J. Mol. Biol. (1993) **229**, 797–801

SEQUENCE NOTE

Characterization of a Multigene Family Encoding an Exopolygalacturonase in Maize

A. Barakate, W. Martin†, F. Quigley and R. Mache‡

Laboratoire de Biologie moléculaire végétale
associé au CNRS (URA 1178), Université J. Fourier
BP53X, 38041 Grenoble, France

(Received 17 June 1992; accepted 9 October 1992)

Genes coding for exopolygalacturonase in plants are abundantly expressed during the development of the male gametophyte (pollen). We have analysed genomic and cDNA clones for several representatives of the small multigene family encoding exopolygalacturonase from *Zea mays*. Structures for both actively transcribed genes and non-transcribed pseudogenes are reported. Comparisons of the nucleotide sequences for coding and flanking regions of different members of the gene family reveal surprisingly few base substitutions, suggesting that the exopolygalacturonase gene family of maize arose through very recent multiple duplication events. The pseudogenes are shown to possess an 80 bp insertion within the coding region, which may represent a relictual intron that has been lost in the active genes. We estimate that 12 exopolygalacturonase genes exist in maize. None appear to be expressed at a detectable level in tissue other than those associated with pollen development.

Keywords: multigene family; polygalacturonase; pollen-expressed genes; maize

Exopolygalacturonase (exoPG§), also termed polygalacturonase (PG), catalyses the degradation of highly polymeric galacturonate, a major component of pectin in plant cell walls, into individual polygalacturonic acid residues. exoPG, as other enzymes involved in pectin degradation which are specifically synthesized in pollen, has been suggested to be instrumental in pollen tube growth and wall synthesis (Niogret *et al.*, 1991; McCormick, 1991). In both *Oenothera* and maize, several cDNAs encoding exoPG have been characterized suggesting the presence in the nuclear genome of a small multigene family (Brown & Crouch, 1990; Niogret *et al.*, 1991). We wished to characterize this gene family in maize and to know whether individual exoPG genes are differentially expressed.

We constructed a genomic library in λEMBL4 with *Mbo*I, partially digested DNA from an inbred breeder's line of maize (MO17) and screened it with the 350 bp *Xba*I–*Eco*RI fragment from the 3′ non-coding region of the cDNA clone PGc1 coding for exoPG (Niogret *et al.*, 1991). From roughly 400,000 recombinants 30 positively hybridizing clones were obtained. Of these, 20 were selected for sequence analysis. To obtain an indication of the size of the genes we performed a PCR amplification of the gene of each clone using the primers a and c located as indicated on Figure 1A. For all genomic clones except one (PGg5), the size of the amplified fragment corresponds to that derived from the coding region of the PGc1 cDNA, indicating that these genomic clones do not contain intervening sequences. For PGg5, the amplification product is longer than the cDNA sequence due to the presence of an 80 bp insertion in the reading frame gene (Fig. 2). The 80 bp insertion lacks the consensus elements typical of plant introns (Brown, 1986). More than one copy of this insertion-containing class of PG genes exists in the MO17 maize line, yet these are not expressed at detectable levels in any maize tissue studied (see below). We will refer to them as the exoPG pseudogene subfamily. Those genes that are completely colinear in the coding region with the expressed mRNAs will be referred to collectively as the active exoPG gene subfamily, although expression has been demonstrated for only a subset of these (see below).

† Present address: Institut für Genetik, Technische Universität Braunschweig, D-3300 Braunschweig, F.R.G.

‡ Author to whom reprint requests should be addressed.

§ Abbreviations used: exoPG, exopolygalacturonase; PG, polygalacturonase; bp, base-pair(s); PCR, polymerase chain reaction.

