0135	Ribosomal protein, 60S, L3		96	54	S	P17094	C
ATTS0797	Rubisco, ssu ^a		97	65	P	A29280	F
ATTS0704	Rubisco, ssu		100	45	S	P10798	F
ATTS0797	Rubisco, ssu		97	42	P	S03719	F
ATTS0375	Rubisco, ssu, 1A		97	38	S	P10795	S
ATTS0402	Rubisco, ssu		98.1	257b	G	X13611	L
ATTS0246	S-adenosylmethionine synthase		100	116	S	P23686	C; 2S
ATTS0164	Seed storage protein, 12S		100	53	S	P15455	2S
ATTS0228	Seed storage protein, 12S type B		100	121	S	P15456	S
ATTS0212	Seed storage protein, 2S, 1, precursor		98	53	S	P15457	2S
ATTS0168	Seed storage protein, 2S, 2, precursor		100	53	S	P15458	7S
ATTS0154	Seed storage protein, 2S, 3, precursor		100	62	S	P15459	47S
ATTS0171	Seed storage protein, 2S, 4, precursor		100	87	S	P15460	11S
ATTS0616	Tonoplast intrinsic protein, gamma		100	115	S	P25818	E
ATTS0464	Tubulin, alpha-2		98	193b	G	M84696	C
ATTS0635	Tubulin, alpha-3		97	342b	G	M17189	E
ATTS0883	Tubulin, alpha-5		100	192b	G	M84698	C
ATTS0750	Tubulin, alpha-6		97.5	162b	G	M84699	F
ATTS0696	Tubulin, beta-3		96.9	224b	G	M84701	F; C
ATTS0538	Tubulin, beta-5		100	68	G	M84702	E
ATTS0752	Ubiquitin, precursor		96	49	P	S01425	F; C; 2E
(b) cDNAs showing similarity to other <i>Arabidopsis</i> proteins (new family members)							
ATTS0057	Actin		87	109	S	P10671	2C; S
ATTS0366	Peroxidase, basic, precursor		38	47	S	P24102	S
ATTS0393	Catalase isozyme-1		76	227b	E	X64271	C
ATTS0732	Chlorophyll <i>a/b</i> binding protein, CAB-165 Lhcb1		78	23	P	S06765	F
ATTS0553	Cor47: cold-regulated protein		68	89	G	X59814	E
ATTS0613	Dehydrin		41	123	S	P25863	E
ATTS0075	Heat-shock protein 83		87	77	G	M62984	C
ATTS0231	Lipid body-associated protein		48	89	S	P21641	S
ATTS0176	Meri-5		91	132	S	P24806	2S; E; C
ATTS0663	Metallothionein		38	28	S	P25860	2E
ATTS0225	Nitrilase		89	102	G	X63445	S
ATTS0312	Oleosin		51	73	E	X62353	S
ATTS0120	Poly(A)-binding protein		45	68	G	M97657	C
ATTS0475	Receptor-like protein kinase (putative)		45.6	136	E	M84660	C; S
ATTS0595	S-adenosylmethionine synthase		81	54	S	P17562	E
ATTS0190	Seed storage protein, 12S type C		68	81	S	P15455	S
ATTS0243	Seed storage protein, 2S, 5, precursor		57	63	S	P15459	S
ATTS0533	Tonoplast intrinsic protein, gamma		89	103	S	P25818	3E

Table 2. (Cont.)

