Expression of MADS-Box genes in developing cotton fibre cells

Damien Lightfoot B. Sc.

Discipline of Genetics
Department of Molecular Biosciences
Adelaide University

A dossier submitted in part fulfilment of the requirements for the degree of Bachelor of Science with Honours
22nd October

Table of contents

Literature review

Dossier

Literature Review

1: Cotton

1.1: The cotton plant

Cotton is an angiosperm of the dicotyledon group belonging to the genus *Gossypium*. The most distinctive feature of the plant is the cotton boll (Fig. 1) which contains the cotton seed with attached fibres (or hairs). Upon ripening, the boll bursts open, exposing a mass of long pearly white fibres (Fig. 1), in which the dark coloured seeds are embedded. Cotton is grown primarily for the use of these fibres, known as lint fibres, which are spun into yarns for textile production. As well as these commercially valuable longer fibres, there are also shorter hairs known as fuzz fibres. The cotton fibres have probably evolved as a protective or dispersal mechanism for the seeds (Burrows, 1975).

1.2: Development of the cotton fibre

Cotton fibre development can be divided into four distinct but overlapping phases: (a) initiation, (b) elongation, (c) secondary cell wall thickening and (d) maturation (Naithani et al., 1982).

A) initiation

The process of fibre initiation starts in the period ranging from a day before to two days after anthesis (flowering), with the initiating cells entering elongation immediately (Basra and Malik, 1984). Both the lint and fuzz types of fibre are derived from individual epidermal cells of the outer wall of the ovules in the developing cotton fruit (Beasley, 1975; Fig. 2). About 10% of the epidermal cell population of the cotton ovule develop into fibres (Stewart, 1975). When fibre development begins, the differentiating cells of the epidermal surface enlarge and become rounded. Fibre initials first appear at the bottom of the ovule and

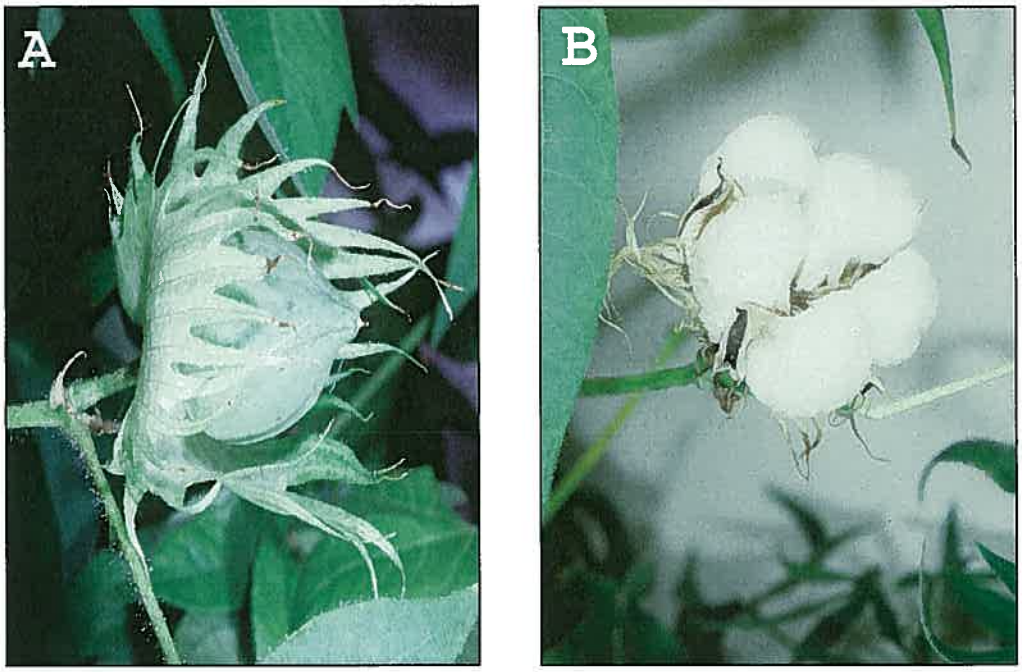


Figure 1: Maturation of the cotton boll: (A) A maturing cotton boll. Inside the boll (or fruit) are the developing fibres attached to the cotton seeds. **(B)** Cotton fibres. After maturation, the cotton boll bursts open, to expose the white cotton fibres.

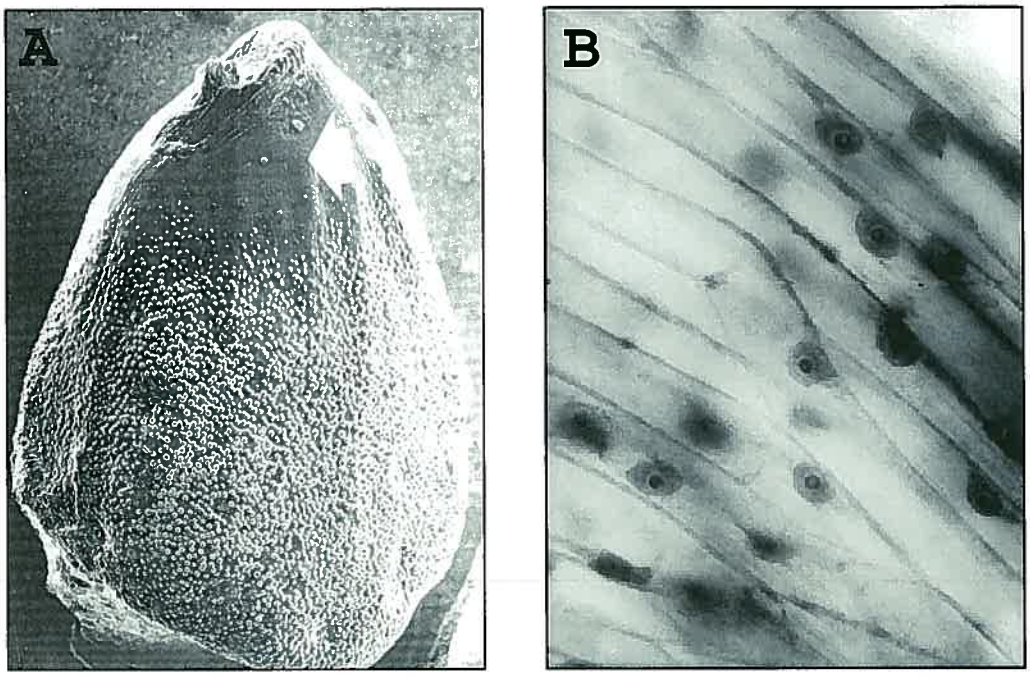


Figure 2: Fibre initiation and elongation: (A) Cotton ovule (50x magnification). Fibres (small white dots) initiate at all areas of the ovule except the micropylar end (white arrow - top of ovule). The fibre initials first appear at the crest of the funiculus (bottom of the ovule) and then around the sides of the ovule **(B)** Elongating cotton fibre cells from four days post-anthesis fibres showing the nuclei and nucleoli, within the vacuole of elongated fibre cells (1000x magnification) (Stewart, 1975).

initiation then progresses around the sides (Fig. 2). The emergence of fibre initials is delayed by up to four days at the top of the ovule (Beasley, 1975).

B) elongation

Fibres that initiate elongation on the day of anthesis develop into lint fibres, while those epidermal cells initiating on subsequent days develop into fuzz fibres attaining lengths of between 1.5 and 3.3 millimetres (Beasley, 1975). Elongation continues for 16 to 25 days with lint fibres reaching lengths of 20 to 60 mm (Schubert *et al.*, 1974; Fig. 2). Both final fibre length and duration of elongation are dependant upon environmental and genetic factors (Basra, 1999).

C) secondary cell wall thickening

Plants have two types of cell wall with the thinner primary cell wall being deposited during fibre elongation (Beasley, 1979). The thicker and stronger secondary cell wall is deposited when elongation ceases and this deposition continues for several weeks. The secondary cell wall, composed largely of cellulose, is deposited inside the primary wall (Schubert *et al*, 1974).

D) maturation

Maturation of the fibre occurs at about 50 to 60 days post-anthesis (DPA) when the fruit capsule opens. The mature fibres desiccate and collapse to form twisted ribbon-like structures (Basra, 1999).

1.3: The cotton genome

There are more than 50 species of the *Gossypium* genus including diploids and tetraploids (Endrizzi, 1985). *G. hirsutum*, which accounts for 90% of world cotton production, is a tetraploid with a haploid chromosome number of 26 (Fehr and Hedley, 1980). The tetraploid species are thought to have formed through the hybridisation, and chromosome doubling, of two ancestral diploid species 1 to 1.2 million years ago (Wendel, 1989). The diploid species are categorized into seven distinct groups, A-G, based on chromosome pairing relationships (Reinisch *et al.*, 1994). *G. hirsutum* is classified as having an AADD genome as a result of the hybridisation of ancestral A and D genome species. *G. herbaceum* (A genome) and *G. raimondii* (D genome) are the extant diploids most similar to the genomes of the tetraploid species (Wendel, 1989).

1.4: The future of the cotton fibre

Currently, the qualities that are being engineered into cotton relate to crop management issues such as insect-resistance (Frutos *et al.*, 1999), herbicide-tolerance (Schmidt, 1995) and stress-resistance (John and Stewart, 1992). However, in the future the aims are to modify the fibre to introduce new properties or to enhance existing properties such as length, strength and/or fineness in order to produce a more desirable end product. Before this can be accomplished, a fundamental understanding of how the fibre quality traits are biologically regulated must be obtained (John and Stewart, 1992).

The developing cotton fibre is an attractive experimental system because the fibre cells originate and end as a single cell and thus elongation can be studied free from the complications that arise from cell division (Basra and Malik, 1984). Therefore, in addition to

its economic importance as a natural textile fibre, the developing cotton fibre is an excellent model system for unravelling the fundamental processes of plant cell growth.

2: MADS-box genes

2.1: MADS-box genes act as homeotic genes

The term 'MADS-box gene' was first used ten years ago to describe a family of genes that encode DNA-binding proteins that is conserved in plants, fungi and animals (Schwarz-Sommer *et al.*, 1990). The name was derived from the four founding members of the family: MCM1 (*Saccharomyces cerevisiae*), AGAMOUS (*Arabidopsis thaliana*), DEFICIENS (*Antirrhinum majus*) and SRF (Serum Response Factor from *Homo sapiens*). The proteins were found to possess DNA-binding and dimerization functions and acted as transcription factors (Schwarz-Sommer *et al.*, 1990).

In plants, some mutations in MADS-box genes were observed to cause phenotypic changes in which one type of floral organ (sepals, petals, stamen and carpels) would develop in the place usually occupied by a different type of floral organ (Riechmann and Meyerowitz, 1997). This type of mutation is analogous to homeotic gene mutations observed previously in animals where mutations alter the identity of body segments (Theissen *et al.*, 2000).

2.2: Structure of MADS-box proteins

Within each protein of the MADS-box family there is a region of 56 amino acids, termed the MADS-box, which is highly conserved between all members, with nine amino acid residues identical in all species (Fig. 3). The MADS box is the major determinant of DNA binding, but also performs a role in dimerization (Shore and Sharrocks, 1995).

Another conserved region in MADS-box proteins is the K-box, which is particularly well

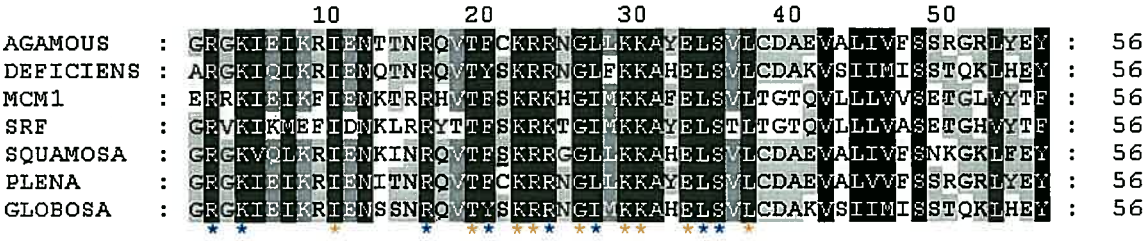


Figure 3: MADS-box domain alignment. The MADS-box domains of the four founding members of the MADS-box family (AGAMOUS from *Arabidopsis thalina*; DEFICIENS from *Antirrhinum majus*; MCM1 from *Saccharomyces cerevisiae* and SERUM RESPONSE FACTOR from *Homo sapiens*) as well as MADS-box regions from three other proteins (SQUAMOSA, PLENA and GLOBOSA from *Antirrhinum majus*). The nine identical residues in all MADS-box proteins are indicated by a red asterix (*), while eight other residues positions, which are highly conserved are indicated by a blue asterix (*). At each position, the residues shaded in black are those that are identical or chemically conserved, while lighter shading of residues at one position indicates a weaker level of conservation.

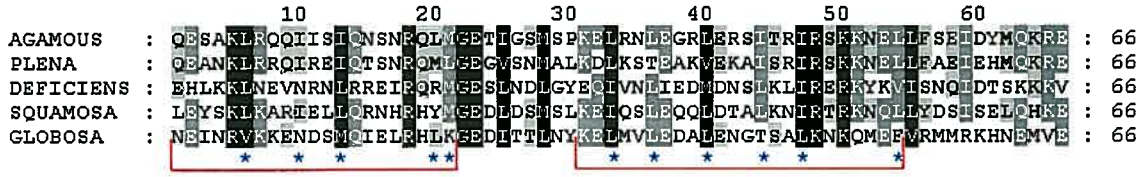


Figure 4 – K-box domain alignment. The K-box domains of five plant MADS-box genes (AGAMOUS, PLENA, SQUAMOSA and GLOBOSA from *Arabidopsis thalina* and DEFICIENS from *Antirrhinum majus*). Hydrophobic residues that are highly conserved are indicated by a blue asterix while red brackets indicate the amino acids that form the putative amphipathic alpha-helices. At each position, the residues shaded in black are those that are identical or chemically conserved, while lighter shading of other residues indicates a weaker level of conservation at that position.

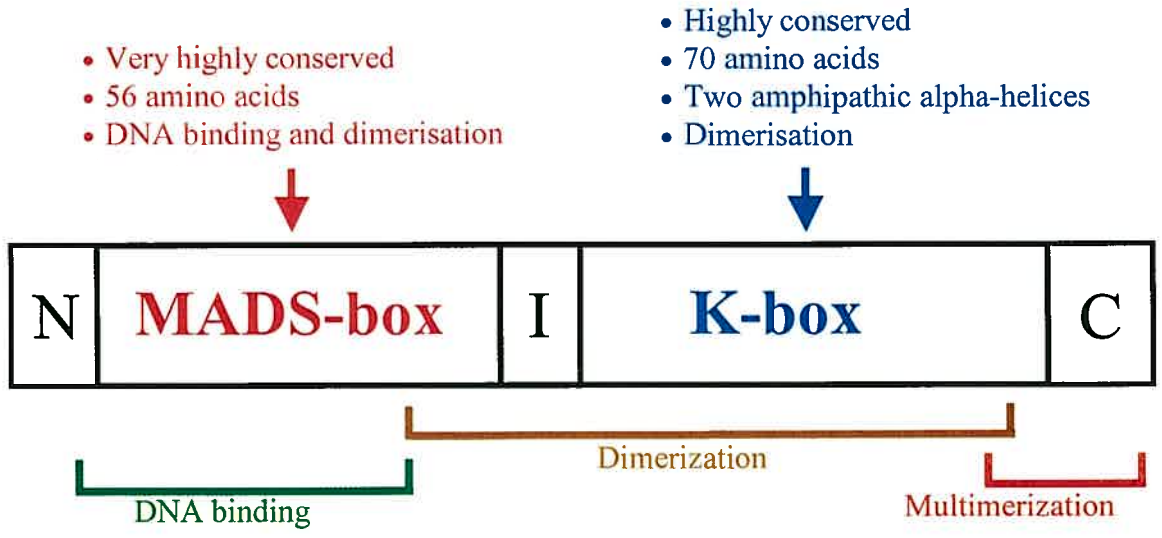


Figure 5 -MADS-box protein representation: A typical MADS-box protein with five regions. The MADS-box region is the most highly conserved region followed by the K-box. The intervening (I), N-terminal (N) and C-terminal (C) vary in sequence as well as length between different MADS-box proteins (Vergara-Silva *et al.*, 2000).

conserved in a region containing hydrophobic residues (Fig. 4) which have a spatial arrangement that may permit the formation of α–helices as seen in the keratin coiled coil motif (Ma *et al.*, 1991). It is believed that this region allows MADS-box proteins to interact with each other to form dimers (Riechmann and Meyerowitz, 1997).

Between the MADS-box and the K-box is a region termed the intervening, or I, region (Fig. 5) which is much more variable in both sequence and length, being only weakly conserved. In plants, the I region is important for the specificity of dimerization (Riechmann *et al.*, 1996).

The MADS-box is usually located at the amino terminus of the protein, but some members of the family have an amino terminal extension (N-region) that is variable in length and sequence (Riechmann and Meyerowitz, 1997). The carboxy-terminal region (C-region) of the MADS-box protein is also weakly conserved and variable in length, and has a suggested role in multimerization (Reichmann and Meyerowitz, 1997).

2.3: ABC(DE) model of floral organ identity

Studies of floral development in two plants, *Arabidopsis thaliana* and *Antirrhinum majus*, led to the derivation of the 'ABC model' of floral organ identity. The model was built around the observation that homeotic mutations in plants changed the identity of floral organs in two adjacent whorls of a flower (Coen and Meyerowitz, 1991). Flowers have four concentric regions (whorls) with each being occupied by a different type of floral organ (Fig. 6). There are four types of floral organ, sepals, petals, stamen and carpels, with the outermost whorl being dictated as whorl one (sepals) (Riechmann and Meyerowitz, 1997).

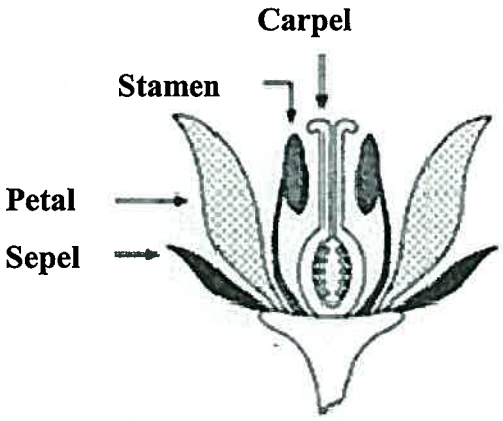


Figure 6: Schematic representation of a flower: The four concentric whorls of a flower are represented above and are conventionally numbered from whorl one (sepals) to whorl four (carpels). (Adapted from Saedler and Huijser, 1995).

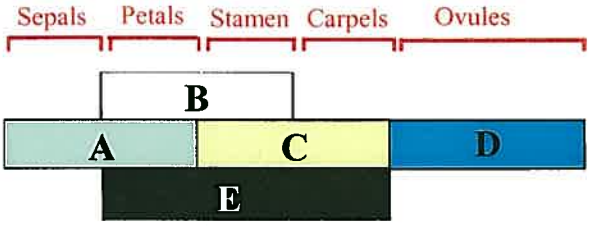


Figure 7 – ABC(DE) model of floral organ identity: A schematic longitudinal section through a flower. The action of class A MADS-box genes alone determines the development of sepals, A, B and E that of petals, B, C and E that of stamen and C and E that of carpels. Class A genes repress class C genes in whorls one and two and C class genes repress A class genes in whorls three and four. Class D genes dictate ovule identity (Adapted from Coen and Meyerowitz, 1991).

Three classes of mutant were observed to affect the morphology of the whorls - designated as class A (affecting whorls one and two), class B (affecting whorls two and three) and class C (affecting whorls three and four). According to the ABC model, specification of whorl one (sepal identity) is therefore dictated by class A genes, whorl two (petal identity) is dictated by a combination of class A and B gene activities, whorl three (stamen identity) by a combination of class B and C gene activities and whorl four (carpel identity) by class C genes (Coen and Meyerowitz, 1991; Fig. 7).

However, the ABC model has recently been amended to include D and E function specifying genes. Class D genes were identified in petunia (*Petunia hybrida*) as specifying ovule identity (Cheng *et al.*, 2000). However, ectopic expression of these genes (ABCD) was not sufficient to result in transformation of leaves into floral organs (Pelaz *et al.*, 2000). The functions of the B and C class genes were shown to require the activities of three redundantly functioning MADS-box genes, *Sepallata 1* to *Sepallata 3*, designated as having E function (Palez *et al.*, 2000). Ectopic expression studies, using either A, B and E or B, C and E function genes were able to transform leaves into petal- or stamen-like organs respectively (Honma and Goto, 2001). All of the genes identified as being involved in the ABC(DE) model of floral organ identity have been characterized as MADS-box genes.

2.4: MADS-box gene function

MADS-box genes were first extensively characterised for their roles in reproductive organs, where they control floral meristem and floral organ identity (see section 2.3). However, MADS-box genes have since been identified that are expressed predominantly or exclusively in tissues other than the flower. The differential expression of some MADS-box genes has

indicated that they may have roles in trichome development (Alvanez-Buylla, 2000), root development (Rounsley *et al.*, 1995; Alvanez-Buylla, 2000), the transition from vegetative to reproductive development (Immink *et al.*, 1999), ovule development (Angenent *et al.*, 1995; Lopez-Dee *et al.*, 1999), fruit development (Gu *et al.*, 1998; Yao *et al.*, 1999; Sung *et al.*, 2000), embryogenesis (Heck *et al.*, 1995; Perry *et al.*, 1999; Fernandez *et al.*, 2000) and seed coat development (Buchner, 1998). These studies have not yet demonstrated the role of the MADS-box gene product in the relevant tissue but they suggest that the genes may be involved in the morphogenesis of plant organs other than flowers (Vergara-Silva *et al.*, 2000).

MADS-box gene transcripts have been detected in a number of different cotton tissues, including elongating fibres and whole ovules (Orford, unpublished). This is of interest when considered in combination with the likely involvement of MADS-box genes in developmental processes other than floral organ identity. This suggests that MADS-box genes may play a role in the development of the cotton fibre.

3: Alternative splicing

3.1: mRNA processing

Precursor RNAs, or pre-mRNAs, of eukaryotic organisms are processed after transcription. The transcripts have 5' cap and 3' poly A tail structures added and they undergo removal of introns and subsequent joining of the flanking exonic regions (Goodall *et al.*, 1991). Intron excision is mediated by a ribonucleoprotein complex, the spliceosome, which assembles on each intron. Spliceosomes recognise an intron by consensus sequences at its 5' and 3' ends, and also by the different base composition of introns compared with exons (reviewed by Simpson and Filipowicz, 1996). Splicing involves two transesterification reactions and can be simplified in two steps:

(1) The 5' end of the intron is cleaved and attached to an adenosine nucleotide within the intron to form an intron lariat structure. The adenosine is within a semi-conserved region in the intron known as the branchpoint (Fig. 8).

(2) The 3' end of the intron is cleaved and the 3' and 5' exon ends are ligated together. The intron is released as a lariat and degraded (Fig. 8).

