© 1996 Oxford University Press
Nucleic Acids Research, 1996, Vol. 24, No. 21 **4313–4318**

Sequence analysis of an 81 kb contig from *Arabidopsis thaliana* chromosome III

Françoise Quigley*, Patrick Dao, Annick Cottet and Régis Mache

Laboratoire de Génétique Moléculaire des Plantes, Université Joseph Fourier et Centre National de la Recherche Scientifique, BP 53, 38041 Grenoble cedex 9, France

Received June 18, 1996; Revised and Accepted September 12, 1996

DDBJ/EMBL/GenBank accession nos[+]

ABSTRACT

The nucleotide sequence of a 81 493 bp contig from *Arabidopsis thaliana* chromosome III has been determined together with 11 corresponding cognate cDNAs. Analysis of the finished sequence and comparison with public databases indicates a gene density of one gene per 4527 bp and identifies 17 novel genes, 10 of which are totally unknown or have no well-defined function. In addition, the contig contains part of a non-LTR retrotransposon and large direct and inverted repeats. Contig analysis also provides information on the structure and genomic organization of plant genes.

INTRODUCTION

The biological importance of genome sequencing is now well recognized. Information concerning new gene identity and organization, gene density and the structure of intergenic regions and specific genetic elements is being obtained. Comparison of sequence information across species brings new insights into the evolution of organisms. At the present time, total genome sequencing of the three model organisms *Haemophilus influenzae* (1) *Mycoplasma genitalium* (2) and yeast has been achieved. The next logical step is sequencing of higher eukaryote complex genomes and in this respect the flowering plant *Arabidopsis thaliana*, a member of the Cruciferae (Brassica) family, is an excellent candidate as a model organism. *Arabidopsis thaliana* is particularly well suited to large-scale systematic sequencing, with one of the smallest genomes observed in higher plants and a low content of interspersed repetitive DNA (3). The haploid genome of this plant is distributed into five chromosomes.

A European project, ESSA (European Scientists Sequencing *Arabidopsis*), has been established to sequence large contigs of chromosome IV and other regions dispersed throughout the genome. As part of this project we sequenced a 81 493 bp contig from chromosome III.

A major problem in sequence analysis of a large contig is gene identification and the determination of exons and introns and especially of their exact position. To solve this problem we have used sequence information from cognate cDNAs and expressed sequence tags (ESTs) together with informatics programs trained for the *Arabidopsis* genome. We thus identified 18 genes and the remnants of a non-LTR retrotransposon on the 81 kb contig. Results of this analysis are presented here.

MATERIALS AND METHODS

Chromosome walking

The λ clones covering the contig were generated from two *A.thaliana* ecotype Columbia genomic libraries in EMBL3 and λ GEM11 vectors (the latest from J.T.Mulligan). Selected λ clone DNAs were purified and subcloned into pUC or pBluescript using standard methods.

Cognate cDNAs

Cognate cDNAs were obtained by screening λ ZAPII cDNA libraries prepared from young shoots (a gift from M.Kreis, Orsay), cell culture (a gift from B.Lescure, Toulouse) or by RT–PCR using 20mer synthetic primers covering the putative gene length followed by cloning into pUC.

Computer assisted analysis

Sequence assembly was performed using the assembly program from the Genetics Computer Group (GCG, University of Wisconsin), version 7.2. Completed contigs were analyzed at the Martinsried Institute for Protein Sequences (MIPS) with GENFINDER trained for *Arabidopsis* (C.Wilson and S.Klosterman), FINDORF from Dr S.Liebl and other gene identifying algorithms.

Specific analysis and databank searches were performed with a program package available at INFOBIOGEN (French National Centre for Bioinformatics). The G+C content was calculated with FRAQCES from the SQX package. FASTA (4) and BLAST (GCG, version 8.1) were used in combination with public, daily updated databases. Exon–intron boundaries were examined when necessary with the program NetPlantGen from the Center for Biological Sequence Analysis (Denmark).