EST-identifier	Putative identification	Organism	% Id (aa)	Overlap (aa)	Dbase	Acc. No.	No. cDNAs / library
(c) cDNAs showing similarity to other plant proteins							
ATTS0167	(S)-2-hydroxy-acid oxidase	<i>Spinacea</i>	84	88	S	P05414	S
ATTS0444	Enolase	<i>Zea</i>	80,9	270b	E	X55981	L
ATTS0127	Actin depolymerizing factor	<i>Lilium</i>	62	79	G	Z14110	C
ATTS0442	Alcohol dehydrogenase-1	<i>Solanum</i>	80.2	263b	G	M25154	L
ATTS0432	ATP synthase alpha subunit	<i>Spinacea</i>	88.3	309b	G	X05916	2L
ATTS0631	ATP synthase beta chain	<i>N. tabacum</i>	80	25	S	P17614	E
ATTS0357	ATP synthase protein 9, mitoch.	<i>Oenothera</i>	100	20	S	P14572	S
ATTS0122	Auxin-induced protein (parC)	<i>N. tabacum</i>	84	58	P	S16636	C
ATTS0672	Auxin-induced protein AUX28	<i>Glycine</i>	76	85	S	P13089	3E; C
ATTS0306	Catalase isozyme A	<i>Gossypium</i>	80	69	S	P17598	S; E
ATTS0779	Chaperonin 62.5K beta chain	<i>Brassica</i>	94	79	P	PW00007	F
ATTS0700	Chlorophyll <i>a/b</i> binding protein	<i>Lycopersicon</i>	98	52	P	517737	4F; C; L; 2E
ATTS0792	Cyc07 protein	<i>C. roseus</i>	80	62	E	JQ0939	F
ATTS0362	Cysteine proteinase	<i>C. papaya</i>	78	59	S	P05993	2S; 3E
ATTS0192	Cysteine proteinase inhibitor	<i>O. sativa</i>	56	94	S	P09229	2S
ATTS0558	Cytochrome P450	<i>Persea</i>	32	135	S	P24465	E
ATTS0203	Desiccation-related protein clone PCC13-62 precursor	<i>Craterostigma</i>	49	173	S	P22242	3S
ATTS0834	Disulphonide-isomerase	<i>Mesocricetus</i>	73	22	P	S19656	F
ATTS0536	E4: DNA-bind. factor, ethylene induced	<i>Lycopersicon</i>	73	30	P	JQ0988	E
ATTS0570	E8, fruit-ripening protein, ethylene responsive	<i>Lycopersicon</i>	48	98	S	P10967	E; 2S
ATTS0424	Elongation factor 2	<i>Chlorella</i>	72.6	259b	E	M68064	L
ATTS0349	Ethylene forming enzyme	<i>Lycopersicon</i>	73	79	S	P24157	S
ATTS0666	Extensin	<i>Daucus</i>	71	99	S	P06599	2E
ATTS0281	Extensin-like protein	<i>N. tabacum</i>	42	119	S	P13983	2S
ATTS0104	Ferredoxin	<i>Raphanus</i>	96	27	S	P14937	C
ATTS0655	Ferritin	<i>Phaseolus</i>	50	60	S	P25699	E
ATTS0153	Flavonol O-(3)-glucosyl transferase	<i>Zea</i>	40	93	S	P16166	S
ATTS0248	Germin precursor	<i>Triticum</i>	44	213	S	P15290	S
ATTS0543	Glyceraldehyde-3-phosphate dehydrogenase	<i>Sinapis</i>	89	87	S	P04796	E; L
ATTS0303	Glutamine synthase PR-2	<i>Phaseolus</i>	87	95	S	P04771	S
ATTS0693	Glycine-rich RNA-binding protein	<i>Sorghum</i>	66	47	P	S12311	F
ATTS0807	Heat-shock protein 70	<i>Lycopersicon</i>	92	68	P	S14950	F; C; S
ATTS0835	Heat-shock protein 82	<i>N. tabacum</i>	83	36	S	S18865	F
ATTS0262	Histone H2A	<i>Lycopersicon</i>	74	46	S	P25469	S
ATTS0707	Histone H2B(2)	<i>Triticum</i>	68	120	P	S00622	F
ATTS0760	Hypothetical protein 1	<i>Pisum</i>	81	68	S	S11861	3F
ATTS0334	Immunoglobulin-binding protein (BIP) homologue	<i>Zea</i>	69	45	S	P24067	S
ATTS0885	Initiation factor 5A	<i>N. plumbaginifolia</i>	88,6	88	S	P24921	2C; F
ATTS0541	Isocitrate lyase	<i>Ricinus</i>	90	97	S	P15479	E
ATTS0307	Light-regulated protein	<i>Gossypium</i>	46	107	P	S03277	S
ATTS0669	Lipid-transfer protein (nonspecific)	<i>Zea</i>	53	50	S	P19656	2E; F; 2S
ATTS0380	Myb-like transcription factor	<i>Zea</i>	71.0	217b	E	M73028	C
ATTS0451	NADH dehydrogenase	<i>N. tabacum</i>	86.2	304b	G	Z00044	L
ATTS0351	Oxygen-evolving enhancer protein 3 precursor	<i>Spinacea</i>	71	65	S	P12301	S
ATTS0711	Pathogenesis-related protein	<i>Hordeum</i>	46	68	P	S09604	F
ATTS0062	Pectin esterase	<i>Phaseolus</i>	67	105	G	X68028	C; F
ATTS0359	Peptidyl-prolyl <i>cis-trans</i> isomerase	<i>Brassica</i>	89	52	S	P24525	S
ATTS0592	Peroxidase anionic (lignin forming)	<i>N. tabacum</i>	31	120	S	P11965	E
ATTS0561	Peroxidase cationic	<i>A. hypogaea</i>	39	137	S	P22196	E
ATTS0284	Phospho-2-dehydro-3-deoxyheptonate aldolase prec.	<i>Solanum</i>	96	26	S	P21357	3S; F
ATTS0162	Photosystem I reaction centre subunit 5 precursor	<i>Spinacea</i>	95	22	P	S00452	S
ATTS0166	Photosystem II 10 kDa polypeptide precursor	<i>Solanum</i>	100	68	P	S00411	S
ATTS0210	Photosystem II 7 kDa protein	<i>Spinacea</i>	100	12	P	S01121	S
ATTS0748	Photosystem II oxygen-evolving complex-33K	<i>Pisum</i>	60	35	P	S04132	F
ATTS0706	Plastocyanin	<i>Capsella</i>	93	29	P	A00304	F
ATTS0744	Polygalacturonase	<i>Zea</i>	33	32	P	S16998	3F

Table 2. (Cont.)

