Plant Physiol. (1992) 99, 779–782
0032-0889/92/99/0779/04/$01.00/0

Received for publication January 28, 1992
Accepted March 5, 1992

Plant Gene Register

Cotton *Mat5-A* (C164) Gene and *Mat5-D* cDNAs Encoding Methionine-Rich 2S Albumin Storage Proteins[1]

Glenn A. Galau*, Helen Y.-C. Wang, and D. Wayne Hughes
Department of Botany, University of Georgia, Athens, Georgia 30602

Several abundant mRNAs are coordinately expressed specifically during the maturation stage of embryogenesis (6, 7). The cotton *Mat* mRNAs include those of the major vicilin and legumin storage protein genes (6), and representatives of these genes have been sequenced (2, 3). To help define the temporal and spatial regulation of the maturation program of gene expression, we isolated an additional cotton *Mat* gene, *Mat5*, which encodes mRNAs represented by cDNA clone C164. The mRNAs from the two *Mat5* alloalleles in the allotetraploid genome of *Gossypium hirsutum* make up 2% of the mRNA in maturation stage embryos (6), and the sequences show that they encode methionine-rich 2S albumin storage proteins.

Three genomic fragments were recovered, but restriction site analysis and sequencing showed that they contain the same alloallele, *Mat5-A*, present in the A genome. One was extensively sequenced, as was all of cDNA C164 and extensive portions of five other cDNA clones (Table I). Figure 1 shows the sequence of clone GC164–24RC of *Mat5-A*. All six cDNAs are identical with each other in the regions sequenced, but they differ significantly from the sequence of *Mat5-A*. The gene and cDNA sequences are colinear, with eight nucleotide differences, until *Mat5-A* nucleotide 3661 and C164 nucleotide 514. There is no obvious similarity for the next 62 nucleotides until the site of polyadenylation of the cDNAs (Fig. 1). Because there are only two alloallelic genes in *G. hirustum* (6), we presume that all the sequenced cDNAs are transcribed from the other alloallele, *Mat5-D*.

The two *Mat5* alloalleles encode proteins that differ in only three of 139 amino acid residues. Their sequences are very similar to those of the 2S albumin storage protein family (1, 8, 9), and an alignment of the cotton sequences with other 2S albumin preproproteins is unambiguous in its important features (Fig. 2). There is complete identity at the cysteine and leucine residues that are diagnostic of the family and about 44% of the cotton residues are identical with those of *Bertholletia* (1) and *Arabidopsis* (9) 2S albumins. The predicted mature form (9; Fig. 2) of the cotton 2S albumin shares with that of *Bertholletia* a high methionine content. It is 10% in cotton and 19% in *Bertholletia*, compared with 2 to 3% in other 2S albumins (1, 8, 9). Four of the 10 methionine residues in the cotton protein are at novel positions that may be sites at which other 2S albumins could be engineered for higher methionine content.

Because none of the six cDNAs that were examined are transcribed from *Mat5-A*, it is possible that this gene is not active. However, it is more likely that the apparent absence of *Mat5-A* cDNAs is due to failure to process or polyadenylate its transcripts. In the sequence in Figure 1 and in overlapping phage isolate GC164–27 (Table I), there is a class I-like retrotransposon (10) in reverse orientation to the *Mat5-A* transcription unit (data not shown). The retrotransposon's putative internal domain/3′-long terminal repeat boundary is at *Mat5-A* nucleotide 3791/92, and the 130 nucleotides between this boundary and the *Mat5-A/Mat5-D* cDNA divergence point can contain only a highly truncated long terminal repeat, if any part of it. We predict that the element, or element-induced rearrangements of *Mat5-A* sequence, interferes with the processing of the 3′ end of *Mat5-A* transcripts. Rearrangement of *Mat5-A* in this region is suggested by the presence of long direct repeats, part of one of which is present in *Mat5-D* cDNAs (Fig. 1).