0022-2836/93/030797-05 $08.00/0

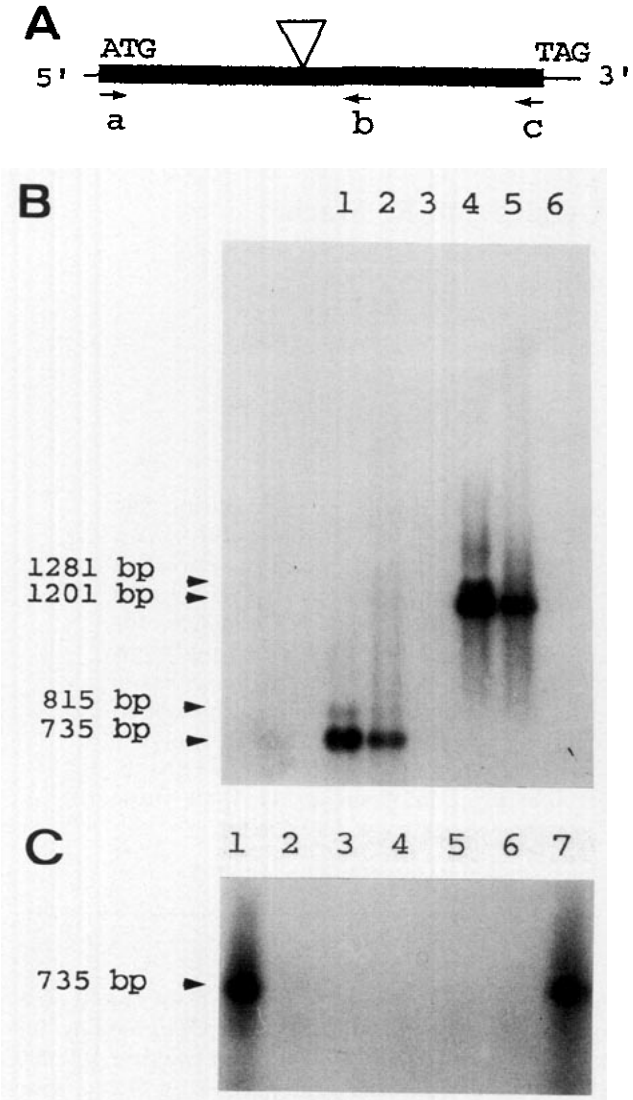

Figure 1. Identification and expression of 2 exoPG gene subfamilies by PCR analysis of genomic DNA. A, Schematic localization of primers a (5'-TTGCACAAACA-ATGCGATGAGAGCC-3'), b (5'-GATGTTCACCTTGGAG-GTCCC-3') and c (5'-CCCTTGGTGCTGCCCTTGGCG-3') used for PCR and of the 80 bp insertion. B, PCR amplification of genomic DNA. PCR amplification was performed using primers a and b (1, 2, 3) or a and c (4, 5, 6), with genomic DNA (1, 4), or the PGc1 cDNA (2, 5) or without DNA (3, 6). PCR products were electrophoresed and blotted onto nylon membrane and hybridized with a radiolabelled DNA probe from PGc1 cDNA. Autoradiography of the blot is shown. Length of amplified fragments is indicated. C, PCR amplification of exoPG cDNAs. RNA (1 μg) from different organs was used for first strand cDNA synthesis and PCR amplification following the method of Buck & Axel (1991). Total RNA from microspores was isolated as described (Jepson *et al.*, 1991) at the S1-S2 stages of microspore development (Mandaron *et al.*, 1990). Primers a and b were used for amplification. DNA products were size fractionated and revealed as indicated in B. 1, pollen; 2, roots; 3, mesocotyle; 4, coleoptile; 5, shoot inside of coleoptile; 6, leaves of plantlets; 7, PGc1 DNA as a size control. Each 25 μl PCR reaction contained 1 μmol each of 2 opposing primers, 1·25 mM each of 4 dNTPs, and 2 units AmpliTaq DNA polymerase (Perkin-Elmers/Cetus). PCR amplifications were performed with an initial denaturation followed by 35 cycles of denaturation (94°C, 1 min), annealing (65°C, 2 min), and extension (72°C, 10 min). Controls included omitting template DNA, omitting the reverse transcription reaction, and adding DNase-free RNase before the extension reaction.