EST-identifier	Putative identification	Organism	% Id (aa)	Overlap (aa)	Dbase	Acc. No.	No. cDNAs / library
ATTS0239	Protease inhibitor P322 precursor (Bowman-Birk), putative	<i>Solanum</i>	53	74	S	P20346	S; F
ATTS0881	Protein kinase C inhibitor (putative)	<i>Oenothera</i>	85.6	90	G	X62838	C
ATTS0294	Protein P59 precursor	<i>Lycopersicon</i>	65	104	S	P15722	S
ATTS0283	Protein, 14 kDa	<i>Daucus</i>	68	81	S	P14009	S
ATTS0197	Pyrroline-5-carboxylate reductase	<i>Glycine</i>	71	117	S	P17817	S
ATTS0876	Receptor-like protein kinase (putative)	<i>Zea</i>	52.2	69	S	P17801	C
ATTS0329	Ribosomal protein S13, mitoch.	<i>Daucus</i>	63	79	S	P23209	S
ATTS0708	Ribosomal protein, 30S, S19, mitoch.	<i>Petunia</i>	67	79	P	S16665	F
ATTS0332	Ribosomal protein, 50S, L2, chloroplast	<i>N. tabacum</i>	100	25	S	P06379	2S
ATTS0017	Ribosomal protein, 60S, L12	<i>Triticum</i>	57	40	P	S00295	C
ATTS0317	Ribosomal protein, 60S, L5, mitoch.	<i>Marchantia</i>	50	34	G	M68929	S
ATTS0443	Sucrose-UDP glucosyl-transferase	<i>Solanum</i>	73.0	230b	G	M18745	L
ATTS0252	Thioesterase mRNA	<i>Umbellularia</i>	63	147	G	M94159	S
ATTS0795	Thioglucosidase	<i>Sinapis</i>	76	42	P	S19149	4F
ATTS0234	Thiol protease precursor	<i>Lycopersicon</i>	76	119	S	P20721	S
ATTS0078	Thioredoxin	<i>N. tabacum</i>	61	95	P	S16590	2C
ATTS0421	Translation initiation factor 2	<i>N. plumbaginifolia</i>	81.1	259b	E	X61205	L
ATTS0617	TriosePho.Isomerase	<i>Zea</i>	76	99	S	P12863	E
ATTS0762	Tumour protein (fragment)	<i>Medicago</i>	83	42	P	S19895	F
ATTS0019	Ubiquitin precursor	<i>Helianthus</i>	95	65	G	S17436	C; F
ATTS0889	Ubiquitin-conjugating enzyme	<i>Triticum</i>	67.4	46	S	P25868	C
ATTS0751	Vegetative storage protein 27K (precursor)	<i>Soja</i>	47	55	P	JA0139	F
(d) cDNAs showing similarity to non-plant proteins							
ATTS0845	AAC-rich mRNA (AAC3) protein	<i>Dictyostelium</i>	28.6	70	S	P14197	C
ATTS0636	Adenosylhomocysteinase	<i>H. sapiens</i>	61	121	S	P23526	E; C
ATTS0010	ADP-ribosylation factor	<i>H. sapiens</i>	79	29	S	P18085	2C; E
ATTS0709	Aldolase	<i>D. melanogaster</i>	52	69	P	S06439	F
ATTS0894	Annexin IV	<i>Sus</i>	54.5	66	S	P08132	C; E
ATTS0694	AWD protein	<i>D. melanogaster</i>	60	76	P	S01908	F
ATTS0867	Carbamoyl-phosphate synthetase, large chain	<i>Methanosarcina</i>	52.6	95	S	P13258	C
ATTS0571	Elongation factor 1(β ?)	<i>H. sapiens</i>	26	113	S	P24534	E
ATTS0055	Expressed sequence tag, unknown	<i>H. sapiens</i>	53	60	G	M62199	C
ATTS0844	Expressed sequence tag, unknown	<i>H. sapiens</i>	37.4	115	E	M79223	C
ATTS0069	Fibrillarin	<i>H. sapiens</i>	80	96	S	P22087	C
ATTS0136	Glutamic acid-rich protein	<i>Plasmodium</i>	41	56	P	P13816	C
ATTS0370	Hemoglobinase precursor	<i>Schistosoma</i>	67	33	S	P09841	S
ATTS0546	HSP/groes: chaperonin	<i>Mycobacterium</i>	50	78	S	P09621	E
ATTS0263	Hypothetical 17.1 kDa protein in BIOA 3' region	<i>E. coli</i>	35	92	S	P12994	2S
ATTS0046	Laminin receptor (or factor P40)	<i>D. melanogaster</i>	69	71	G	M77133	4C
ATTS0740	Laminin receptor	<i>Mus</i>	56	48	P	A30317	F
ATTS0848	Laminin receptor	<i>H. sapiens</i>	74	77	P	JN0273	C
ATTS0619	Mitochondrial fusion target protein	<i>S. cerevisiae</i>	54	70	S	P23248	E
ATTS0368	MTHDH ^a	<i>Mus</i>	63	71	S	P18155	S
ATTS0161	NADH ubiquinone oxidoreductase subunit	<i>Bos</i>	35	67	S	P23934	S
ATTS0846	Nucleic acid binding protein	<i>H. sapiens</i>	54.5	22	S	P20694	C
ATTS0319	O-acetylserine sulfhydrylase A	<i>Salmonella</i>	56	50	S	P12674	S
ATTS0661	Phenylalanine:tRNA synthetase	<i>S. cerevisiae</i>	36	111	S	P15624	E
ATTS0195	Pre-mRNA processing RNA helicase	<i>S. cerevisiae</i>	37	51	S	P21372	S
ATTS0471	Proteasome alpha subunit	<i>D. melanogaster</i>	67.1	79	G	M57712	C
ATTS0470	Proteasome subunit RC5	<i>Rattus</i>	45.7	81	G	X52783	C
ATTS0184	Protein in DNAX 5' region, 17 kDa	<i>Bacillus</i>	30	149	S	P21335	S
ATTS0056	Pyruvate dehydrogenase E1 component, beta subunit precursor	<i>Ascaris</i>	55	52	S	P26269	C
ATTS0839	Ribosomal protein 40S, S22	<i>Xenopus</i>	64.9	57	S	P23403	C
ATTS0874	Ribosomal protein 40S, S4	<i>H. sapiens</i>	59.3	27	S	P12750	C
ATTS0872	Ribosomal protein 40S, S5	<i>Rattus</i>	61.8	68	S	P24050	C
ATTS0836	Ribosomal protein 60S, L1	<i>Xenopus</i>	56.8	81	S	P08429	C

Table 2. (Cont.)

