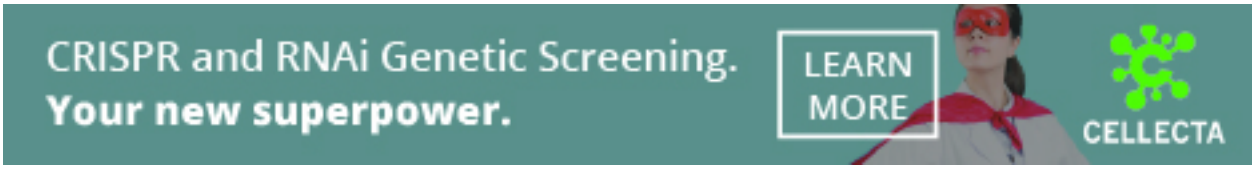
Genome DNA sequencing around the EF-1 alpha multigene locus of Arabidopsis thaliana indicates a high gene density and a shuffling of noncoding regions.

D Tremousaygue, C Bardet, P Dabos, et al.

Genome Res. 1997 7: 198-209
Access the most recent version at doi:10.1101/gr.7.3.198



RESEARCH

Genome DNA Sequencing Around the *EF-1α* Multigene Locus of *Arabidopsis thaliana* Indicates a High Gene Density and a Shuffling of Noncoding Regions

Dominique Tremousaygue,[1] Claude Bardet, Patrick Dabos, Farid Regad, Florence Pelese, Rana Nazer, Eugene Gander, and Bernard Lescure

Laboratoire de Biologie Moleculaire des relations Plantes-Microorganismes, unite mixte de recherche (UMR) 0215, Centre National de la Recherche Scientifique (CNRS)–Institut National de la Recherche Agronomique (INRA), BP 27, 31326 Castanet Tolosan, France

In *Arabidopsis thaliana*, EF-1α proteins are encoded by a multigene family of four members. Three of them are clustered at the same locus, which was positioned 24 cM from the top of chromosome 1. A region of DNA spanning 63 kb around these locus was sequenced and analyzed. One main characteristic of the locus is the mosaic organization of both genes and intergenic regions. Fourteen genes were identified, among which only four were already described, and other unidentified are most likely present. Functionally diverse genes are found at close intervals. Exon and intron distribution is highly variable at this locus, one gene being split into at least 20 introns. Several duplications were found within the sequenced segment both in coding and noncoding regions, including two gene families. Moreover, a sequence corresponding to the 5′ noncoding region of the *EF-1α* genes and harboring a 5′ intervening sequence is duplicated and found upstream of several genes, suggesting that noncoding regions can be shuffled during evolution.

[The sequence data described in this paper have been submitted to GenBank under accession no. U63815.]

The model plant *Arabidopsis thaliana* has been chosen to elucidate plant genome organization at nucleotide level in an international sequencing effort (Goodman et al. 1995). Partial cDNA sequencing was initiated first to access the transcribed part of the genome (Höfte et al. 1993; Newman et al. 1994; Cooke et al. 1996). This approach proved to be extremely powerful and supplied to date >20,000 expressed sequence tags (EST) to public databases such as dbEST (Boguski et al. 1993). In addition to providing tags for new genes (Hervé et al. 1996), it indicates clearly that many *A. thaliana* proteins are encoded by multigene families with tissue or developmental-specific expression patterns and demonstrates that genetic redundancy is a common occurrence in plants. Nevertheless, tags for genes expressed at very low levels or in some specific conditions may escape detection by this approach. After completion of the yeast genome, sequencing of the entire *A. thaliana* genome (no more than five times bigger than *Saccharomyces cerevisiae*) was conceivable. A global view of the organization of coding and noncoding regions of the genome undoubtedly would enhance our understanding of the biological processes involved in its architecture, expression, and evolution. The European biotechnology program ESSA (for European Scientists Sequencing Arabidopsis) supported this approach for the analysis of large regions of *A. thaliana* chromosome 4 and some other loci of the genome. Herein, we extend our interest from EST (Höfte et al. 1993; Cooke et al. 1996) to genomic DNA sequencing; we have determined the sequence around a region containing three duplicated genes coding for the translation elongation factor-1α (EF-1α).

A. thaliana EF-1α is a cytoplasmic eukaryotic counterpart of EF-tu described in bacteria, a protein that binds and delivers aminoacyl tRNA to the ribosome. These genes provide a model for the understanding of the mechanisms involved in the regulation of the expression of genes encoding components of the translation apparatus (Axelos et al. 1989; Liboz et al. 1989; Curie et al. 1991, 1993; Re-

[1]**Corresponding author.**
E-MAIL tremou@toulouse.inra.fr; FAX 33-561-28-50-61.

7:198–209 ©1997 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/97 $5.00

gad et al. 1995). In *A. thaliana,* they are encoded by a small family of four genes residing in two different loci. In the present work the locus containing the three *A1–A3* genes has been mapped, and 63.1 kb of the DNA sequence around them has been determined. A high gene density is observed. Some genes are split into numerous exons, and a number of partially conserved repeated sequences are found interspersed in the noncoding regions. The implications of this kind of pilot program are evaluated.

RESULTS

Mapping of the Locus

Specific primers were used in PCR amplification experiments with two different yeast artificial chromosome (YAC) libraries to select clones carrying *EF-1α A1* gene sequences. The Erwin Grill (EG) library (Grill and Somerville 1991) was screened in our laboratory, and the Ceph/Inra/Cnrs (CIC) library (Creusot et al. 1995) was screened by David Bouchez's group in Versailles. A positive clone was found in each library and yeast DNA from these clones was analyzed on pulse-field gel electrophoresis (PFGE) as shown in Figure 1A. YAC clones EG22C6 and CICXC9 are 120 and 480 kb in length and both hybridized to an EF-1α-specific probe in a Southern blotting experiment (Fig. 1B). The YAC EG22C6 had been shown to contained the PAI1 (phosphoribosylanthranilate isomerase) probe, which was mapped previously in position 24 on chromosome 1, 3 cM above marker m488 (Li et al. 1995). We have verified that the *A1* and *PAI1* genes were also associated in YAC CICXC9 (Fig. 1C). Because the two YAC clones selected for the presence of *A1* gene also contain the *PAI1* gene, and YAC EG22C6 is only 120 kb in length, our results indicate that the *A1* and *PAI1* genes are separated by <120 kb on chromosome 1 (see Fig. 1D).

Sequence Determination and Basic Analysis

Two λ clones containing the *A1–A3* genes have been selected previously in our laboratory and partially sequenced (Axelos et al. 1989; Liboz et al. 1989). A λ contig of five clones was constructed step by step by screening a genomic library with subfragments corresponding to the ends of each newly selected clone. To permit rapid progress in walking between λ clones, the ends of the selected clones were sequenced, then new oligonucleotide pairs allowing amplification of DNA fragments at the clone extremities were chosen, thereby minimizing clone overlap in the subsequent steps. Using both strands from five λ clones (see Methods), a contiguous sequence of 63,093 bp was read, with at least a fourfold redundancy. The presence of both ends of the λ contig in YAC CICXC9 was checked by PCR amplification with specific primers, suggesting that there was no recombination in the selected λ clones (data not shown). A preliminary analysis of the sequence data was performed and the results are shown below. The 63,093-bp sequence presented here has been deposited in the GenBank database and assigned accession number U63815.