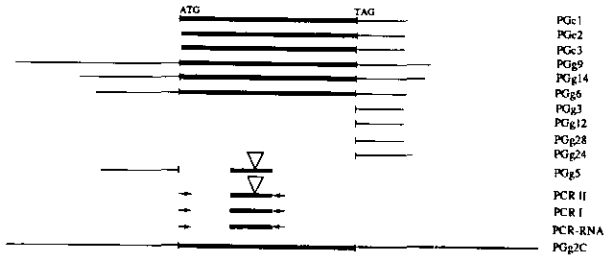

Figure 2. Schematic representation of the sequenced regions of PGg genes, PGc cDNA and of their flanking regions. Coding sequences are represented by filled rectangles. 5' and 3' flanking regions are represented by thin lines. The divergent part of the 3' non-coding region (PGg24) is represented by a middle size line. Position of insertion sequence is indicated by an open triangle. The localization of primers (a and b) used for PCR amplification and sequencing is represented by arrows. Sequences of DNA fragments resulting from PCR amplification of the 735 and 815 bp (see Fig. 1) genomic DNA (PCRI and PCRII) or of cDNA synthesized from pollen mRNAs (PCR-RNA) are indicated. The PGc 2 sequence is entirely identical to the PGg14 sequence. All cDNA and genomic clones are from the MO17 inbred line of maize with the exception of the PGg2C gene, which is from the W22 maize line. EMBL accession numbers are: X65844 (PGg6 gene), X65845 (PGg14 gene), X65846 (PGg5, 5' flanking region), X65847 (PGg5 gene), X65848 (PGg3, 3' flanking region), X65849 (PGg12, 3' flanking region), X65850 (PGg24, 3' flanking region), X65851 (PGg28, 3' flanking region), X64408 (PGg9 gene), X65852 (PG PCRI gene), X65853 (PG PCRII gene), X65854 (PG PCRI mRNA), X66422 (PGg2C gene). For construction and isolation of genomic clones, genomic DNA was isolated from light-grown, one-week-old maize MO17 seedlings and cloned into λEMBL4 as described (Schwarz-Sommer *et al.*, 1985). Replica filters from 400,000 recombinants were screened with the 350 bp *Xba*I–*Eco*RI fragment from the 3' non-coding region of PGc1 cDNA (Niogret *et al.*, 1991). By the same methods, the genomic clone PGg2C was isolated from seedlings of a W22 maize line. Restriction fragments of interest from lambda clones were subcloned into pUC18 and sequenced as double-stranded DNA by the dideoxynucleotide chain termination method of Sanger *et al.* (1977), using either pUC/M13 universal primers or exoPG specific primers. PCR-amplified double-stranded DNAs were amplified asymmetrically as described by Allard *et al.* (1991) prior to dideoxy sequencing.

We searched for the presence of other exoPG gene structures in the total genomic DNA by PCR amplification using two pairs of primers (a and b or a and c, see Fig. 1A). The amplified products were of two different sizes (Fig. 1B), which corresponded to the two subfamilies observed in the cloned genes, as described above. The selected clones therefore

appear to be a representative subset of the entire exoPG gene family. Comparison of the relative intensity of the two bands (Fig. 1B) shows that the exoPG active subfamily occurs more frequently than the exoPG pseudogene subfamily. The two subfamilies were further characterized by sequencing DNA fragments in a subset of eight different clones chosen among the 20 clones analysed by PCR as described above. The DNA regions of each of the genes that were sequenced are schematically represented in Figure 2 together with the previously characterized cDNAs (Niogret *et al.*, 1991). DNA of the PGg2C gene, which has been isolated from a W22 line of maize, is also represented.

To determine the degree of divergence between coding regions of the cloned genes, the PGg6, PGg9 and PGg14 genes were selected for sequencing. Only a few base substitutions are observed between the sequenced genes, about 1·4% between genes on average (Table 1). Only six (in PGg6) or seven (in PGg14) nucleotide substitutions result in amino acid replacements relative to PGg9 and three of these replacements result in functionally equivalent amino acids. Therefore, the very low variation that was previously observed in the sequence of three cDNAs (PGc1, PGc2 and PGc3; see Niogret *et al.*, 1991) is also reflected in the coding sequences deduced from the three active genes that have been analysed.