3.2: Mechanisms of alternative splicing

Organisms use a mechanism known as alternative splicing, whereby introns can be processed in alternative ways to give different mRNA products and therefore different proteins. Alternative splicing can be accomplished by a number of different methods:

- The skipping of entire regions that are usually exons;
- The inclusion of regions that are usually introns;
- Selection of different 5' and 3' splice sites to include regions that are usually intronic or to exclude regions that are usually exonic.

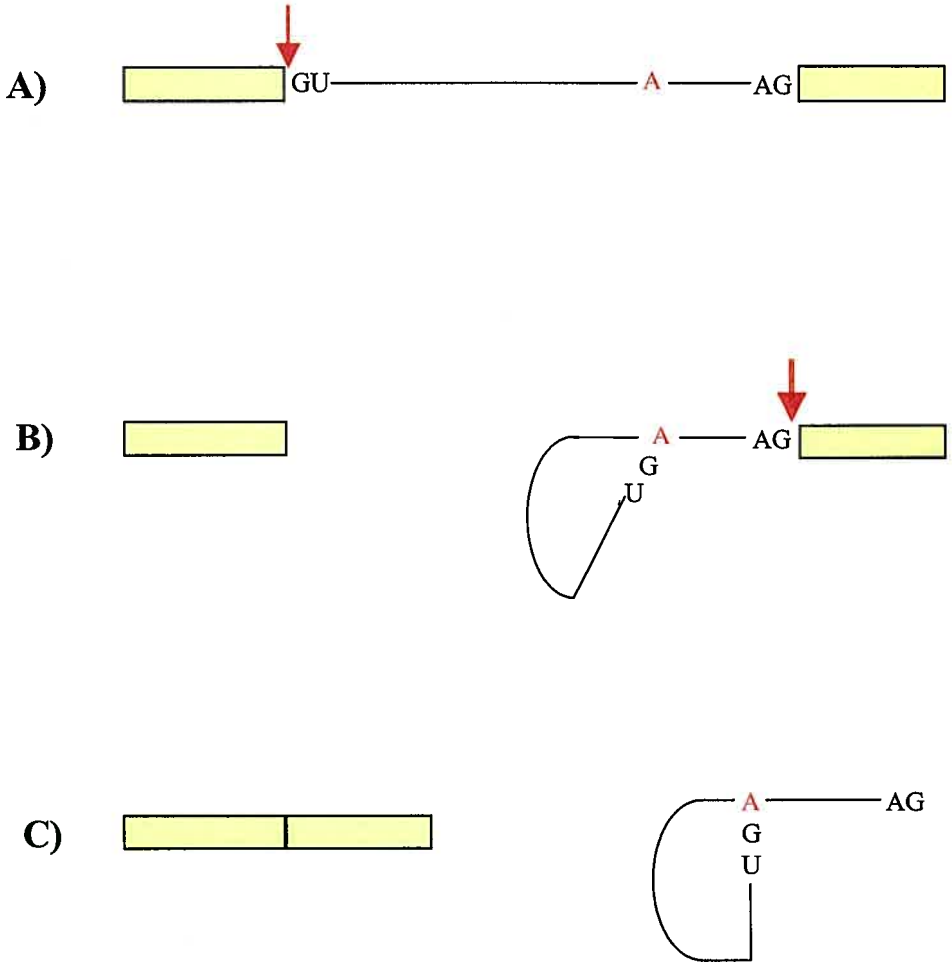


Figure 8: Mechanism of pre-mRNA splicing: Introns are excised in a two step process. The first step involves cleavage at the 5' splice site with formation of an intron lariot at an adenosine nucleotide by the formation of a 2'-5' phosphodiester bond. The adenosine nucleotide is within a weakly conserved region called the branchpoint, usually 18-40 nucleotides upstream of the 3' splice site (A to B). Secondly, the 3' splice site is cleaved and the exons are ligated together. The intron lariot is released to be debranched and degraded (B to C). (Adapted from McKeown, 1992)

3.3: The role of alternative splicing

The selection of different combinations of splice sites within the precursor mRNA generates different mature mRNAs and therefore different proteins, from a single gene (Smith and Valcarel, 2000). Alternative splicing is often tightly regulated in either a cell-type or developmental-stage-specific manner (Chabot, 1996). Some alternative splicing events are constitutive, with the variants being maintained at constant ratios, while others are regulated by cues that are given in either a spatial or temporal manner. Structural and functional differences can be introduced into the protein product by altering the nucleotides that are included in the fully processed mRNA (Lopez, 1998).

Transcription factors, including the MADS-box family, are modulated proteins with distinct regions playing specialized roles in different processes (Mitchell, 1989). It makes sense that the alternative splicing mechanism has evolved to regulate the inclusion or exclusion of sections of pre-mRNA that encode specific domains of a protein to allow different functions to be conferred upon the potential proteins (Lopez, 1995). Alternative splicing has been found to be a feature of MADS-box gene expression in both plants and animals (Schmitz *et al.*, 2000).

3.4: Alternative splicing of MADS-box transcripts in mammals

There are many well-characterized alternatively spliced mammalian MADS-box transcripts such as murine *SRF* (serum response factor). Compared to the major SRF transcript, an alternatively spliced transcript, SRFΔ5, lacks a portion of the C-terminal region and was shown to play an important role as a dominant negative regulator in muscle differentiation (Belaguli *et al.*, 1999). The protein was able to form dimers and bind DNA but was deficient

for transactivation and therefore, in specific tissue types, acted to repress SRF promoter activity (Belaguli *et al.*, 1999).

Muscle myocyte enhancer factor-2A, 2C and *2D* are further examples of MADS-box genes that undergo extensive alternative splicing to give both muscle-specific and ubiquitous expression of different protein isoforms. Mutually exclusive exons are used in different muscle tissues to confer different functions to proteins in different tissue locations (reviewed by Olson *et al.*, 1995).

3.5: Alternative splicing of MADS-box transcripts in plants

In barley, the MADS-box gene BM1 is spliced alternatively at the 3' splice site to give two transcripts, BM1 and BM1-2 of which BM1-2 was found to be preferentially expressed in all tissues (Schmitz *et al.*, 2000). The two putative proteins differ by two amino acids with the BM1 protein having an extra leucine and glutamine at the end of the second amphipathic α–helix (Fig. 4). It was suggested that the leucine and glutamine in BM1-2 would modify the amphipathic nature of the predicted α-helix by extending the hydrophobic region by one amino acid. This may modify the specificity of protein/protein interactions for the second protein (Schmitz *et al.*, 2000).

In maize, alternatively spliced genes such as *ZEMα*, which encodes transcripts ZEM1 to ZEM5, have been identified (Montag *et al.*, 1995). Differentially spliced forms were shown to be preferentially expressed in different tissues. The ZEM1 to ZEM3 transcripts were found to encode putative proteins with very short regions C-terminal of the MADS-box and consequently lacked the entire K-box (Fig. 5). This region is involved in dimerization and

tertiary protein interactions, and as such, the proteins may have different dimerization activities (Montag *et al.*, 1995).

Another example of an alternatively spliced MADS-box gene was found in the wild rose (*Rosa rugosa*), where six MADS-box containing cDNA clones were identified, MASAKO C1 to MASAKO C6, arise *via* alternative splicing from a single transcript. MASAKO C2 and C4 were found to have an additional serine residue, immediately C-terminal of the MADS-box, due to different 3' splice sites selection compared with the other four transcripts (Kitahara and Matsumoto, 2000). At the corresponding position another MADS-box gene, MEF2C, is known to have a serine residue that was shown to enhance DNA-binding and transcription activity when phosphorylated (Molkentin *et al.*, 1996). Therefore, the differently spliced MASAKO C2 and MASAKO C4 protein products may be phosphorylated and have different activities compared with the other variants (Kitahara and Matsumoto, 2000).

MADS-box cDNA clones PrMADS 4 to PrMADS 9 were isolated from *Pinus radiata* and shown to arise *via* alternative splicing from one transcript (Walden *et al*, 1998). While both proteins would possess the C-terminal region, the putative protein derived from PrMADS 8 would have 14 residues missing at the 3' end of the K-box and PrMADS 7 would completely lack the K-box. The authors suggested that the proteins lacking the K-box might regulate the activity of specific proteins by a dominant negative mechanism.

4: Previous lab work

Whilst MADS-box genes were originally identified as regulators of floral organ identity, they have been implicated in a number of other developmental processes such as root development, embryogenesis, trichome development, the transition from vegetative to reproductive development, ovule development, fruit development and seed coat development in many different plant species (reviewed by Krogan and Ashton, 2000; Jack, 2001).

MADS-box gene transcripts were first discovered in cotton fibre cells when a sample of cDNA from 12 DPA cotton fibre was used as a positive control in an unrelated experiment (Donati, 1997). Four different MADS-box transcripts (6.1, 6.2, 12.2 and 24.11) were detected in elongating cotton fibres (Orford, unpublished), and full-length cDNAs corresponding to these transcripts were obtained (Malone, 2000).

Two of the transcripts were shown to correspond to mRNA splice variants, namely 24.11A and 24.11B. The cDNAs were very similar to each other except that 24.11B was found to contain an extra 120 bp section of sequence flanked on either side by characteristic intron splice sites. It was suggested that the 120 bp of extra sequence represented an unspliced intron (Malone, 2000). This supposition was supported by the presence of several bands of different sizes after RT-PCR of fibre mRNA, indicating different lengths of mRNA between the primers. A PCR product from the corresponding region of cotton genomic DNA, 24PCR1, was sequenced and shown to contain the 120 bp insertion that was present in 24.11B. When aligned against 24.11A and 24.11B, the genomic sequence had three other likely intronic regions that were spliced out to give 24.11A and 24.11B (Malone, 2000).

The presence of differently sized transcripts amplified by the RT-PCR and the intron-like sequences in the genomic PCR clone led to the proposal that the differently sized bands were the result of alternative splicing events in which intronic regions were not being removed (Malone, 2000). Of particular interest was the observation, based on a preliminary result, that the levels of some of the larger sized transcripts reduced over time in the developing cotton fibre cells (Malone, 2000). This suggested specifically regulated differential splicing of this MADS-box gene in developing cotton fibres that was not observed for the other MADS-box transcripts (6.1, 6.2 and 12.2) that were identified.

5: Project aims

The broad aim of this project is to confirm and characterise the differently sized 24.11 transcripts observed after RT-PCR of cotton fibre mRNA and to determine whether or not they are alternatively spliced mRNAs from a single gene. This will be done by repeating the RT-PCR experiment and using Southern analysis to verify that the differently sized variants are similar in sequence to 24.11A, 24.11B and 24PCR1 and may, therefore, be derived from a single transcript. The PCR products will be cloned and sequence data from the different transcripts compared with the previously sequenced transcripts (24.11A and 24.11B) and the genomic PCR clone to investigate the presence and extent of alternative splicing in the fibre tissue. Then, if time permits, a genomic library screening method or a genome walking approach will be used to find the gene that corresponds to the cDNAs.

References

Alvarez-Buylla, E.R., Lijegren. S.J., Palez, S., Gold, S.E., Burgeff, C., Ditta, G.S., Vergara-Silva, F., Yanofsky, M.F. (2000) MADS-box gene evolution beyond flowers: expression in pollen, endosperm, guard cells, roots and trichomes. *The Plant Journal* 24: 457-466.

Angenent, G.C., Fraken, J., Busscher, M., van Dijken, A. Van Went, J.L., Dons, H., van Tumen A.J. (1995) A novel class of MADS-box genes is involved in ovule development in petunia. *The Plant Cell* 7:1569-1582.

Basra, A.S. (1999) Cotton fibres: *Developmental biology, quality improvement and textile processing*, The Haworth Press, USA.

Basra, A.S., Malik, C.P. (1984) Development of the Cotton Fiber. *International Review Of Cytology* 89: 65-113.

Beasley, C.A. (1975) Developmental morphology of cotton flowers and seed as seen with the scanning electron microscope. *American Journal Of Botany* 62: 584-592.

Beasley, C.A. (1979) Cellulose content in fibres of cotton which differ in their lint lengths and extent of fuzz. *Physiologia Plantarum* 45: 77-82.

Belaguli, N., Zhou, W., Trinh, T., Majesky, M., Schwartz, R. (1999) Dominant negative murine Serum Response Factor: alternative splicing within the activation domain inhibis transactivation of Serum Response Factor binding targets. *Molecular and Cellular Biology*: 19: 4582-4591.

Buchner, P., Boutin, J.P. (1998) A MADS-box transcription factor of the AP1/AGL9 subfamily is also expressed in the seed coat of pea (*Pisum sativum*) during development. *Plant Molecular Biology* 38: 1253-1255.

Burrows, F.M., (1975) Wind-borne seed and fruit movement. *New Phytologist* 75: 405-418.

Chabot, B. (1996) Directing alternative splicing: cast and scenarios. *Trends In Genetics*: 12 472-478.

Cheng, X.F., Wittich, P.E., Kieft, H., Angenent, G., XuHan, X., van Lammeren, A.A.M. (2000) Temporal and spatial expression of MADS box genes, *FBP7* and *FBP11*, during initiation and early development of ovules in wild type and mutant *Petunia hybrida*. *Plant Biology* 2: 693-702.

Coen, E.S., Meyerowitz, E.M. (1991) The war of the whorls: genetic interactions controllong flower development. *Nature* 353: 31-37.

Donati, S. (Unpublished) (1997) Tristyly in *Oxalis spp.*: Aspects of morphology and development. Honours thesis, Department of genetics, The University of Adelaide.

Endrizzi, J.E., Turcotte, E.L., Kohel, R.J. (1985) Genetics, cytology and evolution of *Gossypium. Advances in Genetics* 23: 271-275.

Fehr, W.R., Hedley, H.H. (1980) *Hybridisation of crop plants.* American Society Of Agronomy And Crop Science; Madison, Wisconsin.

Fernandez, D.H., Heck, G.E., Perry, S.E., Patterson, S.E., Bleekler, A.B., Fang, S.C. (2000) The embryo MADS-domain factor AGL15 acts postembryonically: inhibition of perianth senescence and abscission via constitutive expression. *The Plant Cell* 12:183-197.

Frutos, R., Rang C., Royer F. (1999) Managing insect resistance to plants producing *Bacillus thuringiensis* toxins. *Critical Reviews In Biotechnology* 19:227-276.

Goodall, G.J., Kiss, T., Filipowicz, W. (1991) Nuclear RNA splicing and small nuclear RNAs and their genes in higher plants. *Oxford Survey Of Plant, Molecular And Cell Biology* 7: 255-296.

Gu, Q., Ferrandiz, C., Yanofsky, M.F., Martienssen, R. (1998) The *FRUITFULL* MADS-box gene mediates cell differentiation during *Arabidopsis* fruit development. *Development* 125: 1509-1517.

Heck, G.E., Perry, S.E., Nichols, K.W., Fernandez, D.E. (1995) AGL15, A MADS-domain protein expressed in developing embryos. *The Plant Cell* 7: 1271-1282.

Honma, T., Goto, K. (2001) Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. *Nature* 409: 525-529.

Immink, R.G.H., Hannapel, D.J., Ferrario, S., Busscher, M., Franken, J., Lookeren Campagne, M.M., Angenent, G.C. (1999) A petunia MADS-box gene involved in the transition from vegetative to reproductive development. *Development* 126: 5117-5126.

Jack, T. (2001) Plant development going MADS. *Plant Molecular Biology* 46: 515-520.

John, M.E., Stewart, S.M. (1992) Genes For Jeans: Biotechnological Advances In Cotton. *Trends in Biotechnology* 10: 165-170.

Kitahara, K., Matsumoto, S. (2000) Rose MADS-Box Genes 'MASAKO C1 and D1' homologous to class C floral identity genes. *Plant Science* 151: 121-134.

Krogan, N.T., Ashton, N.W. (2000) Ancestry of plant MADS-box genes revealed by Bryophyte (*Physcomitrella patens*) homologues. *New Phytology* 147: 505-517.

Lopez, A.J. (1995) Developmental role of transcription factor isoforms generated by alternative splicing. *Developmental Biology* 172: 396-411.

Lopez, A.J. (1998) Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annual Review Of Genetics* 32: 279-305.

Lopez-Dee, Z.P., Wittich, P., Enrico Pe, M., Rigola, D., Del Buono, I., Sari Gorla, M., Kater, M.K., Colombo, L. (1999) A novel rice MADS-box gene expressed during ovule development. *Developmental Genetics* 25: .237-244.

Ma, H., Yanofsky, M.F., Meyerowitz, E.M. (1991) AGL1-AGL6, An *Arabidopsis* gene family with similarity to floral homeotic and transcription factor genes. *Genes and Development* 5: 484-495.

Malone, K.M., (Unpublished) (2000) Expression of MADS-box genes in developing cotton fibres. Honours dossier, Department of Genetics, The University of Adelaide.

McKeown, M. (1992) Alternative mRNA splicing. *Annual Reviews in Cell Biology* 8:133-155.

Mitchell, P., Tollervey, D. (1989) mRNA turnover. *Current Opinion in Cell Biology* 13: 320-325.

Molkentin, J.D., Li, L., Olson, E.N. (1996) Phosphorylation of the MADS-box transcription factor MEF2C enhances its DNA binding activity. *Journal of Biological Chemistry* 271:17199-17204.

Montag, K., Salamini, F., Thompson, R.D. (1995) *ZEMa*, a member Of A Novel Group Of MADS-Box Genes, Is Alternatively Spliced In Maize Endosperm. *Nucleic Acids Research* 23: 2168-2177.

Naithani, S.C., Rama, R., Singh, Y.D. (1982). *Physiologia plantarum* 54: 225-29.

Olson, E.N., Perry, M., Schulz, A. (1995) Regulation of muscle differentiation by the MEF2 family of MADS box transcription factors. *Developmental Biology* 172: 2-14.

Pelaz, S., Ditta, G.S., Baumann, E., Wisman, E., Yanofsky, M.F. (2000) B and C floral organ identity functions require *SEPALLATA* MADS-box genes. *Nature* 405: 200-203.

Perry, S.E., Lehti, M.D., Fernandez, D,E. (1999) The MADS-domain protein AGAMOUS-like 15 accumulates in embryonic tissues with diverse origins. *Plant Physiology* 120: 121-129.

Reinisch A.J., Dong, J., Brubaker, C.L., Stelly, D.M., Wendel, J.F. Paterson, A.H. (1994) A detailed RFLP map of cotton, *Gossypium hirsutum* x *Gossypium barbadense*; Chromosome organization and evolution in a disomic polyploid genome. *Genetics* 138: 829-847.

Riechmann, J.L, Meyerowitz, E.M. (1997) MADS domain proteins in plant development. *Biological Chemistry* 378: 1079-1101.

Riechmann, J.L, Wang, M., Meyerowitz, E.M. (1996) DNA-binding properties of *Arabidopsis* MADS domain homeotic proteins APETALA1, APETALA3, PISTILLATA and AGAMOUS. *Nucleic Acids Research* 24: 3134-3141.

Rounsley, S.D., Ditta, G.S., Yanofsky, M.F. (1995) Diverse roles for MADS-box genes in *Arabidopsis* development. *Plant Cell* 7: 1259-1269.

Saedler, H., Huijser, P. (1993) Molecular biology of flower development in *Antirrhinum majus* (Snapdragon). *Gene* 135 239-243.

Schmidt, K. (1995) Whatever happened to the gene revolution? *New Scientist* 145: 21-25.

Schmitz, J., Frazen, R., Ngyuen, T.H., Garcia-Maroto, F., Pozzi, C., Salamini, F., Rohde, W. (2000) Cloning, mapping and expression analysis of barley MADS-box genes. *Plant Molecular Biology* 42: 899-913.

Schubert, A.M., Benedict, C.R., Berlin, J.D., Kohel, R.J. (1974) Cotton fiber development - kinetics of cell elongation and secondary wall thickening. *Crop Science* 13: 704-709.

Schwarz-Sommer, Z., Huijser, P., Nacken, W., Saedler, H., Sommer, H. (1990) Genetic control of flower development by homeotic genes In *Antirrinum majus*, *Science* 250: 931-936.

Shore, P., Sharrocks, A.D. (1995) The MADS-box family of transcription factors. *European Journal Of Biochemistry* 229: 1-13.

Simpson, G.G., Filipowicz, W. (1996) Splicing of precursors to mRNA in higher plants: mechanisms, regulation and sub-nuclear organisation of the splicing machinery. *Plant Molecular Biology* 32: 1-41 1996.

Smith, C.W.J., Valcarel, J. (2000) Alternative pre-mRNA splicing: the logic of co-ordinatorial control. *Trends In Biochemical Science* 25: 381-387.

Stewart, J.M. (1975) Fiber initiation on the cotton ovule (*Gossypium hirsutum*). *American Journal Of Botany* 62: 723-730.

Sung, S.K., Yu, G.H., Nam, J., Jeong, D.H., An, G. (2000) Developmentally regulated expression of two MADS-box genes, MdMADS3 and MdMADS4, in the morphogenesis of flower buds and fruits in apple. *Planta* 210: 519-528.

Theissen, G., Becker, A., Rosa, A.D., Kanno., A., Kim, A., Munster, T., Winter, K., Saedler, H. (2000) A short history of MADS-box genes in plants. *Plant Molecular Biology* 42: 115-149.

Vergara-Silva, F., Martinez-Castilla, L., Alvarez-Buylla, E.R. (2000) MADS-box genes: development and evolution of plant body plans. *Journal Of phycology* 36: 803-812.

Walden, A.R., Wang, D.Y., Walter, C., Gardner, R.C. (1998) A large family of TM3 MADS-box cDNAs in *Pinus radiata* includes two with deletions of the conserved K domain. *Plant Science* 138: 167-176.

Wendel, J.F. (1989) New world tetraploid cottons contain old world cytoplasm. *Proceedings Of The National Academy Of Sciences Of The USA* 86: 4132-4136.

Yao, J.L., Dong, Y.H., Kvarnhedon, A., Morris, B. (1999) Seven MADS-Box genes in apple are expressed in different parts of the fruit. *Journal Of The American Society For Horticultural Science* 124: 8-13.

Dossier

Introduction

Background

The MADS-box family of proteins consist of transcription factors that are defined by the presence of a 56 amino acid DNA binding domain termed the MADS-box (Fig. 1). This domain is well conserved between all members of the MADS-box family across a diverse range of species. The MADS-box domain is involved in DNA binding and dimerization and is the most conserved region of the MADS-box protein (Shore and Sharrocks, 1995). Another region of the protein, the K-box, is also well conserved, but much less so than the MADS-box. The K-box is approximately 70 amino acids long and plays a role in MADS-box protein dimerization. The dimerization between MADS-box proteins is believed to be mediated by two putative amphipathic α-helices within the K-box region (Riechmann and Meyerowitz, 1997). Also present in the typical MADS-box protein are the intervening, N-terminal and C-terminal regions. The C-terminal region is involved in multimerization of MADS-box proteins, while the intervening region plays a minor role in determining the specificity of dimerization (Riechmann *et al.*, 1996)).