EST-identifier	Putative identification	Organism	% Id (aa)	Overlap (aa)	Dbase	Acc. No.	No. cDNAs / library
ATTS0853	Ribosomal protein 60S, L17	<i>Rattus</i>	84.7	85	S	P24048	C
ATTS0861	Ribosomal protein 60S, L26	<i>Rattus</i>	69	87	S	P12749	C
ATTS0893	Ribosomal protein L35A	<i>S. cerevisiae</i>	46	52	P	S18431	C
ATTS0516	Ribosomal protein S8	<i>Xenopus</i>	47.5	101	S	P02362	C
ATTS0185	Ribosomal protein, 30S, S13	<i>Bacillus</i>	53	120	S	P15757	S
ATTS0333	Ribosomal protein, 40S, S12	<i>Tetrahymena</i>	65	107	S	P06147	S
ATTS0358	Ribosomal protein, 40S, S13	<i>Rattus</i>	82	107	S	P19116	S
ATTS0264	Ribosomal protein, 40S, S14	<i>H. sapiens</i>	95	84	S	P06366	S
ATTS0365	Ribosomal protein, 40S, S15	<i>H. sapiens</i>	94	48	S	P11174	S
ATTS0150	Ribosomal protein, 40S, S2	<i>Rattus</i>	71	70	S	P27952	2C; L
ATTS0143	Ribosomal protein, 40S, S28	<i>Rattus</i>	80	61	S	P25112	C
ATTS0011	Ribosomal protein, 40S, S3	<i>H. sapiens</i>	87	48	S	P23396	C
ATTS0196	Ribosomal protein, 40S, S4	<i>H. sapiens</i>	79	149	S	P14036	2S
ATTS0020	Ribosomal protein, 40S, S4 (S7)	<i>S. cerevisiae</i>	70	83	S	P05753	C
ATTS0060	Ribosomal protein, 40S, S7	<i>Rattus</i>	67	53	S	P23821	C
ATTS0053	Ribosomal protein, 40S, S16	<i>H. sapiens</i>	67	102	S	P17008	C
ATTS0594	Ribosomal protein, 60S, KD4/L2	<i>S. pombe</i>	54	122	S	P14067	E
ATTS0330	Ribosomal protein, 60S, L12	<i>Rattus</i>	57	60	S	P23358	S
ATTS0038	Ribosomal protein, 60S, L12	<i>S. pombe</i>	48	92	S	P08094	C
ATTS0523	Ribosomal protein, 60S, L18	<i>H. sapiens</i>	57.5	87	E	M77998	C
ATTS0123	Ribosomal protein, 60S, L18	<i>Rattus</i>	64	62	S	P12001	C
ATTS0465	Ribosomal protein, 60S, L19	<i>Mus</i>	68.8	80	G	M62952	C; 2F
ATTS0163	Ribosomal protein, 60S, L1A	<i>Xenopus</i>	52	174	S	P08429	S
ATTS0610	Ribosomal protein, 60S, L1B	<i>Xenopus</i>	71	139	S	P02385	E
ATTS0111	Ribosomal protein, 60S, L21	<i>Rattus</i>	58	74	S	P20280	C
ATTS0745	Ribosomal protein, 60S, L23	<i>Rattus</i>	80	168b	P	JH0418	F
ATTS0091	Ribosomal protein, 60S, L23A	<i>Rattus</i>	70	30	P	S19992	C
ATTS0145	Ribosomal protein, 60S, L27	<i>Rattus</i>	84	72	S	P18445	2C; S
ATTS0142	Ribosomal protein, 60S, L30A	<i>S. cerevisiae</i>	49	61	S	P04449	C
ATTS0313	Ribosomal protein, 60S, L32	<i>H. sapiens</i>	68	110	S	P02433	S
ATTS0089	Ribosomal protein, 60S, L37	<i>Rattus</i>	63	74	G	X66369	C
ATTS0112	Ribosomal protein, 60S, L46	<i>S. cerevisiae</i>	69	49	S	P04650	C
ATTS0270	Ribosomal protein, 60S, L5	<i>Gallus</i>	78	58	S	P22451	S; 2C
ATTS0301	Ribosomal protein, 60S, L9	<i>Rattus</i>	67	46	S	P17077	S
ATTS0862	RNA polymerases I and III, 40 kDa protein	<i>S. cerevisiae</i>	52.2	69	S	P07703	C
ATTS0819	Succinyl-coA ligase -alpha chain	<i>Rattus</i>	69	72	P	P13086	F
ATTS0692	Suppressor protein 1	<i>S. cerevisiae</i>	87	71	P	A23642	F
ATTS0627	Translation initiation factor	<i>H. sapiens</i>	60	156	S	P20042	E
ATTS0705	U2 small nuclear ribonucleoprotein A	<i>H. sapiens</i>	64	110	P	P09661	F
ATTS0266	Ubiquitin-conjugating enzyme	<i>S. cerevisiae</i>	80	112	S	P15731	S; C
ATTS0849	Ubiquitin-conjugating enzyme	<i>Drosophila</i>	73.2	82	S	P25867	C
ATTS0324	Unp gene	<i>Mus</i>	35	140	G	L00681	S
ATTS0735	Urease	<i>Klebsiella</i>	58	62	E	F36138	F

Columns refer respectively to: (1) the EST identifier in dbEST (NCBI), (2) the identity of the protein with the highest similarity score, (3) the organism, (4) the % sequence identity and (5) the overlap in amino acids or in nucleotides (the latter indicated by 'b'), (6) the database and (7) the accession number of the database entry and (8) the libraries from which the ESTs were derived (first letter) and in which additional matches were found (numbers indicate the number of matches observed). Libraries were made from: C: suspension cultured cells; L: cultured leaf strips to induce cell division; E: etiolated seedlings; F: flower buds; S: developing siliques. When several cDNA sequences matched the same entry, one representative match is shown. ^aMTHDH, methylenetetrahydrofolate dehydrogenase; Rubisco, ssu, ribulose-bisphosphate carboxylase, small subunit.

were translated in the six open reading frames and compared with protein sequence databases using the programs FASTA (Pearson and Lipman, 1988) or BLASTX (Altschul *et al.*, 1990), except for a number of cases in which nucleotide sequences were directly compared with

DNA data banks. Significant matches were found for 32% (375) of the sequences (PAM120 scores above 80, see Experimental procedures). When more than one EST matched the same gene, one representative sequence was chosen for the table and the number of matches for

each library indicated in the last column. Unfortunately, no comparisons could be made with the rice ESTs as these have not been submitted to the public databases. The number of database matches is comparable with that observed for 1517 ESTs from *C. elegans* (31%, Waterston *et al.*, 1992) but surprisingly high compared with the 8% provisionally identified sequences among 830 ESTs from rice (Uchimiya *et al.*, 1992). This difference might be explained by the rather high cut-off score (160) used for the database comparisons in the latter study as a result of which more distant sequence homologies probably remained undetected. On the other hand, only 17% of the human ESTs showed database matches. This might be explained in part by the observation that an estimated 50–85% of the human ESTs without homologies to known proteins may be non-coding (Adams *et al.*, 1992) which does not seem to be the case for *C. elegans* (Green *et al.*, 1993) and still needs to be investigated in plants.