The 3′ non-coding regions of the PGg14 gene and of the PGc2 cDNA are identical, suggesting that the cDNA corresponds to the mRNA from this gene. PGg6 differs from PGg1 by only two base substitutions. The 3′ regions of all exoPG genes characterized are very closely related with the exception of the PGg24 gene, which is highly divergent from other members of the active family in the first 180 bp following the stop codon, but realigns with high sequence downstream from that point. It is noteworthy that the two polyadenylation site sequences are highly conserved. Taken together, the nucleotide sequence of the coding region and of the 3′ non-coding region of different clones has allowed the identification of nine different exoPG genes or cDNAs: PGc1, PGc2 (or PGg14), PGc3, PGg3, PGg6, PGg9, PGg12, PGg24 and PGg28 (Fig. 2), where PGc indicates cDNA and PGg indicates genomic clones.

In order to obtain information on the diversity within the coding sequence of all the genes that belong to the active subfamily, PCR amplification followed by sequencing was performed. DNA from the 735 bp PCR band (Fig. 1B) was amplified asymmetrically and sequenced. At several positions in the sequencing gels two bands reproducibly appeared at the same level, some of these corresponding to nucleotide changes observed in sequence comparisons of individual clones. A total of 16 nucleotide changes were observed relative to the corresponding 285 bp segment of the PGc1 cDNA. These changes are dispersed throughout the 285 bp fragment. The sequence of the same 285 bp fragment in four different clones (PGg9, PGc1, PGc2, PGc3) revealed a total of five nucleotide substitutions. From these data and assuming that the number of base changes is roughly equally distributed in the different genes we can estimate at 12 or 13 the total number of genes present in the subfamily. This estimate is in good agreement with the results of genomic Southern blots (Niogret, 1991). We conclude that all active genes thus identified belong to a very closely related gene family. Several small gene families previously characterized from plants were shown to possess a higher degree of divergence (Rottmann *et al.*, 1991; Harada *et al.*, 1990; Marks *et al.*, 1985; Russell & Sachs, 1989). We suggest that the exoPG genes have arisen recently, probably by successive duplication of an ancestor gene. If nuclear substitution rates in plants are similar to those of vertebrates, as has been suggested (Martin *et al.*, 1989), the low degree of divergence between 3′ non-coding regions of active exoPG genes (<2%) studied here would suggest that these arose through successive duplications within the last 2 million years or so. exoPG of maize shares only 44% amino acid identity with the exoPG of *O. organensis* (Niogret *et al.*, 1991). The divergence between these exoPG gene sequences may correspond to the separation of monocotyledon (maize) and dicotyledon (*O. organensis*) lineages.

With the exception of the 80 bp insertion, the PGg5 pseudogene is highly homologous to the nucleotide sequence of the other genes of the active

Table 1

Comparison of three exoPG gene coding regions

Codon position starting from the first Met	Codon and amino acid changes in PGg9	PGg6	PGg14
2	GCG (A)	=	GCA (A)
5	GAC (D)	AAC (N)*	AAC (N)*
81	GGC (G)	GGT (G)	=
117	GAA (E)	GAC (D)*	GAC (D)*
126	CGC (R)	CGT (R)	=
144	GCC (A)	=	GCT (A)
155	TAT (Y)	TAC (Y)	TAC (Y)
191	CAG (Q)	CGG (R)*	CGG (R)*
194	AAC (N)	GAC (D)*	GAC (D)*
223	ACC (T)	ACG (T)	=
228	GTC (V)	GTG (V)	GTG (V)
232	GGC (G)	GGT (G)	=
248	ACT (T)	ACC (T)	ACC (T)
254	CCC (P)	CCT (P)	CCT (P)
261	GGC (G)	=	GGA (G)
285	ACA (T)	ACG (T)	ACG (T)
333	ACC (T)	ACT (T)	ACT (T)
334	GCC (A)	GCT (A)	GCT (A)
351	ACT (T)	ACC (T)	ACC (T)
359	GCC (A)	=	GCT (V)*
360	ATT (I)	GTT (V)*	GTT (V)*
373	GCT (A)	GTC (V)*	GTC (V)*