RESULTS

Assembly of the sequence

The 81 kb sequence was determined from a set of eight partially overlapping λ clones obtained by chromosome walking. The starting point of the walk was a λ clone containing the gene coding for eRF1-3, a eukaryotic peptide chain early release factor (5). At each walking step, several independent overlapping λ clones were obtained and their relative position mapped by restriction analysis. The mapping data was confirmed by end sequencing of both end and common restriction fragment subclones of the λ

*To whom correspondence should be addressed. Tel: +33 76 51 48 93; Fax: +33 76 51 43 36; Email: fquigley@grenet.fr

[+]X97484–X97488, X97616, X97826–X97829, X97970, X98130

Table 1. List of ORFs/genes and genetic element identified in the 81 kb contig and short description of their homologies

ORF/genes	Size (aa)	Introns no.	Cognate cDNA acc.	Similar database entries			% identity / aa	Score* (FASTA)
				Acc. no.	Name	Description		
ORF 01	1122	1	X97970	P24384	PRP22	Pre-mRNA splicing factor RNA helicase *Saccharomyces cerevisiae*	47.4 / 980	2458
ORF 02	587	2	X97484	P45268	HI1604	Phosphate permease *Haemophilus influenza*	35.4 / 195	410
ORF 03	350	3	X97485				-	-
ORF 04	446	7	X97826	Q10085		Hypothetical protein 49,1 KD *Saccharomyces cerevisiae*	27.8 / 370	427
ORF 05	616	0	X97616				-	-
ORF 06	475	6	X97828	P35336	PGA	Polygalacturonase precursor (EC 3.2.1.15) *Actinidia chinensis* (Kiwi)	41.6 / 344	743
ORF 07	910	0		P14381	TX1ORF2	Non-LTR retrotransposon Reverse transcriptase *Xenopus laevis* transposon TX1 (ORF2)	23.1 / 835	591
ORF 08	435	0	X97486	P46055	eRF1	Eukaryotic peptide chain early release factor subunit1 -human-	74.6 / 426	1638
ORF 09	455	0	X97487				-	-
ORF 10	346	0	X97488			Beta transducin repeats protein (TRP-ASP repeats)	-	-
ORF 11	350	4	M64114	P25856	GAPDH	GAPDH glyceraldehyde-3-phosphate dehydrogenase subunit A (EC 1.2.1.13) *Arabidopsis thaliana*	100 / 450	1649
ORF 12	484	8	X97827	P30620	SNM1	DNA cross-link repair protein *Saccharomyces cerevisiae*	37.7 / 130	237
ORF 13	>59						-	-
ORF 14	374	6		Q06548	APK1A	Protein kinase *Arabidopsis thaliana* (EC 2.7.1.-)	33.3 / 354	518
ORF 15	217	3					-	-
ORF 16	990	26		P34098	MANA	Alpha-mannosidase precursor (EC 3.2.1.114) - *Dictyostelium discoideum*	39.9 / 627	1082
ORF 17	441	0				Zinc finger signature	-	-
ORF 18	191	4					-	-
ORF 19	141	2	X97829	Q03200	LIR1	Light regulated protein precursor *Oryza sativa* (rice)	46.2 / 104	249

*Highest scoring homologue from SwissProt. Only scores > 200 were considered.

inserts, in order to make sure that each region of the contig was covered by at least two independent λ clones. The collinearity of the contig was finally confirmed by hybridization to independent YAC clones (CIC5H4, CIC6B11 and CIC11D3) covering this region of chromosome III (data not shown).

The sequencing strategy evolved from directed sub-cloning and manual sequencing at the start of the project to a random approach (shotgun cloning into M13) followed by automated fluorescent sequencing after DNA nebulization of overlappping λ inserts subcloned into pUC. At this stage sequence assembly was performed using the assembly program of the GCG package, making sure that both DNA strands were covered. Eventual ambiguities were solved by resequencing or directed sequencing with specific primers.

The cognate cDNAs identification and sequencing which was undertaken during this work allowed us to verify the accuracy of the genomic sequencing. A 23.8% coverage of the 81 kb contig was realized by sequencing of the cognate cDNAs on one strand at least. The overall sequence accuracy achieved was 99.98%, a value of the same order as that reported in similar studies.

At this stage, the finished sequence was sent to MIPS, where gene identification analyses were performed. More specific sequence analyses and similarity searches of the public sequence databases as described in Materials and Methods were done locally.

