Characterisation of Putative Cotton Fibre-Specific Transcription Factors

Joanna Sundstrom B. Biotech
Honours Thesis

Discipline of Genetics
School of Molecular & Biomedical Sciences
University of Adelaide

November 2005

Declaration

This work does not contain any material written by another person, except where due reference is given in the text, and the work has not been presented previously as a component of any other academic course.

Joanna Sundstrom
November 2005

Word Count: 13,695

Acknowledgments

Firstly, I would like to thank my supervisor Jeremy Timmis for taking me on as an honours student in 2005. Thank you Jeremy for providing me with the opportunity to work on plant genetics whilst still on the familiar North Terrace campus.

I'd also like to take this opportunity to express my appreciation to the Cotton Research and Development Corporation for providing financial assistance to both myself and the project.

A very big thank you also goes out to all the members of the Timmis lab; Sven Delaney, John Humphries, Eun-Lee Jeong, Damien Lightfoot, Sharon Orford and Anna Sheppard. Thanks for making me feel welcome as one of the lab and for all your efforts in teaching me the 'hands on' skills of science.

In particular I would like to thank Anna, Damien and John for the never-ending laughs throughout the year. Thank you Anna for your patience, wisdom and company on walks to get chocolate or muffins. Thanks Damo for providing a listening ear, some good (and not so good) games of pool and making it look like I get in early in the mornings! Thank you John for your generous and caring nature, taped Survivor episodes and footy tipping chats on Friday and Mondays. I hope I've provided you all with some good entertainment in return!

Thanks also to all the Genetics honours and PhD students who helped make my year both enjoyable and memorable.

I'd like to acknowledge the help provided by Donna Crack and Louise O'Keefe of the Richards lab, for assistance with my Frontiers in Genetics assignment.

A special mention to Velta Vingelis for her many hours of support and guidance throughout my undergraduate years.

Finally, I would also like to acknowledge the ongoing support and encouragement provided by mum and dad throughout my time at university.

Abbreviations

5’ RACE	5’ rapid amplification of cDNA ends
AD	activation domain
CaMV	*cauliflower mosaic virus*
cDNA	deoxyribonucleic acid complementary to ribonucleic acid
DNA	deoxyribonucleic acid
dNTP	2’-deoxynucleotide-5’-triphosphate
DPA	days post anthesis
GFP	green fluorescent protein
GSP	gene specific primer
GUS	β Glucuronidase
HMG	high mobility group
LB	luria Broth
LTP	lipid transfer protein
$MgCl_2$	magnesium chloride
mRNA	messenger ribonucleic acid
MULE	*Mutator* transposase-like protein
ORF	open reading frame
PCR	polymerase chain reaction
RNA	ribonucleic acid
RT-PCR	reverse transcription polymerase chain reaction
TAE	tris-acetate-EDTA
UV	ultra violet

Contents

Pg

1. Summary

Cotton fibres are differentiated single cells, originating from the epidermal layer of the cotton ovule. Cotton fibre quality is determined by its length, strength and fineness properties, which are largely controlled by genetic factors. Genes that are expressed predominantly in fibre cells are likely to control development of the fibre and contribute to these fibre characteristics.

The promoter sequence of the fibre-specific gene *FSltp6* was isolated (Orford, unpublished) and the region necessary for fibre-specific expression was identified through successive 5' deletions of the promoter (DELANEY 2005). An 84 bp region identified, was subsequently used in a yeast-one-hybrid assay to identify transcription factors that may regulate expression of *FSltp6* (DELANEY 2005). Three classes of novel cotton proteins were identified; HMGA-like proteins, *Mutator* transposase-like proteins and an AT-hook protein.

The main objective of this project was to analyse the *Mutator* transposase-like and the HMGA-like partial cDNAs to verify that these proteins do contribute to the regulation of the *FSltp6* promoter. The initial aim was to isolate full-length cDNA sequences for each protein. A further aim was to perform functional analysis to determine the sub-cellular localisation and expression profiles of each of the proteins. A full-length sequence was identified for the *Mutator* transposase-like cDNA that encodes a predicted 771 amino acid peptide sequence. Preliminary functional analysis was also performed on both of the partial cDNA sequences and each showed expression in all of the cotton tissues tested. Expression of both mRNAs were higher in the ovule and 6 DPA fibre compared to 12 DPA fibre, suggesting a potential role for these genes in early fibre development.

Results from this project contribute to the field of cotton fibre development. More specifically, this work contributes to the understanding of cotton promoters and their roles in regulating fibre-specific expression. The full-length cDNA isolated in this project could encode a transcription factor required for fibre-specific expression, but further functional analysis must still to be performed for confirmation. Information about cotton fibre-specific promoters and the transcription factors that regulate them, such as the ones investigated in this study, could be used in the development of transgenic cotton with improved fibre quality and yield.

2. Introduction

2.1 The Cotton Plant

Cotton belongs to the genus *Gossypium* within which there are between 40 and 50 species. The majority of *Gossypium* species are diploids (2n = 26), but there are a number of allotetraploid species (2n = 52) (ENDRIZZI *et al.* 1985; WENDEL 1989). The diploid species have been divided into seven genomic groups (A to G) based upon relationships such as chromosome number and nuclear DNA content (EDWARDS and ENDRIZZI 1975; ENDRIZZI *et al.* 1985). The allotetraploid group is believed to have originated around 1–2 million years ago (WENDEL 1989) via a hybridisation and polyploidisation event between an Old World diploid of group A and a New World diploid of group D (ENDRIZZI *et al.* 1985).

Only four species of cotton are used for fibre production. These include the Old World diploid species *G. arboreum* and *G. herbaceum* along with the New World allotetraploids *G. barbadense* and *G. hirsutum* (ENDRIZZI *et al.* 1985; WENDEL 1989; WENDEL *et al.* 1992). The allotetraploids account for the majority of cultivated cotton, with *G. hirsutum* contributing around 90% of world production and *G. barbadense* around 8% (WENDEL *et al.* 1992). The allotetraploids have become the most widely cultivated due to their relatively long fibre length. Although *G. barbadense* has a longer and stronger fibre than *G. hisutum* it is not as widely cultivated due to its lower yields (ABDALLA *et al.* 2001; WENDEL *et al.* 1992).

2.2 The Cotton Industry

Cotton accounts for the greatest share of the world's natural textile fibre consumption, however a fall in cotton demand has been steady since the 1960s due to the introduction of synthetic fibres such as nylon and polyester. For cotton to regain its popularity, improvements need to be made in both fibre quality and the diversity of its uses.

The majority of Australia's cotton is grown in regional New South Wales and Queensland where the climates are most suitable. In 1999/2000 Australia's exported cotton accounted for 16% of the world's market share making Australia the third largest exporter (behind the USA and Uzbekistan) (GORDON 2001). It was also reported that 93% of Australia's cultivated cotton is exported, making the cotton industry an important one for the Australian economy.

The primary use of cotton is for fibre production although by-products also contribute to the crops value. Once the seed has been separated from the lint, the oil can be extracted from the seed kernels and used in many products including vegetable oils and margarines. The outside of the seed (hull) and leftover kernels can also be used for stock feed.

2.3 The Cotton Fibre

Cotton fibres are differentiated single cells, originating from the epidermal layer of the ovule. There are two forms of cotton fibre, lint and fuzz. Lint fibres are used in the manufacture of textiles due to their increased length compared to fuzz fibres.

During ginning lint fibres are removed from the seeds whilst the fuzz fibres remain attached.

There are four overlapping phases of fibre development (BASRA and MALIK 1984): initiation, elongation, secondary cell wall synthesis and maturation.

2.3.1 Initiation

The events that determine which of the epidermal cells will differentiate into fibre cells occurs two to three days before flowering (GRAVES and STEWART 1988). Around one in four epidermal cells differentiate into fibre cells, although the specific molecular mechanisms of differentiation are yet to be determined. Epidermal cells that initiate between 0 and 3 days after flowering (or days post anthesis; DPA) will develop into lint fibres (BEASLEY 1975; STEWART 1975; TIWARI and WILKINS 1995), while those that initiate later, between 4 and 12 DPA, are destined to become fuzz fibres (BASRA and MALIK 1984). Therefore the timing of fibre initiation is crucial as it determines whether the fibres will be commercially useful.

At anthesis, those cells that have begun initiation appear as tiny protrusions on the surface of the ovule; these are the fibre initials. Within 12 hours, these cells swell into a sphere above the surface of the ovule and the elongation phase begins (STEWART 1975; TIWARI and WILKINS 1995).

2.3.2 Elongation

By around 1 DPA the fibre initials develop tapered ends and the rate of elongation increases (STEWART 1975; TIWARI and WILKINS 1995). Maximum rates of elongation occur between 6 and 12 DPA (MEINERT and DELMER 1977).

During elongation a thin primary cell wall is deposited. As each fibre grows in length it meets adjacent fibres with which it continues to grow and spiral (STEWART 1975). The fibres continue to elongate for around 16 to 25 days. Final lint fibres are about 1000 to 3000 times longer than their diameter and are typically between 20 and 30 mm long in *G. hirsutum* (BASRA and SAHA 1999; TIWARI and WILKINS 1995). Environmental and genetic factors determine the rate of elongation and final fibre length (BASRA and SAHA 1999).

2.3.3 Secondary Wall Synthesis

Secondary cell wall synthesis begins around 16 to 18 DPA whilst elongation is still in progress and peaks when elongation is near completion, thus preventing further increases in length (BEASLEY 1979).

Secondary cell wall synthesis involves the deposition of cellulose just inside the primary cell wall. The secondary cell wall is much thicker and stronger than the primary cell wall, with the mature cell wall containing around 94% cellulose (BASRA and MALIK 1984). The extent of secondary cell wall deposition determines the final strength of the cotton fibres. Secondary cell wall synthesis ends at approximately 32 DPA (MEINERT and DELMER 1977).

2.3.4 Maturation

At 50-60 DPA the mature cotton boll splits open, exposing the fibres and seeds. The cotton fibres dry and collapse into twisted ribbon-like structures, allowing them to be removed from the seeds by ginning and then spun into yarn (BASRA and MALIK 1984; BASRA and SAHA 1999).

2.4 Cotton Quality

The quality of cotton fibres is determined by a few key factors, namely strength, length and fineness (BASRA and MALIK 1984). Increased length is an advantageous characteristic that is, to a large degree, controlled genetically (BASRA and SAHA 1999). Fibre strength and fineness are determined by genetic and environmental factors along with the maturity of the fibre at the time of harvesting. The more mature the fibre, the greater the degree of secondary cell wall synthesis, resulting in increased strength and hence good spinning characteristics. Immature fibres are weaker and do not twist together well during spinning, causing a decrease in quality of the final product (BASRA and MALIK 1984).