The overall percentage of A + T is ~61.6%,

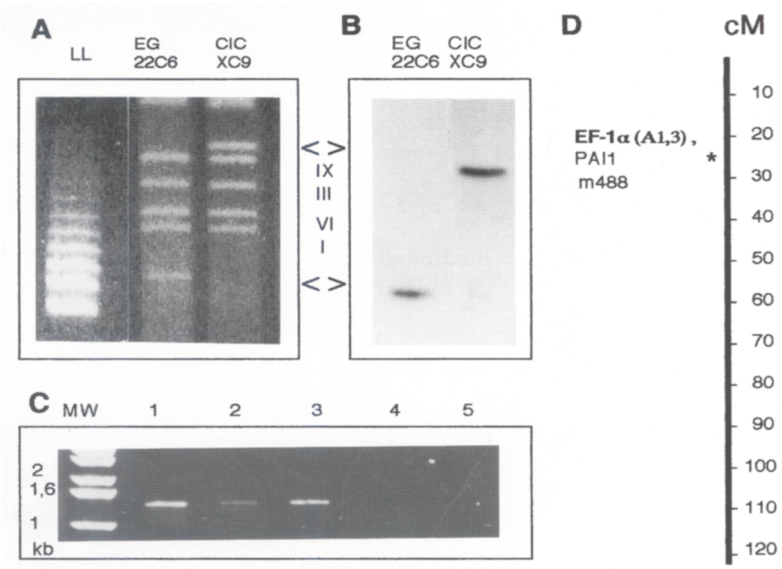


Figure 1 Characterization and mapping of YAC clones containing the *EF-1α A1–A3* genes. EG22C6 has been selected in the EG library (Grill and Somerville 1991). CICXC9 was found by screening the CIC library (Creusot et al. 1995), done by the David Bouchez group (INRA). (*A*) EtBr staining of a 1% agarose PFGE. The gel was run in 0.5× TBE at 12°C for 20 hr with a 50 sec/50 sec pulse. The positions of the YACs are indicated by open arrowheads. The yeast chromosomes are numbered at *right.* (LL) λ ladder. (*B*) Autoradiograph obtained after hybridization of a Southern blot of the gel with a probe containing a specific region of gene *A1*. (*C*) PCR amplification with specific primers to the *PAI 1* gene. (P1, 5′-AGAGGATTGAGCTTAAGGC-3′; P2, 5′-AAAGCAGCACGCGAACC-3′) of DNA from *A. thaliana,* ecotype Columbia (lane *1*); YAC EG22C6 (lane *2*); YAC CICXC9 (lane *3*); yeast strain AB1380 (lane *4*); control without DNA (lane *5*). (*D*) The asterisk (*) indicates the map position on chromosome 1. The scale is in centiMorgans.

which is slightly higher than the estimated content of the genome (58.6%) (Meyerowitz 1994), and includes some very A + T-rich regions (<71.7%). Typical microsatellite sequences described in *A. thaliana* (Depeigne et al. 1995) were searched within the sequence. A 20-bp poly(A) sequence is observed in position 2149, a 15-bp poly(T) sequence in position 5840, and two trinucleotides motifs, $(CTT)_5$ in position 2287 and $(CAC)_5$ in position 61,438.

Gene Identification and Structure

Public databases were searched for DNA similarity or homology with the BLASTN program (Altschul et al. 1990). The predicted genes were then analyzed further, using the BLASTX program, compairing the predicted amino acid sequences to protein databases and to a translation of the nucleic acid database GenBank. In addition to this preliminary work, the analysis carried out at Matinsried Institut für Protein Sequence (MIPS) used BLASTN or BLASTX programs against specialized databases, such as those containing small RNA sequences or corresponding to a translation of *A. thaliana* EST sequences. A BLASTX search was performed specifically for each long open reading frame (ORF) (>100 amino acids) against all the databases. Compilation of all the analyses suggest that, altogether with the three *EF-1*α genes, at least 14 genes are included in the sequenced area. The genes are all positioned on a schematic representation of the locus (Fig. 2). They have been named from AT.I.24-1 to AT.I.24-14, because of their location on chromosome 1 in position 24. Gene prediction programs were also run at MIPS. Among the different programs used to predict coding regions, GENEFINDER was found to be the most adaptable. Figure 3 illustrates the prediction reliability of this program for one gene that has a very complex organization of introns and exons (see the description of *AT.I.24-9* gene, below). From the GENEFINDER program at least nine more genes are proposed to be encoded in the area; however, they were not considered further in our analysis. Details of characterization of the genes and their main structural features are described below.

AT.I.24-1

The product of this gene has a slight similarity to *Ricinus communis* glutaredoxin. Glutaredoxins are member of a superfamily of proteins that catalyze the reduction of disulfide bonds using glutathione as a cosubstrate (Homlgren 1976). In the *AT.I.24-1* gene the two cysteines of the active site are conserved.

AT.I.24-2

The protein encoded by this gene shows similarity to prokaryotic uridylyltransferase, an enzyme that

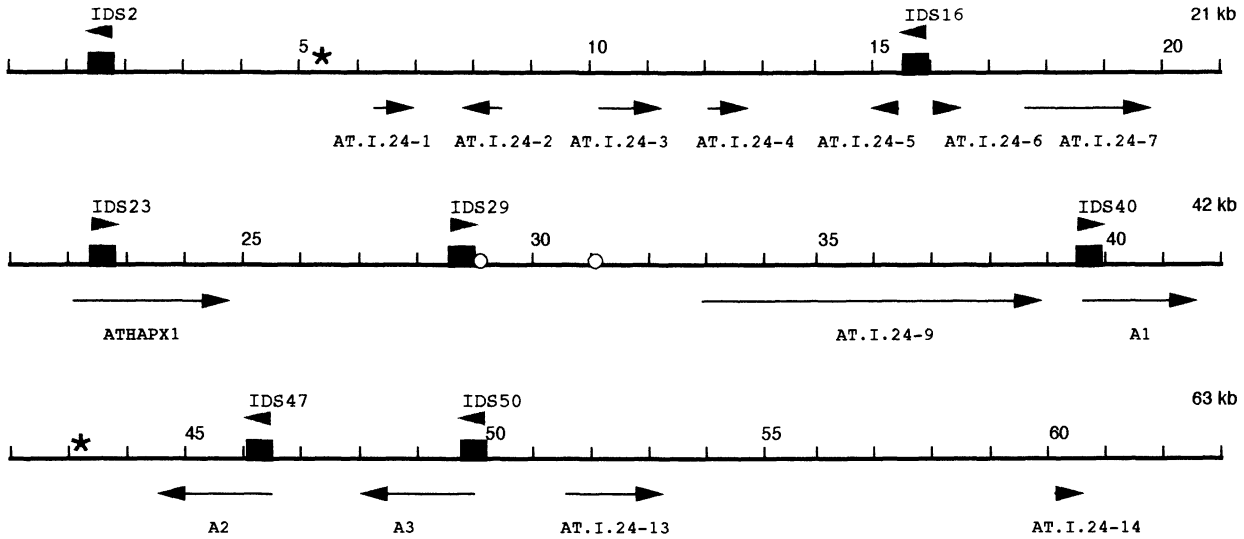


Figure 2 Organization of the 63.1-kb sequenced DNA fragment. Solid arrows indicate the location and orientation of the genes or putative genes. Solid boxes indicate the localization of duplicated sequences related to the leader intron of the *EF-1*α genes (IDS). A number is associated with each IDS, as explained in the text, arrowheads show the orientation of the duplication; open circles underline the inverted repeated 30-bp sequences; the asterisks correspond to the largest degenerated palindromes.