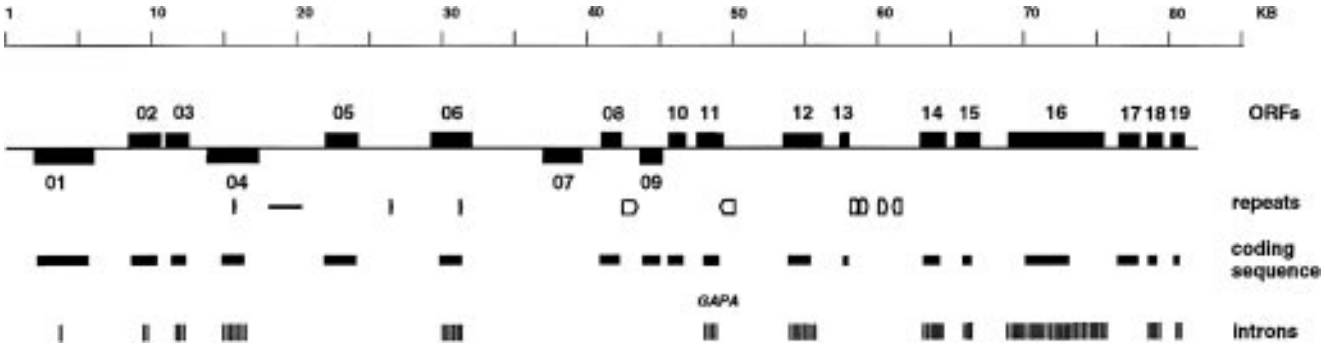

Figure 1. General organization of the 81 kb contig. The contig is oriented 5′→3′ from left to right and constitutive elements, ORFs, repeats and coding sequences are drawn to scale. Genes/ORFs numbered from 01 to 19 are indicated by boxes above the black line for the Watson strand and beneath the line for the Crick strand. Gene coding regions are represented by thick black lines beneath the corresponding genes. Large repeats are represented by open pointed boxes. Short AT repeats are represented on the same line by narrow pointed boxes. The thick black line indicates the region similar to the *RPL*12A–*RPL*12B interspace. The number of introns for each gene is represented by bars underneath the corresponding coding region.

The annotated sequence was submitted to the EMBL databank (accession no. X98130). Cognate cDNA accession numbers are listed in Table 1.

Chromosomal location of the contig

At an early stage in the project similarity searches of the public databases showed a region located ~5000 bp downstream of the sequencing starting point presenting a 100% match with *GAP*A, the gene for subunit A of glyceraldehyde-3-phosphate dehydrogenase (GAPDH). The latest map using RI lines released for chromosome III shows *GAP*A at 56 cM down the chromosome (6). The orientation of the 81 kb contig relative to *GAP*A on chromosome III is not known.

Definition of ORFs and candidate genes

A total of 19 ORFs and putative genes were identified on the 81 kb contig. The identified genes were named ORFs 01–19 as a working nomenclature and will be mentioned indifferently as ORF or gene in the text.

The ORFs and putative genes were identified according to a combination of the following principles: the use of the program GENEFINDER trained for plant sequences, comparison with public databases, localization of expressed sequence tags (ESTs) on the sequence and, finally, examination of exon–intron boundaries with the program NetPlantGen (see Materials and Methods).

An essential source of additional information was the identification and sequencing of the cDNAs corresponding to 11 of the putative genes (Fig. 1). Three of the putative genes were characterized by sequencing the corresponding EST clones: ORF 15 (accession no. T22276); ORF 16 (accession nos R65277 and T46660); ORF 18 (accession no. T43095). The sequence data generated by sequencing the cognate cDNAs and ESTs was used to define or confirm exons and gene boundaries and define splice sites exactly.

One element on the sequence, ORF 07, was identified as a remnant of a non-LTR retrotransposon by comparison with databank entries (Table 1). Consequently, it was not taken into account in the analysis of the contig presented below.

Sequence organization and composition

Genes account for 44% of the sequence (30% for exons alone), with a gene every 4500 bp, which is within the range of 2–7 kb predicted prior to large-scale sequencing data being available (7). The average gene length including introns and the 3′ non-coding region is 2300 bp.