MADS-box genes have been extensively investigated and characterized because of their roles in floral organ identity in plants. Inactivating mutations in MADS-box genes result in phenotypic changes, where one type of floral organ (sepals, petals, stamen or carpels) develops in the location usually occupied by an alternative type of floral organ (Reichmann and Meyerowitz, 1997). This type of mutation is analogous to mutations in homeotic genes in animals which result in altered body segment identity (Theissen *et al.*, 2000).

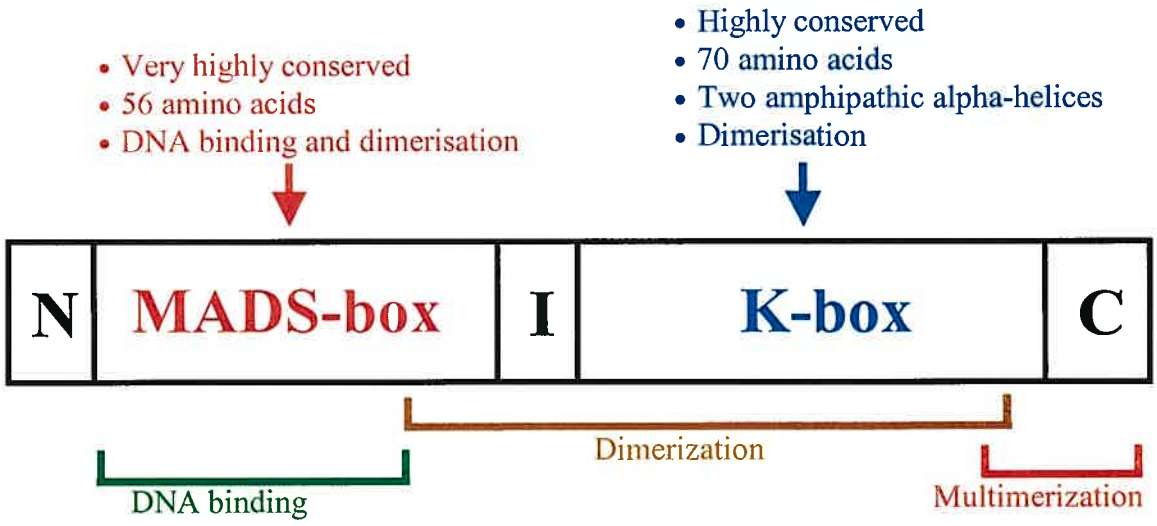


Figure 1: Typical MADS-box protein structure.
The MADS-box domain is the most highly conserved region of the protein, followed by the K-box. The intervening (I), N-terminal (N), and C-terminal (C) regions vary in sequence as well as length between different MADS-box proteins. The regions of protein that are known to be required for various functions (DNA binding, dimerization and multimerization) are indicated by square brackets below the protein.

Although MADS-box genes were initially studied because of their role in floral organ identity, they have since been implicated in a variety of other processes. MADS-box gene transcripts have been detected that are specifically expressed in tissues other than flower, indicating that MADS-box genes may have roles in other developmental processes. Based on their temporal and spatial patterns of expression, MADS-box genes have been implicated in root development, trichome development, the transition from vegetative to reproductive development and seed coat development, (reviewed by Krogan and Ashton, 2000 and Jack, 2001), although no functional studies have been performed to demonstrate the biological significance of this tissue specific expression for these genes. Thus, based on expression studies, MADS-box genes are implicated as transcriptional regulators in many fundamental growth processes and, since MADS-box gene transcripts have been detected in elongating cotton fibre cells (Orford, unpublished), it seems reasonable to predict that they may have a role in the elongation of cotton fibre cells.

Previous work

MADS-box transcripts were detected in cotton fibres from *Gossypium hirsutum* (cultivar Siokra 1-4) fibre cells when a sample of 12 DPA (days post anthesis) cotton fibre cDNA was used as a positive control for a reverse transcription PCR (RT-PCR) experiment using *Oxalis* (Donati, 1997). Two primers, MADSS and MADSQ, which bind to regions in the MADS-box domain were used (Miller *et al.*, 1995; Fig. 2). Because this domain is highly conserved, the primers were expected to amplify the corresponding 150 bp region of MADS-box genes from different plant species. This experiment amplified a 150 bp band from the sample of 12 DPA cotton fibre. In a subsequent RT-PCR experiment using the same primers, four distinct MADS-box transcripts, namely 6.1, 6.2, 12.2 and

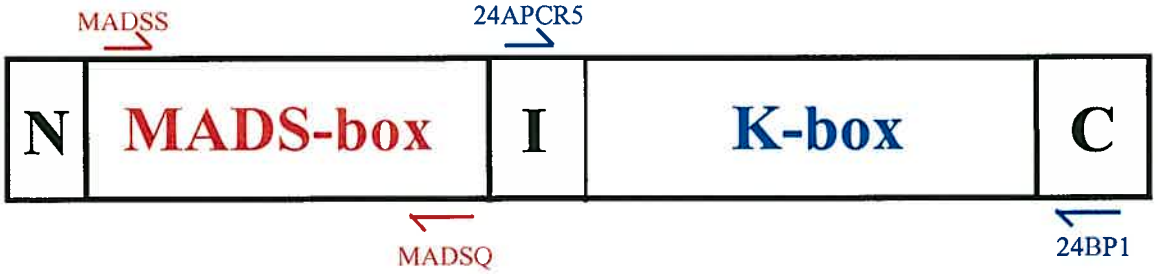
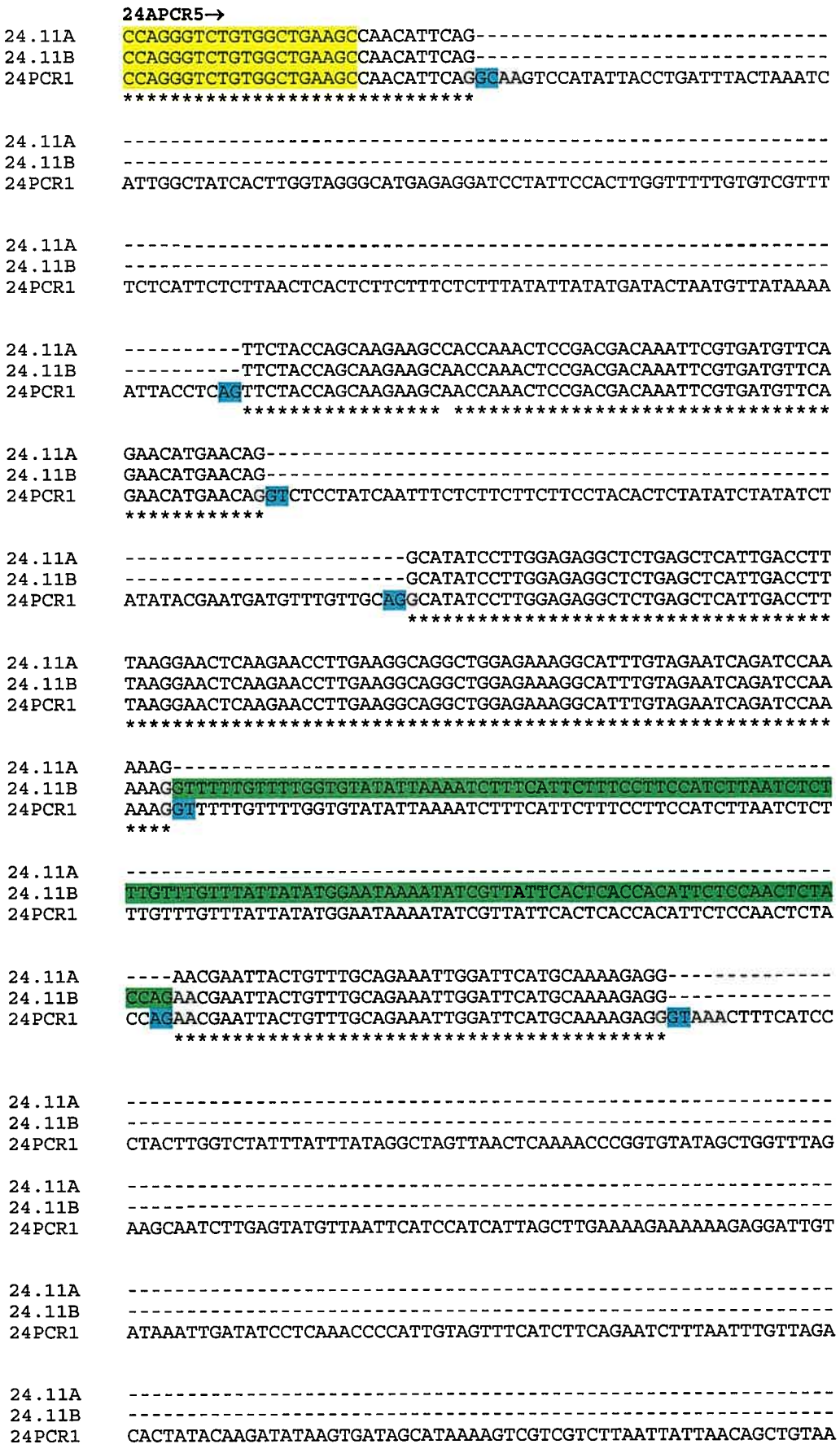


Figure 2: Primer binding sites.
The regions of the MADS-box gene amplified by the primer pairs MADSS and MADSQ; 24APCR5 and 24BP1. In previous studies the MADSS and MADSQ primers have been used to amplify the highly conserved MADS-box region from different species. In this project, 24APCR5 and 24BP1 are used to amplify the K-box as well as flanking regions from the intervening (I) and C-terminal (C) domains.

24.11, were detected in elongating cotton fibre cells of various ages (Orford, unpublished) and full-length cDNAs corresponding to these transcripts were obtained by screening a 12 DPA cotton fibre cDNA library (Malone, 2000). Two cDNAs, 24.11A and 24.11B, were found to correspond to 24.11 mRNA variants (Malone, 2000). These cDNAs were very similar except that 24.11B contained an extra 120 bp section of sequence that was flanked by characteristic intron splice sites (Fig. 3). It was suggested that the 120 bp section of sequence represented an unspliced intron (Malone, 2000).

Gene-specific primers, 24APCR5 and 24BP1 (Malone, 2000; Appendix A; Fig. 2), were then designed to be complementary to more variable regions of the MADS-box gene – the I and C-terminal regions. These primers were designed based on differences between the four identified cotton MADS-box transcripts (6.1, 6.2, 12.2 and 24.11). To test whether the primers were gene specific, a PCR was performed on genomic cotton DNA and the products were analysed by gel electrophoresis (Fig. 4). Extension product of only one size was amplified from genomic DNA with the primers, suggesting that they are specific to only one genomic region, although it is possible that other MADS-box genes of similar size and with similar primer binding sites were also amplified.

RT-PCRs using the same pair of primers were conducted on samples of total RNA from cotton fibre, leaf, flower, stem and root tissues. RNA from fibre at five different stages: 0, 6, 12, 18 and 24 DPA was used. At least five differently sized transcripts were amplified by the primers specific for 24.11 mRNA. These are shown in figure 5, where the major species is a prominant band at 370 bp, which corresponds in size to 24.11A (expected size of 373 bp), while a weaker band of about 490 bp that corresponds to 24.11B (expected size of 493 bp) is also present. There are also larger bands of


24APCR5→
24.11A CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11B CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24PCR1 CCAGGGTCTGTGGCTGAAGCCAACATTCAGGCAAGTCCATATTACCTGATTTACTAAATC
****
24.11A ----------
24.11B ----------
24PCR1 ATTGGCTATCACTTGGTAGGGCATGAGAGGATCCTATTCCACTTGGTTTTTGTGTCGTTT
24.11A ----------
24.11B ----------
24PCR1 TCTCATTCTCTTAACTCACTCTTCTTTCTCTTTATATTATATGATACTAATGTTATAAAA
24.11A ----------TTCTACCAGCAAGAAGCCACCAAACTCCGACGACAAATTCGTGATGTTCA
24.11B ----------TTCTACCAGCAAGAAGCAACCAAACTCCGACGACAAATTCGTGATGTTCA
24PCR1 ATTACCTCAGTTCTACCAGCAAGAAGCAACCAAACTCCGACGACAAATTCGTGATGTTCA
********** **********
24.11A GAACATGAACAG----------
24.11B GAACATGAACAG----------
24PCR1 GAACATGAACAGGTCTCCTATCAATTTCTCTTCTTCTTCCTACACTCTATATCTATATCT
****
24.11A ----------GCATATCCTTGGAGAGGCTCTGAGCTCATTGACCTT
24.11B ----------GCATATCCTTGGAGAGGCTCTGAGCTCATTGACCTT
24PCR1 ATATACGAATGATGTTTGTTGCAGGCATATCCTTGGAGAGGCTCTGAGCTCATTGACCTT
****
24.11A TAAGGAACTCAAGAACCTTGAAGGCAGGCTGGAGAAAGGCATTTGTAGAATCAGATCCAA
24.11B TAAGGAACTCAAGAACCTTGAAGGCAGGCTGGAGAAAGGCATTTGTAGAATCAGATCCAA
24PCR1 TAAGGAACTCAAGAACCTTGAAGGCAGGCTGGAGAAAGGCATTTGTAGAATCAGATCCAA
****
24.11A AAAG----------
24.11B AAAGGTTTTTGTTTTGGTGTATATTAAAATCTTTCATTCTTTCCTTCCATCTTAATCTCT
24PCR1 AAAGGTTTTTGTTTTGGTGTATATTAAAATCTTTCATTCTTTCCTTCCATCTTAATCTCT
****
24.11A ----------
24.11B TTGTTTGTTTATTATATGGAATAAAATATCGTTATTCACTCACCACATTCTCCAACTCTA
24PCR1 TTGTTTGTTTATTATATGGAATAAAATATCGTTATTCACTCACCACATTCTCCAACTCTA
24.11A ----AACGAATTACTGTTTGCAGAAATTGGATTCATGCAAAAGAGG----------
24.11B CCAGAACGAATTACTGTTTGCAGAAATTGGATTCATGCAAAAGAGG----------
24PCR1 CCAGAACGAATTACTGTTTGCAGAAATTGGATTCATGCAAAAGAGGGTAAACTTTCATCC
****
24.11A ----------
24.11B ----------
24PCR1 CTACTTGGTCTATTTATTTATAGGCTAGTTAACTCAAAACCCGGTGTATAGCTGGTTTAG
24.11A ----------
24.11B ----------
24PCR1 AAGCAATCTTGAGTATGTTAATTCATCCATCATTAGCTTGAAAAGAAAAAAGAGGATTGT
24.11A ----------
24.11B ----------
24PCR1 ATAAATTGATATCCTCAAACCCCATTGTAGTTTCATCTTCAGAATCTTTAATTTGTTAGA
24.11A ----------
24.11B ----------
24PCR1 CACTATACAAGATATAAGTGATAGCATAAAAGTCGTCGTCTTAATTATTAACAGCTGTAA

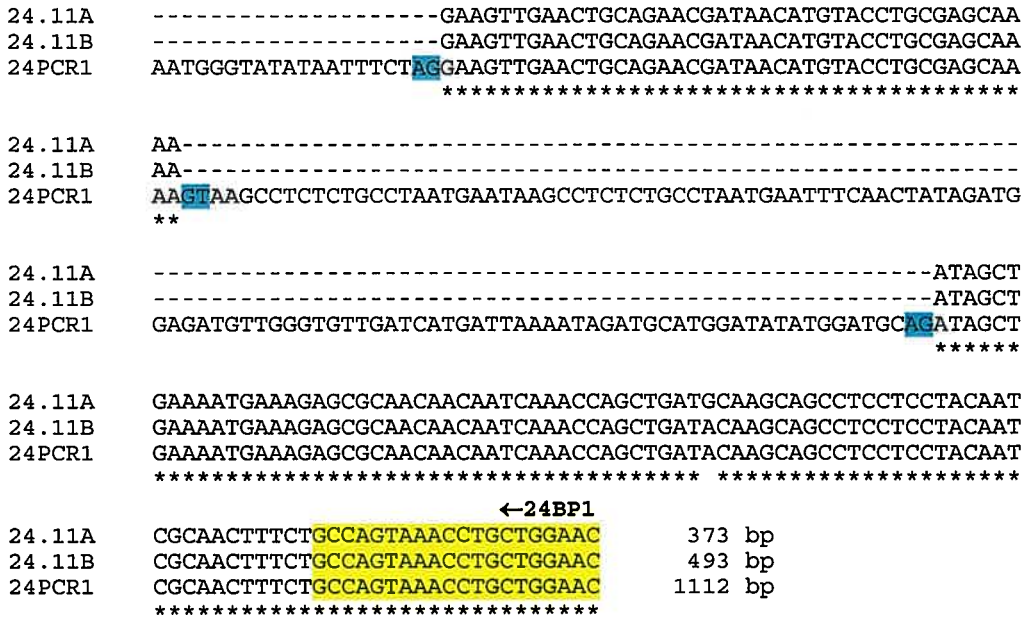


Figure 3: Alignment of the genomic PCR clone, 24PCR1, with MADS-box cDNAs, 24.11A and 24.11B.

The alignment of the previously sequenced cDNA clones, 24.11A and 24.11B, with the genomic PCR clone, 24PCR1. 24.11A appears to contain five putative introns while 24.11B contains only four. The 120 bp section of sequence that is present in 24.11B but absent in 24.11A is illustrated in green. Characteristic splice sites are highlighted in blue and the two primer binding sites are highlighted in yellow.

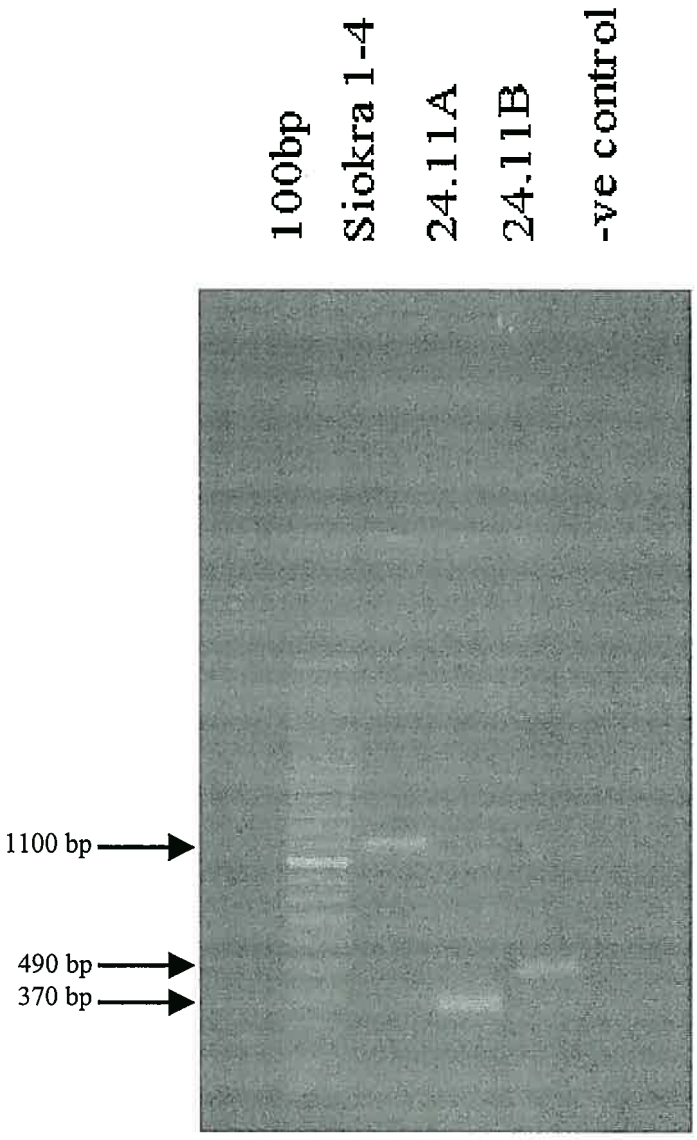


Figure 4: Amplification of genomic DNA and cDNA clones.
The two primers 24APCR5 and 24BP1 that were designed to be specific to the 24.11 transcript amplified a single band from genomic DNA of *G. hirsutum*. The amplification of only one band (1.1 kb) suggests that the primers may be gene specific. Two positive controls, 24.11A and 24.11B, were used and ran at their expected sizes of 373 bp and 490 bp respectively. A negative control for the PCR was performed, and showed no amplification.

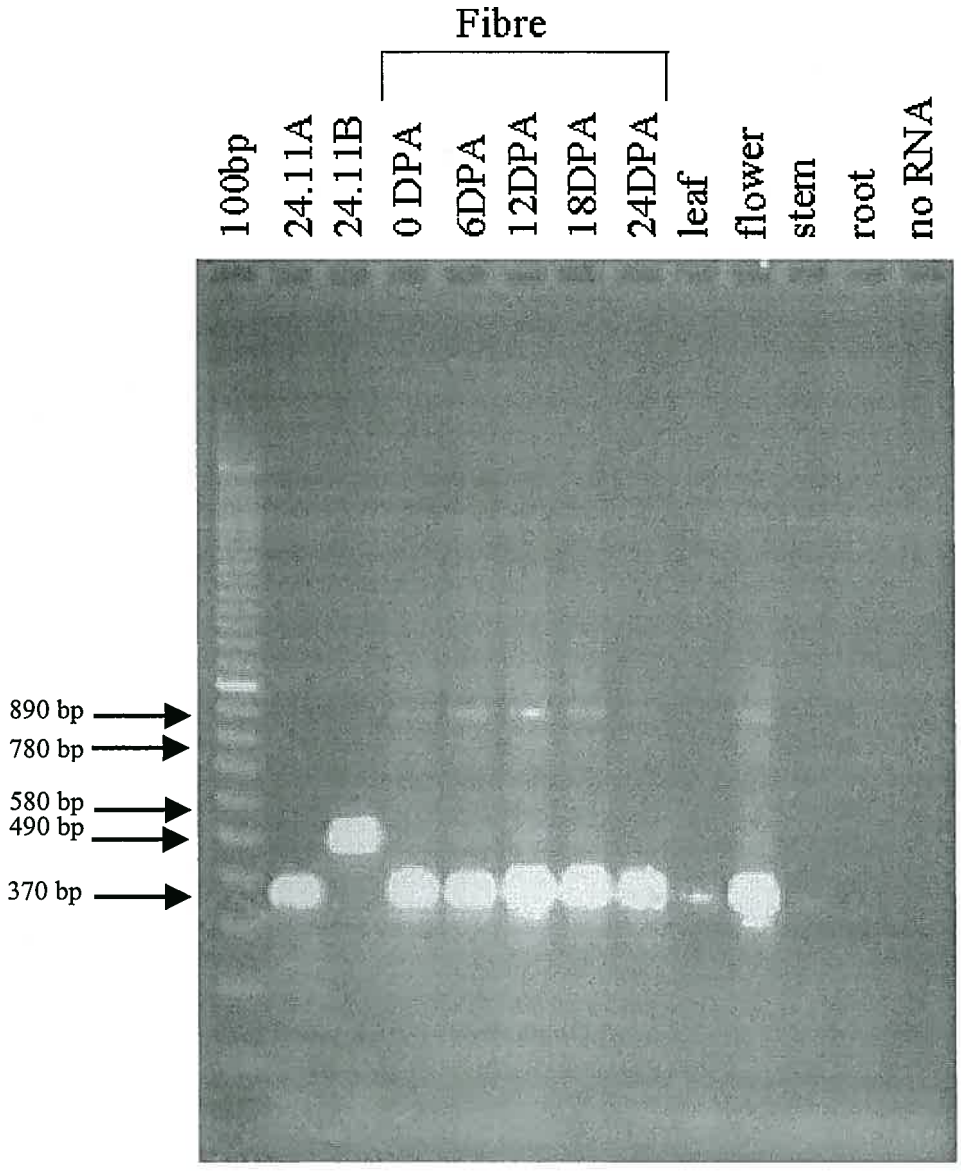


Figure 5: RT-PCR of mRNA from cotton tissues.
24APCR5 and 24BP1 were used to amplify the reverse transcription products of mRNA from cotton fibre (0, 6, 12, 18 and 24 DPA), leaf, flower, stem and root tissues (Malone, 2000). Two positive controls of known size (24.11A and 24.11B), and one negative control to which no DNA was added, were also amplified by PCR.

approximately 580 bp, 780 bp and 890 bp. It is possible that there are also products of other sizes, but they cannot be clearly seen on this gel due to their relative lack of abundance compared with the major 370 bp species.