regulates the uridylylation or deuridylylation of protein PII involved in the adenylation process of glutamine synthetase (Merrick and Edwards 1995).

AT.I.24-3

The amino acid sequence derived from this gene presents similarity with a tomato protein. This type of protein contains two regions of similarity: (1) to the alcohol dehydrogenase family consensus sequence (with conservation of a tyrosine residue known to be involved in catalytic activity), and (2) to similar to several reported glucose dehydrogenases (Picton et al. 1993).

AT.I.24-4

This gene encodes putatively a transcription factor related to the sporamin factor 1 (SPF1) protein cloned from sweet potato (Ishiguro and Nakamura 1994). The region of homology corresponds to the DNA-binding domain of the protein.

AT.I.24-5 and AT.I.24-6

These two genes are obviously members of the same family and present similarity to rice ESTs of unknown function.

AT.I.24-7

The probable protein corresponding to this gene has slight similarity to a hypothetical protein F402 of *Esherichia coli* (Sofia et al. 1994). The cognate cDNA clone was sequenced and complete structure of the gene was obtained. The similarity to protein F402 is not restricted to a localized region of the gene, but spread all along the sequence. This similarity is optimized when a program allowing gaps in the alignment is used (17.3% identity and 37.5% positives on 555 amino acids).

AT.I.24-8

This gene was identified by homology to a previously identified cytosolic ascorbate peroxidase gene of *A. thaliana* (Kubo et al. 1993).

AT.I.24-9

A 3-kb cDNA clone containing the 3′ end of the *AT.I.24-9* gene was selected in the AC16H library (Regad et al. 1993). Twenty exons constitute the 3′ end of the gene, the shortest of them being only 32 bp long. Details about the structural features of this gene are shown in Figure 3. A leucine zipper motif follows a rather basic domain of the protein (Fig. 3), which is reminiscent of transcription factor structure. Northern type hybridization with the cognate cDNA revealed a very low expression of a 4.5-kb-long mRNA in cells at different growing stages (data not shown).

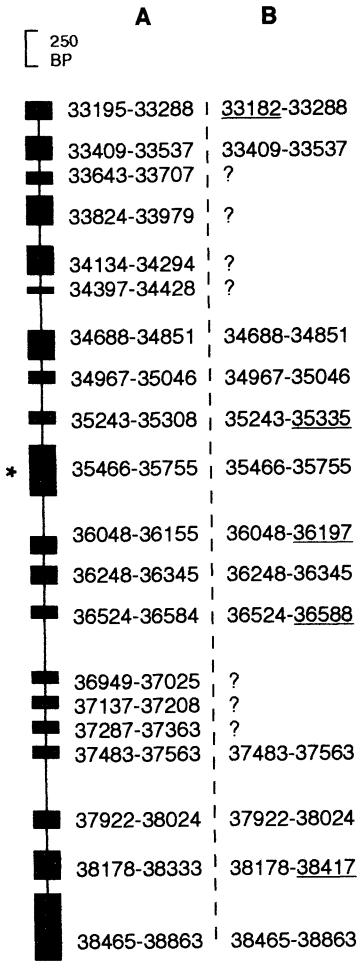


Figure 3 Structure of the *AT.I.24-9* gene. The 20 exons of the gene are represented by shaded boxes. The asterisk shows the position of the leucine zipper motif of the deduced protein. (*A*) The exon position is deduced from comparison of cDNA and genomic sequences. The position within the 63093 bp is indicated for each 5′ and 3′ extremity of exons. (*B*) The exon position is predicted by GENEFINDER. An underlined number indicates a wrong prediction of the exon end. A question mark indicates that an exon is omitted by the prediction program.

AT.I.24-10, AT.I.24-11, and AT.I.24-12

These genes encode different copies of the translation EF-1α.

AT.I.24-13

The protein possibly encoded by *AT.I.24-13* gene has similarity to a hamster protein disulfide isomerase, which catalyzes thiol-disulfide interchange in the folding of many proteins within the endoplasmic reticulum (Shorrosh and Dixon 1991). The amino-terminal end of the protein is very hydrophobic and may correspond to a signal peptide but the carboxy-terminal end of the protein does not show the KDEL signal responsible for the retention in the lumen of endoplasmic reticulum (Denecke et al. 1992).

AT.I.24-14

This gene was localized by homology to an *A. thaliana* EST. Information about the putative peptides that were detected are summarized in Table 1. The genes described in this work display completely different structures, from 2 to >20 exons. The great majority of splicing sites localized within the genes correspond to the consensus sequences (GT–AG). The only exception is for the intron in the 5′ noncoding region of the *AT.I.24-13* gene, where TT is found as splicing acceptor site. Unexpectedly, 5 of the 14 genes possess an intron in their 5′ untranslated region.