The different genes are dispersed evenly along the region, with some appearance of clustering (ORFs 01–04, 08–11 and 14–19). Intergenic space varies from 327 to 5293 bp, with a mean value of 2300 bp. Some differences in intergenic mean values are observed according to gene orientation: terminator followed by promoter (1960 bp), convergent promoters (1500 bp) and terminators (2600 bp). The most striking feature of the gene organization in the 81 kb region is the unequal gene distribution between the Watson and Crick strands, with 15 (35% occupancy) and three genes (11% occupancy) respectively.

The average base composition for the 81 kb sequence is 36.1% G+C. As expected, coding sequences have a higher than average G+C content relative to non-coding regions, with average values of 43.9% (40.1–47%) and 32.6% (29.6–38.1%) respectively. Intron G+C base composition varies from 24.6% in ORF 06 to 36.6% in ORF 03.

Gene structure

A total of 72 introns were identified in 12 genes (Fig. 1), the number of introns per gene ranging from 0 to 26. A surprisingly high number of genes had no introns: ORFs 05, 08, 09, 10 and probably 17. Genes with a high number of small sized introns of average length 149 bp have already been identified in *Arabidopsis* (8). The intron mean AU content of 68% of these introns corresponds to the AU composition found for dicotyledon plant introns (9). Confirming the observations of Goodall and Filipowicz (10), none of the introns contains the consensus *Saccharomyces cerevisiae* branching sequence (UACUAAC). Also, there is no 3′ polypyrimidine region necessary for the branching lariat in the course of vertebrate intron splicing.

Cognate cDNA and EST sequence data allowed precise definition of 45 donor and 47 acceptor intron sites, thus giving the following consensus sequence: CAG/GTAAGT...TKCAG/GTT, with 100% conservation at the GT...AG borders. This data agrees with the consensus sequences established from nearly 900 databank

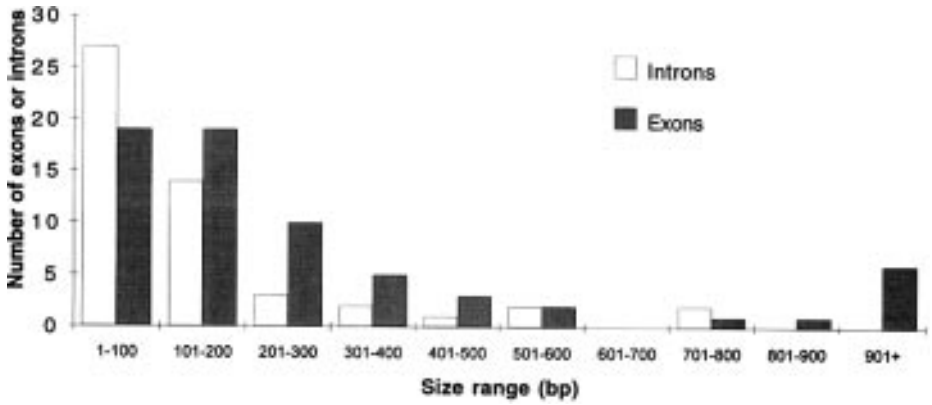

Figure 2. Distribution of exon and intron lengths. A total of 89 exons and 72 introns belonging to 17 genes were analyzed. ORFs 07 and 13 were omitted.

entries (11), with a slight variation in the acceptor site sequence found to be TGYAG/GT.

Exons are relatively small in size, the majority being <67 codons (Fig. 2). The smallest one with 12 codons is found in ORF 06. The larger exon size corresponds to ORF 01, with 991 codons.

Gene function

Comparison of the contig nucleotide sequence and of the predicted protein products with public databases gives indications on the putative functions of the genes (Table 1). Twelve of the 18 genes find significant matches with proteins from *Arabidopsis* or other organisms. The function of seven of these putative proteins has been determined experimentally in *Arabidopsis* or other organisms. The remaining five present a protein signature or are similar to genes of unknown function found in other organisms. The last six genes do not exhibit any significative similarities with databank entries. A variety of different functions are represented, with, however, a greater representation of fundamental functions, but it may just be that such genes are well represented in public databases and thus more likely to be identified (Table 1). There is apparently no clustering of genes with related functions or belonging to the same gene family. The latter could have been expected, as gene family members have been found to be organized in tandem arrays in *Arabidopsis* (12).