2.5 Cotton Biotechnology

In the past 10-20 years there has been a move towards genetic manipulation as a means of improving cotton crop yield and quality through the introduction of transgenes (JOHN and STEWART 1992). This is being achieved in conjunction with traditional plant breeding methods. So far the main achievements have been introducing exogenous genes conferring new traits that protect cotton crops against biotic and abiotic stresses. For example, transgenic varieties were released in the

United States in 1996 and 1997 containing genes that confer pesticide and herbicide resistance (KIM and TRIPLETT 2001; WILKINS *et al.* 2000).

The focus has now shifted towards the development of cotton with improvements in physical characteristics. To date, transgenic cotton crops have had transgenes expressed constitutively throughout the plant, often under the regulation of the *Cauliflower mosaic virus (CaMV) 35S* promoter. However, expressing a transgene in all tissues of a plant may have detrimental side effects to the environment and/or the plant itself. For example, expression of an insecticide gene in all tissues of the plant may harm not only the intended insect, but other beneficial or non-harmful insects. Also the constitutive expression of a transgene may place extra energy requirements on the plant, such that yield and quality are compromised. To avoid these detrimental side effects, attention is now being focussed on modifying gene expression in a particular tissue rather than in the whole plant. An example of this strategy is to express an insecticide gene only in the plant tissue which is attacked by the target insect.

2.6 Tissue-specific Promoters

To facilitate this preferred approach of expressing transgenes in particular plant tissues rather than all plant tissues, tissue-specific genes and their promoters must be identified. Once the expression patterns of tissue-specific genes are understood, promoters controlling expression of these genes could be used to drive expression of a wide variety of potential transgenes in target tissues at particular developmental stages.

To date only a limited number of cotton tissue-specific promoters have been characterised. Examples of these include fibre-specific promoters that drive expression of genes predominantly in cotton fibre cells (HSU *et al.* 1999; LIU *et al.* 2000) and seed-specific promoters that drive expression of genes exclusively in the developing cotton seeds (SUNILKUMAR *et al.* 2002). These promoters have the potential to be used in the future to express transgenes in developing fibre cells to improve fibre characteristics or in cotton seeds to modify seed properties.

Transcription factors are proteins that interact with promoters to control the location, level and timing of gene expression. Knowledge of the transcription factors that act on tissue-specific promoters could therefore also be used to control tissue-specific transgene expression. For example, altering the level or timing of transcription factor expression in particular tissues could be used to modify tissue-specific gene expression. To date, few cotton transcription factors that act on tissue-specific promoters have been identified.

2.7 Enhancing Fibre Characteristics

Research into the enhancement of fibre quality and yield, by direct modification of gene expression in the fibre is being conducted by various research groups. A number of fibre-specific genes have been identified that are specifically expressed in the cotton fibre at particular developmental stages (e.g. ORFORD and TIMMIS 1997). However, limited cotton genome sequence information combined with the difficulty and time required to generate transgenic plants, has led to the use of model systems in most studies into these genes and their promoters.

Arabidopsis and tobacco are advantageous model systems as they are easier to generate transgenic plants and have a shorter generation time compared to cotton. *Arabidopsis* and tobacco leaf trichomes have been used as models for cotton fibre development. Both cotton fibres and leaf trichomes are single-celled structures originating from epidermal tissue. *Arabidopsis* and tobacco leaf trichomes share genetic and developmental similarities to cotton fibres (HUMPHRIES *et al.* 2005; PAYNE *et al.* 1999). However, Payne and co-workers (1999) showed that expression of a cotton gene in *Arabidopsis* and tobacco did not effect the leaf trichomes in a similar mannar as the processes of *Arabidopsis* and tobacco trichome development are regulated differently. These results suggest that any findings in *Arabidopsis* or tobacco must still be tested in a cotton system for confirmation. Some of the fibre-specific promoters isolated have been tested in these model systems to determine whether they direct expression of transgenes specifically in trichomes and hence possibly also in cotton fibres. An example is outlined below.

Lipid transfer proteins (LTPs) are thought to be involved in cutin deposition during fibre development (MA *et al.* 1995). Two LTPs were shown to be expressed specifically during the early stages of fibre cell elongation (MA *et al.* 1997; MA *et al.* 1995). The promoters of these genes (*Ltp3* and *Ltp6*) were identified and analysed in subsequent studies (HSU *et al.* 1999; LIU *et al.* 2000). Various promoter constructs were made consisting of different 5’ deletions of the promoters, fused to the β-Glucuronidase (GUS) reporter gene. These constructs were stably transformed into tobacco and showed GUS expression predominantly in the leaf trichomes. The results from these studies added support for using tobacco trichomes as heterologous models of cotton fibre development (HSU *et al.* 1999; LIU *et al.* 2000). However, before these

LTP promoters can be used for generating transgenic cotton with improved fibre characteristics, the specificity of the promoters for directing expression specifically in cotton fibres will first need to be tested in the cotton system.

2.8 Previous Results that Underpin the Project

Five cDNA clones were isolated corresponding to mRNAs that were preferentially expressed in fibre cells (ORFORD and TIMMIS 1997). One of these cDNAs corresponded to a LTP termed *FSltp6*. Northern analysis showed it was highly and exclusively expressed in fibre cells between 6 and 14 DPA, which indicated a specific role for *FSltp6* in early fibre development. With such a high expression level specifically in fibre cells, the gene that encodes *FSltp6* was expected to have a strong fibre-specific promoter. The full-length *FSltp6* mRNA was characterised using 5' rapid amplification of cDNA ends (5' RACE) (ORFORD and TIMMIS 1997). The promoter sequence of *FSltp6* was subsequently determined using genome walking (Orford, unpublished).

To isolate the region of the *FSltp6* promoter necessary for fibre-specific expression, various 5' deletions of the promoter were generated by PCR and placed into constructs fused to the GUS reporter gene (DELANEY 2005). Each of the constructs were transiently expressed in various cotton tissues by microprojectile bombardment. This technique involved coating gold or tungsten particles with a promoter construct and then bombarding the construct into a plant tissue. Results from the bombardment experiments identified an 84 bp AT-rich region of the *FSltp6* promoter necessary for exclusive fibre-specific expression.

A yeast-one-hybrid assay was performed to identify putative transcription factors that bind to the 84 bp region identified in the *FSltp6* promoter (DELANEY 2005). As the *FSltp6* promoter drives expression predominantly between 6 and 14 DPA (ORFORD and TIMMIS 1997), a cDNA library was constructed from RNA extracted from 8 DPA cotton fibres. Each cDNA was fused to a yeast activation domain (AD) and transformed into a yeast strain containing reporter genes under the control of a promoter that includes a target DNA sequence (in this case three copies of the 84 bp region in tandem). Proteins translated from cDNAs that bind to the target sequence can be identified by activation of the reporter genes.

Three classes of proteins, encoded by the cDNAs, were identified in the yeast-one-hybrid assay: *Mutator* transposase-like, HMGA-like and AT-hook. Proteins from each of these three classes have been identified as transcription factors in other plant species. A brief description of each class is outlined below:

2.8.1 *Mutator* transposase-like proteins

Mutator is a plant transposon identified in maize, with *Mutator* transposase-like proteins (MULEs) identified in many plant species including rice, *Arabidopsis* and barley. MULEs are composed of unique internal sequences flanked by *Mutator* terminal inverted repeats (TIRs) (LISCH 2002). The internal sequences are often derived from the host genome. A study in rice showed that there are over 3000 MULEs, with greater than 1000 of these containing internal sequences homologous to rice genes (JIANG *et al.* 2004). The study also showed that around 20% of the MULEs, containing homologous rice genes, were composed of sequences derived from two or more genes resulting in novel open reading frames. Due to the nature and frequency of MULEs in many plant species the study suggests that they have contributed to the evolution of genes in the plant kingdom.

2.8.2 HMGA-like proteins

High mobility group (HMG) proteins associate with chromatin and other transcription factors, thereby enabling modification of the chromatin structure (GRASSER 2003). Two families of HMG proteins exist in plants; HMGA and HMGB. Plant HMGA proteins usually consist of four copies of an AT-hook DNA binding motif that binds to AT-rich DNA sequences of five bases or more. HMGA-like proteins have also been identified in plants and differ from HMGA proteins in that they are nearly twice as large and contain more AT-hook binding motifs (KLOSTERMAN and HADWIGER 2002). Limited functional studies have been carried out on these plant HMGA-like proteins, but they are thought to still be involved in transcriptional processes due to the presence of the AT-hook binding motifs.

2.8.3 AT-hook proteins

AT-hook proteins contain the same AT-hook DNA binding motifs as HMGA proteins, but consist of only two motifs instead of four. No other similarities exist outside of the AT-hook DNA binding motifs. These proteins are speculated to be transcription factors due to the presence of the AT-hook DNA binding motifs. AT-hook proteins have been identified in *Arabidopsis* and rice genome sequencing projects and other plant species, but no biological function has been ascertained to date (DELANEY 2005).

cDNAs encoding four *Mutator* transposase-like, three HMGA-like and one AT-hook protein activated reporter gene expression in the yeast-one-hybrid assay (DELANEY 2005). Two of the *Mutator* transposase-like cDNAs and one each of the HMGA-like and AT-hook cDNAs were fully sequenced. The *Mutator* transposase-like cDNAs encoded different proteins but both showed similarity to *Mutator* transposase-like genes in *Arabidopsis* and rice. The HMGA-like cDNA showed regions of homology to a tobacco HMGA-like protein, while the AT-hook cDNA showed homology to several predicted *Arabidopsis* and rice proteins with two AT-hook motifs. All of these sequences were shorter than similar full-length cDNAs of each of the classes of protein (668 bp and 820 bp compared to ~3000 bp for the *Mutator* transposase-like proteins, 1225 bp compared to ~2000 bp for the HMGA-like proteins and 1191 bp compared to ~2000 bp for the AT-hook proteins) (DELANEY 2005). This suggested that only partial cDNAs had been isolated.

A full-length cDNA sequence encoding the AT-hook protein was isolated using 5' rapid amplification of cDNA ends (5' RACE) on a cDNA library isolated from 12 DPA fibres (DELANEY 2005). The subcellular localisation of the AT-hook protein was investigated by generation of a fusion protein containing the AT-hook coding sequence and the green fluorescent protein (GFP) coding sequence. A plasmid containing the coding sequence for this fusion protein was bombarded into onion epidermal cells where GFP was primarily localized to the nucleus. The result was consistent with the expected localisation profile of a transcription factor, suggesting that this AT-hook protein may be involved in regulating the *FSltp6* promoter.

2.9 Project Aims

The main objective of this project was to further analyse the *Mutator* transposase-like and the HMGA-like partial cDNAs identified in the yeast-one-hybrid assay and to determine if these cDNAs encoded transcription factors involved in the regulation of the *FSltp6* promoter. Steps involved in achieving this objective are outlined below.