Sequence Duplications

Taking into account their different localizations and the relative divergence of their nucleotide sequences, it has been suggested that the *EF-1α A4* gene and *EF-1α A1, A2,* and *A3* genes constitute two distinct subfamilies within the *A. thaliana* genome (Axelos et al. 1989; Liboz et al. 1989). The duplication leading to the formation of the two loci probably preceded the recent duplication events at the *A1, A2, A3* locus. The coding regions of the *EF-1α A1, A2, A3* genes are indeed highly conserved (>99% similarity). The unique intervening sequence of the *A1* and *A3* genes are identical and nearly homologous to that of the *A2* gene, and even the 5′ and 3′ flanking regions are also strongly conserved. Moreover, the intervening sequences detected within the 5′ noncoding region of the *A1, A2, A3* genes appear to be well conserved (60% on ~500 bp). Surprisingly, a FASTA search on the 63-kb sequence led to the identification of an almost perfect duplicated sequence sharing 92.3% similarity on 405 bp with the leader intron sequence of the *A3* gene. This duplication located 10 kb upstream of the *A1* gene was called IDS29 (for intron duplicated sequence) between positions 28,000 and 29,000 within the 63-kb fragment (see Fig. 2). In addition to the intron duplication, the promoter sequence of the *A3* gene, from the start point of transcription to position −143, is also duplicated in 5′ position of IDS29 (Fig. 4A). Within the duplication a fragment of 57 bp including the TATA box and a putative *cis*-regulatory element of *EF-1α* genes, the telo box, is deleted (Axelos et al. 1989). Conversely, the tef box, a ubiquitous element conferring overexpression of genes in dividing cells (Regad et al. 1995), is still present in the duplication (Fig. 4A). Additional degenerated duplicated sequences, IDS2, IDS16, and IDS23 are observed (see Fig. 2). These additional duplications present 51.8% to 53.4 % similarity with each other in regions of ~400 bp. Strikingly, the position of IDS23 corresponds to the position of the 5′ intron of the gene coding for ascorbate peroxidase (Fig. 4A). Sequences that present an homology to IDS29 were also looked for using the FASTA program in plant sequences from GenBank, but the sequences found showed only weak similarities to IDS29 (data not shown). In Figure 4B, the nucleotidic sequence of IDS29 was compared to the sequences of the leader intron of the *EF-1α A1, A3* and the ascorbate peroxidase genes. To avoid artifacts attributable to a peculiar composition of the intron sequence, a shuffled composition of the same sequence was used against the same databases, but this control did not lead to similar results. Besides the *EF-1α* gene family, another gene family of two copies, *AT.I.24-5* and *AT.I.24-6,* was found. They present significant nucleic acid (63.3%) and amino acid (80.8%) similarities and are related to three similar rice ESTs.

Concerning the intergenic regions, our analysis revealed the presence of 28 palindromic sequences, ranging from 10 bp (half palindrome size) with 90% similarity to 400 bp with 50.2% similarity. The detected repeats are degenerated and their degree of similarity is in general inversely proportional to the length of the repeat. Interestingly one of these elements is located between the *EF-1α A1* and *A2* inverted repeated genes (see Fig. 2). The half palindrome shows 49.3% similarity on 300 bp and could reflect duplication events in this region. Further analysis is required to validate and elucidate the biological significance of this observation. Among nu-

Table 1. Characteristics of Predicted Genes and Deduced Protein Products

Gene name	Coding strand	Coding region (5' end)	Gene length (bp)	Exon no.	Homology Accession no. (BLASTN/BLASTX)	Similarity Accession no. BLASTN/BLASTX [high score, P (no.)]	Referring organism	Main features of deduced protein products
AT.I.24-1	W	6317	>300	>1		T04053/Z49699 (101, 6.6e-16)	RICINUS COMMUNIS	GLUTAREDOXIN
AT.I.24-2	C	8424	>500	>2		H37067/P43919 (94, 5.9e-3)	HAEM. INFLUENZE	URIDYLYL TRANSFERASE
AT.I.24-3	W	10305	>1300	>3	H37486/	/S39508 (133, 2.4e-23)	TOMATO	DEHYDROGENASE
AT.I.24-4	W	12179	>300	>2		F14100/L44134 (145, 8.5e-22)	CUCUMIS SATIVUS	PUTATIVE TRANSCRIPTION FACTOR
AT.I.24-5	C	15097	>200 =	>1		D48916/ D46687/	RICE	
AT.I.24-6	W	16096	>200 =	>1 =		D48916/ D46687/	RICE	
AT.I.24-7	W	17651	2100	3	T44566/	/S47768 (68, 2e-2)	E. COLI	HYPOT. PROT. F402
ATHAPX1	W	22485 *	2200	8	D14442/S28856		A.THALIANA	ASCORBATE PEROXIDASE
AT.I.24-9	W	33195	>6000	>20	Z35021/		A.THALIANA	PUTATIVE TRANSCRIPTION FACTOR
ATEF1A1	W	40037 *	2200	2	X16430/		A.THALIANA	EF-1α
ATEF1A2	C	45820 *	2200	2	X16431/		A.THALIANA	EF-1α
ATEF1A3	C	49258 *	2200	2	X16431/		A.THALIANA	EF-1α
AT.I.24-13	W	52149 *	1500	4	T44189/	/P38660 (61, 9.6e-4)	HAMSTER	PROTEIN DISULFIDE ISOMERASE
AT.I.24-14	C	60318	>300	>1	F14018/		A. THALIANA	

The 5' end of the identified coding region is positioned. An asterisk (*) specifies when the complete gene structure is known. When a BLASTX score was used to propose similarity to a known gene, the high score and the smallest sum probability are indicated. C and W indicate, respectively, on which DNA strand the coding sequences are found.