Quality of the libraries

To evaluate the quality of the libraries we first investigated whether any contamination with DNA from micro-organisms could be detected. ESTs matching sequences of non-plant origin were examined in more detail (Table 2d). Twelve ESTs were found with 80% or more sequence identity to non-plant sequences. Ten of them, found in three independent libraries, were homologous to proteins (fibrillarin and ribosomal proteins) of mammalian origin, an unlikely source of contamination. Only two ESTs showed a high similarity score (80% and 87% amino acid sequence identity, respectively) to yeast sequences ('Ubiquitin-conjugating enzyme' and 'Suppressor protein 1'). However, at the DNA sequence level we observed for the first EST only 68% identity (on a 346 bp overlap), comparable with the 70% identity (on a 221 bp overlap) observed between the same yeast sequence and its *Drosophila* homologue. Similarly, the second EST was 73% identical (200 bp overlap) to the yeast sequence, comparable with the 74% DNA sequence and its *Drosophila* homologue (EMBL accession no. X17493). Similarly, the second EST was 73% identical (200 bp overlap) to the yeast sequence, comparable with the 74% DNA sequence identity (213 bp overlap) between this EST and a human homologue (GenBank accession no. M75715). The high similarity scores therefore indicate evolutionary conservation rather than contamination with yeast-like DNA.

Contamination of cDNA libraries by mitochondrial sequences has been repeatedly observed in human EST projects (up to 15%, Adams *et al.*, 1992). The availability of the complete sequences of chloroplast genomes and

of many mitochondrial genes allowed us to estimate the contamination of the libraries with organelle-derived sequences. Five ESTs, respectively, matched the mitochondrial transcripts for ATP-synthase protein 9, ribosomal proteins S19 and L5, and the chloroplast transcript for the ribosomal protein L2 (two copies). Other ESTs derived from the chloroplast genome would unavoidably be identified if present in the collection. The contamination of the EST sequences by organelle-encoded transcripts is therefore assumed to be lower than 1.8% (5/284 sequences identified in plants, Table 2).

Among the clones from etiolated seedlings that were sequenced from both ends we observed three cases in which the two ends of the same cDNA were similar to different genes, suggesting that these clones were chimaeric. From more recent experiments, we estimate that less than 6% of the clones from etiolated seedlings are chimaeric (five chimaeric clones were found on a total of 91 for which the sequences from both ends showed database matches).

New plant genes

Among the sequences showing significant similarity to sequences in the public databases, 131 (35%) were considered identical to 37 known *Arabidopsis* genes (Table 2a, we define as identical, sequences showing more than 96% identity at the nucleotide level). Twenty-seven cDNAs (7%) represented 17 new *Arabidopsis* gene family members and 124 (33%) were similar to 78 genes from other plants, while 93 (25%) clones corresponded to 76 genes not previously sequenced in plants.

Among the ESTs identifying new plant proteins (Table 2d), at least 46 correspond to new proteins involved in translation and eight to new metabolic enzymes. A number of cDNAs were unexpected and may raise new questions in plant biology. For instance, several ESTs were highly similar to the human laminin receptor, which may point to a new class of cytoskeletal elements for plants. ATTS0069 identifies a homologue of human fibrillarin. In yeast, this highly conserved protein is essential for the processing, methylation of ribosomal preRNAs and the correct assembly of ribosomes (Tollervey *et al.*, 1993). The awd (abnormal wing disc) gene, homologous to ATTS0694, was identified in *Drosophila* by a lethal mutation and codes for a nucleoside diphosphate kinase (Kimura *et al.*, 1990). Annexins (homologous to ATTS0894) are Ca^{2+} -dependent phospholipid-binding proteins which may play a role in intracellular vesicular trafficking. Annexins have been identified in plants biochemically (Blackbourn *et al.*, 1992), however no plant gene sequences were found in

the public databases. Two ESTs matched to two enzymes rather unexpected for non-legume plants: hemoglobinase (ATTS0370) and urease (ATTS0735). Other interesting matches were with three ubiquitin-conjugating enzymes (ATTS0266, 0849 and 0889). It will be worthwhile to see whether any of these map to the locus of the *prt1* mutant, affected in the ubiquitin-dependent protein degradation pathway (Bachmair *et al.*, 1993). Further in relation to ubiquitin-dependent protein degradation, ATTS0471 identifies an α -subunit of the proteasome complex, providing the first demonstration of the existence of such a complex in plants. (The complete sequence of this cDNA was submitted independently under the accession number X66825, Genschik *et al.*, 1992.)

In conclusion, these results underscore the efficiency

with which new plant genes can be identified through partial sequencing of anonymous cDNAs.

Redundancy in and between libraries

The random sequencing approach yielded some interesting information on the structure and expression of gene families. Of the 1152 clones sequenced, 895 appear to be unique cDNAs (representing 31% sequence redundancy, Table 1). This is likely to be an overestimate since non-overlapping sequences derived from the same transcript will be scored as being different. The cDNAs that were sequenced from both ends allow a more precise identification, the 3' sequence providing a unique tag present on all cDNAs derived from the same mRNA even when they are truncated at the 5' end.