The observation of multiple bands in the electrophoresis of the RT-PCR is a very unusual result because RT-PCR experiments do not usually amplify multiple bands. RT-PCRs performed on the other three 150 bp transcripts (6.1, 6.2 and 12.2) revealed only single bands (Malone, 2000). It is therefore very unlikely that the gel electrophoresis is representing a pre-mRNA population that has not been fully processed but instead it is much more likely that alternatively spliced final mRNA are being amplified.

Although RT-PCR is only semi-quantitative, it is likely that some of these transcripts are expressed at different levels in different tissues. The leaf and stem samples showed only the 370 bp band while no extension products were produced from the root tissue. Flower tissue showed expression of a set of bands (370 bp, 490 bp, 580 bp, 780 bp and 890 bp) that were similar to those amplified from the fibre tissue (Malone, 2000).

A genomic PCR clone of this region was obtained by amplifying genomic DNA with the primers 24APCR5 and 24BP1. The hypothesis that 24.11B is a mRNA variant of 24.11, with an extra 120 bp section of sequence is supported by the presence of the 120 bp section in the genomic clone, when it is aligned with 24.11A and 24.11B (Fig. 3). This alignment also indicated that there are other introns that are spliced out to give the 24.11A and 24.11B mRNAs (Malone, 2000). As can be seen in figure 3, there appear to be five introns spliced from the sequence of 24.11A and four introns spliced from 24.11B, all flanked by characteristic intron splice sites when compared to the genomic

sequence. For dicot plants, the first two nucleotides at the 5' end of an intron are 5'-GT-3' or 5'-GC-3' and the last two nucleotides of an intron, at the 3' end, are 5'-AG-3' (Goodall *et al.*, 1991). The presence of these characteristic splice sites flanking each 'gap' further suggests that they are indeed introns.

The presence of differently sized transcripts in the various cotton tissues, as revealed by the multiple RT-PCR products, indicates that there are mRNAs of different sizes present. The genomic clone alignment (Fig. 3) and the presence of differently sized RT-PCR bands (Fig. 5) suggest that the presence of differently sized transcripts may be due to the inclusion of the intronic regions in a series of RNA molecules related to 24.11A and 24.11B. It is possible that some sort of alternative splicing mechanism may be acting to differentially splice the pre-mRNA to include some regions that are spliced from 24.11A and 24.11B. Such alternative splicing is a feature of other MADS-box genes in both animals and plants (Theiβen *et al.*, 1996).

Aims of the project

The broad aim of this project was to investigate the possibility of alternative splicing of a cotton MADS-box gene. Specifically, the first aim of the project was to repeat the RT-PCR and then clone and characterise the differently sized transcripts to determine whether they may be alternatively spliced mRNA. If time allowed, the second aim of the project was to obtain the gene from which these transcripts arise.

Part A: RT-PCR analysis

Previous work detected four distinct MADS-box transcripts expressed in cotton fibre cells (Orford, unpublished). To investigate these transcripts, reverse transcription was performed on total RNA from several different tissues. The tissues used were 0, 6, 12, 18 and 24 DPA fibres, as well as leaf, flower, stem and root. The region studied in this project is the K-box with some flanking intervening (I) and C-terminal (C) regions. With the use of the 24APCR5 and 24BP1 primers (Fig. 2) this region was amplified by PCR and the products analysed by gel electrophoresis.

A1: RT-PCR conditions

Previously, total RNA had been isolated and DNase treated to digest any genomic DNA. Total RNA from each tissue was reverse transcribed using 45 μM oligo (dT) and M-MLV (Moloney Murine Leukemia Virus) reverse transcriptase (Promega). PCR was performed on 100 ng of reverse transcribed template from the different cotton tissues using the gene-specific primers 24APCR5 and 24BP1. PCR conditions used were denaturation at 94°C for two minutes followed by 30 PCR cycles of denaturation at 94°C for 30 seconds, primer annealing at 65°C for 30 seconds and extension at 72°C for 1 1/2 minutes with final extension at 72°C for 4 minutes.

Two negative controls were used for the RT-PCR and for future experiments. The first was a PCR control containing no DNA template, which it was anticipated should yield no amplification products during PCR. The second was an RT-PCR control which consisted of a 6 DPA cotton fibre RNA template to which reverse transcriptase had not been added. The purpose of this control was to test whether there was contaminating genomic DNA present in the sample due to incomplete DNase treatment. Three positive controls of

known sequence and size were also used throughout the experiment. 24.11A, 24.11B and 24PCR1 were amplified under the same conditions and analysed by gel electrophoresis to verify that the PCR was successful and that the products ran at their expected sizes of 373 bp, 493 bp and 1112 bp respectively.

A2: Actin control

Actin mRNA was used as a control for the PCR because there is evidence it is expressed at a relatively constant level throughout all plant tissues (Shimizu *et al.*, 1997). This allowed a 'calibration' of the RT-PCR experiment: the actin control was used to indicate the relative proportions of RNA present in the tissue samples, and this was then taken into account when the RT-PCR products amplified from the tissues were compared.

The templates from the reverse transcription were amplified by PCR using the actin-specific primers ACTF and ACTR (Appendix A). PCR conditions were: denaturation at 94°C for two minutes followed by 25 PCR cycles of denaturation at 94°C for 30 seconds, primer annealing at 65°C for 30 seconds and extension at 72°C for 1 minute and then final extension at 72°C for 4 minutes. The products were analysed by gel electrophoresis on a 1.5% agarose gel (Fig. 6).

Analysis of the products of the actin control by gel electrophoresis revealed that 120 bp products from each tissue sample were amplified as anticipated (Shimizu *et al.*, 1997; Fig. 2). The gel electrophoresis revealed high levels of actin mRNA in the 6, 12 and 18 DPA fibres and in the stem tissues. Slightly reduced transcription was observed in the 0 and 24 DPA fibre, flower and root tissues. Significantly reduced amplification of actin mRNA was apparent in the leaf tissue sample.

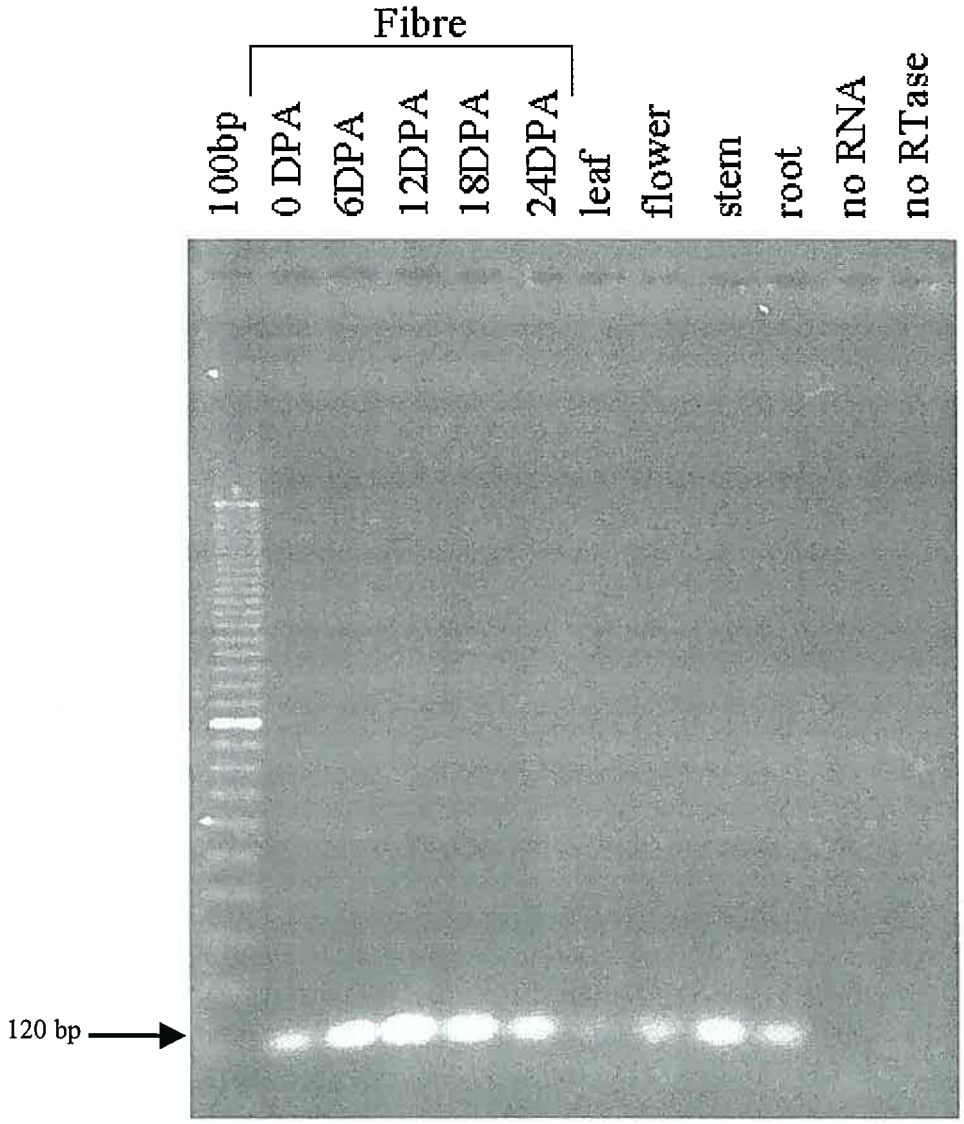


Figure 6: Amplification of actin mRNA in different tissues.
The reverse transcription products amplified by PCR using the actin specific primers ACTF and ACTR, analysed by gel electrophoresis on a 1.5% agarose gel. Tissues used in the RT-PCR were fibre (0, 6, 12, 18 and 24 DPA), leaf, flower, stem and root. Two negative controls were used: 'no RNA' was a negative control for the PCR, with no DNA added, and 'no RTase' was a negative control for the reverse transcription process to which no reverse transcriptase was used. Analysis of the gel electrophoresis indicates reduced RNA levels in the 0 and 24 DPA fibre, leaf, flower and stem tissues compared to the other tissue samples.

A3: RT-PCR

Once the actin control had been performed, PCR was used to amplify 24.11 transcripts from the reverse transcription products with the primers, 24APCR5 and 24BP1. The products were resolved on a 1.5% agarose gel (Fig. 7). No bands were visible in the negative control lanes, verifying that no contamination had occurred during the PCR or from genomic DNA during the reverse transcription. The positive controls 24.11A, 24.11B and 24PCR1 produced bands of the anticipated sizes of 370 bp, 490 bp and 1100 bp respectively. Gel electrophoresis of the RT-PCR demonstrated that no amplification of any transcripts from the stem or root tissues had occurred, while the leaf tissue showed amplification of two bands at 370 bp and 500 bp. The major species in the flower and fibre lanes was a 370 bp band, but larger bands were also evident confirming previous results (Malone, 2000). There appeared to be distinct bands at 780 bp and 890 bp and possibly bands at approximately 490 bp and 580 bp.

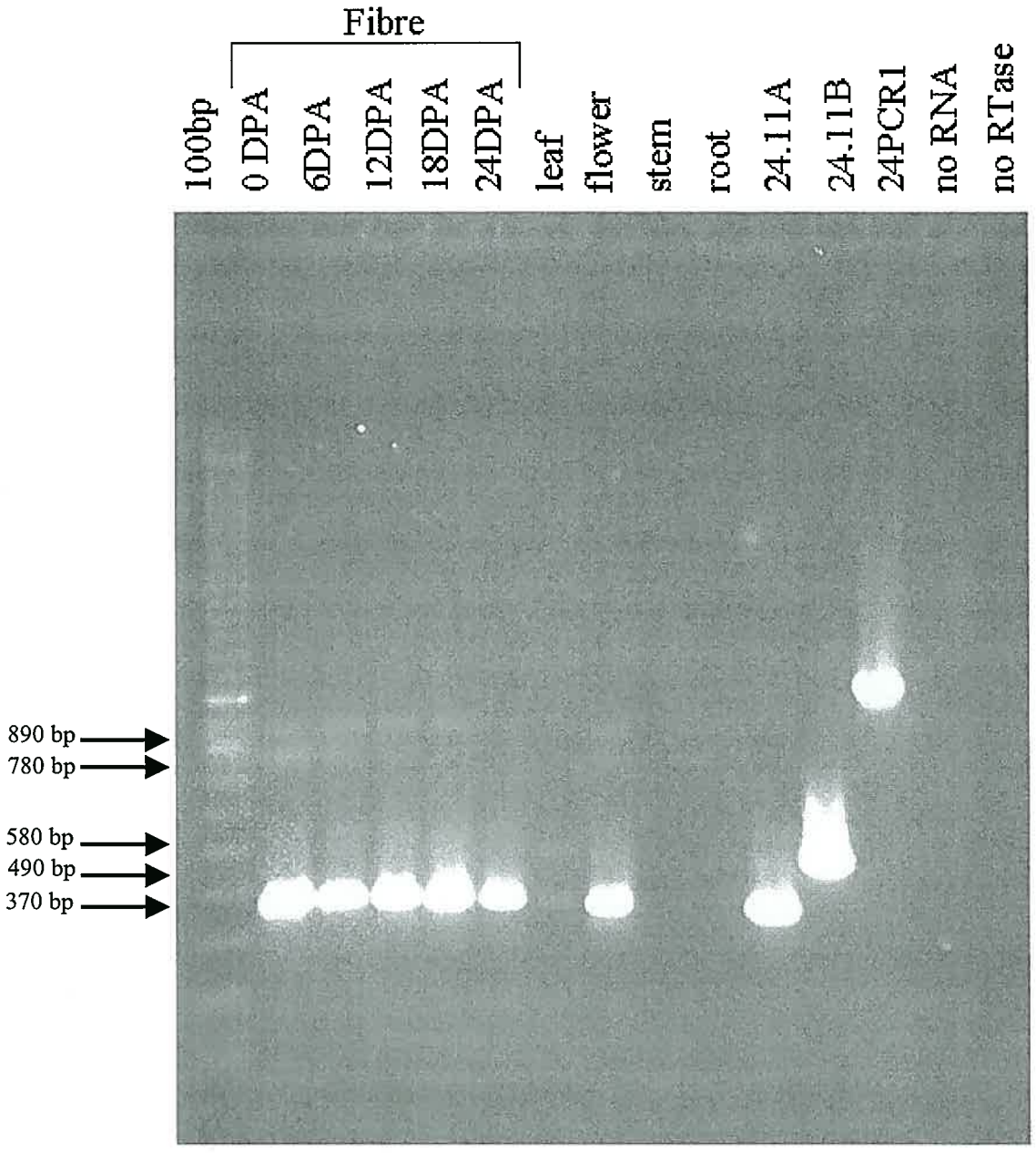


Figure 7: Amplification of 24.11 transcripts.
The PCR products of the reverse transcription were analysed on a 1.5% agarose gel. The gene specific primers 24APCR5 and 24BP1 were used to amplify the reverse transcription products of total RNA from various cotton tissues. Tissues used were fibre (0, 6, 12, 18 and 24 DPA), leaf, flower, stem and root. Three positive controls of known size were used (24.11A, 24.11B and 24PCR1). Two negative controls were used where 'no RNA' was a negative control for the PCR with no DNA added, and 'no RTase' was a negative control for the reverse transcription reaction with no reverse transcriptase added.

Part B: Southern analysis

If the different bands produced by RT-PCR correspond to alternatively spliced mRNA, they must be derived from one pre-mRNA transcript and must share a very high degree of sequence similarity. As an initial step to investigate this possibility, a Southern was performed.

B1: Probe preparation

24.11AS, the full-length cDNA of 24.11A (Appendix B), was used as a probe. The cDNA was excised from pBluescript by *Eco*RI digestion and separated from the plasmid by gel electrophoresis. The 1076 bp band was excised from the agarose gel and purified using a QIAGEN Gel Extraction Kit. 200 ng of 24.11AS (from the gel extraction) was labelled with α-32P-dATP according to Hodgson and Fisk (1987)

B2: Southern transfer

The DNA on the RT-PCR agarose gel was transferred onto Hybond-N+ membrane according to Southern (1975). The membrane was pre-hybridised for one hour at 65°C and then probed with the labelled 24.11AS overnight at 65°C. The membrane was washed with increasing levels of stringency to remove unbound or weakly bound probe using a series of solutions with increasing SDS concentration and decreasing SSC concentration. A series of four washes were conducted at 65°C for 30 minutes each (2xSSC, 0.1% SDS; 1xSSC, 0.1%; 0.5xSSC, 0.1% SDS; 0.1xSSC, 1% SDS).

The filter was exposed to X-ray film for three different exposure conditions with enhancer screens: 12 minutes at room temperature, 30 minutes at room temperature and

30 minutes at -80 °C. As can be seen in figure 8 (12 minutes at room temperature), the labelled probe has bound to most of the bands that were visible in the ethidium bromide-stained gel (Fig. 7) and to the three positive controls, 24.11A, 24.11B and 24PCR1. For the fibre samples (0, 6, 12, 18 and 24 DPA) and the flower sample, the major species at 370 bp hybridised strongly with the probe. The probe also hybridised to discrete bands at 780 bp and 890 bp in these tissues and, although not as clear, hybridisation was observable for bands at approximately 490 bp and 580 bp. The stem and root tissues both showed bands corresponding to the 370 bp species. The probe bound weakly to two bands in the leaf tissue at 370 bp and at 500 bp.

Both of the negative controls, no PCR template and no reverse transcriptase, showed a small amount of contamination at a size corresponding to the major 370 bp band from the cotton tissues. The fact that both negative controls were contaminated, showing the same sized product, and that the product was the same size as the major species in the other lanes, indicates that the contamination occurred during the loading of the gel or in the PCR and not during the reverse transcription. If contamination of genomic DNA had occurred from the reverse transcription, only the RT-PCR control should have amplified a band, not the PCR negative control. The contamination is less abundant that any experimental bands and, as such, does not appear to be a significant concern.

B3: Interpretation of the RT-PCR and Southern hybridisation

The stem and root tissues demonstrated a small amount of amplification of only a 370 bp product, while an additional 500 bp product was amplified from the leaf tissue. The products amplified from the different fibre tissues (0, 6, 12, 18 and 24 DPA) resembled the products from the flower tissue. The six fibre and the leaf tissue samples all showed

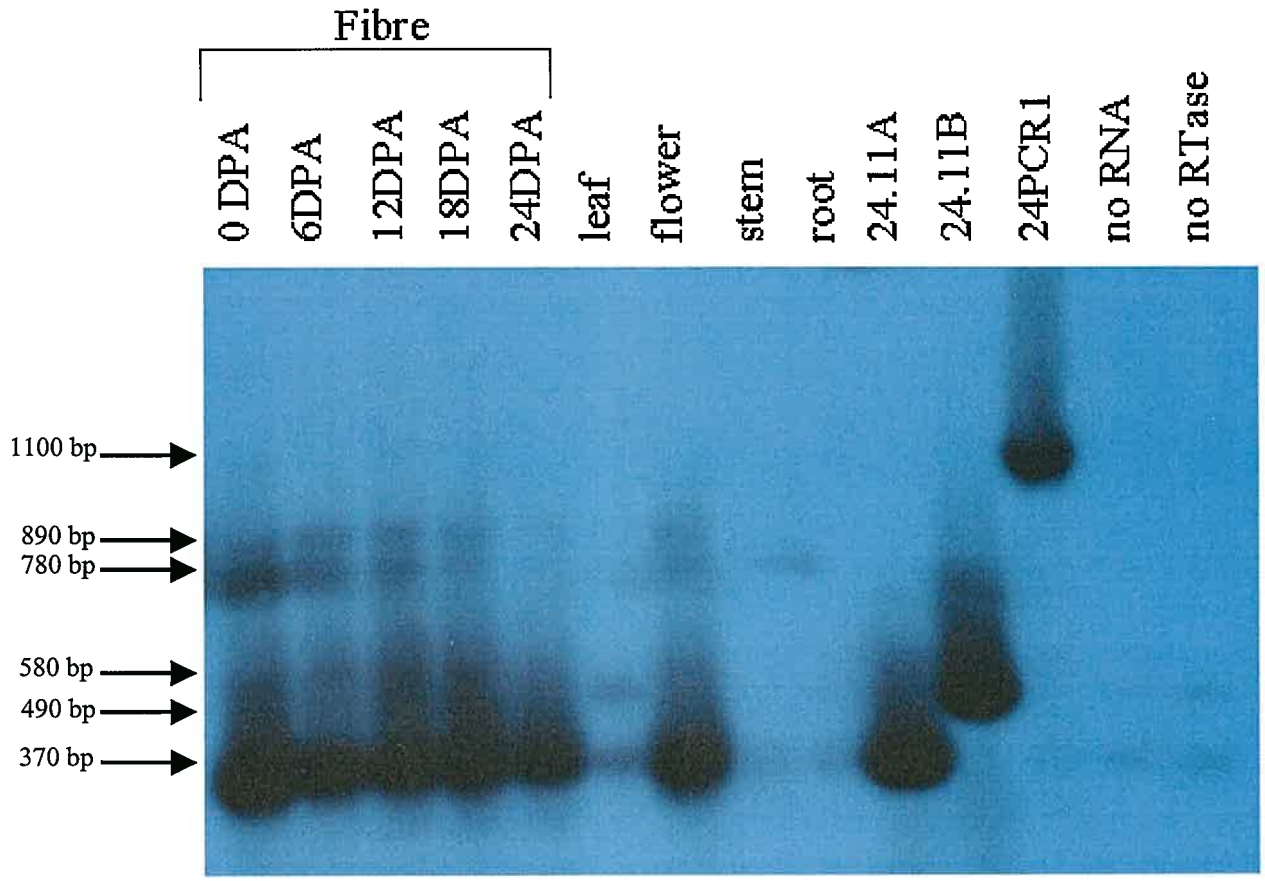


Figure 8: Southern transfer of the RT-PCR gel.
The Southern transfer of the PCR products of the reverse transcription was probed with 24.11AS (the full length cDNA corresponding to 24.11A). The filter was exposed to X-ray film for 12 minutes at room temperature. Tissues used were fibre (0, 6, 12, 18 and 24 DPA), leaf, flower, stem and root. Three positive controls were used, 24.11A, 24.11B and 24PCR1, whose sequences were known to be complementary to the probe. Two negative controls were used: 'no RNA' was a negative control for the PCR, with no DNA added and 'no RTase' was a negative control for the reverse transcription process with no reverse transcriptase added.

the 370 bp product as their major species together with several larger bands. By analysing the RT-PCR and the Southern transfer, bands were visible at 490 bp, 580 bp, 780 bp and 890 bp. The 780 bp and 890 bp bands are very faint in figure 8 for the 24 DPA fibre sample, but the longer exposures of the Southern membrane showed that they specifically hybridised with the probe (data not shown). In both the ethidium bromide-stained gel (Fig. 7) and the Southern analysis (Fig. 8) the larger products (780 bp and 890 bp) decrease in intensity as the fibres age.