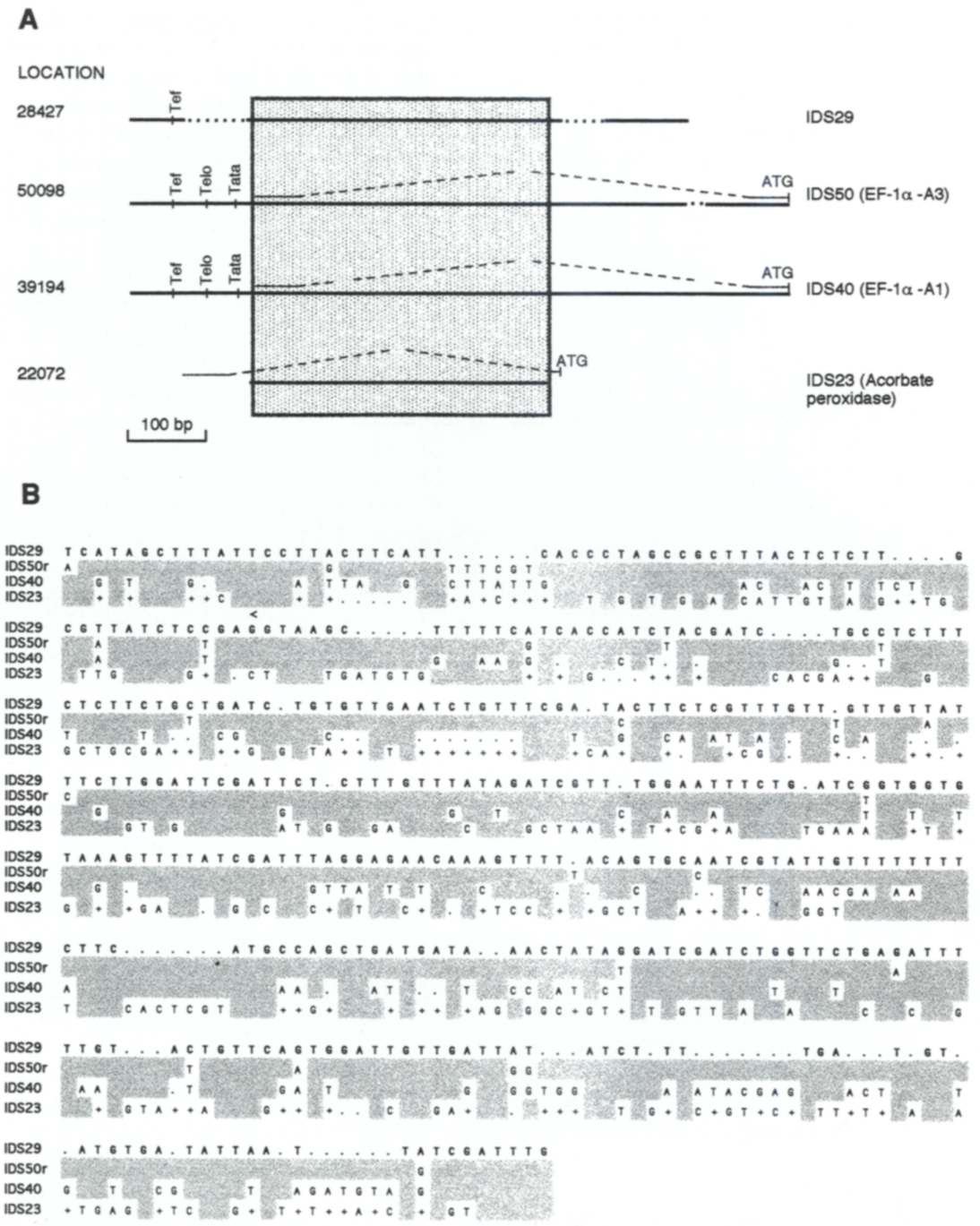


Figure 4 Comparison of IDSs. Alignment of the IDS29 sequence with related elements corresponding to the sequences of the *EF-1α A3* gene (IDS50) (Liboz et al 1989), the *A1* gene (IDS40) (Axelos et al. 1989), and the *ascorbate peroxidase* gene (Kubo et al. 1993). (*A*) A schematic illustration of the duplicated elements is represented with solid lines. The location of each element within the 63-kb sequence is indicated. Broken lines indicate a deleted region within the duplication. When it is known, the structure of the gene is shown from the start point of transcription; the location of the leader intron is dotted. The shaded area corresponds to part of the duplications compared in *B*. (*B*) Similarity among IDS29 sequence and the intron sequences of the *EF-1α A3* gene (IDS50) (Liboz et al. 1989), the *A1* gene (IDS40) (Liboz et al. 1989), and the *ascorbate peroxidase* gene (Kubo et al. 1993). The sequence of IDS29 is shown. Solid boxes indicate conserved bases. The mismatched bases are shown, and gaps are indicated by dots. In many cases, the sequence of IDS23 is identical to that of IDS40 and not to that of IDS29; the + symbol indicates these identities. The open arrowhead indicates the position of the splicing donor site within IDS40 and IDS50.

merous small repeated sequences, we observed a 30-bp inverted repeat downstream to IDS29 (see Fig. 2). Hybridization experiments and databases searches with the 2-kb sequence lying between the two 30-bp repeats did not reveal significant homology with other sequences of the *Arabidopsis* genome (data not shown), suggesting that this structure could be attributable to a local rearrangement rather to a mechanism implicating transposition of a mobile element.

DISCUSSION

A 63-kb fragment containing three of the *EF-1α* genes from the *A. thaliana* genome was cloned and fully sequenced. This work was done during the pilot program for genomic sequencing of the *A. thaliana* genome, which is now initiated on a large scale both by the European Community programs and at the Stanford genome center (USA). This program will be informative about coding regions, but the sequencing of model organisms will also increase our knowledge about plant genome structure, its expression, and evolution. Comparing genetic linkage maps, based on a common set of markers, has allowed the direct identification of colinear chromosome segments in different species (Ahn and Tanksley 1993; Moore et al. 1993; Kurata et al. 1994). Short range comparative mapping in *A. thaliana* and *Brassica* has been initiated recently using a 1.5-Mb contig of *A. thaliana* DNA surrounding the *CO* locus on chromosome 4 (Lagercrantz et al. 1996). Striking colinearity has been revealed at this locus, suggesting that gene organization within important portions of chromosomes might be conserved.

The results obtained in this study predict that a high proportion of the sequenced region codes for protein products. Fourteen genes were positioned within the sequence and mapped on chromosome 1. Considering the haploid genome size of 120 Mb, including heterochromatin, and the presence of 15,000 to 33,000 genes (Gilson and Somerville 1993), the coding capacity of the locus is probably underestimated. According to the gene prediction program GENEFINDER, nine additional genes could be encoded in the fragment, but only genes with cognate EST sequences were positioned. Intergenic distances can be very short. For example, the polyadenylation site of the *AT.I.24-9* gene and the transcription start point of the *EF-1α A1* gene are only separated by 300 bp. This small region contains *cis* elements involved in the regulation of *EF-1α* gene expression (Curie et al. 1993; Regad et al. 1995). Compared to mammals, genes in plants are very compact. Introns, as well as 5′- and 3′-untranslated regions, are short and promoter elements are usually located near the transcription initiation site (Luehrsen et al. 1994). Sensitivity to breakage is proportional to gene length. Consequently, a condensation of functional regions would result in a higher degree of tolerance of plant chromosomes to breaks and translocations (Walbot 1996). The locus analyzed in this study is in favor of such a situation.

The complementarity of the two approches of cDNA and genomic DNA sequencing is illustrated by the complex mosaic structure of the *AT.I.24-9* gene. The latter approach provides access to intron sequences, whereas the former permits their localization. It would have been impossible to position the 20 exons forming this gene without the identification and characterization of a cognate cDNA clone. Clearly, they are not all accessible by the use of gene prediction programs (Fig. 3). Such split genes are already described in *A. thaliana;* for example, the second largest subunit of RNA polymerase II contains 24 introns, although the homologous genes from yeast and *Drosophila* contain no intron and three introns, respectively (Larkin and Guifoyle 1993). This observation seems to be quite common in *A. thaliana,* as reported by other participants in the ESSA project. Alternative splicing of messengers might be involved in expression regulation of this kind of genes, as already demonstrated for some genes in plants (Kopriva et al. 1995; Montag et al. 1995).

BLASTN or BLASTX results were analyzed to attribute a putative function to all of the genes located on this locus. Nevertheless no putative function could be found for *AT.I.24-5, AT.I.24-6, AT.I.24-7, AT.I.24-9,* and *AT.I.24-14* genes. There is no obvious functional grouping of genes in this area. However, two enzymes similar to those encoded by *AT.I.24-1* and *AT.I.24-13* genes (glutaredoxin and protein disulfide isomerase) have a synergistic effect in *E. coli* (Lundstrom-Ljung and Holmgren 1995). Unlike their bacterial counterparts, *EF-tu* genes, which in *E. coli* belong to large transcription units containing other genes coding for ribosomal protein or tRNA (Jaskunas et al. 1975; Hudson et al. 1981), there is apparently no association of such genes with the *EF-1α* genes. The *AT.I.24-5* and *AT.I.24-6* genes are members of the same multigene family (63.3% similarity at DNA level), as they encode distinct but similar proteins (80.8% similarity). Such genes would be difficult to clone by conventional DNA hybridization techniques. Indeed, even low stringency washes would not allow cross-hybridization

between the two genes (McGrath et al. 1993). Existence of small multigene families with overlapping functions leads to genetic redundancy and could be a compensatory mechanism to chromosome alterations. Redundancy has been demonstrated between homologous genes arising from sequence duplication, but also between nonhomologous genes that can affect the same developmental process, whereas in other cases they possess independent function (Pickett and Meeks-Wagner 1995 and references therein). Increased knowledge of gene structures and of their genomic environment will help to evaluate the level of redundancy in the genome and to approach molecular mechanisms underlying the process of duplication. The transcription levels of all the genes are not related. For example, the highly expressed housekeeping *EF-1α* genes are closely linked to the gene *AT.I.24-9,* which is specifically regulated as it shows a very low basal expression in cell cultures (data not shown).