Codons of the PGg6 and PGg14 genes in which a nucleotide change occurs are reported. The PGg9 gene is taken as a reference. Numbers indicate the position of codons from the translational initiation start. Amino acids corresponding to codons are indicated with the one-letter code. Identical codons are indicated (=). Equivalent amino acids are indicated by an asterisk.

subfamily. The degree of divergence estimated after the comparison of the 285 bp homologous fragments of PGg5 and PGc1 is 3·8%, i.e. a slightly higher degree of divergence than for the active genes. The diversity within the exoPG pseudogene subfamily was analysed by using PCR as for the other gene subfamily. DNA from the 815 bp band was isolated, asymmetrically amplified and sequenced. Fifteen nucleotide changes were observed, 11 of them are present in the PGg5 gene. No substitutions in the 80 bp insertion sequence were detected. We conclude that at least two, perhaps more, insertion-containing genes are present in the MO17 genome. The insertion is located at 562 bp from the first base of the initiation codon and contains stop codons in all three reading frames. The 3′ end of the insertion shows modest similarity to 3′ acceptor sites (Goodall & Filipowicz, 1991), yet no corresponding cryptic 5′ donor site exists within the coding region or insertion which would preserve the reading frame of any putatively spliced product. The lack of 5′ or 3′ intron border consensus sequences (Brown, 1986) suggests that the insertion is very probably not a spliceable intron. The 80 bp insertion may be derived from an intron since it has a high A+U content (58%), which corresponds to the nucleotide composition of monocotyledon introns (Goodall & Filipowicz, 1991). Were this the case, we would expect exoPG genes from graminaceous monocotyledons other than maize to possess an intron at this position.

Examination of sequence divergence in the promoter region of exoPG genes of maize revealed a very high degree of similarity downstream from position −384. In particular, a DNA stretch (AAGAATTTATGA) homologous to the tomato LAT56/59 box that modulates gene expression in tomato pollen (Twell *et al.*, 1991) is entirely conserved. In contrast, a GGTGGTT tandem repeat, similar to the binding sites for transcription factors present in many inducible or developmentally regulated genes (Gilmartin *et al.*, 1990) is not well conserved. The number of nucleotide substitutions relative to the PGg9 sequence and from −384 to the first codon is six (PGg14), ten (PGg6) or 15 (PGg5) corresponding to an average of 1·4% nucleotide substitutions. Thus, the degree of divergence of the promoter regions of three of the active subfamily is very similar to that found within the coding sequence and 3′ non-coding regions. The divergence in the PGg5 pseudogene promoter is slightly higher than for the other active genes. Upstream from position −384, the nucleotide sequence of the PGg5 pseudogene is quite dissimilar to the other sequences. In order to obtain an indication of the length of the enhancer region (far upstream region of the promoter), which is homologous in different maize exoPG genes, we have compared the upstream non-transcribed sequences of genes isolated from two different lines of maize. The PGg2C gene (Fig. 2) was isolated from a genomic library constructed by using DNA isolated from a W22 maize line. The other genes are those isolated from the MO17 line. The two gene promoters become quite dissimilar upstream from −339 from the start codon. Interestingly, the homologous LAT56/59 box (see above) is not well conserved in the genes from the two inbred lines.