The first aim was to isolate the full-length cDNAs encoding one each of the *Mutator* transposase-like and HMGA-like partial cDNAs identified in the yeast-one-hybrid assay, using 5' RACE. Reverse primers were designed to bind specifically to cDNA sequences matching the 5' regions of the partial cDNAs. Forward primers designed to bind to the 5' adaptor sequences already ligated to the 5' ends of each cDNA in a 12 DPA adaptor ligated cDNA library were provided in the kit. A PCR technique using the forward and reverse primers was performed in an attempt to isolate the full-length cDNA sequences. Another round of PCR was used when multiple PCR products were obtained resulting from reverse primer binding to similar sequences. This involved the same process but with alternative reverse primers (also designed to bind specifically to the 5' regions of the partial cDNAs).

Once full-length sequences had been obtained, the second aim was to perform functional analysis on the full-length cDNAs. This would firstly involve construction of vectors, encoding fusion proteins between the full-length cDNA sequence and GFP, using the Gateway™ cloning system. The initial step would involve cloning of each full-length cDNA into an entry vector. Homologous recombination, promoted by appropriate recombinase enzymes, would occur between the entry vector and a destination vector (containing the GFP gene) to generate an expression vector.

Upon successful generation of two expression vectors (one each containing the full-length cDNA sequences of the *Mutator* transposase-like protein and the HMGA-like protein respectively), microprojectile bombardment into onion epidermal cells would be performed. Onion cells are typically used due to their negligible autofluorescence and large nuclei that can be easily visualized compared with cotton fibres. GFP would be observed to determine the sub-cellular localisation of the *Mutator* transposase-like and HMGA-like proteins. It was expected that if these two proteins were transcription factors, they would most likely localise to the nuclei of the onion cells.

A further aim of the functional analysis of the cDNAs, was to use reverse transcription PCR (RT-PCR) on various cotton tissue cDNA samples, in order to determine the tissues that express the *Mutator* transposase-like and HMGA-like mRNAs. It was expected that if these cDNAs did represent the coding sequences for transcription factors that act on the *Fsltp6* promoter, then expression would be seen in the early fibre tissues.

3. Materials and Methods

3.1 Primers

Primer Name	Sequence (5’ - 3’)	Use
ActinF	CTT ATG TTG CCC TGG ACT ATG AGC	RT-PCR Positive Control
ActinR	GGA CAA CGG AAT CTC TCA GCT CC	RT-PCR Positive Control
HMGF1	TAC TCC TTC CAA CCC TCC CGA TG	RT-PCR
HMGR1	ACA AAC GAA CAC ACA CCT CAA CAA C	RT-PCR
MutF1	GTT ATC AGT TGC ATT GGA CGA AGC	RT-PCR
MutR1	CCA ATC TCT TTC TGA TAA TAC AGT AAC	RT-PCR
M13F	GTA AAA CGA CGG CCA G	Colony PCR and sequencing pGEM-T Easy vectors
M13R	CAG GAA ACA GCT ATG AC	Colony PCR and sequencing pGEM-T Easy vectors
HMG GSP	CCA CGG CTT CGT TTG AAT CCA GG	Primary 5’ RACE Round I
HMG NESTED	GGG AGG GTT GGA AGG AGT AGA ATC G	Nested 5’ RACE Round I
MUT GSP	CTG TAG CTC TCG GTT GTA AAG TAT CTG G	Primary 5’ RACE Round I and amplification of *Mutator* transposase-like coding sequence
MUT NESTED	GCT TCG TCC AAT GCA ACT GAT AAC	Nested 5’ RACE Round I
Mut GSP 2	AGC AGC CCC TTC ACA AGT ATG TGT TG	Primary 5’ RACE Round II and amplification of *Mutator* transposase-like coding sequence
Mut Nested 2	CCT TCT GCT CTG CAC TTA ACA GTC ACG	Nested 5’ RACE Round II
MUTFLF	AGG CGT CGA CAT GAC TAC CAA AAG GGT GAT AGC GGT TTG	Amplification of *Mutator* transposase-like coding sequence
MUTFLRXhoI	CCG CTC GAG CCA ACA CTC TAC AAG AGC TCT TTG C	Amplification of *Mutator* transposase-like coding sequence
MUTFLRNotI	ATA AGA ATG CGG CCG CCC AAC ACT CTA CAA GAG C	Amplification of *Mutator* transposase-like coding sequence

MUT6F	CCG ATG GGA TTG CAG TTG TAA AGG TTG GC	Amplification of *Mutator* transposase-like coding sequence
MUT6Fc	GCA ACG ATT CAC TGG TGT TCA GGA ATT TCG	Amplification of *Mutator* transposase-like coding sequence
MUTSF	CCG ATG GGA TTG CAG TTG TAA AGG TTG GC	Amplification of *Mutator* transposase-like coding sequence

Primers were designed and checked for suitability and compatibility using NetPrimer, available at http://www.premierbiosoft.com/netprimer/netprlaunch/netprlaunch.html

3.2 PCR

3.2.1 Reverse Transcription-PCR (RT-PCR)

RT-PCR was performed in 25µL reaction volumes containing 1X *Taq* buffer, 2.0mM $MgCl_2$, 0.4mM of each dNTP, 0.4µM of each primer, 1.5 units *Taq* polymerase (Geneworks) and cDNA template supplied by Sharon Orford.

Cycling was typically performed with an initial denaturation step at 95°C for 2 mins, followed by 30 cycles of denaturation at 95°C for 1 min, annealing at 55°C for 30 secs and extension at 72°C for 30 secs. A final extension step was performed for 4 mins.

3.2.2 5' Rapid Amplification of cDNA Ends (5' RACE)

5' RACE was performed using the Marathon™ cDNA Amplification Kit (BD Biosciences, formerly CLONTECH) according to the manufacturer's instructions, with program 1 for hot-lid thermal cyclers and Advantage® 2 Polymerase Mix (BD

Biosciences, formerly CLONTECH) used. The adaptor ligated 12 DPA cDNA template was supplied by Sharon Orford.

3.2.3 Colony PCR

Colony PCR was performed in 25µL reaction volumes containing 1X *Taq* buffer, 2.0mM $MgCl_2$, 0.4mM of each dNTP, 0.4µM of each primer, 1.5 units *Taq* polymerase (Geneworks) and bacterial cells.

Cycling was typically performed with an initial denaturation step at 95°C for 2 mins, followed by 30 cycles of denaturation at 95°C for 1 min, annealing at 55°C or 60°C for 1 min and extension at 72°C for 2 mins. A final extension step at 72°C was performed for 7 mins.

3.2.4 PCR from cDNA

PCR was performed in 25µL reaction volumes containing 1X buffer, 0.2mM of each dNTP, 0.2µM of each primer, 1 unit Advantage® 2 Polymerase Mix (BD Biosciences, formerly CLONTECH) and 6 DPA cDNA as template.

Cycling was typically performed with an initial 95°C for 2mins, followed by 35 cycles of denaturation at 95°C for 1min, annealing at 55°C or 60°C for 1min and extension at 72°C for 2.5mins. A final step of 72°C was performed for 7mins.

3.2.5 Amplification of full-length *Mutator* transposase-like coding sequence

PCR was performed in 25µL reaction volumes containing 1X buffer, 0.2mM of each dNTP, 1 unit Advantage® 2 Polymerase Mix (BD Biosciences, formerly

CLONTECH) and two separate PCR products of partially amplified *Mutator* transposase-like coding sequence as template. Primers were added to the reaction in a final of 0.2µM each after an initial PCR program (see below).

Cycling was performed with an initial denaturation step at 95°C for 2 mins, followed by five cycles of denaturation at 95°C for 1 min and a combined annealing/extension step at 68°C for 3 mins. A final extension step at 68°C was performed for 4 mins. After this initial PCR program primers were added. A second round of cycling followed, with an initial denaturation step at 95°C for 2 mins, followed 35 cycles of denaturation at 95°C for 1 min, annealing at 60°C for 1 min and extension at 72°C for 2.5 mins. A final extension step at 72°C was performed for 7 mins.

For all PCRs a negative control was also performed containing no template DNA.

3.3 cDNA Synthesis

cDNA was synthesised using an Advantage® RT-for-PCR Kit (BD Biosciences Clontech). 1µg of RNA extracted from 6 DPA cotton fibres (supplied by Damien Lightfoot) was used in the cDNA synthesis reaction according to the manufacturer's instructions.

3.4 Agarose Gel Electrophoresis

DNA was run on 1-2% agarose gels in 1X TAE buffer with a 2-log DNA ladder (New England Biolabs). Gels were stained in ethidium bromide and visualised under UV light.

3.5 Restriction Digests

New England Biolabs restriction enzymes and buffers were used in restriction digests according to manufacturer's instructions.

3.6 Purification of PCRs and Restriction Digests

PCR and restriction digests were purified using a QIAquick Gel Extraction Kit (Qiagen) if a particular fragment required separation from other sized fragments. Otherwise, a PCR Purification Kit (Qiagen) was used according to manufacturer's instructions.

3.7 Cloning

3.7.1 Ligations

Cloning of 5' RACE and RT-PCR products into the pGEM-T Easy vector (Promega) was performed according to manufacturer's instructions with the exception that a 10X ligation buffer (Promega) was used instead of the standard 2X ligation buffer and the reaction was incubated overnight at 16°C.

Cloning of the full-length *Mutator* transposase-like coding sequence into pENTR1a (Invitrogen) was performed after double restriction digest of both insert and vector. T4 DNA ligase (Promega) was used to ligate insert and vector together, as with the pGEM-T Easy ligations.

3.7.2 Transformations

Transformation was performed as described in INOUE *et al.* 1990. Screening of colonies was performed using a combination of colony PCR (see section 3.2.3) and restriction digests (see section 3.5).

3.8 Purification of plasmid DNA

Cells containing a desired plasmid were cultured overnight in LB plus antibiotic (final concentration 100ug/mL ampicillin or 50ug/mL kanamycin) at 37°C with shaking. Plasmid DNA was purified from the cultures using a Genelute™ Plasmid Mini-prep Kit (Sigma) according to manufacturer's instructions.

3.9 Sequencing

Sequencing reactions were performed in 20µL volumes with 4µL Big Dye mix (version 3), 0.25µM primer and 300-500ng vector DNA. Cycling and sequence clean up were performed as per the Institute of Medical and Veterinary Science (IMVS) recommendations. Processing was performed by the IMVS.

3.10 Bioinformatics

Nucleotide sequence alignments were performed using GeneDoc version 2.6.002. Nucleotide and protein sequences were analysed using a variety of GCG programs provided by Accelrys and accessed through the Australian National Genomic Information Service (ANGIS, http://www.angis.org.au). PSORT was used to identify putative nuclear localisation signals available at http://psort.nibb.ac.jp/helpwww.html. Sequence comparisons were performed using BLAST algorithms accessed through the National Centre for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/BLAST/).

4. Reverse Transcription PCR (RT-PCR)

Reverse transcription PCR (RT-PCR) was performed to identify in which cotton tissues the HMGA-like and *Mutator* transposase-like genes were expressed and if these were expressed differentially.