A sequence very similar to *Mat5-A* nucleotides 907 to 1765 has been found in reverse orientation in the upstream region of an unrelated cotton *Lea4* gene (4). Portions of this element are also duplicated at *Mat5-A* nucleotides 2289 to 2416 (Fig. 1, repeats 1 and 2). Consequently, the sequences that are responsible for transcription of *Mat5-A* are probably between nucleotide 2615 and the transcription start at nucleotide 3146. In this region the sequence of *Mat5-A* has some similarities with those of other 2S albumin storage proteins, in particular two elements (double underlined in Fig. 1) present in similar locations in many of these genes (8, 9). Other elements suggested to be important in various storage protein genes (5) are not obvious in *Mat5-A*. Finally, *Mat5-A* has been compared with two other cotton *Mat* genes, vicilin (2) and legumin A (3), with which *Mat5-A* is coordinately expressed in cotton (6, 7), but no similarities are compelling in any sequence pair.

[1] Supported by a grant from the National Institutes of Health.

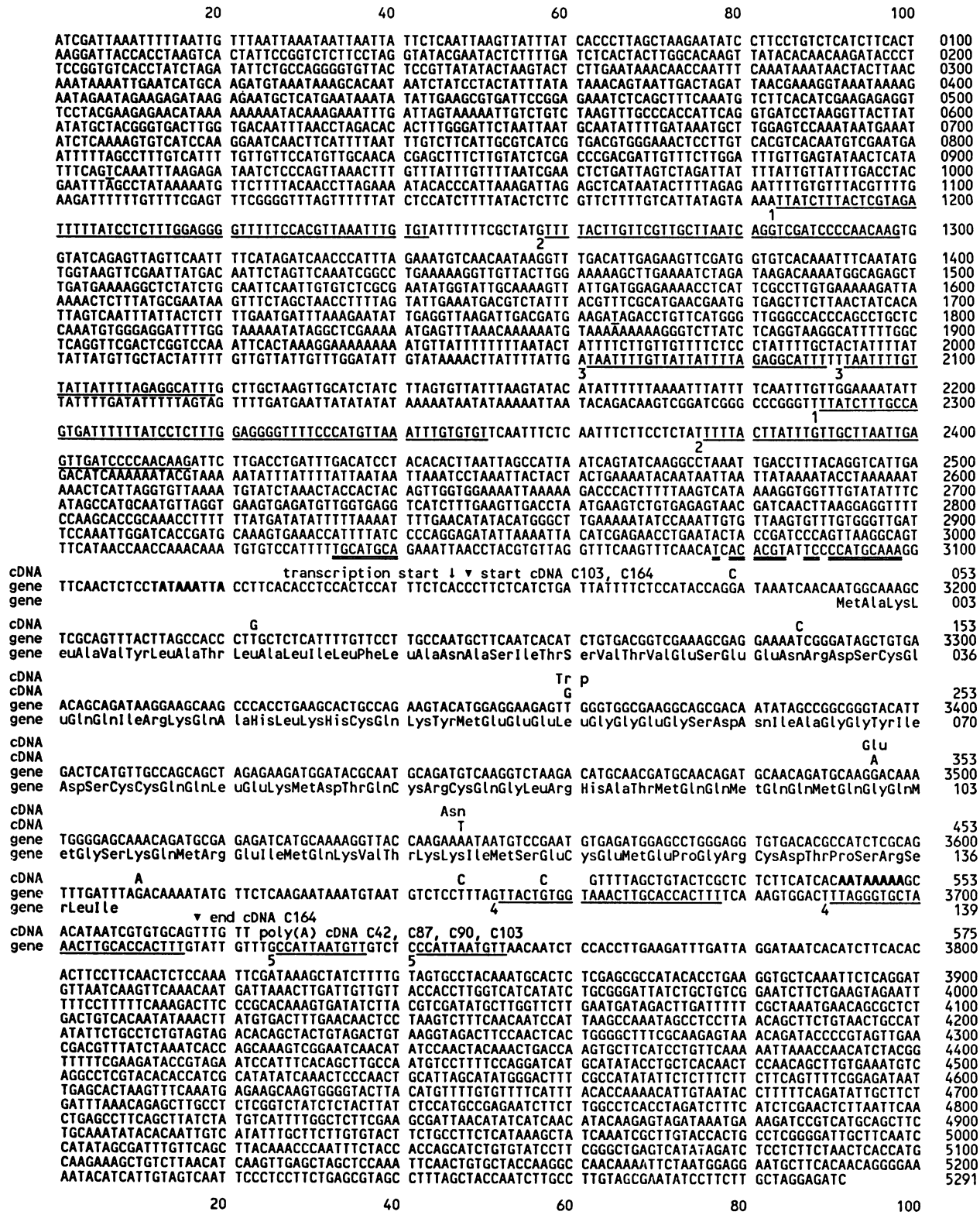

Figure 1. Nucleotide and deduced amino acid sequences of the cotton *Mat5-A* alloallele and the *Mat5-D* alloallele cDNAs encoding methionine-rich 2S albumin storage proteins. The sequence of *Mat5-A* clone GC164–24RC is shown as the reference. Only where the sequences of the cDNAs are different from those of *Mat5-A* are the cDNA sequences shown above those of *Mat5-A*. The probable TATAA sequence and polyadenylation addition sequence are both shown in bold. Pairs of long direct repeats are underlined and numbered below their 5′ end. The limits of a repetitive element found in an unrelated cotton *Lea4* gene are indicated by underlines at nucleotides 907 and 1795. Sequences similar to those found in other 2S albumin storage protein genes (8, 9) are indicated by double underlines.

Table I. *Characteristics of the Mat5-A gene and Mat5-D cDNAs from Gossypium hirsutum*

Organism:
Gossypium hirsutum L cv Coker 201 (Upland cotton), Malvaceae.
Function:
2S albumin storage protein.
Expression:
During the maturation stage in embryo development; coordinately expressed with many other *Mat* genes, including those of vicilin and legumin storage proteins (5, 6).
Source:
Nuclear DNA from embryo cotyledons 20 to 23 d postanthesis (preendoreduplication). Partial *Sau*3AI digest was cloned in λGEM-12 (Promega), and three phage were identified by hybridization with cDNA clone C164. Clone C164 has been described (5).
Genome:
Restriction fragment and sequence analysis indicates that the three phage isolates are from the *Mat5-A* alloallele in the A genome (G.A. Galau, unpublished data). The extensive sequence difference between *Mat5-A* and the sequenced cDNAs indicates that the cDNAs must be transcribed from the other alloallele, *Mat5-D*.
Sequence of *Mat5-A* GC164-24RC:
Cloning of the 7.2-kilobase *Eco*RI/polylinker *Eco*RI fragment of phage isolate 24 into Bluescript (Stratagene), subcloning of a 5.3-kilobase *Cla*I/polylinker *Eco*RI fragment, followed by dideoxy sequencing of both strands with *Taq* polymerase using cloned restriction fragments and unidirectional deletions as double-stranded templates. The polylinker sequence at its 3′ end is not reported. Where checked in the transcribed region, the sequences of isolates GC164-45 and GC164-27 are identical with that of GC164-24RC.
Sequence of *Mat5-D* cDNA C164:
The entire cDNA was sequenced on both strands using cloned restriction fragments as templates. The clone contains an artifact at its 5′ end: a 250-nucleotide-long perfect inverse repeat of a part of its 3′ end (C164 nucleotide 307-556). The artifact is not reported.
Partial Sequence of Other *Mat5-D* cDNA Clones:
Three primers (leftward *Mat5-A* nucleotide 3268–3248 and rightward *Mat5-A* nucleotide 3198–3198 and 3555–3572) were used to sequence most of each of the clones C42, C87, C90, C103, and C115 on at least one strand. Clones C103 and C115 contain artifacts similar to that in C164.
Transcription Start:
The leftward primer *Mat5-A* nucleotide 3268–3248 was used to determine, by primer extension, the transcription start during the maturation stage.
Genbank Accession Nos.:
M86213, *Mat5-A* GC164-24RC; M83301, *Mat5-A* cDNA C164.