The different bands from the RT-PCR and Southern analysis suggest that the differently sized cDNAs to which they correspond may contain different combinations of the putative introns from the 1112 bp genomic PCR clone. There may also be bands other than the five previously described (370 bp, 490 bp, 580 bp, 780 bp and 890 bp) that cannot be clearly seen due to their lower abundance. The Southern transfer has demonstrated that the bands that were visible by ethidium bromide-staining are similar in sequence to the 24.11AS probe and may therefore correspond to differentially spliced 24.11 mRNAs.

Part C: Cloning

The RT-PCR was repeated and analysed by gel electrophoresis with a result that was identical to that depicted in figure 7 and six small segments of the 0 DPA cotton fibre lane were excised from the gel to individually isolate:

(1) the 370 bp product;

(2) the 490 bp and 580 bp products;

(3) the 780 bp product;

(4) the 890 bp product;

(5) any products below 370 bp;

(6) any products above 890 bp.

The leaf lane was also investigated because it showed only two amplification products and this was considered to possibly reflect a situation in which it would be easier to investigate a biological role for each of the differentially spliced mRNAs. Therefore two bands were excised from the leaf lane of the RT-PCR:

(1) the 370 bp product;

(2) the 500 bp product.

DNA was purified from each of the eight excised gel fragments using a QIAGEN Gel Extraction Kit. Overnight ligations were performed with 250 ng of insert DNA (from the gel extraction) and 50 ng of pGEM®T-easy vector (Promega) at 16°C, to individually clone products from the eight different gel regions.

The plasmids were then transformed into DH5aF' *E.coli* cells made competent by the $CaCl_2$ method (Sambrook, 1989) by heat shock at 42 °C. Cells were plated onto L-agar plates supplemented with 100 μg/ml ampicillin with an overlay of L-agar medium containing 100 μg/ml ampicillin, 200 ug/ml X-gal and 160 μg/ml IPTG (LAIX plates). The blue/white colour selection method was used, where successful cloning of an insert disrupts the *LacZ* gene, which encodes β-galactosidase, within the pGEM®T-easy vector. In non-recombinants, IPTG induces expression of β-galactosidase which cleaves X-gal, generating a blue product. Therefore, transformants can be selected as those colonies that are white due to the insertional inactivation of the *LacZ* gene. Both positive and negative controls were used for the ligations. For the negative control, no insert DNA was added and, as expected, only blue colonies were seen on the corresponding LAIX plate. An insert of known ligation efficiency was used for the positive control and that yielded a reproducible result with a large number of colonies, a majority of which were white.

PCR was performed directly upon approximately ten colonies from each of the eight transformations in order to assess the presence and size of inserts. Half of each of the selected colony was picked into 50 μL of water and the cells were lysed at 95°C for five minutes. The cells were then spun at 14,000 rpm for one minute to pellet the genomic DNA and cell debris thereby leaving the plasmid in solution. A 1:10 dilution was performed and 1 uL (~100 ng) was amplified with the T7 and SP6 primers, whose binding sites flank the cloned insert. PCR conditions were: initial template denaturation at 94°C for two minutes followed by 25 PCR cycles of denaturation at 94°C for one minute, primer annealing at 52°C for one minute and extension at 72°C for two minutes with final extension at 72°C for four minutes.

The colony PCRs were analysed on 1.5% agarose gels (data not shown). Several colonies that were found to have inserts of each expected fragment size were selected for sequencing of the insert. The sizes of the inserts that were sequenced can be seen in figure 9 and also in table 1 where the sizes of the PCR products have been derived from the agarose gels.

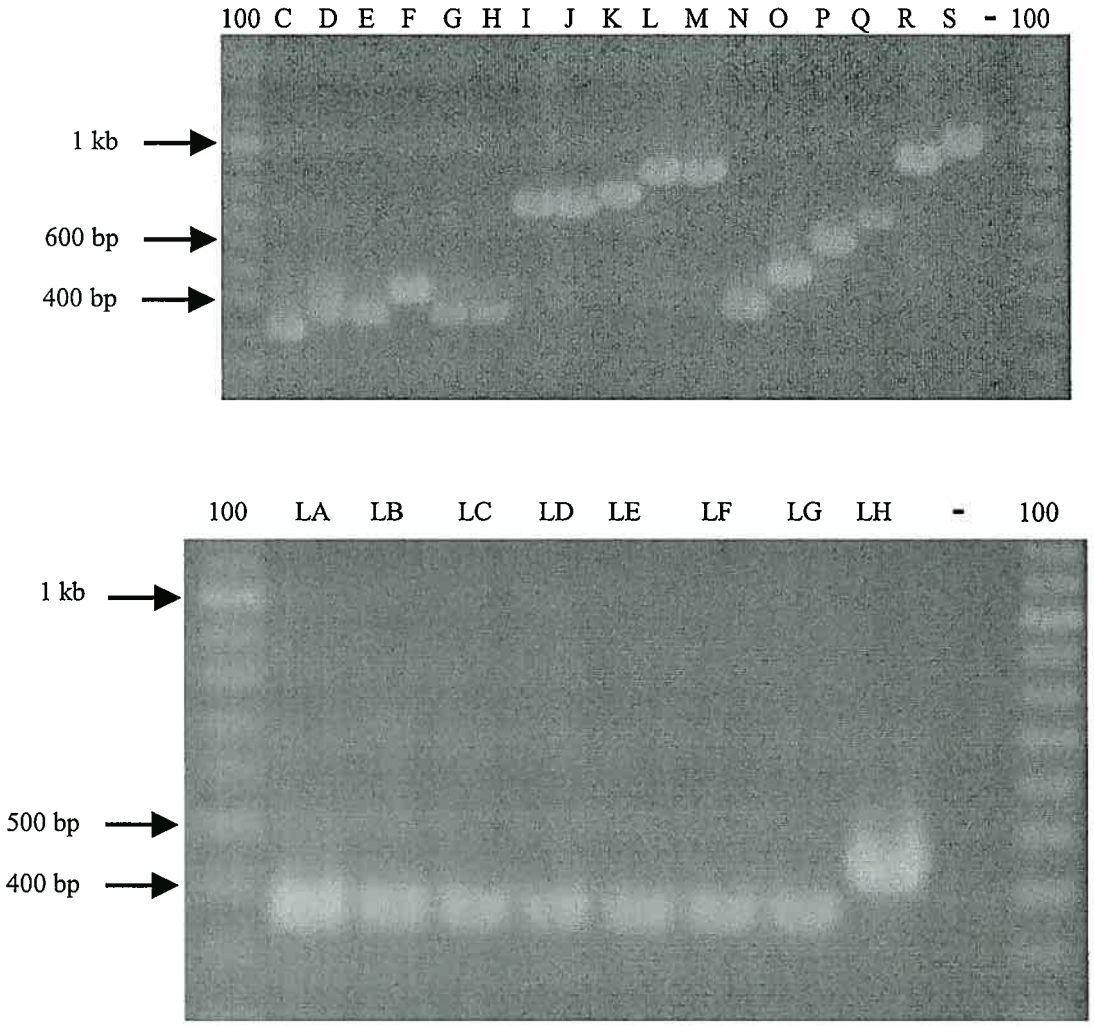


Figure 9: Insert sizes.
Sizes of cloned inserts as checked by PCR amplification using the primers 24PCR5 and 24BP1. Lanes are labelled as '100' (100 bp marker), '-' (negative control with no template) or with one or two letters corresponding to 24.11(labelling letter).

CLONE	SIZE
24.11C	330 bp
24.11D	370 bp
24.11E	370 bp
24.11F	430 bp
24.11G	370 bp
24.11H	370 bp
24.11I	750 bp
24.11J	750 bp
24.11K	760 bp
24.11L	850 bp
24.11M	850 bp
24.11N	370 bp
24.11O	450 bp
24.11P	560 bp
24.11Q	650 bp
24.11R	870 bp
24.11S	890 bp
24.11LA	370 bp
24.11LB	370 bp
24.11LC	370 bp
24.11LD	370 bp
24.11LE	370 bp
24.11LF	370 bp
24.11LG	370 bp
24.11LH	450 bp

Table 1: The sizes of the 25 transcripts.
The left column contains the names of the clones and the right column contains the sizes of the clones as determined by PCR amplification of the insert and gel electrophoresis (Fig. 9).

Part D: Nucleotide sequence analysis

D1: Sequencing methods

Seventeen colonies from 0 DPA cotton fibre tissue and eight colonies derived from leaf tissue were sequenced. Each colony to be sequenced was grown overnight in 10 ml L-broth with 40 ul of 25 mg/ml ampicillin. The plasmid was purified with a BIO-RAD plasmid miniprep kit. 500 ng of DNA from each purification was amplified with either the T7 or SP6 primer and Big Dye Terminator Mix with the following cycling conditions: initial template denaturation at 95°C for five minutes followed by 26 cycles of denaturation at 95°C for 30 seconds, primer annealing at 55°C for 30 seconds and extension at 60°C for four minutes. The DNA was precipitated with isopropanol and sequencing was carried out by the IMVS.

T7 and SP6 primers bind to sequences in the pGEM®T-easy vector that flank the insert, with T7 used to amplify in a 5' to 3' direction, and SP6 to extend the reverse complement. Each strand was sequenced so that the entire length of the insert was sequenced in both directions.

D2: Preliminary sequence analysis

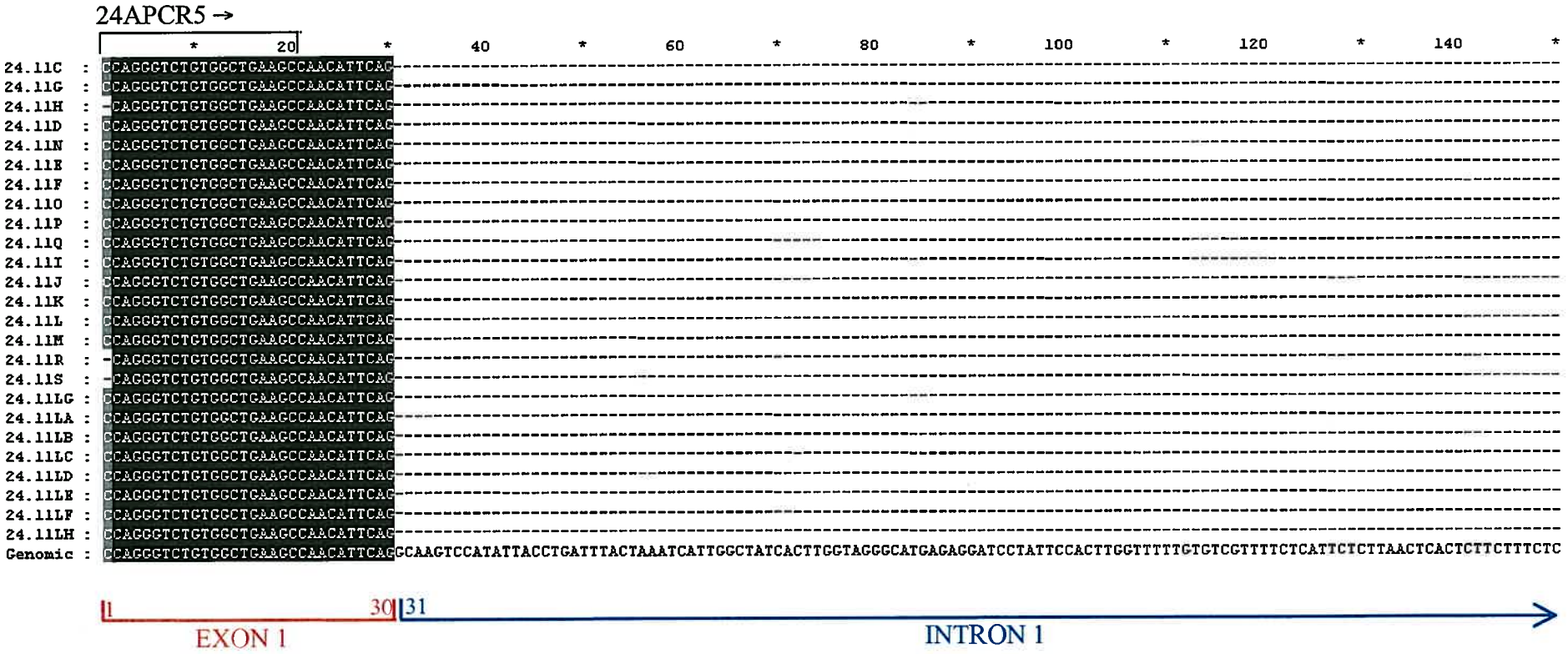
Sequence data was transferred to ANGIS (Australian National Genomic Information Service) for sequence analysis. The chromatograms were inspected with vector ends removed and ambiguous bases interpreted from the chromatogram. The multiple separate sequences for each clone were aligned against each other and the genomic PCR clone, 24PCR1, in the 5' to 3' direction using the program Clustalw (ANGIS) (Appendix C).

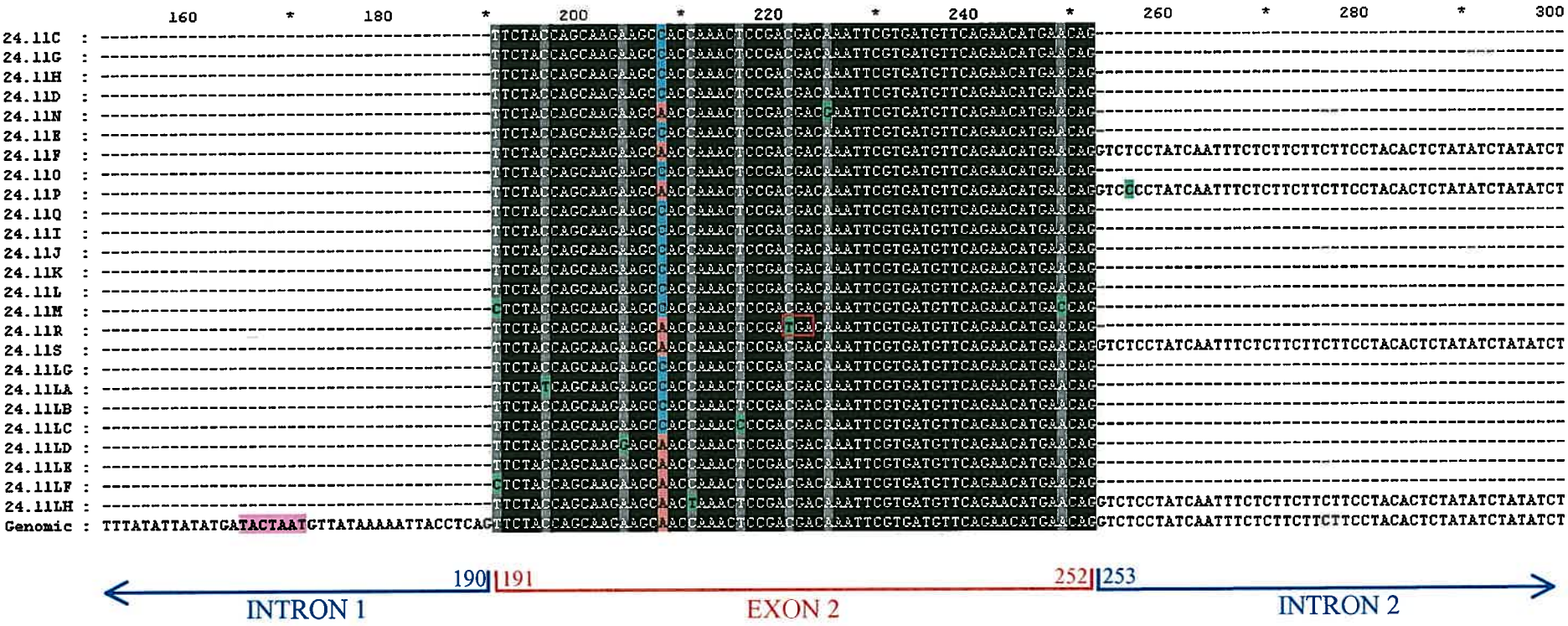
The consensus sequences from each of the 25 clones were aligned against each other, and the previously sequenced genomic PCR clone 24PCR1 (Malone, 2000; Fig. 10). Clustalw (ANGIS) was used to align the sequences, with the alignments being inspected and modified by eye to cope with the large introns. The fibre tissue inserts were designated 24.11C to 24.11S in line with the previous established nomenclature, continuing from 24.11A and 24.11B (Malone, 2000) and the leaf tissue inserts were designated 24.11LA to 24.11LG ('L' standing for leaf). The sequence alignment can be seen in figure 10 with a diagrammatic representation of the sequences in figure 11.

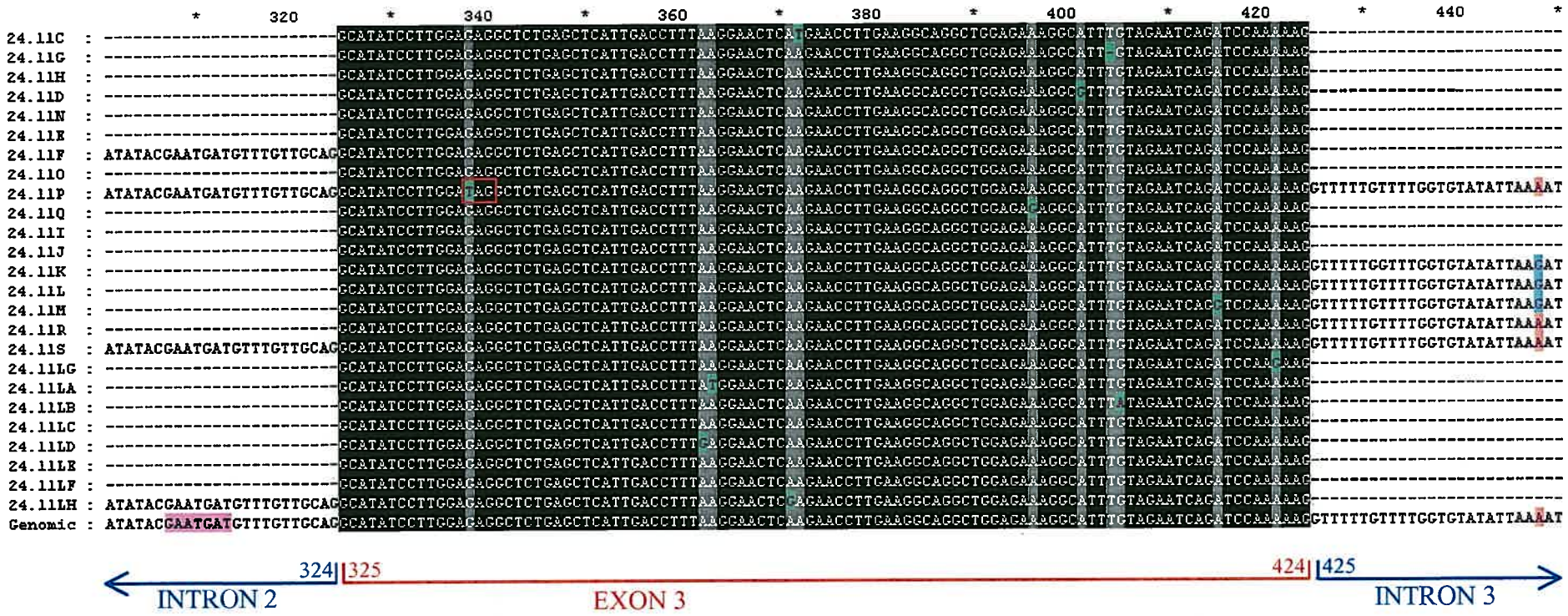
Three clones (24.11H, 24.11R and 24.11S) all lacked the most 5' cytosine when compared to the other clones (Fig. 10). This observation was unexpected as this nucleotide is contained within the primer sequence and should always be present. This nucleotide may have been lost due to activity of a contaminating exonuclease or the presence of truncated primers.