Previous studies have shown that exoPG genes of maize are expressed in pollen (Niogret *et al.*, 1991). An exoPG enzyme activity lower than that found in pollen has been detected in young maize seedlings (Pressey & Reger, 1989). In order to determine if

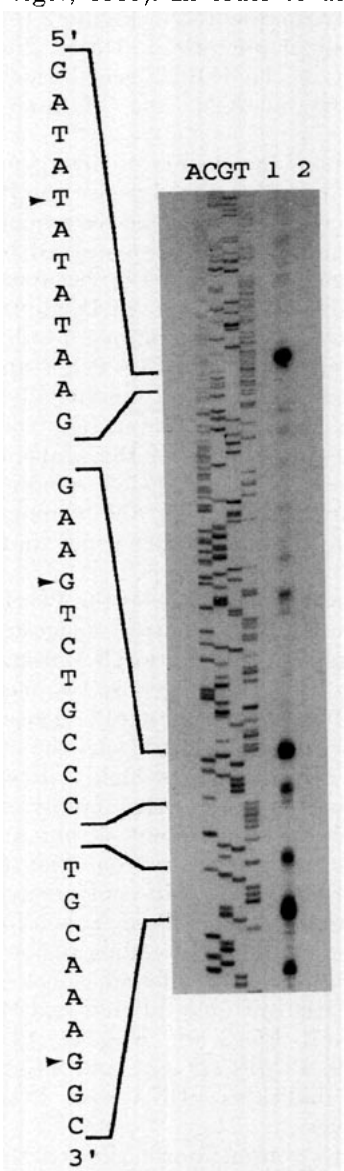

Figure 3. Primer extension mapping of the 5′-end of pollen transcripts of 3 exoPG genes (PGg). 10 μg of total RNA from microspores were used for primer extension mapping (Sambrook *et al.*, 1989) using a 5′ [γ-^{32}P]ATP end-labelled 20-base oligonucleotide (5′-CATTGTTTGTG-CACGCCATC-3′) complementary to the sense strand of the PGc1 cDNA (lane 1). Control (lane 2) was carried out with 10 μg of tRNA. Lanes A, C, G, T show part of the nucleotide sequence of the 5′-flanking region of the PGg9 gene generated with the same primer. Reaction products were fractionated on a 6% (w/v) polyacrylamide/7 M-urea gel and autoradiographed. The nucleotide sequence of the PGg9 gene generated by the same primer was used as a ladder.

the active gene subfamily characterized here is responsible for the exoPG enzyme activity described for young maize seedlings, cDNA was synthesized from these tissues, amplified by PCR using the a and b primers indicated in Figure 1, and hybridized with the PGc1 cDNA probe. No amplification products for exoPG mRNAs could be detected, suggesting that the low PG activity detected in maize seedlings may be due to expression of a class of PG genes which are of sufficient divergence as to be undetectable in our PCR assay. As expected, the active exoPG genes are strongly expressed in pollen. Interestingly, a PCR amplification of the cDNAs synthesized from pollen RNAs (PCR/RNA in Fig 2) shows that only six out of 16 total base changes (PCRI) are found in these mRNAs, suggesting that perhaps only about half of the exoPG gene subfamily is expressed in pollen. This raises the question of whether the other exoPG genes are expressed and if so, in which plant organ.

We determined the start(s) of transcription using the primer extension mapping method and RNAs isolated from maize pollen (Fig. 3). Several starts of transcription were observed, which likely correspond to the expression of different genes in pollen. It is noteworthy that upstream from all the identified initiation start sites no characteristic TATA box is present.

We asked whether the exoPG pseudogenes were transcribed. No transcript corresponding to the size of the exoPG pseudogene family was detected using the PCR method for the amplification of the cDNAs (Fig. 1). The lack of correlation between the nucleotide substitutions found in the amplified cDNAs synthesized from the pollen mRNAs and the PGg5 gene sequence (PGg5) indicates the lack of transcription of this PG pseudogene.

In conclusion we have characterized an active exoPG subfamily which contains more than 10 (possibly 12 or 13) members which arose recently through gene duplication. The expression of about half of these genes has been detected in pollen. A second subfamily of at least two pseudogenes derive from this active subfamily.

We thank Professor P. Mandaron and Dr P. Taberlet for their advice and help, and G. Clabault for his assistance in the utilization of the computer programs.

References

Allard, M. W., Ellsworth, D. L. & Honeycutt, R. L. (1991). The production of single-stranded DNA suitable for sequencing using the polymerase chain reaction. *BioTechniques*, **10**, 24–26.