	Signal Peptide	Discard	Small Subunit	Discard
Gossypium	MA-KL-AVYLATL-ALI-LFLANA	SITSV--TVESEE----N	----RDSC--EQQIRKQAHLKHCQKYME---EEL-GG	EGS--D---
Bertholletia	MA-KI-SVAAAALLVLMALGHATA	FRATVTTTVVEEE----N	----QEECR-EQMQR-QQMLSHCRMYMRQQMEES---	-PYQTM---
Arabidopsis	MANKLFLVC-AAL-ALCFL-LTNA	SIYR--IVVEFEEDDASN	PIGPRQKCRKEFQQ--SQHLRACQKLMRKQMRQGRGG	GPSLDDEFD

	Large Subunit	Discard
Gossypium	NIAG--GYIDS---CC-QQLEKMDTQ-CRCQGLRHATMQQMQQMQGQM-GSKQMREIMQKVIKK-IMSECEMEPGR--CDTPSRSL	I
Bertholletia	PRRG-MEPHMSE--CC-EQLEGMDES-CRCEGLRMMMMR-MQQEEMQPRG-EQMRRMMR--LAENIPSRCNLSPMR--C--PMGGS	IAGF
Arabidopsis	P-QGPQGRPQLLQQCCME-LRQ-EEPVCVCPTLRQAAK--AVSLQGQQHGPEQVRKIYQ--TAKYLPNICKI-PQVGVC--PFQTT	IPFFPS

Figure 2. Alignment of amino acid sequences of *Gossypium, Bertholletia,* and *Arabidopsis* 2S albumin storage proteins. The sequences are *Mat5-A* from *Gossypium*, pHS-3 from *Bertholletia* (1), and a composite of four *Arabidopsis* sequences (9), giving weight to residues that are held in common with either of the other two sequences. The processing sites are those reported in *Arabidopsis* (9). Residues are underlined if they are shared in at least two of the proteins.

LITERATURE CITED

1. **Altenbach SB, Pearson KW, Leung FW, Sun SSM** (1987) Cloning and sequence analysis of a cDNA encoding a Brazil nut protein exceptionally rich in methionine. Plant Mol Biol **8:** 239–250
2. **Chlan CA, Borroto K, Kamalay JA, Dure L III** (1987) Developmental biochemistry of cottonseed embryogenesis and germination. XIX. Sequences and genomic organization of the α globulin (vicilin) genes of cottonseed. Plant Mol Biol **9:** 533–546
3. **Galau GA, Wang HY-C, Hughes DW** (1991) Sequence of the *Gossypium hirsutum* D-genome alloallele of *Legumin A* and its mRNA. Plant Physiol **97:** 1268–1270
4. **Galau GA, Wang HY-C, Hughes DW** (1992) Cotton *Lea4* (D19) and *LeaA2* (D132) Group 1 *Lea* genes encoding water stress-related proteins containing a 20-amino acid motif. Plant Physiol **99:** 738–788
5. **Goldberg RB, Barker SJ, Perez-Grau L** (1989) Regulation of gene expression during plant embryogenesis. Cell **56:** 149–160
6. **Hughes DW, Galau GA** (1989) Temporally modular gene expression during cotyledon development. Genes Dev **3:** 358–369
7. **Hughes DW, Galau GA** (1991) Developmental and environmental induction of *Lea* and *LeaA* mRNAs and the postabscission program during embryo culture. Plant Cell **3:** 605–618
8. **Irwin SD, Keen JN, Findlay JBC, Lord JM** (1990) The *Ricinus communis* 2S albumin precursor: a single preproprotein may be processed into two different heterodimeric storage proteins. Mol Gen Genet **222:** 400–408
9. **Krebbers E, Herdies L, De Clercq A, Seurink J, Leemans J, VanDamme J, Segura M, Gheysen G, Van Montagu M, Vandekerckhove J** (1988) Determination of the processing sites of an *Arabidopsis* 2S albumin and characterization of the complete gene family. Plant Physiol **87:** 859–866
10. **Voytas DF, Ausubel FM** (1988) A copia-like transposable element family in *Arabidopsis thaliana*. Nature **336:** 242–244