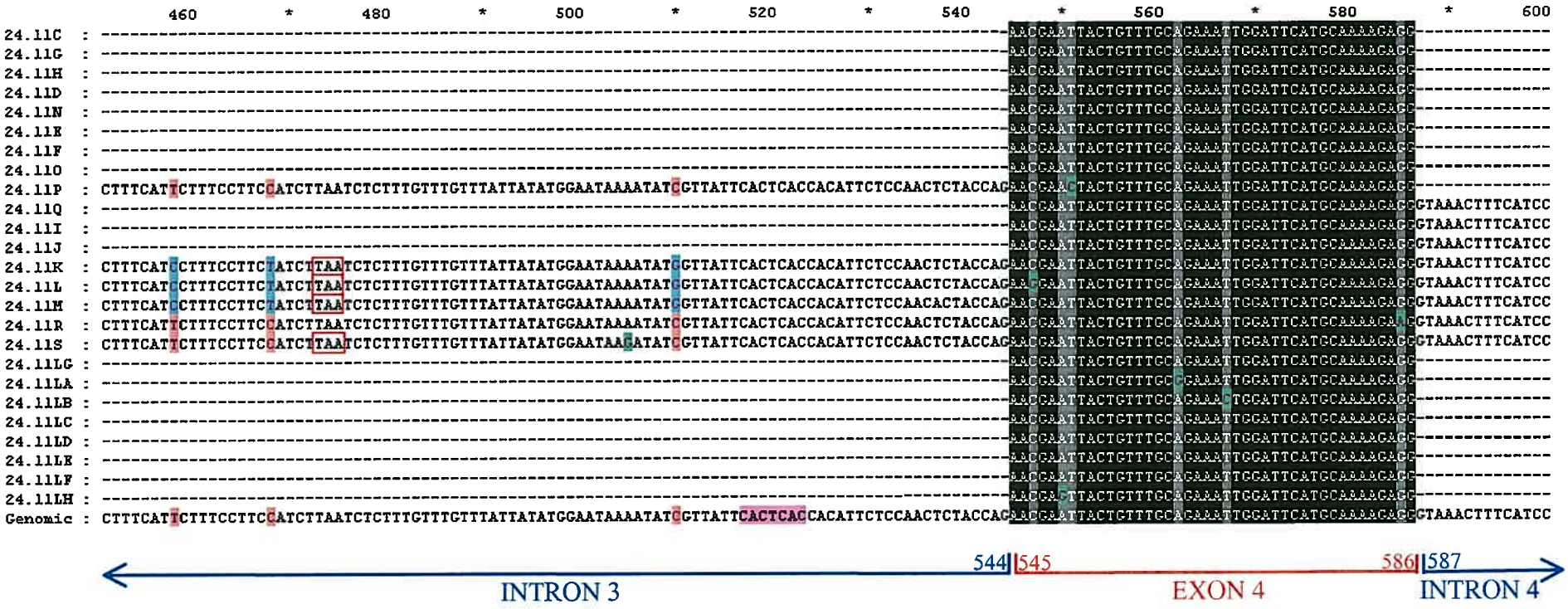
D3: Intron/exon composition

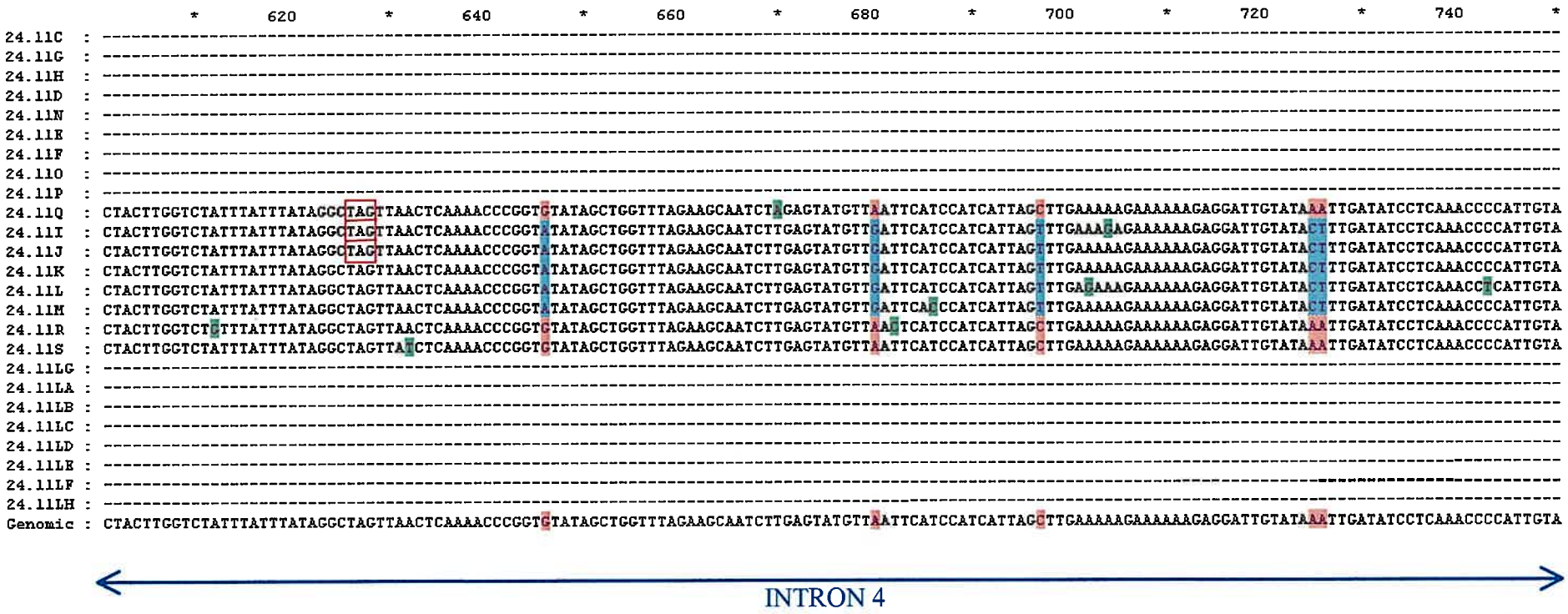
Four regions of sequence (nucleotides 253-324, 425-544, 587-861, 904-1015) (Fig. 10) were variable in terms of their inclusion within the different clones with each being observed in four, six, eight and eight clones, respectively. There was also a 160 bp region of sequence in the genomic clone that was not found in any of the transcripts (nucleotides 31 –190) (Fig. 10). The 5' and 3' ends of these five regions conformed to the consensus intron splice sites for dicot plants, where the first two nucleotides at the 5' end of an intron are always 5'-GT-3' or 5'-GC-3' and the last two nucleotides of an intron, at the 3' end, are always 5'-AG-3' (Goodall *et al.*, 1991). Also, in the most abundant species (the 373 bp product) these five regions were spliced out and therefore,

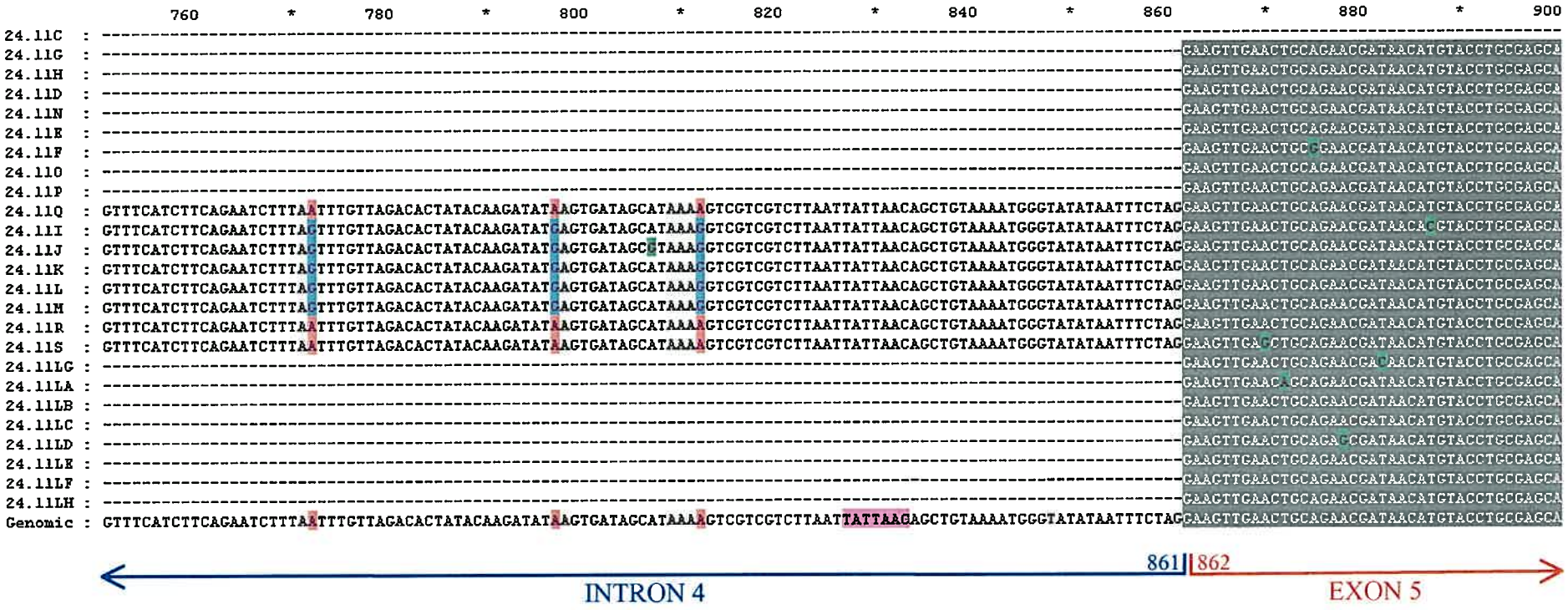

24APCR5 →
* 20 * 40 * 60 * 80 * 100 * 120 * 140 *
24.11C : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11G : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11H : -CAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11D : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11N : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11E : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11F : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11O : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11P : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11Q : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11I : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11J : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11K : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11L : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11M : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11R : -CAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11S : -CAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11LG : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11LA : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11LB : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11LC : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11LD : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11LE : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11LF : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
24.11LH : CCAGGGTCTGTGGCTGAAGCCAACATTCAG----------
Genomic : CCAGGGTCTGTGGCTGAAGCCAACATTCAGGCAAGTCCATATTACCTGATTTACTAAATCATTGGCTATCACTTGGTAGGGCATGAGAGGATCCTATTCCACTTGGTTTTTGTGTCGTTTTCTCATTCTCTTAACTCACTCTTCTTTCTC
1 30 31
EXON 1
INTRON 1

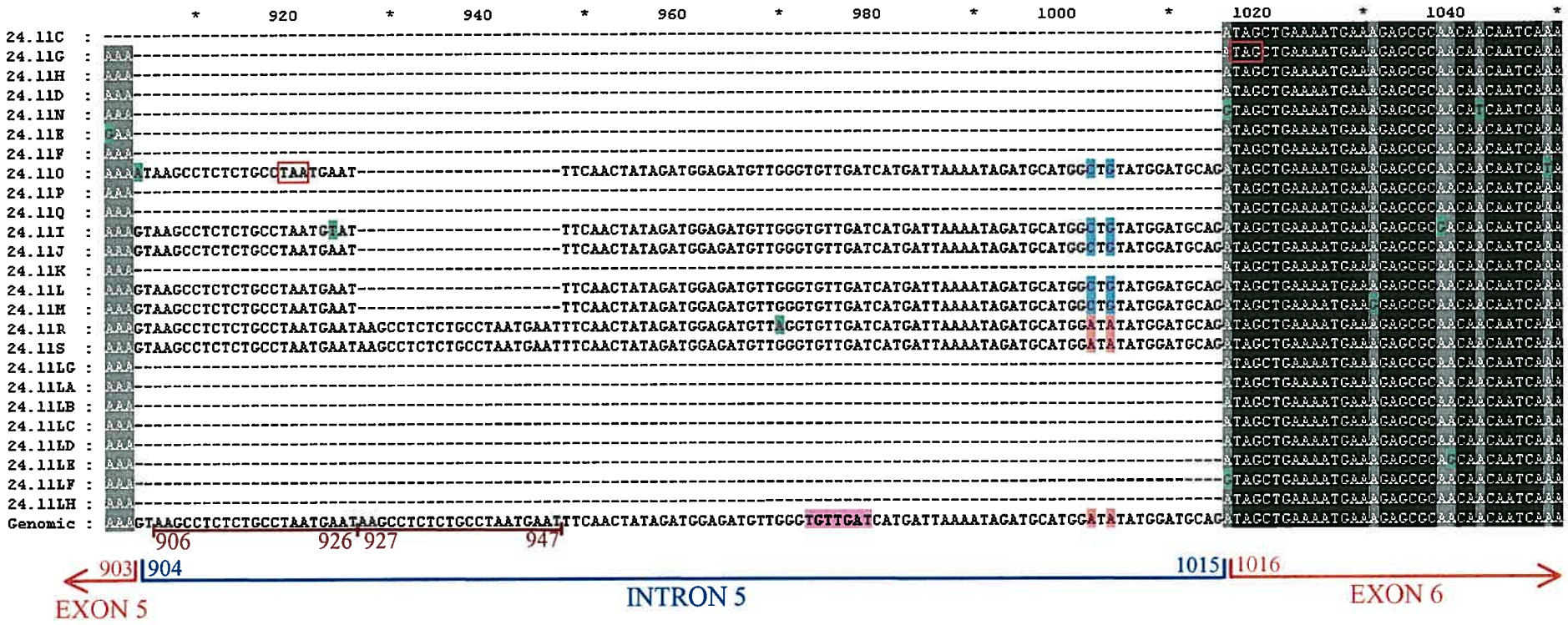

160
180
200
220
240
260
280
300
24.11C :
24.11G :
24.11H :
24.11D :
24.11N :
24.11E :
24.11F :
24.11O :
24.11P :
24.11Q :
24.11I :
24.11J :
24.11K :
24.11L :
24.11M :
24.11R :
24.11S :
24.11LG :
24.11LA :
24.11LB :
24.11LC :
24.11LD :
24.11LE :
24.11LF :
24.11LH :
Genomic : TTTATATTATATGATACTAATGTTATAAAAATTACCTCAG
190
191
252
253
INTRON 1
EXON 2
INTRON 2


320
340
360
380
400
420
440
24.11C
24.11G
24.11H
24.11D
24.11N
24.11E
24.11F
24.11O
24.11P
24.11Q
24.11I
24.11J
24.11K
24.11L
24.11M
24.11R
24.11S
24.11LG
24.11LA
24.11LB
24.11LC
24.11LD
24.11LE
24.11LF
24.11LH
Genomic
324
325
424
425
INTRON 2
EXON 3
INTRON 3


460
480
500
520
540
560
580
600
24.11C :
24.11G :
24.11H :
24.11D :
24.11N :
24.11E :
24.11F :
24.11O :
24.11P : CTTTCATTCTTTCCTTCCATCTTAATCTCTTTGTTTGTTTATTATATGGAATAAAATATCGTTATTCACTCACCACATTCTCCAACTCTACCAC
24.11Q :
24.11I :
24.11J :
24.11K : CTTTCATCCTTTCCTTCTATCTTAATCTCTTTGTTTGTTTATTATATGGAATAAAATATGGTTATTCACTCACCACATTCTCCAACTCTACCAC
24.11L : CTTTCATCCTTTCCTTCTATCTTAATCTCTTTGTTTGTTTATTATATGGAATAAAATATGGTTATTCACTCACCACATTCTCCAACTCTACCAC
24.11M : CTTTCATCCTTTCCTTCTATCTTAATCTCTTTGTTTGTTTATTATATGGAATAAAATATGGTTATTCACTCACCACATTCTCCAACACTACCAC
24.11R : CTTTCATTCTTTCCTTCCATCTTAATCTCTTTGTTTGTTTATTATATGGAATAAAATATCGTTATTCACTCACCACATTCTCCAACTCTACCAC
24.11S : CTTTCATTCTTTCCTTCCATCTTAATCTCTTTGTTTGTTTATTATATGGAATAAGATATCGTTATTCACTCACCACATTCTCCAACTCTACCAG
24.11LG :
24.11LA :
24.11LB :
24.11LC :
24.11LD :
24.11LE :
24.11LF :
24.11LH :
Genomic : CTTTCATTCTTTCCTTCCATCTTAATCTCTTTGTTTGTTTATTATATGGAATAAAATATCGTTATTCACTCACCACATTCTCCAACTCTACCAC
544
545
586
587
INTRON 3
EXON 4
INTRON 4


620
640
660
680
700
720
740
24.11C
24.11G
24.11H
24.11D
24.11N
24.11E
24.11F
24.11O
24.11P
24.11Q
24.11I
24.11J
24.11K
24.11L
24.11M
24.11R
24.11S
24.11LG
24.11LA
24.11LB
24.11LC
24.11LD
24.11LE
24.11LF
24.11LH
Genomic
INTRON 4


760
*
780
*
800
*
820
*
840
*
860
*
880
*
900
24.11C :
24.11G :
24.11H :
24.11D :
24.11N :
24.11E :
24.11F :
24.11O :
24.11P :
24.11Q : GTTTCATCTTCAGAATCTTTAATTTGTTAGACACTATACAAGATATAAGTGATAGCATAAAAGTCGTCGTCTTAATTATTAACAGCTGTAAAATGGGTATATAATTTCTAG
24.11I : GTTTCATCTTCAGAATCTTTAGTTTGTTAGACACTATACAAGATATGAGTGATAGCATAAAGGTCGTCGTCTTAATTATTAACAGCTGTAAAATGGGTATATAATTTCTAG
24.11J : GTTTCATCTTCAGAATCTTTAGTTTGTTAGACACTATACAAGATATGAGTGATAGCGTAAAGGTCGTCGTCTTAATTATTAACAGCTGTAAAATGGGTATATAATTTCTAG
24.11K : GTTTCATCTTCAGAATCTTTAGTTTGTTAGACACTATACAAGATATGAGTGATAGCATAAAGGTCGTCGTCTTAATTATTAACAGCTGTAAAATGGGTATATAATTTCTAG
24.11L : GTTTCATCTTCAGAATCTTTAGTTTGTTAGACACTATACAAGATATGAGTGATAGCATAAAGGTCGTCGTCTTAATTATTAACAGCTGTAAAATGGGTATATAATTTCTAG
24.11M : GTTTCATCTTCAGAATCTTTAGTTTGTTAGACACTATACAAGATATGAGTGATAGCATAAAGGTCGTCGTCTTAATTATTAACAGCTGTAAAATGGGTATATAATTTCTAG
24.11R : GTTTCATCTTCAGAATCTTTAATTTGTTAGACACTATACAAGATATAAGTGATAGCATAAAAGTCGTCGTCTTAATTATTAACAGCTGTAAAATGGGTATATAATTTCTAG
24.11S : GTTTCATCTTCAGAATCTTTAATTTGTTAGACACTATACAAGATATAAGTGATAGCATAAAAGTCGTCGTCTTAATTATTAACAGCTGTAAAATGGGTATATAATTTCTAG
24.11LG :
24.11LA :
24.11LB :
24.11LC :
24.11LD :
24.11LE :
24.11LF :
24.11LH :
Genomic : GTTTCATCTTCAGAATCTTTAATTTGTTAGACACTATACAAGATATAAGTGATAGCATAAAAGTCGTCGTCTTAATTATTAAGAGCTGTAAAATGGGTATATAATTTCTAG
GAAGTTGAACTGCAGAACGATAACATGTACCTGCGAGCA
861
862
INTRON 4
EXON 5


920
940
960
980
1000
1020
1040
24.11C :
24.11G :
24.11H :
24.11D :
24.11N :
24.11E :
24.11F :
24.11O :
24.11P :
24.11Q :
24.11I :
24.11J :
24.11K :
24.11L :
24.11M :
24.11R :
24.11S :
24.11LG :
24.11LA :
24.11LB :
24.11LC :
24.11LD :
24.11LE :
24.11LF :
24.11LH :
Genomic :
906
926 927
947
903
904
1015
1016
EXON 5
INTRON 5
EXON 6

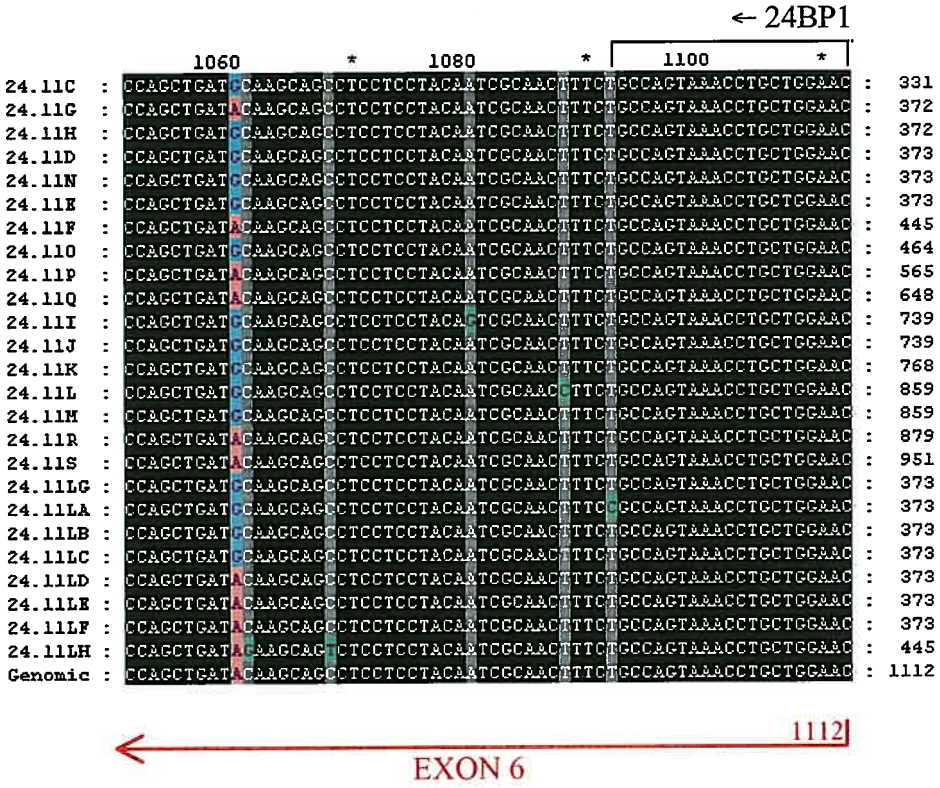


Figure 10: Alignment of the consensus sequences of all 25 clones.
The 17 transcripts that were cloned from the fibre tissue are shown as the top sequences, the genomic PCR clone is the bottom sequence and the transcripts cloned from the leaf tissue are in the middle. The regions that are 'intronic' and 'exonic' are indicated with brackets below the sequence alignment in blue and red respectively. Each intron and exon, from 5' to 3', is numbered with respect to this 1112 bp region of sequence. The numbers within the brackets indicate the nucleotide position with respect to the genomic PCR clone. Nucleotides highlighted in green correspond to errors introduced during RT-PCR, PCR or cloning. At positions where several nucleotides are red and several are blue, the red nucleotides correspond to the genomic nucleotide, while those highlighted in blue correspond to another common nucleotide (i.e. non-genomic). Red boxes around three adjoining nucleotides indicates a putative in-frame stop codon, within that particular transcript, when the transcript is translated. Nucleotides shaded in pink, within the genomic sequence, are putative branchpoint sequences with the penultimate nucleotide (adenosine) being the branchpoint nucleotide. The two copies of the perfect repeat (section D3) are underlined in brown. Primer binding sites are shown at the 5' and 3' ends of the sequence with black brackets.