4.1 Primer Design and Testing

Forward and reverse primers were designed from the sequences of each of the partial HMGA-like and *Mutator* transposase-like cDNAs identified in the yeast-one-hybrid assay (DELANEY 2005) (Appendices 1 & 2). To test the primer pairs, PCR was performed using the yeast vectors containing the partial cDNAs as template. The primers were expected to amplify a 998 bp fragment for the HMGA-like sequence and a 561 bp fragment for the *Mutator* transposase-like sequence. Results showed amplification of the expected fragments indicating that appropriate primer pairs had been designed (Figure 1, lanes 2 and 4). The band in the HMGA-like negative control (lane 3) was believed to be overflow from lane 2 due to the faintness and size of the band. The primer pairs were subsequently used in RT-PCR on various cotton tissues.

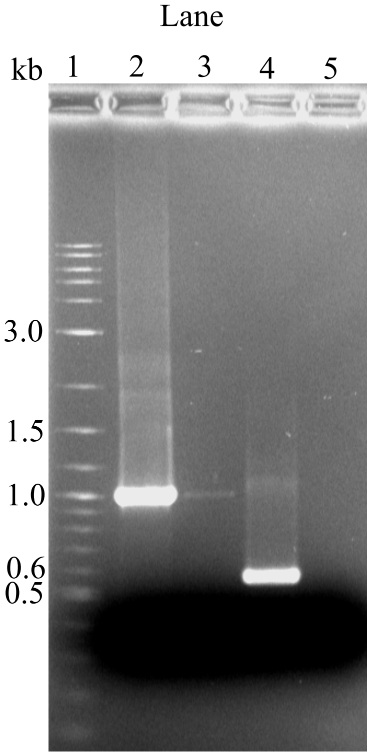


Figure 1: Agarose Gel of PCR on yeast vectors containing the HMGA-like and *Mutator* transposase-like partial cDNA sequences. Lane 1: 2-log DNA size markers in kb. **Lane 2:** Amplification of HMGA-like partial cDNA using primers HMGF1 and HMGR1. **Lane 3:** HMGA-like no template negative control. **Lane 4:** Amplification of *Mutator* transposase-like partial cDNA using primers MutF1 and MutR1. **Lane 5:** *Mutator* transposase-like no template negative control.

4.2 RT-PCR on Cotton Tissue Samples

The first step of an RT-PCR experiment involves reverse transcription using mRNA as template to yield cDNA samples (+RT). This step had already been performed in the Timmis lab with RNA from ovule, 6 DPA fibre, 12 DPA fibre, leaf, flower, stem and root tissues. As a negative control no reverse transcriptase was added to mRNA samples from the same tissues (-RT). These –RT samples were used to test for the presence of genomic DNA contamination that could have occurred during the mRNA extraction step.

4.2.1 The Actin Control

A control was performed to check the amount of cDNA template present in each +RT sample. ActinF and ActinR primers were used to amplify actin transcripts by PCR, with an expected fragment size of 124 bp. Actin is expressed at constant levels in all of the cotton tissues tested. Results showed that actin transcripts were amplified in all samples and the intensity of the bands was similar in each, thus the amount of cDNA template per sample was similar in all tissues tested (Figure 2). This meant that when comparing expression levels between tissue samples in the RT-PCR any major differences would be due to differential expression rather than different amounts of template present in the cDNA.

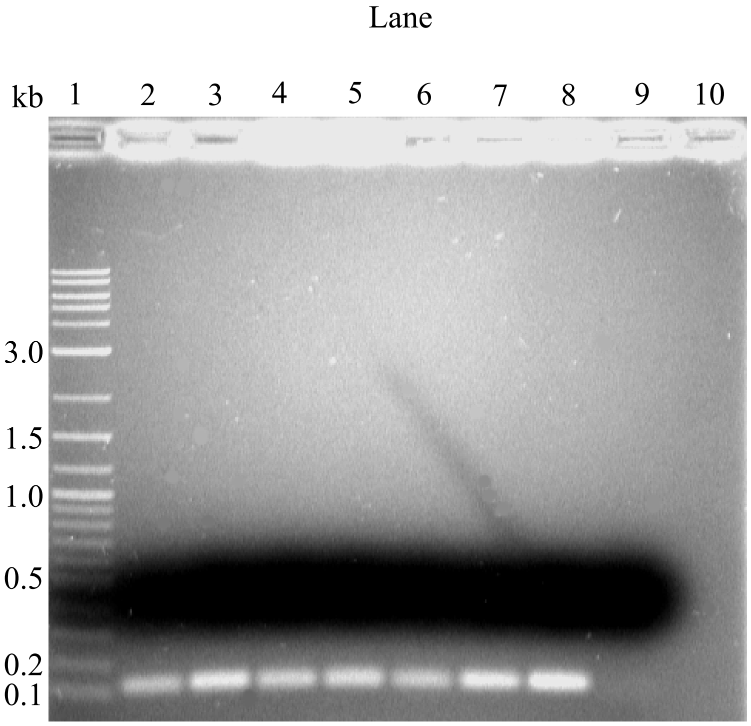


Figure 2: Agarose gel of RT-PCR amplification of actin transcripts from various cotton tissues. Lane 1: 2-log DNA size markers in kb. **Lane 2:** Ovule. **Lane 3:** 6 DPA fibre. **Lane 4:** 12 DPA fibre. **Lane 5:** Leaf. **Lane 6:** Flower. **Lane 7:** Stem. **Lane 8:** Root. **Lane 9:** No template, negative control. **Lane 10:** No sample loaded.

4.2.2 RT-PCR of HMGA-like and *Mutator* transposase-like transcripts in Cotton

PCR was performed twice on each of the +RT and -RT tissue samples, once with the HMGA-like primers (Figures 3 & 4) and once with the *Mutator* transposase-like primers (Figures 5 & 6). Results in Figure 3 showed a fragment of just over 1.0 kb was amplified in all cotton tissues, however this was slightly larger than the expected 1.0 kb fragment amplified from the yeast vector containing the HMGA-like sequence (Figure 1). To determine if the correct transcript was being amplified from the cDNA samples, cloning and sequencing were performed on the bands amplified in the HMGA-like RT-PCR (see section 4.3). Results in Figure 5 showed the amplification of a fragment at approximately 550 bp indicating that the *Mutator* transposase-like mRNA was expressed in all of the cotton tissues tested. The absence of bands in the -RT-PCRs showed that there was no genomic DNA contamination present (Figures 4 and 6).

Faint bands at 1.5 kb in the ovule, stem and root tissue samples and 0.7 kb bands in the ovule and root tissues were amplified in the HMGA-like RT-PCR (lanes 2, 7 and 8, Figure 3), whilst faint bands at 1.0 kb in the ovule, leaf and root lanes were amplified in the *Mutator* transposase-like RT-PCRs (lanes 2, 5 and 8, Figure 5). These bands may result from alternatively spliced mRNAs or represent other genes in the same family as HMGA-like and *Mutator* transposase-like proteins that contain conserved sequences to which the primers were able to bind.

The intensity of the HMGA-like RT-PCR fragments suggests that there is higher expression of the gene in the ovule, 6 DPA, flower, stem and root tissues (lanes 2, 3, 6, 7 and 8, Figure 3) compared to the 12 DPA and leaf tissues (lanes 3 and 4, Figure

3). Similarly, there is likely to be higher expression of the *Mutator* transposase-like gene in ovule, leaf and root tissues (lanes 2, 5 and 8, Figure 5) compared to the 6 DPA, 12 DPA, flower and stem (lanes 3, 4, 6 and 7, Figure 5). These results suggest a potential role for the HMGA-like and *Mutator* transposase-like proteins in early fibre development due to the higher expression identified in the ovule and 6 DPA tissues compared to the 12 DPA tissues. However, these results would need to be confirmed using quantitative real time PCR as reverse transcription PCR is only semi-quantitative.

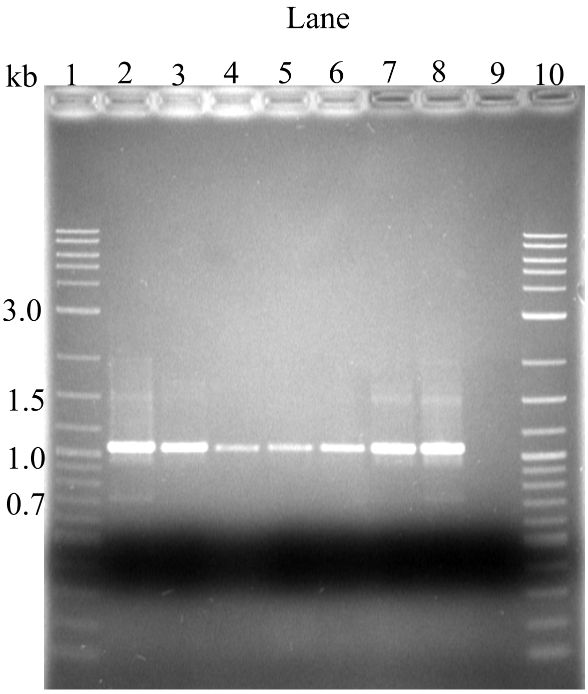


Figure 3: Agarose gel of RT-PCRs directed towards the HMGA-like partial cDNA sequence in various cotton +RT cDNA tissue samples. Lane 1: 2-log DNA size markers in kb. **Lane 2:** Ovule. **Lane 3:** 6 DPA fibre. **Lane 4:** 12 DPA fibre. **Lane 5:** Leaf. **Lane 6:** Flower. **Lane 7:** Stem. **Lane 8:** Root. **Lane 9:** No template, negative control. **Lane 10:** 2-log DNA size markers in kb.

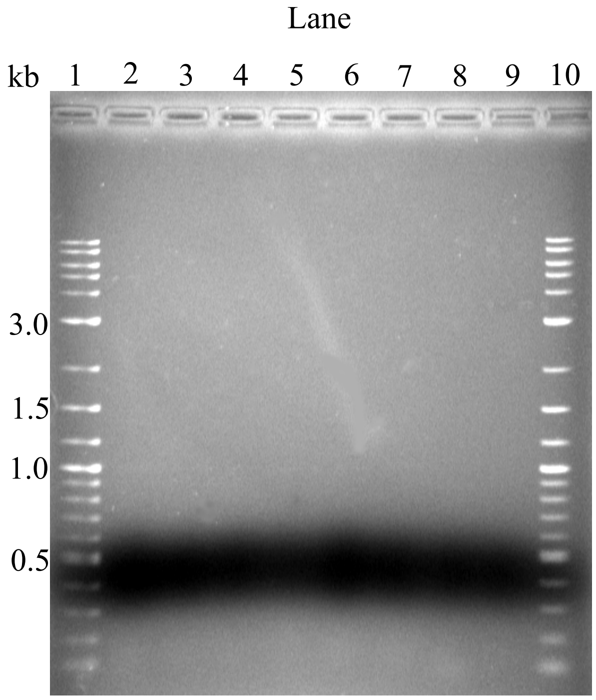


Figure 4: Agarose gel of control RT-PCRs directed towards the HMGA-like partial cDNA sequence in various cotton -RT tissue samples. Lane 1: 2-log DNA size markers in kb. **Lane 2:** Ovule. **Lane 3:** 6 DPA fibre. **Lane 4:** 12 DPA fibre. **Lane 5:** Leaf. **Lane 6:** Flower. **Lane 7:** Stem. **Lane 8:** Root. **Lane 9:** No template, negative control. **Lane 10:** 2-log DNA size markers in kb.