Brown, J. W. S. (1986). A catalogue of splice junction and putative branch point sequences from plant introns. *Nucl. Acids Res.* **14**, 9549–9907.

Brown, S. M. & Crouch, M. L. (1990). Characterization of a gene family abundantly expressed in *Oenothera organensis* pollen that shows sequence similarity to polygalacturonase. *The Plant Cell*, **2**, 263–274.

Buck, L. & Axel, R. (1991). A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell*, **65**, 175–187.

Gilmartin, P. M., Sarokin, L., Memelink, J. & Chua, N. H. (1990). Molecular light switches for plant genes. *The Plant Cell*, **2**, 369–378.

Goodall, G. J. & Filipowicz, W. (1991). Different effects of intron nucleotide composition and secondary structure on pre-mRNA splicing in monocot and dicot plants. *EMBO J.* **10**, 2635–2644.

Harada, J. J., Spadoro-Tank, J., Maxwell, J. C., Schnell, D. J. & Etzler, M. E. (1990). Two lectin genes differentially expressed in *Dolichos biflorus* differ primarily by a 116-base pair sequence in their 5′ flanking regions. *J. Biol. Chem.* **265**, 4997–5001.

Jepson, I., Bray, J., Jenkins, G., Schuch, W. & Edwards, K. (1991). A rapid procedure for the construction of PCR cDNA libraries from small amounts of plant tissue. *Plant Mol. Biol. Reporter*, **9**, 131–139.

McCormick, S. (1991). Molecular analysis of male gametogenesis in plants. *Trends Genet.* **7**, 298–303.

Mandaron, P., Niogret, M. F., Mache, R. & Monéger, F. (1990). In vitro protein synthesis in isolated microspores of *Zea mays* at several stages of development. *Theoret. Appl. Genet.* **80**, 134–138.

Marks, M. D., Lindell, J. S. & Larkin, B. A. (1985). Quantitative analysis of the accumulation of zein mRNA during maize endosperm. *J. Biol. Chem.* **260**, 16451–16459.

Martin, W., Gierl, A. & Saedler, H. (1989). Molecular evidence for pre-Cretaceous angiosperm origins. *Nature (London)*, **339**, 46–48.

Niogret, M.-F. (1991). Etude de l'expression du génome haploïde au cours de la microsporogénèse chez le maïs: isolement et caractérisation d'un gène spécifique de la phase tardive. *Thèse de Doctorat*, University J. Fourier, Grenoble.

Niogret, M. F., Dubald, M., Mandaron, P. & Mache, R. (1991). Characterization of pollen polygalacturonase encoded by several cDNA clones in maize. *Plant Mol. Biol.* **17**, 1155–1164.

Pressey, R. & Reger, B. J. (1989). Polygalacturonase in pollen from corn and other grasses. *Plant Science*, **59**, 57–62.

Rottmann, W. H., Peter, G. F., Oller, P. W., Keller, J. A., Shen, N. F., Nagy, B. P., Taylor, L. P., Campbell, A. D. & Theologis, A. (1991). 1-Aminocyclopropane-1-carboxylate synthase in tomato is encoded by a multigene family whose transcription is induced during fruit and floral senescence. *J. Mol. Biol.* **222**, 937–961.

Russell, D. A. & Sachs, M. M. (1989). Differential expression and sequence analysis of the maize glyceraldehyde-3-phosphate dehydrogenase gene family. *The Plant Cell*, **1**, 793–803.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain terminating inhibitors. *Proc. Nat. Acad. Sci., U.S.A.* **74**, 5463–5467.

Schwarz-Sommer, Z., Cuypers, H., Gierl, A., Peterson, P. A. & Saedler, H. (1985). Plant transposable elements generate the DNA sequence diversity needed in evolution. *EMBO J.* **4**, 591–597.

Twell, D., Yamaguchi, J., Wing, R. A., Ushiba, J. & McCormick, S. (1991). Promoter analysis of genes that are coordinately expressed during pollen development reveals pollen-specific enhancer sequences and shared regulatory elements. *Genes Develop.* **5**, 496–507.

Edited by N.-H. Chua