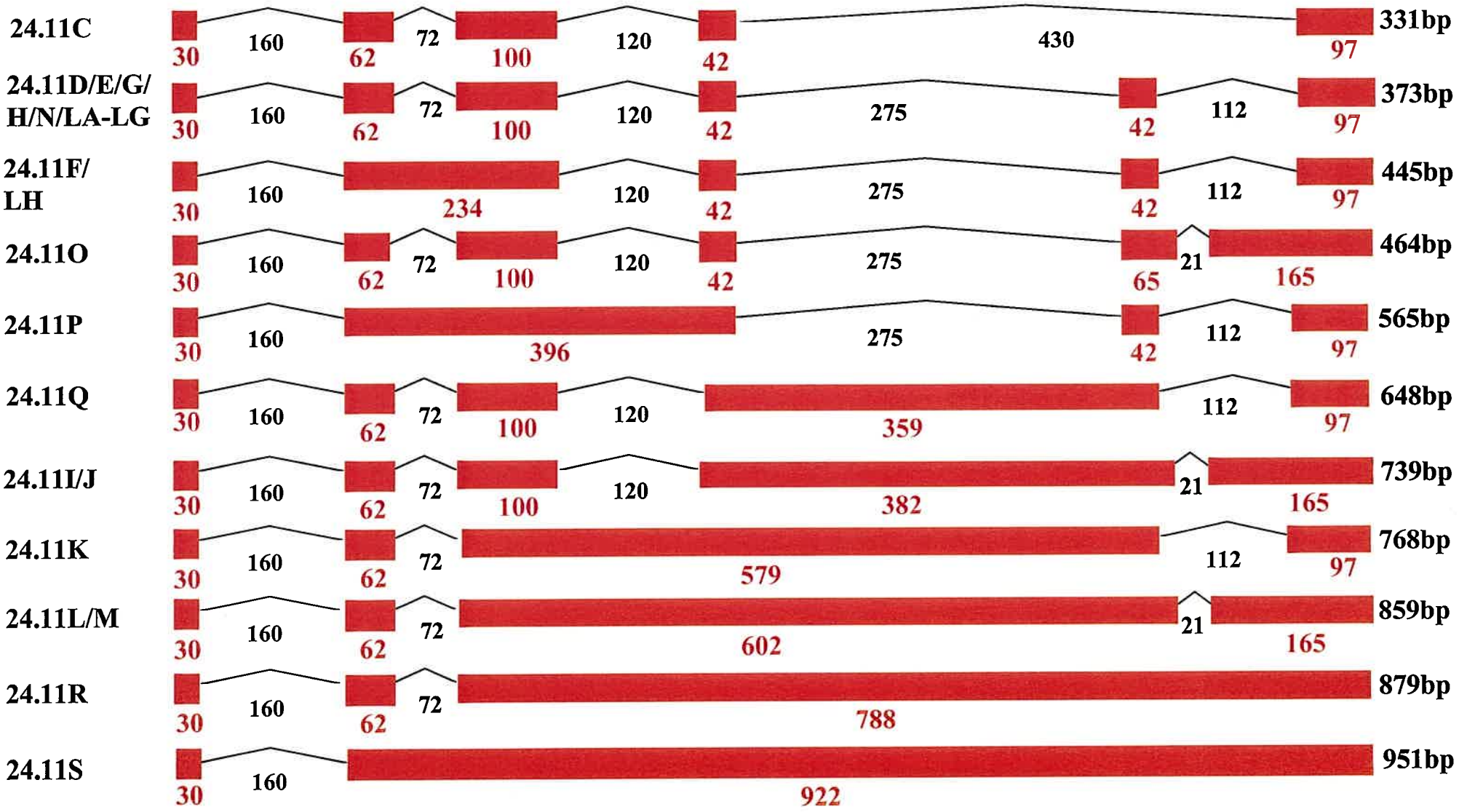


Figure 11: Schematic representation of the different intron/exon compositions found within the 25 clones.
The names of the clones are shown on the left and the length of the clones is shown on the right. Exons within each clone are represented by red boxes and introns within each clone are represented by a thin black line. The length of each region (intron/exon) is represented by numbers below the relevant region, with intron lengths in black and exon lengths in red.

these five regions are referred to as 'intronic', for simplicity, in this paper, despite the fact that they may be included in some transcripts.

When compared with the genomic PCR clone, six regions of sequence (nucleotides: 1-30, 191-252, 325-424, 545-586, 862-903 and 1016-1112; 30 bp, 62 bp, 100 bp, 42 bp, 42 bp and 97 bp long regions respectively; Fig. 10) were contained within every clone except for 24.11C which lacked the 862-903 bp region. Since these regions were contained within almost every clone, and they were flanked by characteristic intron splice sites, they were designated as being 'exonic' regions.

As expected, the most 5' section of sequence (the 30 bp region) and the most 3' section of sequence (the 97 bp region) are necessarily present in every clone because these are the regions where the 5' primer (24APCR5) and the 3' primer (24BP1) bind. Examining the entire region in a 5' to 3' direction, the 160 bp 'intron' from nucleotides 31-190 is spliced out in each of the clones, while the following 62 bp exon (191-252 bp) is not removed from any clone. The following 72 bp 'intron' (253-324 bp) is removed from the transcripts of all clones except for 24.11F, 24.11P, 24.11S and 24.11LH. All clones contain the next 100 bp exon (325-424) while only six clones (24.11K-M, 24.11P, 24.11R and 24.11S) contain the next 120 bp region (425-544). This is the putative intron that is contained within the sequence of 24.11B but not 24.11A (Malone, 2000). All 25 clones have the next 42 bp exon in their transcripts (545-586 bp). Following this, the eight longest clones, 24.11I-M and 24.11Q-S, contain a large, 275 bp, 'intron' (587-861). All clones except 24.11C then contain a small 42 bp 'exon' that is located between nucleotides 862-903. Most clones have the next 112 bp region (904-1015 bp) absent from their transcripts, while two clones (24.11R and 24.11S) contain the entire 112 bp

region. However when clones 24.11I, 24.11J, 24.11L, 24.11M and 24.11O are aligned against the genomic PCR clone, they appear to contain 91 nucleotides within this region with a 'gap' of 21 bp (927-947) (Fig. 10). This 21 bp gap in the sequence, when aligned against 24PCR1, is far too small to be an intronic region as the minimal plant intron length is between 70 and 73 nucleotides (Goodall and Filipowicz, 1990). Also the 'gap' is not flanked by the consensus intron splice sites of dicotyledonous plants. Interestingly, the nucleotides from the genomic PCR clone corresponding to the 21 bp gap (927-947) are a perfect repeat of the 21 preceding nucleotides (906-926) (Fig. 10).

In terms of length and intron/exon composition there are 11 unique transcripts that were cloned from the fibre tissue (Fig. 11). Many different possible combinations of 'intron' inclusion were observed with no correlation observable for combinatorial splicing out or retention of multiple introns. From the leaf tissue, transcripts of only two sizes were cloned. There are multiple transcripts of 373 bp cloned (24.11LA-G) with the same intron/exon structure as each other and also one clone of 445 bp (24.11LH) that contains the 72 bp 'intron', in addition to the five 'exonic' regions (1-30, 191-252, 325-424, 545-586, 862-903 and 1016-1112).

A number of single base pair changes were observed in the sequences of some clones when compared with the other clones and the genomic PCR clone (for example at nucleotide 225 in clone 24.11N, Fig. 10). Many of these differences were only observed in one out of the 25 clones and were thought to be substitutions that were introduced during reverse transcription, PCR or during cloning (green bases in figure 10). The rate of nucleotide misincorporation of MMLV reverse transcriptase has been estimated at one mismatched base every 500 bp (Sambrook *et al*, 1989), while the mismatch error rate of

Taq polymerase is approximately one in every 10,000 bp (Tindall and Kunkel, 1988). Within the sequences of the 25 clones, there were a total of 56 of these 'errors' out of a total of 14032 nucleotides. This is one error every 250 nucleotides, which is within the expected range of substitution frequencies when reverse transcription (MMLV) and PCR (*Taq* polymerase) are both performed.

D4: Conflicting sites

There were 16 sites within the alignment where several clones possessed a different nucleotide from that of the genomic nucleotide at the corresponding position (nucleotide positions: 208, 448, 458, 468, 510, 646, 680, 697, 725, 726, 772, 797, 812, 1002, 1004, 1060: Fig. 10). At these sites, several clones possessed the same nucleotide as the genomic clone (indicated in red in Fig. 10), while several clones possessed a single common alternative – that is a nucleotide that did not correspond to the genomic nucleotide (indicated in blue in Fig. 10). This finding is represented in table 2 where these 16 positions from each clone are shown in the relevant colour (red for genomic or blue for non-genomic).

D5: Sequence interpretation

When all 16 of these sites are considered, clones 24.11F, 24.11P, 24.11R, 24.11S, 24.11LD-LF and 24.11LH contained the nucleotides corresponding to the genomic sequence at all sites (table 2 - in red). Clones 24.11C-E, 24.11H-M, 24.11O, 24.11LA-LC and 24.11LG possessed the non-genomic nucleotide alternatives (table 2 – blue). In

	208	448	458	468	510	646	680	697	725	726	771	797	812	1002	1004	1060	
24.11C	C															G	
24.11D	C															G	
24.11E	C															G	
24.11H	C															G	
24.11I	C					A	G	T	C	T	G	G	G	C	G	G	
24.11J	C					A	G	T	C	T	G	G	G	C	G	G	
24.11K	C	G	C	T	G	A	G	T	C	T	G	G	G			G	
24.11L	C	G	C	T	G	A	G	T	C	T	G	G	G	C	G	G	1
24.11M	C	G	C	T	G	A	G	T	C	T	G	G	G	C	G	G	
24.11O	C													C	G	G	
24.11LA	C															G	
24.11LB	C															G	
24.11LC	C															G	
24.11LG	C															G	
24.11G	C															A	2
24.11Q	C					G	A	C	A	A	A	A	A			A	
24.11N	A															G	3
24.11F	A															A	
24.11P	A	A	T	C	C											A	
24.11R	A	A	T	C	C	G	A	C	A	A	A	A	A	A	A	A	
24.11S	A	A	T	C	C	G	A	C	A	A	A	A	A	A	A	A	4
24.11LD	A															A	
24.11LE	A															A	
24.11LF	A															A	
24.11LH	A															A	
Genomic	A	A	T	C	C	G	A	C	A	A	A	A	A	A	A	A	5

Table 2: Conflicting sites.
The 16 nucleotide positions (nucleotides 208, 448, 458, 468, 510, 646, 680, 697, 725, 726, 771, 797, 812, 1002, 1004 and 1060 from figure 10) at which there are significant differences between the 25 clones. The nucleotides that correspond to the genomic sequence are shown in red and the other common nucleotide is shown in blue. Cells in the table that have no nucleotides correspond to regions in each of the transcripts that are spliced out. The table is horizontally divided into five sections with black horizontal lines. From regions 1 to 5 (top to bottom of the table):

(1) those clones corresponding to the genomic nucleotides;
(2) those clones corresponding to the genomic nucleotide at only their 3' ends;
(3) the clone corresponding to the genomic nucleotides at the 5' end;
(4) those clones corresponding to the non-genomic nucleotides;
(5) the genomic sequence.

the three other clones, 24.11G, 24.11N and 24.11Q, a combination of nucleotides corresponding to both genomic and non-genomic nucleotides was observed. 24.11G and 24.11Q have nucleotides corresponding to non-genomic nucleotides at their 5' ends but their sequences correspond to the genomic sequence at the 3' ends. This is the opposite for 24.11N where there is a nucleotide corresponding to the genomic PCR clone at the 5' end and a non-genomic nucleotide at the 3' end.

The transcripts can be divided into four groups on the basis of comparison to the genomic PCR clone (Table 2):

(1) those corresponding to the genomic nucleotides at all 16 positions;

(2) those not corresponding to the genomic nucleotides at any of the 16 positions;

(3) those corresponding to the genomic nucleotides at only their 5' end;

(4) those corresponding to the genomic nucleotides at only their 3' end.

Since these 16 sites cannot be dismissed as simple errors introduced during RT-PCR, PCR or cloning and because they occur multiple times in several clones, they must be derived from genomic sequences other than 24PCR1.

D6: Implications of sequence data

One explanation is that the transcripts are from alleles and alloalleles of the same gene. *G. hirsutum*, a tetraploid resulting from the fusion of two ancestral diploid species 1 to 1.2 million years ago (Wendel, 1989), would contain two alleles and two alloalleles of each gene. Perhaps the four transcript types correspond to two alleles of two MADS-box genes, one from each of the diploid ancestors. It is difficult to know how likely this

explanation is as there have been few studies to show whether all four alleles of a tetraploid gene are transcribed or if some may be silenced.

Another explanation for these transcripts is that they arose from a family of different genes. Considering that more than 45 MADS-box genes have been identified in *Arabidopsis* (Ng and Yanofsky, 2001), and that MADS-are highly conserved, the primers (24APCR5 and 24BP1) may bind to multiple MADS-box genes. If there had been a recent duplication of a gene, with subsequent divergence, then the sequences of each of the genes may be similar enough to allow amplification of transcripts from both genes. The sequence of the transcripts from each of the genes would be slightly different. To explain the four different types of transcript (Section D5) there might be (a) two genes, each with two alleles or possibly (b) four genes with one allele each whose transcripts are being detected by the RT-PCR.

D7: Splicing Analysis

The efficiency of splicing, or removal, of introns from pre-mRNA is dependant upon a number of factors (reviewed by Brown and Simpson, 1998). Most important are:

(1) the consensus splice sites at the 5' and 3' ends of introns;

(2) the contrast of guanine and cytosine composition of exons compared with the uracil composition of introns (since this project is analysing cDNAs, uracil is replaced by thymine);

(3) the presence of an intron branchpoint sequence to allow intron lariat formation.

As previously discussed, all introns in the region being studied conform to consensus intron splice sites (Goodall *et al.*, 1991; Fig. 3). In all organisms, when the 5' end of an intron is cut, the most 5' base is covalently linked to an adenosine nucleotide that is between 10 and 50 nucleotides from the 3' end of the intron, to form an intron lariat structure (Tolstrup *et al.*, 1997). This adenosine is known as the branchpoint nucleotide and is within a conserved region known as the branchpoint. While the branchpoint sequence is well conserved in yeast and mammals, it is only weakly conserved in dicot plants, having a loose consensus sequence of 5'-WWCURAW-3' (where: 'W' represents adenosine or uracil; 'R' represents adenosine or guanine, the underlined adenosine is the branchpoint nucleotide and uracil corresponds to thymine in the alignment of cDNAs). Putative branchpoints were found between 10 and 50 nucleotides from the 3' end of each intron in the alignment (Fig. 10 – highlighted in figure 10). Although some of the branchpoints appear only weakly conserved, they generally agree with the positions that are best conserved within the branchpoint (i.e. in order of importance A (position 6), U (position 4) and C (position 3)) (Tolstrup *et al.*, 1997; Fig. 12).

General GC richness of exons compared to general AU richness of introns in known to be important for efficient splicing of introns by the splicing machinery (Latijnhouvers *et al.*, 1999). Of particular importance is the contrast of uracil composition of the 50 nucleotides on each side of the intron/exon boundary. For the majority of dicotyledonous plant introns, the 50 nucleotides on the intron side of the intron/exon boundary are typically 15% more uracil rich than the 50 nucleotides on the exon side of the intron/exon boundary. Conversely, the 50 nucleotides on the exon side of the intron/exon boundary are typically 15% more guanine and cytosine rich than the 50 nucleotides on the intron side of the intron/exon boundary (Latijnhouvers *et al.*, 1999).

The nucleotide composition contrast was calculated for the 24.11 transcript by counting the number of each of the nucleotides in the 50 bases on each side of every potential intron/exon boundary. The numbers of each of the four nucleotides on both sides of the intron/exons boundaries were compared and expressed as a percentage (Fig. 12). Most of the introns in the region being studied have ~15% or more thymine residues in the 50 nucleotides on the intron side of intron/exon boundary compared to the 50 nucleotides on the exon side of the intron/exon boundary (Fig. 12 – in blue). However at some intron/exon boundaries the contrast of nucleotide compositions is weak and a comparison of the 50 nucleotides on each side of the boundary did not conform to ~15% difference in thymine composition. For example, the fifth exonic region (nucleotides 862-903) is 3% less GC rich than the 50 nucleotides at the start of the fifth intron at nucleotides 904-953 when it (as an intronic region) would be expected to have a higher GC content. Another intron/exon boundary that does not conform is that between the third intron and the fourth exon. The 50 nucleotides on the intron side of this boundary are only 2% more thymine rich than the 50 nucleotides on the exon side. This exon is also only 2% more guanine and cytosine rich than the 50 nucleotides of the intron at the intron/exon boundary.

The majority of transcripts, as indicated by the fact that the major RT-PCR band is 373 bp, have the five 'intronic' regions spliced out. However, some of the transcripts cloned in this experiment have retained one or more 'introns'. The 160 bp intron (31-190), which has a branchpoint conforming to the consensus sequence (Fig. 10) and a large contrast in uracil, guanine and cytosine content at its 3' intron/exon boundary, is removed from every clone. The 5' exon/intron boundary cannot be analysed because the sequence does not extend 50 bp in the 5' direction of the first exon/intron junction. The

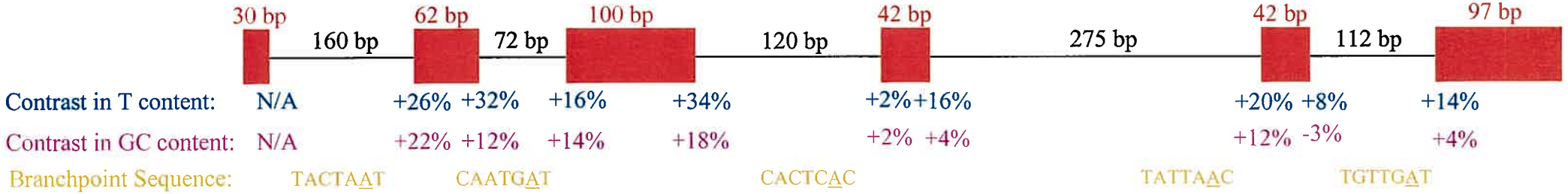


Figure 12: Intron/exon boundary composition and branchpoint sequences.
As an example, a transcript corresponding to the 373 bp major species is shown. 'Exonic' regions are represented as red boxes and 'intronic' regions are represented as thin black lines. The lengths of each 'exonic' and 'intronic' region are shown above the relevant region in red and black text respectively. For efficient intron splicing, the 50 nucleotides on each side of the intron/exon boundary are important, with introns typically being ~15% more uracil rich (thymine in cDNA) and exons being ~15 % more guanine and cytosine rich in these 50 nucleotides (Tolstrup *et al.*, 1997). The contrast of thymine composition between the 50 nucleotides on each side of the intron/exon boundary is shown in blue, below each intron/exon boundary, with the percentage representing thymine composition of the intron compared to the exon. The contrast in guanine and cytosine composition is shown in pink with the percentage representing guanine and cytosine composition of the exon compared to the intron. Comparisons cannot be made at the first intron/exon boundary because there is only 30 bp of sequence on the exon side of the intron/exon boundary. The putative branchpoints are shown below each intron in orange. The consensus branchpoint is 5'-WW**CU**R**A**W-3' (where: 'W' represents adenosine or uracil; 'R' represents adenosine or guanine, the second last position (adenosine) is the branchpoint nucleotide.

other introns, which have less conserved branchpoint sequences (Fig. 10) and, in general, lower nucleotide contrasts at intron/exon boundaries (Fig. 12) are not spliced out in every transcript. The 72 bp (253-324), 120 bp (425-544), 275 bp (587-861), and 112 bp (904-1015) introns are found within four, six, eight and eight clones respectively. Although plant intron splicing is not well understood, in this instance there appears to be some correlation between the removal of introns and the three factors that have been identified as being important for splicing efficiency.

It is not fully understood how the three sequence characteristics interact with each other for efficient splicing. Furthermore, it is also not known if there is any compensation between the three factors so that, for example, an intron with a weakly conserved branchpoint sequence but with a high uracil composition will still be efficiently spliced. Some intron/exon junctions have weak contrasts in nucleotide compositions but have branchpoint sequences close to the consensus. It is possible that some of the introns are not being spliced out of every transcript because the 'signals' specifying splicing are weak. Whether or not this is the case, there still may be some biological function for the proteins encoded by the alternatively spliced transcripts, indeed these sequences might be expected to have been selected in evolution to favour rare alternative splicing.

D8: Alternative splicing

The sequencing data presented indicate that the differently sized MADS-box transcripts which are apparent in the gel electrophoresis of the RT-PCR products may arise *via* alternative splicing. The different transcripts are extremely similar in sequence and the gaps (i.e. introns) are only present such that their 5' and 3' ends correspond to the

consensus splice sequences in the genomic PCR clone. The alternative splicing is of the form where regions that are usually 'intronic' are included in some of the stable mRNA.

D9: Future work

The obvious next experiment to perform is to obtain more genomic PCR clones to investigate the gene/s that correspond to the region being studied. Two previously obtained genomic clones (Malone, 2000) were sequenced during this project but the sequences were identical to 24PCR1. This could be achieved by performing a PCR, using 24APCR5 and 24BP1 as the primers, on genomic DNA prepared from *G. hirsutum*. This should provide genomic sequences that correspond to the cDNA clones that cannot be derived from 24PCR1. Although it seems safe to assume that other genomic clones would correspond to these transcripts (24.11C-E, 24.11G-O, 24.11Q 24.11LA-C and 24.11LG), it would be worthwhile to find these clones to remove any ambiguity about the variable nucleotides within the sequence alignment.

G. hirsutum is classified as having an AADD genome due to the hybridisation of ancestral A and D genome species. It would be interesting to analyse the region studied in this project by PCR amplification and automated sequence analysis, from the extant species most resembling the diploid ancestors, namely *G. herbaceum* (A genome) and *G. raimondii* (D genome) (Wendel, 1989). This would help to determine whether these transcripts are derived from one ancestral diploid or from both of them. By determining whether one or both diploids contained similar regions and by comparing the sequences of these genomic PCR clones it may be possible to determine whether the transcripts are derived from one or both of the diploid ancestors. However, there may be problems with this approach because it is not known how much *G. hirsutum* has diverged from the two

ancestral diploid genomes. It is possible that the primer binding sites might have diverged to the extent that no amplification would be possible.

Part E: Putative proteins

E1: Translation of 24.11 clones

While analyses of the nucleotide sequences were informative it is important to examine the protein sequences to determine what function the proteins encoded by these transcripts may have. BLASTX (ANGIS) was used to search a protein database (SWISS-PROT All) in all six potential frames of translation, with the sequences of the 25 clones, to determine the frame in which the transcripts were likely to be translated (Appendix D). In every case, only one frame returned significant matches: frame 1 (i.e. the first three nucleotides in the alignment (5'-CCA-3') represents the first amino acid position, in this region, for translation). When the clones were aligned against the full-length cDNAs of 24.11A and 24.11B, translation of the putative proteins from 24.11A and 24.11B in the region corresponding to the region being studied was also in the first frame. Using a program called 'Frames' (ANGIS) the sequences were searched for open reading frames. Frame number one gave full-length open reading frames for most clones, whereas the other five frames gave only short open reading frames. The proteins were therefore translated in the first frame using the 'Translate' program from ANGIS.

E2: Removal of redundant putative protein sequences

The nucleotide differences that were explained as being due to RT-PCR, PCR or cloning errors (Section D3) were altered to conform to the nucleotide at the corresponding position in the other clones. Ten amino acid sequences were identified as being unique and, as such, the redundant protein sequences removed. There were several sets of redundant putative proteins:

- 24.11K, 24.11L and 24.11M;
- 24.11I, 24.11J and 24.11Q;

- •24.11P and 24.11S;
- •24.11D, 24.11E, 24.11H, 24.11N, 24.11LA, 24.11LB, 24.11LC and 24.11LG;
- •24.11F and 24.11LH;
- •24.11LD, 24.11LE and 24.11LF.