In comparison with cDNA and genomic sequences, our data show that five of the identified genes possess a leader intron within their 5′ untranslated region. In plants, such a configuration appears much more frequently than initially expected. Many reports have described the critical role played by leader introns that modulate the expression of eukaryotic genes by influencing transcriptional or post-transcriptional processes (Luehrsen and Walbot 1991; Kuhlemeier 1992; Bovy et al. 1995). Thus, in *A. thaliana,* a 5′ intervening sequence contains *cis*-acting elements involved both in quantitative and tissue-specific expression of the *EF-1α A1* gene (Curie et al. 1991). The presence of a perfect duplication of a large part of this intron, IDS29, located 10 kb upstream of the *EF-1α A1* gene (Figs. 2 and 4) is of particular interest. Considering the slight divergence between this sequence and sequence from *A3* gene, probably IDS29 arose from a recent duplication event that took place after duplication of the *EF-1α* genes. In addition to this perfect duplication, other duplicated sequences showing a weaker homology to the same intron sequence have been detected. Strikingly, one of them, IDS23, corresponds perfectly to the intron in the 5′ untranslated region of the gene coding for an ascorbate peroxidase. This observation raises the question of a possible functional and evolutionary role of this type of duplication. Interestingly, among weak similarities of IDS29 found to other plant sequences we observed an overlap of 147 bp to the enhancer sequence 295-6 of *A. thaliana* (Ott and Chua 1990). Shuffling of nuclear protein-binding domains by transposable elements has been proposed (White et al. 1994; Oosumi et al. 1995). For example, "Tourist" and "Stowaway" elements have well-defined structures of mobile elements (Bureau and Wessler 1994a,b) being located mainly in 3′ ends of cDNAs or in 5′ upstream regions of several genes. Degenerated duplications of these elements have been identified by computer-assisted sequence similarity searches in databases. The structure of the duplicated element characterized in this study is not related to any known mobile element. The putative mechanisms involved in mobility remain unknown. Accumulation of data provided in the *A. thaliana* genome-sequencing project should generate additional information to extend our observation. In many cases, introns are very likely involved in exon shuffling during evolution of eukaryotic genes (Patthy 1991). They might be implicated in promoter and enhancer shuffling and therefore, contribute to the great adaptability of plant genomes.

The current report of the multinational *Arabidopsis* Steering Committee (Sommerville 1996) predicts completion of the total sequence of *Arabidodpsis* genome by 2004. The preliminary results we obtained by sequencing and analyzing a 63-kb fragment encourage us to predict that this project will be of great importance for our understanding of plant biology. Our results suggest that sequence duplication and subsequent divergence, supplying genes with regulatory sequences and facilitating gene duplications arise frequently in the clearly dynamic plant genome. They confirm that an international effort in plant research on *Arabidopsis* including genomic sequencing in coordination with cDNA sequencing and mapping, will largely increase our knowledge in plant genome organization and evolution. It is particularly exciting to imagine that new mechanisms involved in the genome plasticity typical of plants might be revealed by this kind of approach.

METHODS

YAC Manipulations

The EG YAC library (Grill and Somerville 1991) was screened by PCR amplification of clone pools as described by Green and Olson (1990; Kwiatkowski et al. 1990) using specific primers for the *EF-1α A1* gene (P1, 5′-ATTATTTTATGTTAAACCCTAA-3′, P2, 5′-AACGATCAGCAGACG-3′). High-quality yeast DNA was prepared for PFGE analysis in agarose plugs. Yeast cells were grown in liquid culture for 30 hr and harvested. They were washed once in 50 mM EDTA (pH 7.5), resuspended at 3×10^8 to 5×10^8 cells/ml in 20 mM EDTA, then mixed 1:1 with a solution of 20 mM EDTA, 1% agarose

(Seakem) containing 0.08 mg/ml of Zymolyase 20T at 40°C. The cells were transferred to a plug mold to set. The plugs were transferred to one volume of 500 mM EDTA (pH 7.5), 7.5 % β-mercapto-ethanol overnight. Plugs were rinsed twice in two volumes of 50 mM EDTA (pH 9.5) and incubated for 2 days at 45°C in 500 mM EDTA (pH 9.5), 1% lauroylsarcosine 1 mg/ml of proteinase K. DNA in plugs was stored at 4°C for subsequent use. Southern blot hybridization on YAC DNA were performed according to standard protocols (Sambrook et al. 1989).

Selection and Preparation of λ Clones

A λ contig was constructed by screening a commercially available library from Clontech (*A. thaliana* ecotype Columbia, no. FL1002J) using an EF-1α probe and conventional hybridization procedures (Sambrook et al. 1989); then, from putative positives, isolation of a clone was performed by PCR amplification of λ phage supernatant with specific primers. Lambda DNA was prepared with "Promega wizard prep" (no. A7290) according to manufacturer's recommendations.

Sequencing Methodology

The λ clones were mapped using usual restriction enzymes and subcloned in Bluescript KS according to standard protocols (Sambrook et al. 1989). Exonuclease III digestions were used to create nested deletions from one end of the insert. Subclones in Bluescript KS were organized according to length, and clones overlapping by ~200 bp of insert were chosen for PCR amplification from DNA of *E. coli* colonies. Template preparation of selected clones was done using the Promega wizard prep kit (no. A7500). Cycle sequencing using *Taq* polymerase was performed according to ABI protocols with Perkin Elmer dye terminator kits (part 402079) and M13 universal primers. One strand of the DNA sequence was thereby sequenced and a second strand was analyzed using primers deduced from the first strand sequence. Sequence information across the cloning sites was obtained by direct sequencing of PCR fragments amplified from the phage insert. The gel was run on an ABI 373 A. Reading on both strands with at least a fourfold redundancy has been obtained, inconsistencies between the two strands have been reassessed systematically.