Table 3. Redundant ESTs

EST-number	Putative identification	F	L	C	E	S	Total
ATTS0154	Seed storage protein, 2S	0	0	0	0	67	67
ATTS0402	Rubisco, ssu	46	1	0	0	1	48
ATTS0156		0	0	0	0	21	21
ATTS0046		1	0	4	0	0	5
ATTS0019		2	0	2	1	0	5
ATTS0493		1	0	3	0	0	4
ATTS0649		0	0	0	3	0	3
ATTS0150	Ribosomal protein, 40S, S2	0	1	2	0	0	3
ATTS0474	Elongation factor 1 α -A1,-A3	0	0	3	0	0	3
ATTS0051		0	0	3	0	0	3
ATTS0754		3	0	0	0	0	3
ATTS0080	Ribosomal protein, 60S, L5	0	0	2	0	1	3
ATTS0095	Ribosomal protein, 60S, L27A	0	0	2	0	1	3
ATTS0344		0	0	0	1	2	3
ATTS0166	Photosystem II 10 kDa polypeptide precursor	1	0	0	0	2	3
ATTS0465	Ribosomal protein, 60S, L19	2	0	1	0	0	3
ATTS0782		3	0	0	0	0	3
ATTS0749		3	0	0	0	0	3
ATTS0464	Tubulin α -2	2	0	1	0	0	3
ATTS0744	Polygalacturonase	3	0	0	0	0	3
ATTS0885	Initiation factor 5A	1	0	2	0	0	3
ATTS0795	Thioglucosidase	3	0	0	0	0	3
ATTS0806		3	0	0	0	0	3
ATTS0519		2	0	1	0	0	3
ATTS0495		0	0	3	0	0	3
ATTS0196		0	0	2	0	1	3
ATTS0010	ADP-ribosylation factor	0	0	2	1	0	3
ATTS0141	H4 histone	1	0	2	0	0	3
ATTS0652		1	0	1	1	0	3
Total number of ESTs overlapping with one other EST		33	7	44	19	31	134
Total		111	9	80	26	127	353

ESTs overlapping with two or more other ESTs are shown. Overlapping was defined as showing greater than 90% nucleotide sequence identity over a stretch of more than 50 nt. The putative identification is shown when a database match was found. Libraries are as in Table 1. The totals of the columns represent the sum of all ESTs overlapping with at least one other EST. To save space we did not include the details for the individual ESTs overlapping with only one other EST.

Varying levels of sequence redundancy were observed among the different libraries, with the highest level in flower buds, which mainly results from the high level of ribulose-bisphosphate carboxylase (Rubisco), small subunit cDNAs (Table 3, 17% for the remaining ESTs) and the lowest for cultured leaf cells and etiolated seedlings. The most highly represented gene family in the immature silique cDNA library is that of the 2S seed storage proteins, the napins. Among the (68) napin cDNAs sequenced, 67 correspond to the previously described classes (Krebbers *et al.*, 1988), providing an estimation of the relative abundance of the corresponding transcripts (Table 2a), while one sequence defines a new fifth class (Table 2b).

For a number of other previously identified *Arabidopsis* gene families new members could be identified as well. For instance, new members of the SAM synthetase (Table 2b) and the enolase families (Table 2c) were detected in etiolated seedlings (E) and induced leaf cells (L), respectively. Similarly, ATTS0475 and ATTS0372 identify two new proteins of a receptor protein kinase family.

The high redundancy observed globally (>31%) indicates that normalization procedures will be required (Patanjali *et al.*, 1991) for a more efficient isolation of less abundant sequences in the future. However, this approach, generally involving hybridization techniques, may also eliminate transcripts similar to, but different from certain abundant mRNA species.

Tissue-specific expression

The use of cDNA libraries derived from different tissues should eventually allow us to draw a global picture of the tissue distribution of abundant and moderately expressed gene products. Despite the relatively small sample size, some preliminary observations can be made:

(i) Abundant cDNAs: ESTs that overlap with two or more other ESTs are shown in Table 3. Extremely abundant cDNAs were observed in flower buds (the photosynthetic proteins Rubisco, ssu (46) and chlorophyll *a/b* binding protein (7)) and in developing siliques (2S seed storage protein (68)). Interestingly, about half of the abundant EST groups in Table 3 (14/29) did not show any database match; ATTS0156 was found eight times in developing siliques only and represents a new family of proteins which are synthesized during embryo development (Delseny, unpublished observations), whereas others appeared in several libraries (ATTS0019 was found in F, C and E) and therefore seem to represent some general cellular function specific to plants. These results indicate that a large number of major plant proteins are still to be identified.

(ii) Known genes at new locations: E4 and E8 are two tomato genes with an ethylene-induced fruit-ripening-related expression pattern (Cordes *et al.*, 1989; Deikman and Fisher, 1988). ESTs homologous to E8 were identified in developing *Arabidopsis* siliques (ATTS0221; 0274) but, unexpectedly, also in etiolated seedlings (ATTS0570). In the same library we also found an E4-homologue. These observations strongly suggest, in agreement with physiological studies, a role for ethylene in etiolated development. cDNAs encoding chlorophyll *a/b* binding proteins (for example, ATTS0657 and ATTS0587) were found in etiolated tissue, indicating a light-independent expression pattern of these photosynthetic genes.

(iii) ESTs and cell elongation: one class of cDNAs was highly represented in the library from etiolated seedlings but not in other libraries ('Tonoplast intrinsic protein', represented by ATTS0616 and ATTS0533 in Table 2). The ESTs identify a family of H₂O-conducting channel proteins in the vacuolar membrane. The expression pattern of at least one family member (γ -TIP) seems to be associated with cell elongation during etiolated development (Ludevid *et al.*, 1992; Höfte, unpublished observations).

Conclusions

Systematic cDNA sequencing led to the identification of a large number of new expressed *Arabidopsis* genes. Random sequencing provides an overall view of genes expressed in the different tissues. However, the degree of redundancy observed for the first 1152 sequences indicates that a pre-screening or a normalization of the libraries will be required in the future to identify less abundant cDNAs. cDNA sequencing from libraries derived from different tissues or tissues under different physiological conditions will eventually shed more light on the expression patterns of groups of abundant and moderately expressed genes. We have recently begun RFLP mapping of a selection of these ESTs on the five *Arabidopsis* chromosomes. In a later stage we hope to use an ordered YAC library to simplify the mapping of ESTs. While awaiting the completion of the genomic sequence, this programme will rapidly lead to a detailed transcript map and thus provide an important tool for the *in vivo* analysis of gene function in *A. thaliana*.