Closely related genes can be identified by searching dbEST for sequences showing a high score with the query sequence. Visual inspection of the resulting data allows distinction between ESTs identical to the genomic sequences and ESTs which differ slightly and are indicative of closely related genes. The summary of this search for the contig coding sequences is shown in Table 2 and it can be concluded that nearly half of the genes found on the contig show similarity to other genes. As an example, the gene encoding the eukaryotic peptide chain release factor eRF1-3 (ORF 08) is closely related to two other genes, one of them, coding for eRF1 2, being highly similar, with a 87% nucleotide identity where comparing partial (915 bp) cDNA sequences (5). The gene corresponding to eRF1-2 has been linked to marker m219 on the top arm of chromosome I (P.Carol, personal communication) and is an example of gene redundancy between two chromosomes.

Database searches at the protein level also give interesting information on the conservation of gene functions between organisms at different stages of evolution (Table 2; see Discussion).

Repeats

Several types of repeated sequences are present in the 81 kb region. Figure 1 indicates the locations of the larger repeats. Three types of repeats have been found: inverted, tandem and interdispersed. Only repeats >30 bp were considered. The most striking repeats are two large inverted repeats located in the *GAP*A vicinity. The two repeated segments show 96% identity in a 792 bp overlap. One repeat is located in the intergenic region between ORF 08 and ORF 09 and includes the ORF 09 putative promoter region. The corresponding inverted segment is located downstream of *GAP*A and starts within the 3′ non-coding trailer of that gene. The 5798 bp loop thus contains *GAP*A, ORF 09 and ORF 10 and the total region corresponds to 7645 bp end-to-end.

The second inverted repeat is located in the intergenic region between ORFs 13 and 14. The repeated segments are 76.5% identical over a 503 bp overlap with a putative loop of 362 bp. Upstream of this inverted repeat, still in the same intergenic region, is located the only large tandem repeat found in the 81 kb contig. The two repeated segments are next to each other and present 68.1% identity in a 503 bp overlap. The intergenic region between ORFs 13 and 14 is relatively G+C poor (29.6%). The presence of these two features and the relatively large AT content of this region may be related.

Another feature similar to those described above is also observed for a region located between ORFs 04 and 05. This region shows 69.5% identity in a 1922 bp fragment with a region corresponding to part of the intergene space of two genes, *RPL*12A and *RPL*12B, coding for chloroplast ribosomal proteins (13). The duplicate region is likewise AT-rich (G+C = 28.3%). It is not known whether *RPL*12A and *RPL*12B are located on chromosome III or another chromosome. Such long range duplications/inversions have been described for large genomes, in particular *Caenorhabditis elegans* (14).

Short repetitive sequences of the dinucleotide repeat type defined for microsatellite motifs are found scattered in the intergenic and intronic regions of the 81 kb contig. The three most significant of these $AT_{(15–17)}$ repeats are indicated in Figure 1. Two of these AT repeats are located in ORF 04 and ORF 06 introns respectively. The different motifs with a length of over six repeats were AA/TT, AT/TA and CT/GA. This is in agreement with published data (15). These motifs were found in relatively equal abundance.

Table 2. Overview of identical and similar *Arabidopsis* ESTs and cross-phylum matches for *Arabidopsis* novel genes

ORFs/genes	*Arabidopsis* ESTs		*Haemophilus influenzae*	yeast	*Caenorhabditis elegans*	human
	identical	similar				
ORF 01	0	1	P45018 (1124)	P24384 (2458)	P34498 (2124)	A56236* (3677)
ORF 02	0	0	P45268 (410)	P38361 (317)	U50312* (261)	L20559+ (379)
ORF 03	4	0	-	-	-	-
ORF 04	4	3	P45272 (297)	Q10085 (427)	-	-
ORF 05	0	1	-	-	-	-
ORF 06	0	3	-	-	-	-
ORF 07	0	5	-	-	-	-
ORF 08	0	6	P12385 (1502)	P46055 (1638)		
ORF 09	0	0	-	-	-	-
ORF 10	0	3	-	-	-	-
ORF 11	120	0	/	/	/	/
ORF 12	1	0	-	P30620 (237)	-	D42045+ (997)
ORF 13	1	0	-	-	-	-
ORF 14	0	4	-	-	-	-
ORF 15	1	0	-	-	-	-
ORF 16	5	0	-	-	U63998* (282)	D63998+ (453)
ORF 17	0	0	-	S53414* (249)	-	-
ORF 18	1	0	-	-	-	-
ORF 19	43	0	-	-	-	-