Data Analysis

A Sun station was used for data analysis (X terminal). Staden package programs were used to assemble sequences and control quality (Bonfield et al. 1995). GCG and Staden package programs were used to format the sequence and visualize ORFs, restriction cutting sites, and repeated sequences. To detect homology or similarity to known sequences, BLASTX and BLASTN (Altschul et al. 1990) searches were done on the NCBI server against a nonredundant public library. In addition, the 63-kb sequence was cut arbitrarily into 1-kb pieces and internal repeats within the whole sequence were searched for and visualized using the FASTA program (Pearson and Lipman 1988) for each subfragment against a database containing all the 1-kb subfragments. Links to AATDB were made through the World Wide Web. A systematic analysis was performed by the informatic coordination center (MIPS, Martinsried Germany, under the direction of Stephan Klosterman and Nicolas Chalwaszis). In addition to public databases, specific databases were searched. For example, BLASTN was used against a database including small RNAs and BLASTX against TRES_at (a six-frame translation of dbEST_at). Regions predicted to code for a gene were evaluated by GENEID, GRAIL, GENMARK, and GENEFINDER, specifically adapted from *Caenorhabditis elegans* to *A. thaliana.*

Cognate cDNAs Characterization

Cognate cDNAs, identified by BLASTN homology to ESTs, were obtained from the DNA stock center at National Center for Biotechnology Information (NCBI). In the absence of homology or similarity the cDNA clones were selected by conventional screening (Sambrook et al. 1989) of the AC16H library constructed in our laboratory (Regad et al. 1993), using a probe corresponding to a putative exon of the gene. Then, the cDNA insert was sequenced from both ends and internal primers were used to sequence the clone completely.

ACKNOWLEDGMENTS

We thank J. Gouzy and A. Moisan for guidance in computer programs,Y. Marco and N. Grimsley for critical reading of the manuscript, and Nicolas Chalwatsis at the MIPS and the Ohio Stock center for sending us the cDNA clones. We express our gratitude to D. Bouchez for screening the CIC library. We gratefully acknowledge financial support from the Groupe de Recherche et d'Etudes des Génomes (GREG) and the European Community (ESSA). F. Regad holds a grant from the Ministère de la Recherche et de l'enseignement supérieur during this work.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Ahn, S. and S. Tanksley. 1993. Comparative linkage maps of the rice and maize genomes. *Proc. Natl. Acad. Sci.* **90:** 7980–7984.

Altschul, S.F., W. Gish, W. Miller, E. Myiers, and D. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Axelos, M., C. Bardet, T. Liboz, A. Le Van Thai, C. Curie, and B. Lescure. 1989. The family encoding the *A. thaliana* translation elongation factor EF-1α: Molecular cloning, characterization and expression. *Mol. Gen. Genet.* **219:** 106–112.

Boguski, M.S., T.M.J. Lowe, and S.H. Toltoshev. 1993. dbEST—database for "expressed sequences tags". *Nature Genet.* **4:** 332–333.

Bonfield, J., K. Smith, and R. Staden. 1995. A new DNA sequence assembly program. *Nucleic Acids Res.* **23:** 4992–4999.

Bovy, A., C. Van Den Berg, G. de Vrieze, W.F. Thompson, P. Weisbeek, and S. Smeekens. 1995. Light regulated expression of the *A. thaliana* ferredoxin gene requires sequences upstream and downstream of the transcription initiation site. *Plant. Mol. Biol.* **27:** 27–39.

Bureau, T. and S. Wessler. 1994a. Mobile inverted repeat elements of the *tourist* family are associated with the genes of many cereal grasses. *Proc. Natl. Acad. Sci.* **91:** 1411–1415.

———. 1994b. *Stowaway:* A new family of inverted repeat elements associated with the genes of both Monocotyledonous and Dicotyledonous plants. *Plant Cell* **6:** 907–916.

Cooke, R., M. Raynal, M. Laudie, F. Grellet, M. Delseny, P.C. Morris, D. Guerrier, J. Giraudat, F. Quigley, G. Clabault, et al. 1996. Further progress towards a catalogue of all *A.* genes: Analysis of a set of 5000 non-redundant ESTs. *Plant J.* **91:** 101–124.

Creusot, F., E. Fouilloux, M. Dron, J. Lafleuriel, G. Picard, A. Billault, D. LePaslier, D. Cohen, M.E. Chabouté, A. Durr, et al. 1995. The CIC library: A large insert Yac library for genome mapping in *A. thaliana*. *Plant J.* **85:** 763–770.

Curie, C., T. Liboz, C. Bardet, E. Gander, C. Medale, M. Axelos, and B. Lescure. 1991. *cis* and *trans* acting elements involved in the activation of *A. thaliana* A1 gene encoding the translation elongation factor EF-1α. *Nucleic Acids Res.* **196:** 1305–1310.

Curie, C., M. Axelos, C. Bardet, N. Atanassova, N. Chaubet, and B. Lescure. 1993. Modular organization and developmental activity of an *A. thaliana* EF-1α gene promoter. *Mol. Gen. Genet.* **238:** 428–436.

Denecke, J., R. DeRycke, and J. Botterman. 1992. Plant and mammalian sorting signals for protein retention in the endoplasmic reticulum contain a conserved epitope. *EMBO J.* **11:** 2345–2355.

Depeignes, A., C. Goubely, A. Lenoir, S. Cocherel, G. Picard, M. Raynal, F. Grellet, and M. Delseny. 1995. Identification of the most represented repeated motifs in *A. thaliana* microsatellite loci. *Theor. Appl. Genet.* **91:** 160–168.

Gilson, S. and C. Somerville. 1993. Isolating plant genes. *Trends Biotechnol.* **11:** 306–313.

Goodman, H.M., J.R. Ecker, and C. Dean. 1995. The genome of *A. thaliana*. *Proc. Natl. Acad. Sci.* **92:** 10831–10835.

Green, E.D. and M.V. Olson. 1990. Systematic screening of yeast artificial chromosome libraries by use of the polymerase chain reaction. *Proc. Natl. Acad. Sci.* **87:** 1213–1217.

Grill, E. and C. Somerville. 1991. Construction and characterization of a yeast artificial chromosome library of *A.* which is suitable for chromosome walking. *Mol. & Gen. Genet.* **226:** 484–490

Hervé, C., P. Dabos, J.P. Galaud, P. Rougé, and B. Lescure. 1996. Characterization of an *A. thaliana* gene that defines a new class of putative plant receptor kinases with an extracellular lectin like domain. *J. Mol. Biol.* **2855:** 778–788.

Höfte, H., T. Desprez, J. Amselem, H. Chiapello, M. Caboche, A. Moisan, M.F. Jourjon, J.L. Charpenteau, P. Berthomieu, D. Guerrier, et al. 1993. An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNA from *A. thaliana*. *Plant J.* **4:** 1051–1061.

Homlgren, A. 1976. Hydrogen donor system for *Escherichia coli* ribonucleoside diphosphate reductase dependent upon glutathione. *Proc. Natl. Acad. Sci.* **73:** 2275–2279.

Hudson, L., J. Rossi, and A. Landy. 1981. Dual function transcripts specifying tRNA and mRNA. *Nature* **294:** 422–427.

Ishiguro, S. and K. Nakamura. 1994. Characterization of a cDNA encoding a novel DNA binding protein SPF1, that recognizes SP8 sequences in the 5′ upstream regions of genes coding for sporamin and β amylase from sweet potato. *Mol. & Gen. Genet.* **244:** 563–571.