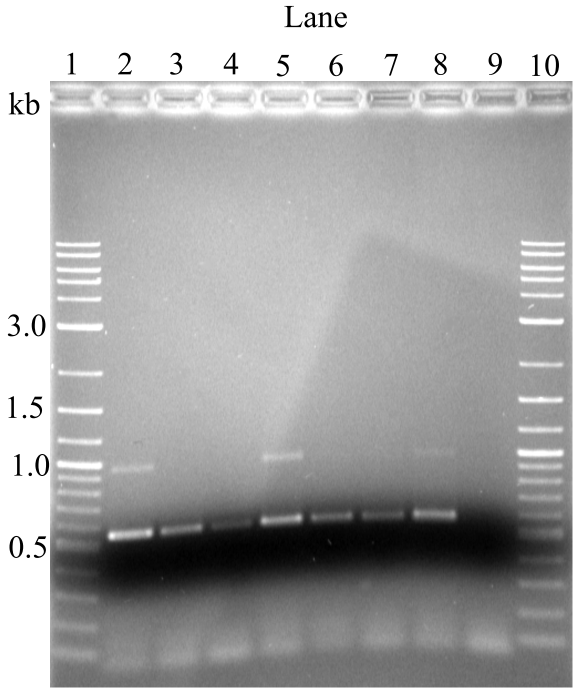


Figure 5: Agarose gel of RT-PCRs directed towards the *Mutator* transposase-like partial cDNA sequence in various cotton +RT cDNA tissue samples. Lane 1: 2-log DNA size markers in kb. **Lane 2:** Ovule. **Lane 3:** 6 DPA fibre. **Lane 4:** 12 DPA fibre. **Lane 5:** Leaf. **Lane 6:** Flower. **Lane 7:** Stem.

Lane 8: Root. **Lane 9:** No template, negative control. **Lane 10:** 2-log DNA size markers in kb.

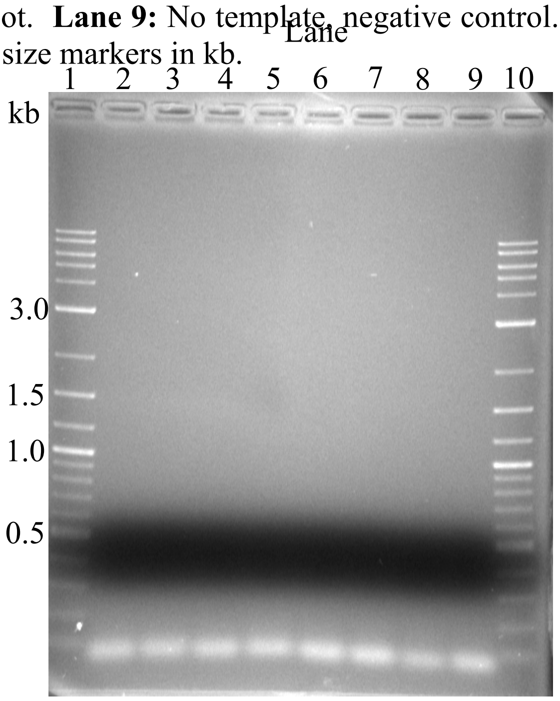


Figure 6: Agarose gel of control RT-PCRs directed towards the *Mutator* transposase-like partial cDNA sequence in various cotton -RT tissue samples. Lane 1: 2-log DNA size markers in kb. **Lane 2:** Ovule. **Lane 3:** 6 DPA fibre. **Lane 4:** 12 DPA fibre. **Lane 5:** Leaf. **Lane 6:** Flower. **Lane 7:** Stem. **Lane 8:** Root. **Lane 9:** No template, negative control. **Lane 10:** 2-log DNA size markers in kb.

4.3 Analysis of HMGA-like RT-PCR Fragments

Each of the bands slightly above 1.0 kb in the HMGA-like RT-PCRs were cut from the gel and DNA extracted. A small amount of each was run on a gel to check that DNA had been retained after gel extraction (Figure 7). Ligations were set up between either the HMGA-like ovule or stem gel extractions and the pGEM-T Easy vector. All ligations were transformed into *E.coli* cells with colonies selected on the basis of blue/white colour selection, with white colonies containing an insert and blue colonies lacking any insert.

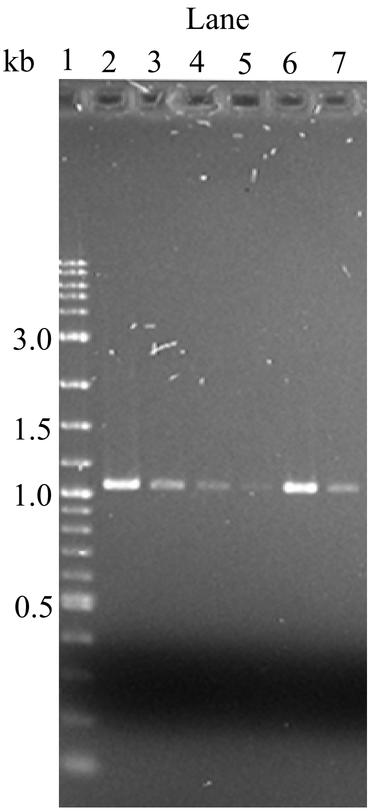


Figure 7: Agarose gel showing a sample of each gel extraction from the HMGA-like RT-PCR bands from various cotton tissues. Lane 1: 2-log DNA size markers in kb. **Lane 2:** Ovule. **Lane 3:** 6 DPA. **Lane 4:** 12 DPA. **Lane 5:** Flower. **Lane 6:** Stem. **Lane 7:** Root.

Colony PCR was performed using M13 primers that are directed towards the sequences flanking the multiple cloning site of pGEM-T Easy. The expected sized fragment for a pGEM-T Easy vector with no insert is 239 bp, while a colony with an insert will have a fragment amplified that is 239 bp larger than the size of the insert. Therefore, fragments were expected at approximately 1.3 kb if a successful ligation had occurred with the HMGA-like RT-PCR insert of approximately 1.1 kb. Colony PCR was performed on five white colonies and four white colonies from the HMGA-like ovule and stem tissue ligation samples respectively (Figure 8). Results showed that all of the colonies, except clone 9 had an insert of around 1.1 kb, indicating a successful ligation. One each of the HMGA-like ovule and stem colonies (clones 3 and 10) were sequenced using the M13 forward and reverse primers.

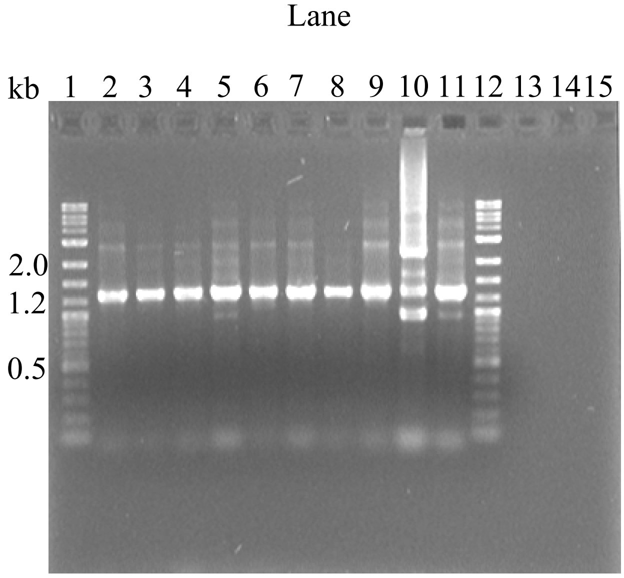


Figure 8: Agarose gel of colony PCRs using M13 primers. Colonies obtained from ligations between gel extractions of HMGA-like ovule or stem RT-PCRs and pGEM-T Easy vectors. Lane 1: 2-log DNA size markers in kb. **Lanes 2-11:** Colonies 1-10 (Colonies 1-6 contain ovule RT-PCR insert, colonies 7-10 contain stem RT-PCR insert). **Lane 12:** 2-log DNA size markers in kb. **Lanes 13-15:** No samples loaded.

Full-length HMGA-like RT-PCR sequences for clones 3 and 10 were obtained by aligning the overlapping forward and reverse sequences in the GeneDoc computer program. The sequence quality of clone 3 was not suitable for analysis, so clone 3 was not used further. A good quality sequence for clone 10 was obtained and aligned to the HMGA-like partial cDNA sequence obtained in the yeast-one-hybrid experiment (DELANEY 2005) (Appendix 3).

Results from the alignment revealed that the HMGA-like stem RT-PCR and the partial cDNA sequences did align, but the RT-PCR sequence has an additional sequence of 54 bp. This explains why the fragments amplified in the RT-PCRs (Figure 3) were slightly larger than the expected 1.0 kb fragment (Figure 1).

5. Isolation of full-length cDNA Sequences

Previous work in the lab identified cDNAs that bound an 84 bp region of the *FSltp6* promoter (required to drive gene expression exclusively in 6 to 14 DPA cotton fibres) by a yeast-one-hybrid assay (DELANEY 2005). These cDNAs could be divided into three classes of proteins based on sequence homology to known plant proteins: *Mutator* transposase-like, HMGA-like and AT-hook proteins. Comparison to known protein sequences and analysis of the cDNA sequences indicated that only partial cDNAs had been isolated. These partial cDNAs were all truncated at the 5' end. The full-length AT-hook cDNA was isolated and characterised but the HMGA-like and *Mutator* transposase-like partial cDNAs were not (DELANEY 2005).

To isolate the full-length HMGA-like and *Mutator* transposase-like cDNA sequences, 5' rapid amplification of cDNA ends (5' RACE) was used. This technique initially involves extraction of mRNA from a tissue of interest followed by reverse transcription to obtain a double stranded cDNA library. Double stranded nucleotide adaptors are then ligated to the 5' ends of all cDNAs present in the library. These initial steps of 5' RACE were previously performed in the Timmis lab, such that an adaptor ligated cDNA library from 12 DPA fibres was available for use.

5.1 Primer Design

Initially, a gene-specific primer (HMG GSP and MUT GSP, section 3.1) and a nested primer (HMG NESTED and MUT NESTED, section 3.1) were designed towards each of the HMGA-like and *Mutator* transposase-like partial cDNA sequences (Appendices 1 & 2). These reverse primers could then be used in conjunction with forward primers (AP1 and AP2, section 3.1) targeted towards the adaptor sequence (Figure 9). Both the gene-specific and nested primers were designed such that a region of overlap would be present between the known partial cDNA sequence and the new cDNA sequence determined from the 5' RACE reaction. This allowed for a comparison between the sequences to check that the desired cDNA was amplified in the 5' RACE reaction instead of another cDNA in the library through non-specific binding of the reverse primers.