dbEST search was performed using BLASTN. Only scores > 400 were considered. The highest scoring database entry is shown for each organism. FASTA scores are indicated beneath each entry. Scores < 200 were not considered significant and are indicated by a dash. Database entries are from SwissProt, PIR* and GenBank+. ORF 11 (*GAP*A), a gene of eubacterial origin (16) is omitted from the cross-phylum matches list.

Expression

The number of *Arabidopsis* ESTs present in the databank is at this time >28 000 (release 052396). The data represented by the collection of ESTs can be used to give a first insight into the expression of a gene, by looking at the number of identical EST hits found for each gene candidate (Table 2). From this data it can be seen that five of the genes are relatively well expressed. One of these genes, *GAP*A, coding for a key enzyme in chloroplast metabolism, is highly expressed, as expected. Four genes are hit by only one EST. The last nine genes are presumably rarely expressed under the conditions under which the cDNA libraries were made. A pattern for trancription according to the location of genes on the contig is not evident and physical closeness of the genes does not apparently mean the same pattern of expression. The two genes *GAP*A (17) and ORF 19 (18,19), which have been experimentally demonstrated to be regulated by light, are >30 kb apart.

DISCUSSION

The sequencing presented here of a 81 kb contig from chromosome III gives a first insight into the organization of a relatively large fragment of the *A.thaliana* genome.

The gene density of one gene every 4500 bp confirms estimations derived from other data for *Arabidopsis* (7) and is of the same order as one gene every 5 kb found in *C.elegans*, which has a genome size (100 kb) comparable with that of *Arabidopsis*. The gene density could reach a higher value on a small fragment. As an example, four genes (ORFs 8–11) are concentrated within a 9000 bp fragment in the 81 kb contig. A similar density has also been found in other parts of the *Arabidopsis* genome (19).

Many of the features described here for the *Arabidopsis* 81 kb contig and relative to the contig structure and gene pattern have been observed on a larger scale in the *Caenorhabditis* genome, of which 2.2 Mb have been sequenced and analysed. The similarities include the organization of gene families, the presence of

remnants of genetic elements and, more importantly, the presence of complex duplicated and inverted genomic pieces. The 81 kb fragment sequenced presents, however, its own characteristics. All the genes identified as members of gene families are of the dispersed type: no close family member is in direct vicinity on the contig. Another characteristic is the apparently high number of genes without introns (five out of 18).

At this point, the EST information generated by the EST sequencing effort and the EST assembly done at TIGR (The Institute for Genome Research) can be used to estimate the total number of genes in *Arabidopsis*. At the moment, there are 28 007 ESTs deposited in dbEST (release 052396). The latest assembly (release 1.1) was performed on 25 282 ESTs, which after assembly gave a total of ~13 600 different sequences. Of the 18 genes identified in the contig, nine were hit by ESTs, which means that by extrapolation 27 200 genes could be present in the *Arabidopsis* genome. However, this estimation has to be taken with caution, due to the probable over-representation in the EST database of highly expressed genes. The number obtained in this way is close to the most recent estimation of 25 000 genes (7).

As mentioned earlier, most of the genes presenting matches with databases entries represent fundamental functions or mechanisms and present cross-phylum similarities. Some are highly conserved in distant phyla. ORF 01, an RNA helicase, is such an example, with 47 and 61.5% amino acid identity to its yeast and human genes respectively.