Jaskunas, S.R., L. Lindahl, M. Nomura, and R.R. Burgess. 1975. Identification of two copies of the gene for the elongation factor EF–Tu in E. coli. *Nature* **257:** 458–462.

Kopriva, S., R. Cossu, and H. Bauwe. 1995. Alternative splicing results in two different transcripts for H-protein of the glycine cleavage system in the C4 species *Flaveria trinervia*. *Plant J.* **8:** 435–441.

Kubo, A., H. Saji, K. Tanaka, and N. Kondo. 1993. Genomic DNA structure of a gene encoding cytosolic ascorbate peroxidase from *A. thaliana*. *FEBS Lett.* **315:** 313–317.

Kuhlemeier, C. 1992. Transcriptional and post transcriptional regulation of gene expression in plants. *Plant Mol. Biol.* **19:** 1–14.

Kurata, N., G. Moore, Y. Nagamura, T. Foote, M. Yano, Y. Minobe, and M. Gale. 1994. Conservation of genome structure between rice and wheat. *Bio/Technology* **12:** 276–278.

Kwiatkowski, T.J., H.Y. Zoghbi, S.A. Ledbetter, K.A. Ellison, and A.C. Chinault. 1990. Rapid identification of yeast artificial chromosome clones by matrix pooling and crude lysate PCR. *Nucleic Acids Res.* **1823:** 7191–7192.

Lagercrantz, U., J. Putterill, G. Coupland, and D. Lydiate. 1996. Comparative mapping in *A.* and *Brassica,* fine scale genome collinearity and congruence of genes controlling flowering time. *Plant J.* **9:** 13–20.

Larkin, R. and T. Guifoyle. 1993. The second largest subunit of RNA polymerase II from *A. thaliana*. *Nucleic Acids Res.* **21:** 1038.

Li, J., J. Zhao, A. Rose, R. Schmidt, and R. Last. 1995. *A.* phosphoribosylanthranilate isomerase: Molecular genetic

analysis of triplicate tryptophan pathway genes. *Plant Cell* **7:** 447–461.

Liboz, T., C. Bardet, A. Le Van Thai, M. Axelos, and B. Lescure. 1989. The four members of the gene family encoding the *A. thaliana* translation elongation factor EF-1α are actively transcribed. *Plant Mol. Biol.* **14:** 107–110.

Luehrsen, K.R. and V. Walbot. 1991. Intron enhancement of gene expression and the splicing efficiency of intons in maize cells. *Mol. & Gen. Genet.* **225:** 81–93.

Luehrsen, K.R., S. Taha, and V. Walbot. 1994. Nuclear pre-mRNA processing in higher plants. *Proc. Nucleic Acid Biochem. Mol. Biol.* **47:** 149–193.

Lundstrom-Ljung, J. and A. Holmgren. 1995. Glutaredoxin accelerates gluthathione dependent folding of reduced ribonuclease A together with protein disulfide isomerase. *J. Biol. Chem.* **270:** 7822–7828.

McGrath, J., M. Jansco, and E. Pichersky. 1993. Duplicate sequences with a similarity to expressed genes in the genome of *A. thaliana*. *Theor. Appl. Genet.* **86:** 880–888.

Merrick, M.J. and R.A. Edwards. 1995. Nitrogen control in bacteria. *Microbiol. Rev.* **59:** 604–622.

Meyerowitz, E.M. 1994. Structure and organization of the nuclear *A. thaliana* nuclear genome. In *Arabidopsis,* pp. 21–36. (eds. E.M. Meyerowitz and C.R. Somerville) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Montag, K., F. Salamini, and R.D. Thompson. 1995. ZEMa, a member of a novel group of MADS box genes, is alternatively spliced in maize endosperm. *Nucleic Acids Res.* **23:** 2168–2177.

Moore, G., M. Gale, N. Kurata, and R. Flavell. 1993. Molecular analysis of small grain nuclear genomes—Current status and prospects. *Bio/Technology* **11:** 584–589.

Newman, T., F. deBruijn, P. Green, K. Keegstra, H. Kende, L. McIntosh, J. Ohlrogge, N. Raikhel, S. Somerville, M. Thomashow, et al. 1994. Genes galore: A summary of methods for accessing results from large-scale partial sequencing of anonymous *A.* cDNA clones. *Plant Physiol.* **106:** 1241–1255.

Oosumi, T., B. Garlick, and W. Belknap. 1995. Identification and characterization of putative transposable DNA elements in solanaceous plants and *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci.* **92:** 8886–8890.

Ott, R. and N. Chua. 1990. Enhancer sequences from *A. thaliana* obtained by library transformation of *Nicotiana tabacum*. *Mol. Gen. Genet.* **223:** 169–179.

Patthy, L. 1991. Exons—Original building blocks of proteins? *BioEssays* **13:** 187–192.

Pearson, W.R. and D.J. Lipman. 1988. Improved tools for bilogical sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.

Pickett, F.B. and D.R. Meeks-Wagner. 1995. Seeing double: Appreciating genetic redundancy. *Plant Cell* **7:** 1347–1356.

Picton, S., J. Gray, S. Barton, U. Abubakar, A. Lowe, and D. Grierson. 1993. cDNA cloning and characterization of novel ripening related mRNAs with altered patterns of accumulation in the *ripening inhibitor rin* tomato ripening mutant. *Plant. Mol. Biol.* **23:** 193–207.

Regad, F., C. Bardet, D. Tremousaygue, A. Moisan, B. Lescure, and M. Axelos. 1993. cDNA cloning and expression of an *A.* GTP binding protein of the ARF family. *FEBS Lett.* **316:** 133–136.

Regad, F., C. Hervé, O. Marinx, C. Bergounioux, D. Tremousaygue, and B. Lescure. 1995. The *tef1 box,* a ubiquitous *cis*-acting element involved in the activation of plant genes that are highly expressed in cycling cells. *Mol. & Gen. Genet.* **248:** 703–711.

Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. *Molecular cloning.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Shorrosh, B.S. and R.A. Dixon. 1991. Molecular cloning of a putative endomembrane protein resembling vertebrate protein disulfide isomerase and a phosphatidylinositol specific phospholipase C. *Proc. Natl. Acad. Sci.* **88:** 10941–10945.

Sofia, H.J., V. Burland, D.L. Daniels, G. Plunket III, and F.R. Blattner. 1994. Analysis of the *Escherichia coli* genome. V. DNA sequence of the region from 76.0 to 81.5 minutes. *Nucleic Acids Res.* **22:** 2576–2586.

Sommerville, C. 1996. The physical map of an *A.* chromosome. *Trends Plant Sci.* **1:** 2.

Walbot,V. 1996. Sources and consequences of phenotypic and genotypic plasticity in flowering plants. *Trends in Plant Sci.* **1:** 27–32.

White, S., L. Habera, and S. Wessler. 1994. Retrotransposons in the flanking regions of normal plant genes: A role for copia like elements in the evolution of gene structure and expression. *Proc. Natl. Acad. Sci.* **91:** 11792–11796.

Received August 23, 1996; accepted in revised form January 10, 1997.