Figure 9: 5' RACE primer binding sites. The AP1 and gene-specific primers are used in the initial primary 5' RACE reaction (shown in black and blue respectively). The AP2 and nested primers are used in the nested round of 5' RACE if required (shown in green and pink respectively). AP1 and AP2 are designed to bind to the adaptor sequences present on all 5' ends of cDNAs. Gene-specific and nested primers are designed to bind to specific cDNAs in a cDNA library.

5.2 Primary 5' RACE-PCR Reactions

Primary (1°) RACE involves a PCR reaction using the forward AP1 primer and the reverse gene-specific primer with the adaptor ligated 12 DPA cDNA library as template. 1° RACE was performed using the gene-specific primers HMG GSP and MUT GSP to amplify the remaining 5' region of the HMGA-like and *Mutator* transposase-like cDNAs respectively. Known plant proteins in the HMGA-like and *Mutator* transposase-like classes have been encoded by cDNAs of approximately 2.0

kb and 3.0 kb respectively, whilst the cDNAs used in this project had partial cDNA nucleotide lengths of 1225 bp and 668 bp respectively. Fragments amplified in the 1° RACE would therefore be expected to be around 900 bp for the HMGA-like sequences and 2.5 kb for the *Mutator* transposase-like sequences when the binding sites of the gene-specific primers are considered.

5.2.1 HMGA-like 1° RACE-PCR

Results of the HMGA-like 1° RACE showed multiple fragments upon separation by gel electrophoresis indicating either non-specific binding of the HMG GSP primer to similar sequences and/or specific primer binding to cDNAs of various lengths (Figure 10). To increase the chances of isolating the full-length HMGA-like cDNA sequence a nested RACE was performed (see section 5.3.1).

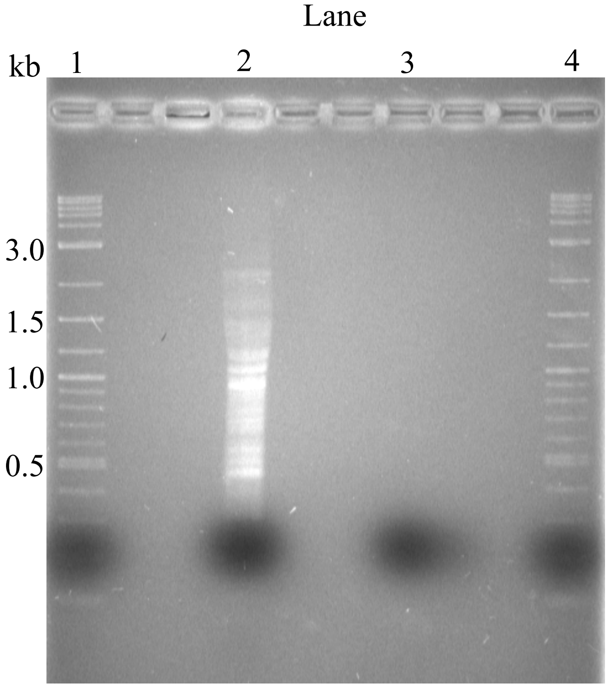


Figure 10: Agarose gel of primary 5' RACE-PCR to isolate the full-length HMGA-like cDNA sequence using the AP1 and HMG GSP primers. Lane 1: 2-log DNA size markers in kb. **Lane 2:** 40µL of HMGA-like primary 5' RACE-PCR product. **Lane 3:** 40µL of no template negative control. **Lane 4:** 2-log DNA size markers in kb. All other lanes contained no loaded samples.

5.2.2 *Mutator* transposase-like 1° RACE-PCR

Gel electrophoresis of the *Mutator* transposase-like 1° RACE reaction showed four prominent bands together with a smear (Figure 11). A sample of the remaining 1° RACE was electrophoresed on another gel and the four prominent bands (2.2 kb, 1.5 kb, 1.2 kb and 1.0 kb) were cut from the gel and DNA extracted.

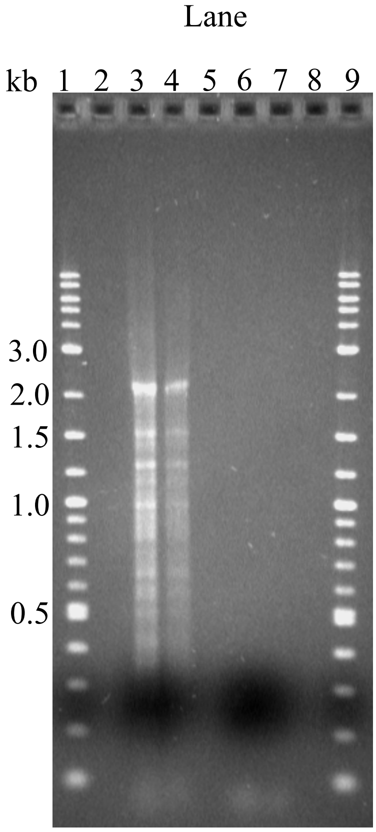


Figure 11: Agarose gel of primary 5' RACE-PCR to isolate the full-length *Mutator* transposase-like cDNA sequence using the AP1 and MUT GSP primers. Lane 1: 2-log DNA size markers in kb. **Lane 2:** No samples loaded. **Lanes 3 & 4:** 15μL of *Mutator* transposase-like 5' RACE-PCR loaded across the two lanes. **Lanes 5:** No sample loaded. **Lanes 6 & 7:** 15μL of no template negative control loaded across the two lanes. **Lane 8:** No sample loaded. **Lane 9:** 2-log DNA size markers in kb.

To increase DNA concentration the four DNA extractions were vacuum centrifuged and a sample of each run on a gel to check that DNA had been retained (Figure 12).

A small concentration of DNA was still present in three of the four samples (lanes 2, 3 and 5). The remaining 2.2 kb (lane 2) and 1.5 kb (lane 5) samples were subsequently used for cloning (see section 5.4.2).

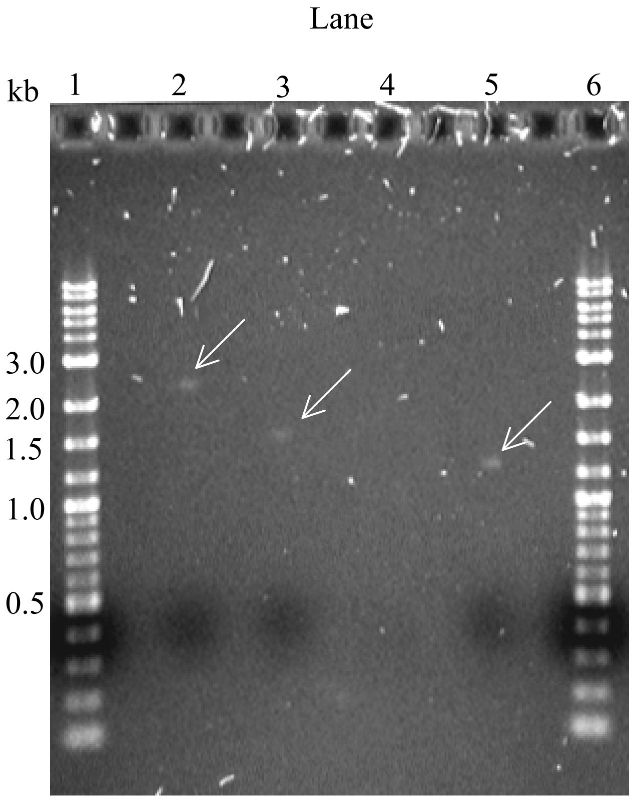


Figure 12: Agarose gel showing samples of each gel extraction of the *Mutator* transposase-like primary RACE-PCR products after vacuum centrifugation. Lane 1: 2-log DNA size markers in kb. **Lane 2:** 2.2 kb 5' RACE product. **Lane 3:** 1.5 kb 5' RACE product. **Lane 4:** 1.0 kb 5' RACE product. **Lane 5:** 1.2 kb 5' RACE product. **Lane 6:** 2-log DNA size markers in kb. All other lanes contained no loaded samples.

5.3 Nested 5' RACE-PCR Reactions

Nested RACE involves dilution of the 1° RACE product followed by a second round of PCR using the forward AP2 primer, the reverse primer and a small amount of the dilution as template. Nested RACE-PCRs are performed to increase specificity of the reaction to a particular cDNA, as it is less likely that cDNAs amplified in the 1°

RACE will also bind non-specifically to a second nested primer. Nested RACE-PCR was performed for the HMGA-like cDNA as no clear fragments were obtained (see section 5.2.1). Nested RACE-PCR was not required for the *Mutator* transposase-like cDNA as four prominent fragments were obtained in the primary RACE-PCR (see section 5.2.2).

5.3.1 HMGA-like Nested RACE-PCR

Results of the HMGA-like nested RACE-PCR showed three different sized fragments had been amplified. These fragments were approximately 1.7 kb, 1.5 kb and 400 bp (Figure 13). This indicated that the numerous fragments amplified in the 1° RACE (Figure 10) were most likely due to non-specific binding of the reverse HMG GSP primer. The remainder of the nested RACE reaction was run out on another gel (Figure 14). Although the fragments did not appear as bright, the 1.7 kb and 1.5 kb bands were cut from the gel and DNA extracted. The 400 bp band was not used further as this would only add 300 bp of extra cDNA sequence, and this would be unlikely to represent a full-length protein (see section 5.2). A sample of the entire reaction was also purified in preparation for cloning.

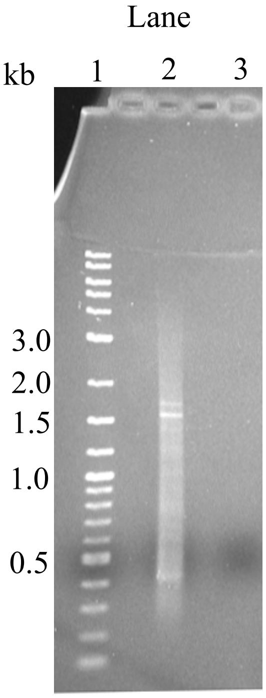


Figure 13: Agarose gel of Nested RACE-PCR using AP2 and HMG NESTED primers. Lane 1: 2-log DNA size markers in kb. **Lane 2:** 15µL of the HMGA-like nested RACE-PCR. **Lane 3:** 15µL of the no template negative control. All other lanes contained no loaded samples.