Experimental procedures

cDNA libraries

Table 1 summarizes information on the cDNA libraries used. *Arabidopsis thaliana* ecotype Columbia was used for all libraries,

except for F for which ecotype C-24 was used. S, F, C, L and E refer to the libraries by indicating the plant material used. Libraries S and C were constructed as described (Giraudat *et al.*, 1992; Regad *et al.*, 1993). For F and L, plants were grown in soil as described (Giraudat *et al.*, 1992). For L, leaf strips were incubated in a medium (Nagy and Maliga, 1976) containing 1 mg l^{-1} 2,4-dichlorophenoxy-acetic acid to induce the G0/S-transition in the cell division cycle. For E, seeds were sown *in vitro* on a solid medium (Estelle and Somerville, 1987) without sucrose, incubated at 4°C for at least 2 days, transferred to the light for 12 h to induce germination and incubated in the dark at 22°C for 4 more days. Seedlings were collected under a green safe light. Plant material was frozen in liquid nitrogen and stored at -70°C until use. Total RNAs were prepared as described (Vervoerd *et al.*, 1989) and for E further purified on a CsCl gradient (Sambrook *et al.*, 1989). Poly A⁺ mRNA was purified on oligo-dT columns as described (Sambrook *et al.*, 1989). For L and F, cDNA was synthesized and used to prepare an oriented library in the vector λ ZAPII using the instructions from the supplier (Stratagene Inc., La Jolla, CA). For E, oligo-dT primed ds cDNA was ligated to *Bst*XI-adaptors (Minet *et al.*, 1992), fractionated into three size classes (0.3–1 kb; 1–2.2 kb; > 2.2 kb) on a 1% agarose gel and ligated to the *Bst*XI-digested plasmid vector pHD1. pHD1 is pBS-SK (Stratagene Inc., La Jolla, CA) with the following modifications: (i) the *Bst*XI site was destroyed by cleavage, exonuclease treatment and religation, (ii) the sequence between bp 85 and 151 has been replaced by a 42 bp *Not*I–*Not*I fragment from pLF61 containing two new *Bst*XI sites (Minet *et al.*, 1992). Ligation reactions were electroporated into *Escherichia coli* strain DH10B (Bethesda Research Laboratories Inc., Gaithersburg, MD), giving rise to libraries VD, VB and VC, respectively, with 2.6×10^5 , 1.2×10^5 and 3×10^4 primary transformants. Plasmid libraries were amplified on 15 cm agar plates, and stored at -70°C in 15% glycerol. VC was not used in this study.

Sequencing and data analysis

Aliquots of λ ZAPII libraries were globally excised *in vivo* (Stratagene Inc., La Jolla, CA). Template preparation was done with liquid cultures prepared from individual clones, using an alkaline lysis procedure and Qiagen TIP20 columns (Qiagen Inc., Chatsworth, CA). Sequencing was done manually (Tabor and Richardson, 1987) with T7 or T3 primers, except for E where a thermal cycling protocol with Taq polymerase and dye-labelled T7 or T3 primers was used (Applied Biosystems, Foster City, CA) using a Perkin Elmer 4800 (Norwalk, CT) thermal cyclor and an ABI 373A sequencer (Applied Biosystems).

Sequences were edited manually to remove vector and ambiguous sequences at the end. The average length of usable sequence was 340 bp and 275 bp for automatically and manually determined sequences, respectively. Sequences were translated in the six reading frames and compared with the protein databases SwissProt and PIR using FASTA, or against the NCBI (National Centre for Biotechnology Information, Bethesda, MD) non-redundant protein database using the NCBI's GenInfo network BLAST (Altschul *et al.*, 1990) server. PAM120 scores greater than 80 were considered to indicate probable hits. Nucleotide sequences were compared with GenBank using BLASTN (cut-off PAM120 score: 150). In many cases the sequences were manually inspected to correct base assignment errors or compressions giving rise to frame shifts in the coding sequence. Comparison of cDNAs against themselves was done using FASTA, sequences with greater than 90% overlap on a minimum stretch of 50 nt were considered to be overlapping.

Acknowledgement

We wish to thank Claude Paoletti, head of the Life Sciences Department of the CNRS for encouraging the development of this initiative.