The first clone, in terms of alphabetical order (24.11K, 24.11I, 24.11P, 24.11D, 24.11F and 24.11LD), of each group, was retained for the protein alignment.

E3: Protein analysis

The gene, and superfamily, to which the region of sequence being studied corresponds, belongs to the AGAMOUS family of MADS-box genes (Malone, 2000). Using Clustalw (ANGIS), the unique putative proteins were aligned against each other and AGAMOUS (Fig. 13). As can be seen, the differently spliced mRNAs could yield different proteins.

In terms of length, and which regions of RNA are translated from each clone, 24.11D, 24.11G and 24.11LD all corresponded closely with AGAMOUS (Fig. 13). They are 124 amino acids long and have the same regions of RNA translated into protein (i.e. they do not contain residues 32-55, 89-104, 119-134 or 146-150 as shown in figure 13).

As previously discussed (Section D3) some clones had different 'intronic' regions included, or 'exonic' regions excluded, in their transcripts. Because of this, the proteins that were encoded by the transcripts were of different lengths and had different regions of RNA translated. The proteins encoded by 24.11C, 24.11F, 24.11I, 24.11K, 24.11O, 24.11P and 24.11R, showed 'extra' regions of protein when aligned with AGAMOUS (i.e. introns had been unspliced from the transcripts). 24.11F and 24.11P contained

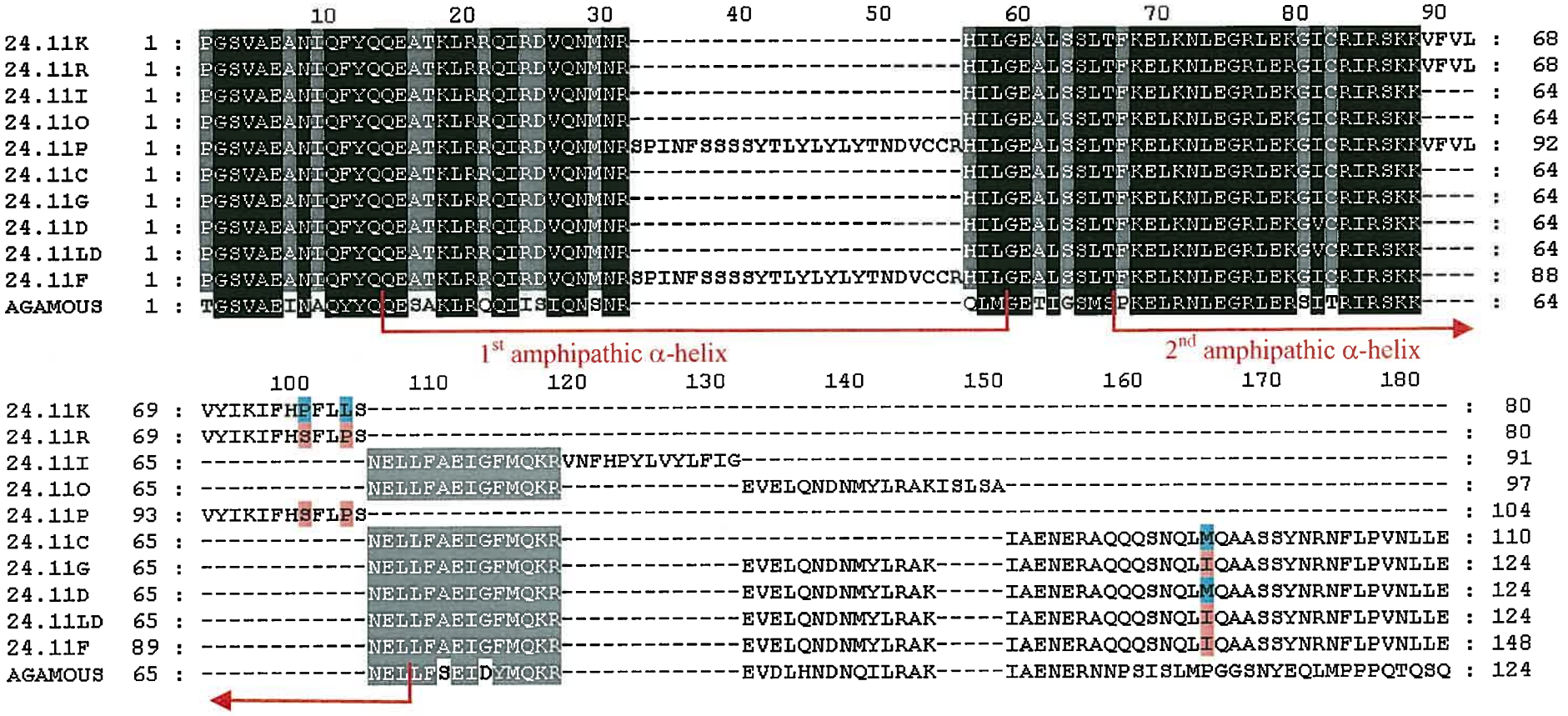


Figure 13: Putative protein alignment.
The 10 unique conceptual amino acids aligned with each other and the amino acid sequence of AGAMOUS, using Clustalw (ANGIS). The two putative amphipathic α-helices are indicated with red brackets below the sequence alignment. At each position, the residues shaded in black are those that are identical or chemically conserved, while lighter shading of residues indicates a weaker level of conservation. In the putative proteins, residues that are different, that are encoded by transcripts of the same length and intron/exon structure, are differentiated from each other by red and blue shading. The residues translated from transcripts that corresponded to the genomic PCR clone are shaded in red, while the residues that do not correspond to the genomic nucleotide are shown in blue

residues 32-55 due to the translation of the second 'intronic' region from figure 10 (nucleotides 253-324). This 72 bp intron was translated to give 13 residues that were not in the other clones. 24.11K, 24.11P and 24.11R encode a 16 residue region (residues 89-104) that was unique to their proteins due to translation of part of the third intronic region (nucleotides 425-475). A stop codon is found at nucleotides 473-475 and the protein terminates leaving the rest of the intron (nucleotides 476-544) untranslated. The protein sequence of clone 24.11I contains residues from 119 to 131 (Fig. 13) corresponding to nucleotides 587-628 from the fourth intronic region (Fig. 10) with nucleotides 626-628 encoding a stop codon leaving nucleotides 629-861 untranslated. 24.11O contains five residues (146-150) corresponding to the first 15 nucleotides of the 112 bp intron (nucleotides 904-918), with nucleotides 919 to 921 representing a stop codon. The alignment is designed to reflect the regions of RNA that the protein would be translated from (i.e. a gap has been placed in 24.11O and not 24.11P at residues 32-55 because 24.11P contains the 72 bp 'intronic' region within its transcript).

E4: Functionality of the proteins

To try to determine possible functions for the 10 unique proteins, a BLASTP (ANGIS) search was used to find other proteins, of known function (SWISS-PROT-All), with similar sequence. The clones that aligned well with AGAMOUS (24.11D, 24.11G and 24.11LD) were very similar to many other MADS-box genes (Appendix E). However, no MADS-box gene from available databases was found that showed homology to the translated intronic regions of protein (i.e. residues 32-55, 89-104, 119-131 or 146-150). Sequence searches with just these 'extra' regions of sequence revealed no significant matches.

The first two regions of translated intron-like sequence (residues 32-55 and 89-104) lie within the two amphipathic α-helices of the K-box (see introduction; Fig. 1). The clones that contain the second 'intronic' sequence (residues 32-55, nucleotides 253-324) would have an altered amphipathic α-helix, with the consensus hydrophobic amino acid positions altered. In the clones similar to AGAMOUS, the last three amino acids of the first α–helix are Histidine, Isoleucine and Leucine, but in 24.11F and 24.11P the last three amino acids of the first α-helix are Serine, Proline and Isoleucine. This results in the loss of one conserved hydrophobic amino acid (Isoleucine) thereby potentially changing the properties of this amphipathic α-helix. The translation of this intron would also change the spacing between the first and second amphipathic α-helix which may also have an effect upon dimerization

Clones 24.11K, 24.11P and 24.11R contain part of the third intronic region (nucleotides 425-475, residues 89-104). This region is within the second amphipathic α-helix and, as such, the second α-helix would be altered. The last three amino acids of the second helix would be changed from Asparagine, Glutamic Acid and Leucine to Valine, Phenylalanine and Valine with the inclusion of the 'intronic' region. Asparagine would be replaced by Valine (addition of one hydrophobic residue), possibly having some effect on the dimerization of the putative protein. This intron contains an in-frame stop codon that would terminate the protein one amino acid after the usual length of the K-box.

Clones 24.11I and 24.11O have segments of either the fourth (24.11I) (nucleotides 587-625, residues 119-131) or the fifth (24.11O) (nucleotides 904-918, residues 146-150) introns translated into protein with in-frame stop codons at nucleotides 626-628 and 919-

921 respectively. These proteins, as well as 24.11K, 24.11P and 24.11R have shortened and altered C-terminal regions. The C-terminal region has been demonstrated to be important for multimerization of MADS-box proteins, (Riechmann *et al.*, 1996) so alterations to this region may be important.

E5: Biological Relevance

It is not known how stable these transcripts are and whether they would encode biologically relevant proteins. Because the 373 bp cDNA species is the major product and it is very similar to AGAMOUS, within the K-box region, it seems likely that it would encode the major protein product from this gene. The role of the other transcripts is more unclear. They are much less abundant, as indicated by the weak bands on the gel electrophoresis of the RT-PCR, and show less homology with known MADS-box proteins, especially within the translated 'intronic' regions.

Although not direct evidence, there are some indications that the 'intronic' regions may be translated. The second 'intronic' region (nucleotides 253-324, residues 32-55) is 72 nucleotides long and therefore its translation would allow the rest of the protein to be maintained in the same frame as a transcript without the intron. Because the rest of the protein would be in-frame, any regions or domains translated after this intron may still retain their function, but the protein would be internally modified. All of the putative proteins are at least 80 amino acids long in terms of the region being studied. This may be important because all of the proteins would be long enough to have a region of sequence that would correspond to the K-box. Also, the introns that would be included for translation, before the end of the K-box, do not contain any in-frame stop codons. Although not proof that the intron would be translated, this lack of stop codons may

indicate that there has been some selection pressure to maintain protein contiguity. This suggests that the transcripts may be translated to give proteins with a biological function.

It must be remembered that the region being studied here corresponds only to the K-box domain of the gene. For each transcript, it is not known what is occurring in regions 5' and 3'of the K-box. However, it is likely that the alternative splicing in the K-box reported here is biologically relevant because full length cDNAs of 24.11A and 24.11B appear to encode protein that extend through the this region (Malone, 2000).

With altered K-box regions, proteins may have different dimerization properties. It may be envisaged that inclusion of an intron within a transcript may lead to a slightly altered protein that may have a different function from the protein encoded by the fully processed mRNA. The inclusion of introns in the 24.11 transcripts could potentially change the specificity of dimerization of the encoded MADS-box protein while maintaining the same DNA binding specificity, thereby generating variation and diversity of function.

The putative proteins with shortened or absent C-terminal regions may also exhibit altered binding specificities. The C-terminal region is known to be important for multimerization of MADS-box genes (Riechmann *et al.*, 1996). The proteins that lack the C-terminal region may therefore be able to bind DNA (MADS-box region) and also form dimers (K-box region), but may be deficient for multimerization.

Another MADS-box gene, murine *serum response factor*, has alternatively spliced C-terminally truncated proteins. These were shown to be competent for DNA binding and dimerization but were deficient for transactivation and had dominant negative effects

(Belaguli *et al.*, 1999). That is, the alternatively spliced form is still competent to form homo- or heterodimers and to bind DNA, but acts negatively to block gene transcription. By lacking the C-terminal region, higher order interactions with multiple proteins are no longer possible. It is possible that some of the proteins encoded by the alternatively spliced 24.11 transcripts act in the same way that dominant negative mutations act to block the usual function or interaction of a protein.

Part F: Genomic library screening

A cotton genomic library, constructed by S. Orford, was probed with the genomic PCR clone, 24PCR1, to achieve the projects second aim of finding the gene corresponding to the 24.11 transcripts. The library was screened on three occasions with no positive result obtained. On the X-ray film, background was clearly evident indicating that the filters had been left long enough to develop and the probe was sufficiently radioactive. Possible explanations are that the filters were old (pressed from the original library plates in 1997) and had been screened and stripped many times which may have caused loss of DNA. It is also possible that the probe labelling was unsuccessful or that the genomic library didn't possess any regions complementary to the probe.

F1: Future Work

For this part of the project, a previously constructed library was screened due to the large time requirement in constructing a new genomic library. To find the gene corresponding to the transcripts a fresh library would have to be constructed. To do this, the stored phage stock could be used to reinfected new bacterial host cells or an entirely new library could be constructed.

Alternatively, a kit such as CLONTECH's GenomeWalker could be used to obtain the gene by a PCR based approach. This approach utilizes a PCR-based method to 'walk' in the 5' and 3' direction of known genomic DNA. In this instance, it would amplify outwards, in both 5' and 3' directions, from the region being studied.

Future work

In terms of future work arising from this project, there are several areas that need to be further investigated. Firstly the gene corresponding to the region studied here should be isolated, as previously described by genomic library screening or by genome walking. It is necessary to characterize the rest of the gene to determine whether there is alternative splicing occurring in other regions of the gene, or if it is only localized within the K-box region.

Of importance is the issue of biological relevance. Do these cloned transcripts actually give rise to proteins and are these proteins able to interact? To investigate this, the different alternatively spliced transcripts could be cloned into expression vectors. The expression vectors can then be transformed into *E.coli* with a reporter plasmid (such as the luciferase system) with MADS-box DNA binding regions. The activity of the reporter system can be measured and the abilities of the different protein isoforms to active transcription of the reporter gene compared. This would show if the alternatively spliced isoforms are competent for DNA binding.

To investigate potential dominant negative functions of the alternatively spliced transcripts an affinity column method could be used. Both the proteins corresponding to the major 373 bp species, and the C-terminally truncated transcripts could be generated by *in vitro* translation. The protein of the major species could be His-tagged and attached to the column. The proteins corresponding to the truncated transcripts could be passed through the column and eluted, and the eluted fractions could be run on denaturing gels to determine whether there has been any binding.

It would also be interesting to perform Southerns of the RT-PCR with different probes. For example, the 'intronic' regions could be individually used to determine which bands on the RT-PCR, in which tissues, contain the various 'intronic' regions

The multiple clones have all been derived from one reverse transcription performed on tissue isolated from one plant. It would add further validity to the work to perform another reverse transcription on tissue from another plant to confirm that the multiple RT-PCR bands are genuine and reproducible. It might be possible that the multiple bands were produced because the reverse transcription was performed on one plant that exhibited poor splicing efficiency at the time of the experiment. For example, it has been previously observed that plant intron splicing efficiency is reduced with increased levels of plant stress (Simpson and Filipowicz, 1996).

In terms of more distant future work, experiments to identify other MADS-box genes from cotton could be undertaken. Since there have been more than 45 MADS-box genes identified from the model organism, *Arabidopsis* (Ng and Yanofsky, 2001), it is reasonable to expect many to be present within the cotton genome. The identification of more genes could be performed by screening cotton genomic libraries with the conserved MADS-box region as a probe or by performing PCR with primers to the ends of the conserved MADS-box domain.

Although not easy to do, it would be interesting to determine the function of the gene that the transcripts correspond to. This could be done by generating a transgenic plant that has an inactivating mutation in this gene. However, this experiment is probably not be

feasible, given current technology, because of the long time delay (~18 months) and the possibility of the gene possessing redundant functions.

This region of sequence has been investigated in this project because it corresponds to the K-box of a MADS-box gene. MADS-box genes have been suggested to have roles in growth and regulation processes, so the discovery of an apparently alternatively spliced gene, with expression in the cotton fibres, is of interest. Further investigation of the gene is warranted because the alternatively spliced forms of the transcripts may have some role in cotton fibre elongation.

Summary

The aims of the project have been partially achieved. The first aim of investigating multiple bands on the gel electrophoresis of the RT-PCR led to the cloning of 17 transcripts from 0 DPA cotton fibre and eight transcripts from leaf tissue. There contained five putative introns, revealed by the consensus 5' and 3' splice sites at the beginnings and ends of introns. In some clones, regions of DNA were present that corresponded to these 'introns', as defined by the 5' and 3' consensus splice sites. Different combinations of 'introns' were shown to be included in some of the clones that corresponded to the larger bands on the gel electrophoresis of the RT-PCR. The high level of sequence similarity of the different transcripts, and the conservation of the 5' and 3' consensus splice sites, indicates that these clones correspond to alternatively spliced transcripts.

Analysis of the alignment of all clones against the genomic PCR clone, 24PCR1 revealed a number of positions in which nucleotides in several clones did not correspond to the genomic PCR clone. This observation was investigated and four distinct types of transcript were detected. Three of the four groups of the transcripts identified could not be derived from the known genomic clone because the sequences of the transcripts were different from the sequence of the gnomic clone. This implied that the transcripts must be derived from either multiple genes or from multiple alleles of one or more genes.

The 25 clones were translated into theoretical proteins and analysed. About half (11/25) of the transcripts were similar to AGAMOUS, in terms of the regions of RNA that were translated. The others were significantly different to AGAMOUS as a result of

translation of regions that appeared to be 'intronic'. No obvious role was able to be attributed to the alternatively spliced transcripts, but it was of note that both the K-box and C-terminal regions would be altered in clones with included 'intronic' regions. In some clones, the conserved amphipathic α-helices within the K-box would be altered, while in other clones, the C-terminal region would be absent. The α–helices are known to be involved in MADS-box dimerization, and the C-terminal region required for multimerization, so it is easy to envisage that the alternatively spliced transcripts (those with 'intronic' regions) may have roles different from those of the clones that are fully processed.

Bibliography

Belaguli, N., Zhou, W., Trinh, T., Majesky, M., Schwartz, R. (1999) Dominant Negative Murine Serum Response Factor: Alternative Splicing within the Activation Domain Inhibis Transactivation of Serum Response Factor Binding Targets. *Molecular and Cellular Biology*: 19 4582-4591.

Brown, J.W.S., Simpson, C.G. (1998) Splice Site Selection in Plant Pre-mRNA Splicing. *Annual Review of Plant Physiology and Plant Molecular Biology* 49: 77-95.

Donati, S. (Unpublished) (1997) Tristyly in *Oxalis spp.*: Aspects of morphology and development. Honours thesis, Department of genetics, The University of Adelaide.

Goodall, G.J., Filipowicz, W. (1990) The Minimum Functional Length of pre-mRNA in Monocots and Dicots (1990) *Plant Molecular Biology* 14: 727-733.

Goodall, G.J., Kiss, T., Filipowicz, W. (1991) Nuclear RNA Splicing And Small Nuclear RNAs And Their Genes In Higher Plants. *Oxford Survey Of Plant, Molecular And Cell Biology* 7:255-296.

Hodgson, C.P. and Fisk, R.Z. (1987) Hybridisation Probe Size Control: Optimised "Oligolabelling". *Nucleic Acids Research* 15:6295.

Jack, T. (2001) Plant Development Going MADS. *Plant Molecular Biology* 46: 515-520.

Krogan, N.T., Ashton, N.W. (2000) Ancestry of Plant MADS-Box Genes Revealed by Bryophyte (*Physcomitrella patens*) Homologues. *New Phytology* 147: 505-517.

Latijnhouwers, M.J., Pairoba, C.F., Brendal, V., Walbot, V., Carle-Urioste, J.S. (1999). Test of the Combinatorial Model of Intron Recognition in a native Maize Gene. *Plant Molecular Biology* 41: 637-644.

Malone, K.M., (Unpublished) (2000) Expression of MADS-box genes in developing cotton fibres. Honours dossier, Department of Genetics, The University of Adelaide.

Miller, H.G., Kocher, T.D., Loy, B. (1995) New MADS box domains in *Asparagus officinalis* L. *Sexual Plant Reproduction* 8: 318-320.

Ng, M., Yanofsky, M.F. (2001) Function and Evolution of the MADS-Box Gene Family. *Nature Reviews Genetics* 2:186-195.

Riechmann, J.L., Meyerowitz, E.M. (1997) MADS Domain Proteins In Plant Development. *Biological Chemistry* 378: 1079-1101.

Riechmann, J.L, Wang, M., Meyerowitz, E.M. (1996) DNA-Binding Properties Of *Arabidopsis* MADS Domain Homeotic Proteins APETALA1, APETALA3, PISTILLATA and AGAMOUS. *Nucleic Acids Research* 24: 3134-3141.

Sambrook, J., Fritsch, E.F., Maniatis, T. (1989) *Molecular Cloning: A Laborarory Manual, 2nd Edition.* Cold Spring Harbour Laboratory Press, Cold Spring Harbour, N.Y.

Shimizu, Y., Aotsuka, S., Hasegawa, O., Kawada, T., Sakuno, T., Sakai, F., Hayashi, T. (1997) Changes in levels of mRNAs for cell wall-related enzymes in growing cotton fibre cells. *Plant Cell Physiology* 38: 375-378

Shore, P., Sharrocks, A.D. (1995) The MADS-Box Family Of Transcription Factors. *European Journal Of Biochemistry* 229: 1-13.

Simpson, G.G., Filipowicz, W. (1996) Splicing of Precursors to mRNA in Higher Plants: Mechanisms, Regulation and Sun-Nuclear Organisation of the Splicing Machinery. *Plant Molecular Biology* 32: 1-41 1996.

Southern, E.M. (1975) Detection of Specific Sequences Among DNA Fragments Separated by Gel Electrophoresis. *Journal of Molecular Biology* 98: 503-517.

Thieβen, G., Kim, J.T., Saedler, H. (1996) Classification and phylogeny of the MADS-box multigene family suggest defined roles of MADS-box gene Subfamilies in the morphological evolution of Eukaryotes. *Journal of Molecular Evolution* 43: 484-516.

Theissen, G., Becker, A., Rosa, A.D., Kanno., A., Kim, A., Munster, T., Winter, K., Saedler, H. (2000) A short History Of MADS-Box Genes In Plants. *Plant Molecular Biology* 42: 115-149.

Tindall, K.R., Kunkel, T.A. (1988) Fidelity of DNA Synthesis by the *Thermus aquaticus* DNA Polymerase. *Biochemistry* 27, 6008-6013.

Tolstrup, N., Rouze, P., Brunak, S. (1997) A branch point consensus from *Arabidopsis* found by non-circular analysis for better prediction of acceptor sites. *Nucleic Acids Research* 253159-3163.

Wendel, J.F. (1989) New World Tetraploid Cottons Contain Old World Cytoplasm. *Proceedings Of The National Academy Of Sciences Of The USA* 86: 4132-4136.