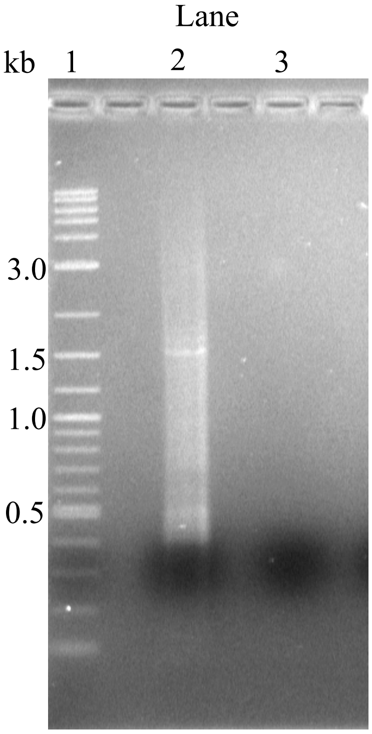


Figure 14: Agarose gel of the remaining HMGA-like nested RACE-PCR. Lane 1: 2-log DNA size markers in kb. **Lane 2:** 35µL of the HMGA-like nested RACE-PCR. **Lane 3:** 35µL of no template negative control. All other lanes contained no loaded samples.

Attempts to clone these RACE products into pGEM-T Easy (see section 5.4) were unsuccessful as no white colonies were obtained. It is possible that only small amounts of the DNA had been retained after the gel extraction. Therefore, only a small amount of DNA would have been present in the ligation reactions thus reducing the efficiency of the ligations. To isolate more cDNA for a new ligation, a new nested RACE-PCR was performed. To ensure enough DNA was present after DNA extraction from a gel, four nested RACE reactions were performed with three and a half reactions electrophoresed in adjacent lanes of a gel (data not shown). The two higher bands were cut from the gel in two separate gel slices. DNA extraction followed by concentration by vacuum centrifugation was performed on each of the two samples, with a small amount of each run on a gel to check that cDNA had been retained (Figure 15). The presence of bands showed that extraction was successful (Figure 15) and the remainder of the samples was subsequently used for cloning (see section 5.4.1).

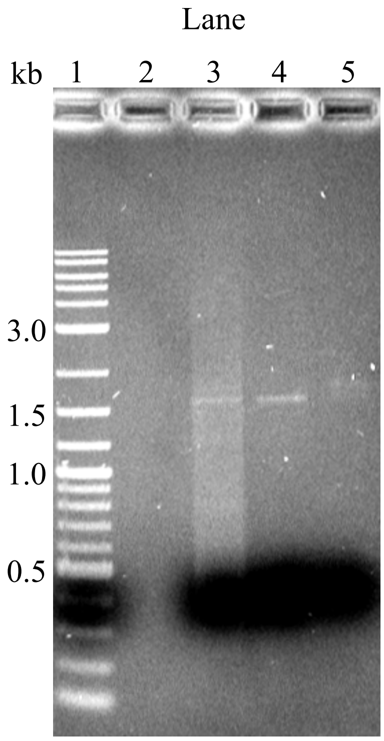


Figure 15: Agarose gel showing a small sample of each HMGA-like nested RACE-PCR after purification or vacuum centrifugation to check retention of cDNA after DNA extraction. Lane 1: 2-log DNA size markers in kb. **Lane 2:** No sample loaded. **Lane 3:** Purified HMGA-like nested RACE-PCR. **Lane 4:** Purified 1.5 kb fragment from nested RACE-PCR. **Lane 5:** Purified 1.7 kb fragment from nested RACE- PCR.

5.4 Cloning of the 5’ RACE-PCR Products

Following amplification and DNA purification of 5’ RACE products for both cDNAs, sequencing of the products was performed. Sequence comparison between the 5’ RACE products and the partial cDNAs, in the region of overlap would determine if the 5’ RACE products isolated were true 5’ extensions of the partial cDNAs. To achieve this the 5’ RACE products were ligated into pGEM-T Easy vectors. Following transformation, colonies were selected on the basis of blue/white colour selection, with white colonies containing an insert and blue colonies lacking any insert. Colony PCR and restriction digests were performed with white colonies to determine whether an insert of the expected size was present.

5.4.1 Cloning of HMGA-like 5' RACE-PCR Products

Ligations were set up using the remaining 1.5 kb and 1.7 kb DNA extractions and the purified nested RACE product (see section 5.3.1) into the pGEM-T Easy vector. Following transformation eight and eleven white colonies were obtained from the 1.7 kb and purified nested RACE transformations respectively. No transformants were obtained from the 1.5 kb ligation. Colony PCR using M13 primers was performed on the seventeen white colonies along with two blue colonies, 12 and 21 (Gel A, lane 13 and Gel B, lane 10) to act as no insert controls (Figure 16). A fragment of around 1.9 kb would indicate an insert into pGEM-T Easy of around 1.7 kb, whilst any sized fragment could be expected from the purified nested RACE colony PCRs. Colonies 7 (Gel A, lane 8) and 9 (Gel A, lane 10) were picked from the purified nested RACE transformants to be analysed further as they had strong amplifications of larger sized fragments compared to the other colonies. The colony PCR of clone 17 (Gel B, lane 6) appeared to have a fragment around 1.9 kb which would indicate a correctly sized 1.7 kb insert. Although a smaller sized fragment was amplified for clone 16 (Gel A, lane 5), this along with clone 17 were chosen to be analysed further.

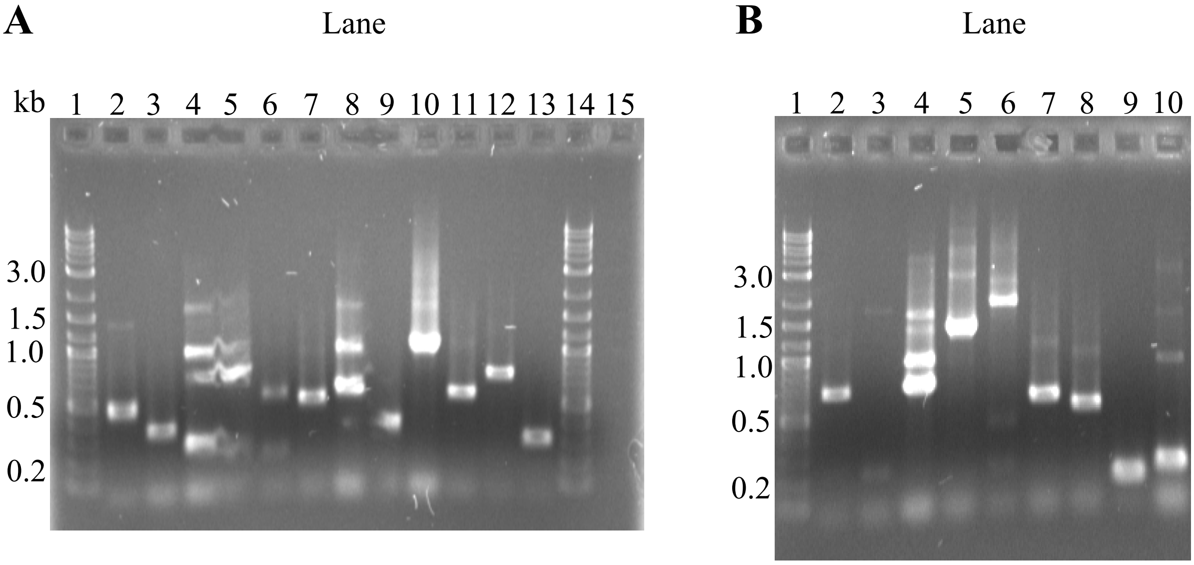


Figure 16: Agarose gels of colony PCRs using M13 primers. Colonies obtained from ligations between gel extractions of HMGA-like nested RACE products and pGEM-T Easy vectors.

- **A.** **Lane 1:** 2-log DNA size markers in kb. **Lanes 2-12:** Colonies 1-11 (obtained with the purified nested RACE product). **Lane 13:** No insert control Colony 12. **Lane 14:** 2-log DNA size markers in kb. **Lane 15:** No sample loaded.
- **B.** **Lane 1:** 2-log DNA size markers in kb. **Lane 2-9:** Colonies 13-20 (obtained with 1.7 kb nested RACE product). **Lane 10:** No insert control Colony 21.

EcoRI restriction digests were performed on colonies 7, 9, 16 and 17 along with the two blue colonies 12 and 21. Results of the digests showed that all colonies contained the 3.0 kb vector fragment (as an *EcoRI* restriciton site is present on either side of the pGEM-T Easy insertion site) and either no insert (e.g. control colonies 12 and 21) or an insert band (Figure 17). Colonies 7, 9, 16 and 17 had inserts of approximately 350 bp, 700 bp, 1100 bp and 1650 bp respectively (Gel A lanes 7, 9 and 11, Gel B lane 2). The insert in clone 7 would only add a small amount of new sequence to the partial HMGA-like cDNA so was not considered further. Clone 16 did not contain the expected 1.7 kb insert so was also not considered further. If the insert in clone 9 proved to be a true extension of the partial HMGA-like cDNA it would extend the 5’ sequence by around 650 bp. Although it was hard to determine if the insert seen in the restriction digest (Figure 17, lane 9) corresponded to the same sized insert in the colony PCR (Figure 16, lane 10) (as the colony PCR was not run far enough to separate the bands sufficiently) the colony was still considered for sequencing. The digest of clone 17 showed two insert fragments, one of around 1400 bp and a faint band of around 250 bp. This is equivalent to an expected sized insert of approximately 1.7 kb containing an internal *EcoRI* site, therefore this clone was marked for sequencing. HMGA-like clones 9 and 17 were sequenced in the forward and reverse directions using M13 primers (see section 5.5).

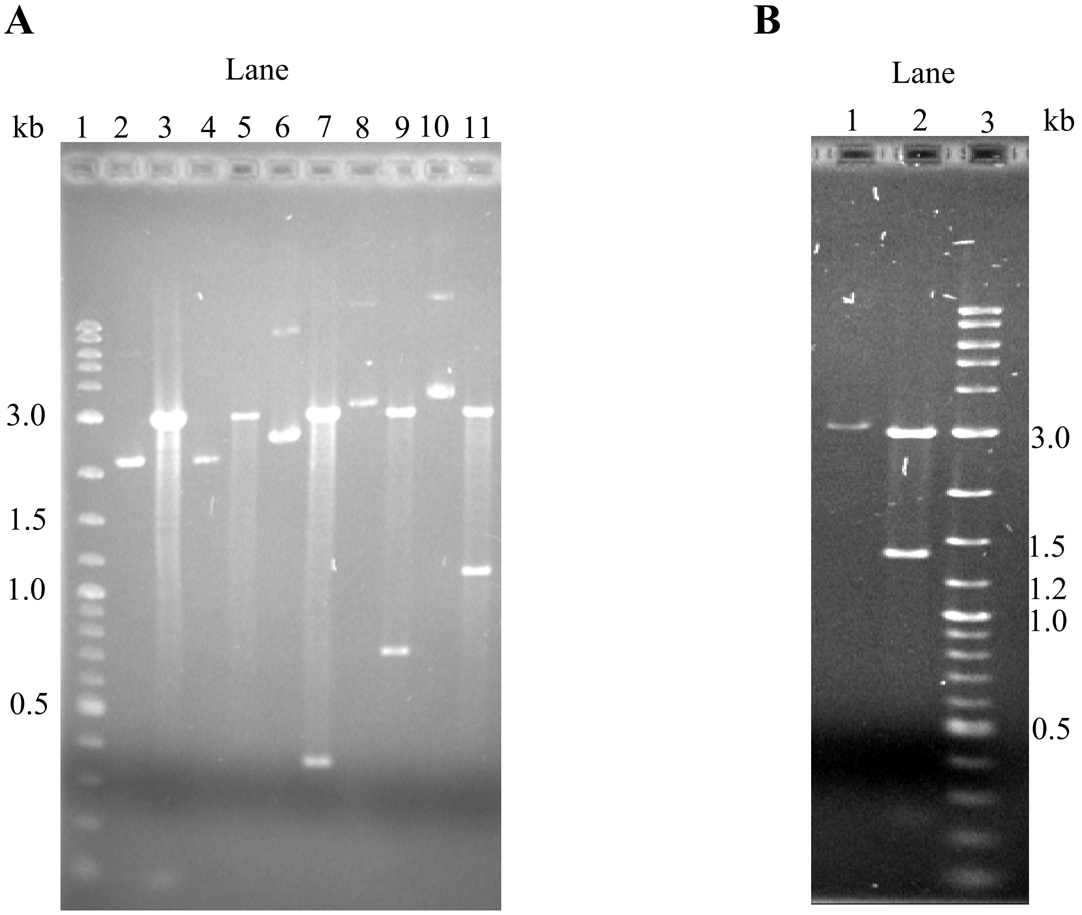


Figure 17: Agarose Gels of *EcoRI* restriction digests on colonies obtained from ligations between HMGA-like nested RACE products and pGEM-T Easy vectors.