References

- Adams, M. D., Dubnick, M., Kerlavage, A. R., Moreno, R., Kelley, J. M., Utterback, T. R., Nagle, J. W., Fields, C. and Venter, J. C. (1992) Sequence identification of 2375 human brain genes. *Nature*, **355**, 632–634.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
- Arondel, V., Lemieux, B., Hwang, I., Gibson, S., Goodman, H. and Somerville, C. R. (1992) Map-based cloning of a gene controlling omega-3 fatty acid desaturation in *Arabidopsis*. *Science*, **258**, 1353–1355.
- Bachmair, A., Becker, F. and Schell, J. (1993) Use of a reporter transgene to generate *Arabidopsis* mutants in ubiquitin-dependent protein degradation. *Proc. Natl Acad. Sci. USA*, **90**, 418–421.
- Blackbourn, H. D., Parker, P., Huskisson, N. S. and Battey, N. H. (1992) Properties and partial protein sequence of plant annexins. *Plant Physiol.* **99**, 864–871.
- Cordes, S., Deikman, J., Margossian, L. J. and Fischer, R. L. (1989) Interaction of a developmentally regulated DNA-binding factor with sites flanking two different fruit-ripening genes from tomato. *Plant Cell*, **1**, 1025–1034.
- Deikman, J. and Fischer, R. L. (1988) Interaction of a DNA binding factor with the 5'-flanking region of an ethylene-responsive fruit ripening gene from tomato. *EMBO J.* **7**, 3315–3320.
- Estelle, M. A. and Somerville, C. R. (1987) Auxin-resistant mutants of *Arabidopsis thaliana* with an altered morphology. *Mol. Gen. Genet.* **206**, 200–206.
- Flavell, R. (1980) The molecular characterization and organization of plant chromosomal DNA sequences. *Ann. Rev. Plant Physiol.* **31**, 569–596.
- Genschik, P., Philipps, G., Gigot, C. and Fleck, J. (1992) Cloning and sequence analysis of a cDNA clone from *Arabidopsis thaliana* homologous to a proteasome α subunit from *Drosophila*. *FEBS Lett.* **309**, 311–315.
- Giraudat, J., Hauge, B. M., Valon, C., Smalle, J., Parcy, F. and Goodman, H. M. (1992) Isolation of the *Arabidopsis* ABI3 gene by positional cloning. *Plant Cell*, **4**, 1251–1261.
- Goldberg, R. B., Hoschek, G. and Kamalay, J. (1978) Sequence complexity of nuclear and polysomal RNA in leaves of the tobacco plant. *Cell*, **14**, 123–131.
- Green, P., Lipman, D. L. H., Waterston, R., States, D. and Claverie, J.-M. (1993) Ancient conserved regions in new gene sequences and protein databases. *Science*, **259**, 1711–1716.
- Hauge, B. M. and Goodman, H. M. (1992) Genome mapping in *Arabidopsis*. In *Methods in Arabidopsis Research* (Koncz, C., Chua, N.-H. and Schell, J., eds). London: World Scientific, pp. 191–223.
- Höög, C. (1991) Isolation of a large number of novel mammalian genes by a differential screening strategy. *Nucl. Acids Res.* **19**, 6123–6127.
- Keith, C. S., Hoang, D. O., Barret, B. M., Feigelman, B., Nelson, M. C., Thai, H. and Baysdorfer, C. (1993) Partial sequence analysis of 130 randomly selected maize cDNA clones. *Plant Physiol.* **101**, 329–332.
- Kimura, N., Shimada, N., Nomura, K. and Watanabe, K. (1990)

- Isolation and characterization of a cDNA clone encoding rat nucleoside diphosphate kinase. *J. Biol. Chem.* **265**, 15 744–15 749.
- Krebbbers, E., Herdies, L., De Clercq, A., Seurinck, J., Lee-mans, J., Van Damme, J., Segura, M., Gheysen, G., Van Montagu, M. and Vandeckerckhove, J.** (1988) Determination of the processing sites of an *Arabidopsis* 2S albumin and characterization of the complete gene family. *Plant Physiol.* **87**, 859–866.
- Ludevid, D., Höfte, H., Himelblau, E. and Chrispeels, M. J.** (1992) The expression pattern of the tonoplast intrinsic protein λ -TIP in *Arabidopsis thaliana* is correlated with cell enlargement. *Plant Physiol.* **100**, 1633–1639.
- Mandel, M.A., Gustafson-Brown, C., Savidge, B. and Yanof-sky, M. F.** (1992) Molecular characterization of the *Arabidopsis* floral homeotic gene APETALA1. *Nature*, **360**, 273–277.
- McCombie, W. R., Adams, M. D., Kelley, J. M., FitzGerald, M. G., Utterback, T. R., Khan, M., Dubnick, M., Kerlavage, A. R., Venter, J. C. and Fields, C.** (1992) *Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues. *Nature Genet.* **1**, 124–131.
- Minet, M., Dufour, M. and Lacroute, F.** (1992) Complementation of *Saccharomyces cerevisiae* auxotrophic mutants by *Arabidopsis thaliana* cDNAs. *Plant J.* **2**, 417–422.
- Nagy, J. I. and Maliga, P.** (1976) Callus induction and plant regeneration from mesophyll protoplasts of *Nicotiana sylvestris*. *Z. Pflanzenphysiol.* **78**, 453–455.
- Patanjali, S. R., Parimoo, S. and Weisman, S. M.** (1991) Construction of a uniform-abundance (normalized) cDNA library. *Proc. Natl Acad. Sci. USA*, **88**, 1943–1947.
- Pearson, W. R. and Lipman, D. J.** (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
- Rédei, G. P. and Koncz, C.** (1992) Classical mutagenesis. In *Methods in Arabidopsis Research* (Koncz, C., Chua, N.-H. and Schell, J., eds). London: World Scientific, pp. 16–82.
- Reed, J. W., Nagpal, P., Pool, D. S., Furuya, M. and Chory, J.** (1993) Mutations in the gene for the red/far-red light receptor phytochrome B alter cell elongation and physiological responses throughout *Arabidopsis* development. *Plant Cell*, **5**, 147–157.
- Regad, F., Bardet, C., Tremousaygue, D., Moisan, A., Les-cure, B. and Axelos, M.** (1993) cDNA cloning and expression of an *Arabidopsis* GTP-binding protein of the ARF family. *FEBS Lett.* **316**, 133–136.
- Sambrook, J., Fritsch, E. F. and Maniatis, T.** (1989) *Molecular Cloning: Laboratory Manual, 2nd Edn.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
- Tabor, S. and Richardson, C. C.** (1987) DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proc. Natl Acad. Sci. USA*, **84**, 4767–4771.
- Tollervey, D., Lehtonen, H., Jansen, R., Kern, H. and Hurt, E. C.** (1993) Temperature-sensitive mutations demonstrate roles for yeast fibrillarin in pre-rRNA processing pre-rRNA methylation and ribosome assembly. *Cell*, **72**, 443–457.
- Uchimiya, H., Kidou, S.-I., Shimazaki, T., Aotsuka, S., Takamatsu, S., Nishi, R., Hashimoto, H., Matsubayashi, Y., Kidou, N., Umeda, M. and 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.
- Vervoerd, T. C., Dekker, M. M. and Hoekema, A.** (1989) A small scale procedure for the rapid isolation of plant RNAs. *Nucl. Acid Res.* **17**, 2362.
- Waterston, R., Martin, C., Craxton, M. et al.** (1992) A survey of expressed genes in *Caenorhabditis elegans*. *Nature Genet.* **1**, 114–123.