The availability of the complete genomes of three lower organisms in the databases (see Introduction) makes it possible to infer which of the candidate genes may be specific to multicellular organisms. Eleven of the genes identified during this work do not show any significant match with either bacterial or yeast ORFs in BLASTP searches (Table 2) and might be good candidates for these functions. Two genes, ORFs 06 and 14, a polygalacturonase and a serine–threonine kinase respectively, which belong to large gene families widespread in plants (see Table 1 for highest scoring similar genes) do not show significant cross-phylum similarity and thus could possibly be considered as specialized for plant functions. More interestingly, eight out of the nine remaining genes do not show any similarities with vertebrate genes and may be putative candidates for plant-specific genes. As an example, ORF 19, of unknown function, has been shown to be regulated by light in rice (20) and in citrus (18) plants and is very likely plant specific. The complete sequence of another organism of similar size to *Arabidospsis* is however needed before assigning plant specificity to an unknown gene.

To summarize, this work has allowed extension of the catalogue of genes detected in *Arabidopsis*, with 17 novel genes and characterization of 11 cognate cDNAs. The function of eight of these genes is totally unknown. Most of the genes described here have no homologous function described in yeast or prokaryotes and might perform differentiated functions in multicellular organisms and are of particular interest for more detailed study.

ACKNOWLEDGEMENTS

We thank M.Bevan for coordinating the ESSA Program, P.Slonimski, Director of the GREG (Groupement de Recherches et d'Etudes sur les Génomes) for his support, H.W.Mewes, S.Klosterman and M.Chalwatzis from the MIPS for computer data analysis, G.Clabault for computer assistance, Y.-F.Li for excellent sequencing support, J.Dangl and J.T.Mulligan for supplying the *Arabidopsis* λ GEM11 library, G.Picard and J.Lafleuriel for helping with the YAC library hybridization, the *Arabidopsis* Biological Resource Center at Ohio State for supplying ESTs clones, D.Tremoussaygue for the cell culture cDNA library and M.Kreis for the green shoots cDNA library. This work was supported by the EU under the ESSA program and by the the French government under the GREG program.

REFERENCES

1 Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,R., Bult,C.J., Tomb,J.F., Doughery,B.A., Menick,J.M. *et al.* (1995) *Science*, **269**, 496–512.
2 Fraser,C.M., Gocayne,J.D., White, O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M. *et al.* (1995) *Science*, **270**, 397–403.
3 Dean,C. and Schmidt,R. (1995) *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, **46**, 395–418.
4 Pearson,W.R. and Lipman,D.S.(1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.
5 Brown,C.M., Quigley,F.R. and Miller,W.A. (1995) *Plant Physiol.*, **110**, 336.
6 Lister,C. and Dean,C. (1995) *Weeds World*, **2**, 11–18.
7 Meyerowitz,E.M. (1994) In Meyerowitz,E.M. and Somerville,C.R. (eds), *Arabidopsis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 89–120.
8 Larkin,R. and Guilfoyle,T. (1993) *Nucleic Acids Res*, **21**, 1038.
9 Goodall,G.J. and Filipowicz,W. (1991) *EMBO J.*, **10**, 2635–2644.
10 Goodall,G.J. and Filipowicz,W. (1989) *Cell*, **58**, 473–483.
11 Korning,P.G., Hebsgaard,S.M., Rouzé,P. and Brunak,S. (1996) *Nucleic Acids Res.*, **24**, 316–320.
12 Chaubet,N., Chaboute,M.-E., Philips,G. and Gigot,C. (1987) *Dev. Genet.*, **8**, 461–473.
13 Wegloehner,W. and Subramanian,A.R. (1994) *J. Biol. Chem.*, **269**, 7330–7336.
14 Wilson,R.K. *et al.* (1994) *Nature*, **368**, 32–38.
15 Lagercrantz,U., Ellegren,H. and Andersson,L. (1993) *Nucleic Acids Res.*, **21**, 1111–1115.
16 Martin,W., Brinkmann,H., Savona,C. and Cerff,R. (1994) *Proc. Natl. Acad. Sci. USA*, **90**, 8692–8696.
17 Conley,T.R., Park,S.C., Kwon,H.B., Peng,H.P. and Shih,M.C. (1994) *Mol. Cell. Biol.*, **14**, 2525–2533.
18 Abied,M.A. and Holland,D. (1994) *Plant Mol. Biol.*, **26**, 165–173.
19 Le Guen,L., Thomas,M. and Kreis,M. (1994) *Mol. Gen. Genet.*, **245**, 390–396.
20 Reimmann,C. and Dudler,R. (1993) *Plant Mol. Biol.*, **22**, 165–170.