A. **Lane 1:** 2-log DNA size markers in kb. **Lane 2:** Uncut colony 12B. **Lane 3:** Digested colony 12B. **Lane 4:** Uncut colony 21B. **Lane 5:** Digested colony 21B. **Lane 6:** Uncut colony 7. **Lane 7:** Digested colony 7. **Lane 8:** Uncut colony 9. **Lane 9:** Digested colony 9. **Lane 10:** Uncut colony 16. **Lane 11:** Digested colony 16.
B. **Lane 1:** Uncut colony 17. **Lane 2:** Digested colony 17. **Lane 3:** 2-log DNA size markers in kb.

5.4.2 Cloning of *Mutator* transposase-like 5' RACE-PCR Products

Ligations were set up using the remaining 2.0 kb and 1.5 kb gel extractions (see section 5.2.2). The transformation yielded one and seven white colonies from the 2.0 kb and 1.5 kb ligations respectively. *EcoRI* restriction digests were performed on all eight colonies to analyse the size of the inserts present (Figure 18). Results in the digested lanes showed that clones 2, 5, 6 and 9 (lanes 5, 18, 20 and 24) may have contained the expected 1.5 kb insert if an internal *EcoRI* restriction site was present. These lanes all contained a 3.0 kb vector band and bands at 1.5 kb, 900 bp and 700 bp. Results of the restriction digest on a positive control (Figure 18, lane 14), containing a pGEM-T Easy vector with an approximate 1.0 kb insert, showed that only a partial digestion had occurred as a linearised 4.0 kb band was present. Therefore clones 2, 5, 6 and 9 showed a pattern consistent with a pGEM-T Easy vector containing a 1.5 kb band that was only partially digested. A modified restriction digest was performed using less DNA, more *EcoRI* enzyme and BSA (Figure 19). The positive control (Figure 19, lane 7) showed a complete digestion whilst the other clones did contain an approximate 1.5 kb insert with an internal *EcoRI* restriction site resulting in bands at 900 bp and 700 bp. Clones 2, 5, 6 and 9 were all sequenced in the forward and reverse directions using M13 primers (see section 5.5).

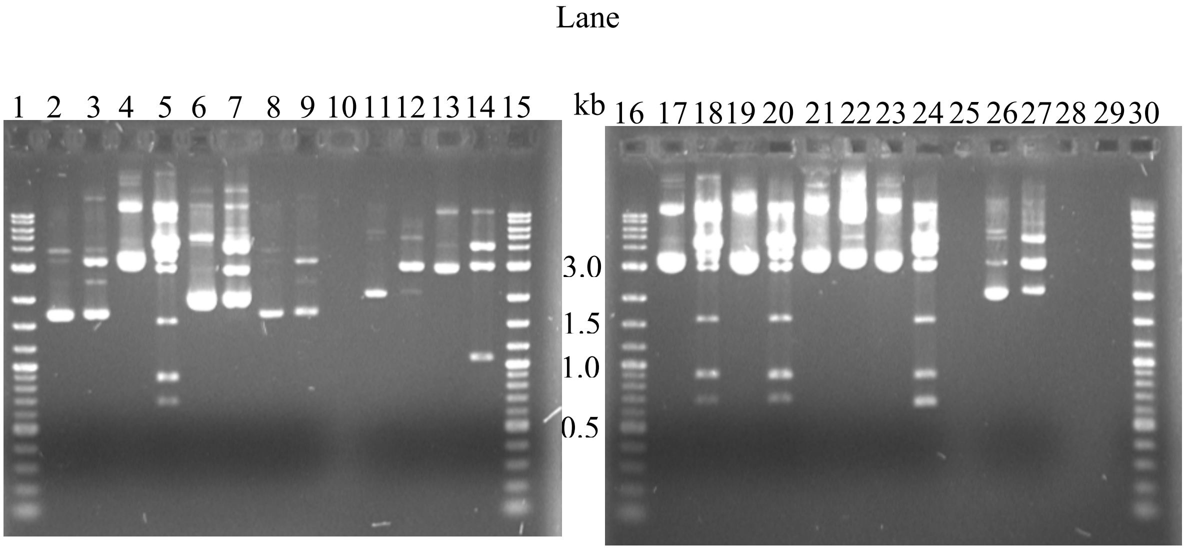


Figure 18: Agarose Gels of *EcoRI* restriction digests on colonies obtained from ligations between *Mutator* transposase-like nested RACE products and pGEM-T Easy vectors. Lane 1: 2-log DNA size markers in kb. **Lane 2:** Uncut clone 1. **3:** Digested clone 1. **Lane 4:** Uncut clone 2. **Lane 5:** Digested clone 2. **Lane 6:** Uncut clone 3. **Lane 7:** Digested clone 3. **Lane 8:** Uncut clone 4. **Lane 9:** Digested clone 4. **Lane 10:** No sample loaded. **Lane 11:** Uncut clone 10. **Lane 12:** Digested clone 10. **Lane 13:** Uncut control clone with 1.0 kb insert. **Lane 14:** Digested control clone with 1.0 kb insert. **Lanes 15 & 16:** 2-log DNA size markers in kb. **Lane 17:** Uncut clone 5. **Lane 18:** Digested clone 5. **Lane 19:** Uncut clone 6. **Lane 20:** Digested clone 6. **Lane 21:** Uncut clone 7. **Lane 22:** Digested clone 7. **Lane 23:** Uncut clone 9. **Lane 24:** Digested clone 9. **Lane 25:** No sample loaded. **Lane 26:** Uncut clone 11. **Lane 27:** Digested clone 11. **Lanes 28 & 29:** No samples loaded. **Lane 30:** 2-log DNA size markers in kb. Clone 1 obtained from 2.0 kb primary RACE product. All other colonies obtained from 1.5 kb primary RACE product.

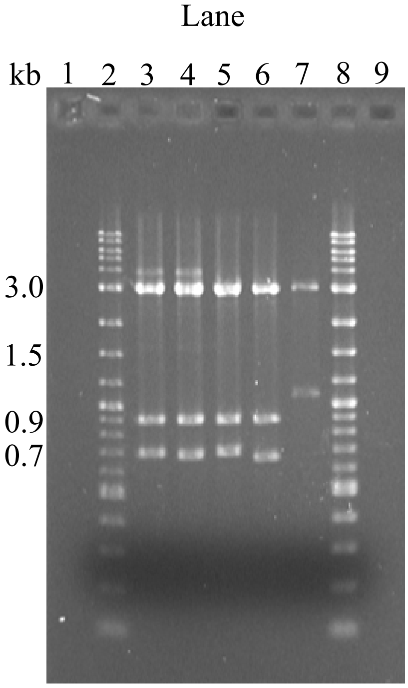


Figure 19: Agarose Gel of modified *EcoRI* restriction digests on colonies obtained from ligation between *Mutator* transposase-like nested RACE products and pGEM-T Easy vectors. Lane 1: No samples loaded. **Lane 2:** 2-log DNA size markers in kb. **Lane 3:** Digested clone 2. **Lane 4:** Digested clone 5. **Lane 5:** Digested clone 6. **Lane 6:** Digested clone 9. **Lane 7:** Digested control clone containing 1.0 kb insert. **Lane 8:** 2-log DNA size markers in kb. **Lane 9:** No sample loaded.

5.5 Sequence Analysis of the 5' RACE-PCR Products

The M13 primers were used to sequence inserts in clones 9 and 17 from the HMGA-like 5' RACE-PCR (section 5.4.1) and clones 2, 5, 6 and 9 from the *Mutator* transposase-like RACE-PCR (section 5.4.2). A full-length insert sequence for each of the clones was obtained by aligning the overlapping forward and reverse sequences for each. Each of the HMGA-like and *Mutator* transposase-like sequences were then aligned to the respective partial cDNA sequences to determine if the clones matched the partial cDNAs in the region of overlap (Appendices 4 & 5, Figure C).

5.5.1 Sequence Analysis of Cotton HMGA-like cDNAs

Clones 9 and 17 were aligned with the partial HMGA-like cDNA such that a 60 bp region of overlap was present (Appendix 4, Figure C). Only a single base pair difference in clone 9 compared to the other two sequences was present in this region. However, clones 9 and 17 did not match each other in the region 5’ of the region of overlap. Upon further analysis runs of poly As and poly T’-s were identified in the sequences (Appendix 4, Figures A & B). This suggested that the inserts in the clones may have been composed of two or more cDNAs joined together (chimeric cDNAs). Such a fusion can occur when adaptors are ligated to the 5’ ends of cDNAs during the 5’ RACE procedure. Therefore, a full-length cDNA sequence was not obtained and attention was shifted towards the *Mutator* transposase-like cDNA.

5.5.2 Sequence Analysis of Cotton *Mutator* transposase-like cDNAs

Clones 2, 5, 6 and 9 and the partial *Mutator* transposase-like cDNA were aligned with a 100 bp region of overlap. All bases matched except for one base pair difference in clone 9, one base pair difference in the partial cDNA sequence and one base where clones 6 and 9 matched the partial cDNA sequence whilst clones 2 and 5 matched one another (Appendix 5, Figure C). Upon further analysis of the alignment it appeared that although the four clones were similar for a large portion of the sequence, there appeared to be two different cDNAs present, one represented by clones 2 and 5, the other represented by clones 6 and 9 (e.g. Appendix 5, Figure B, bases 530-570). These two cDNAs may represent alloalleles of the same gene. Alloalleles are a result of the hybridisation and polyploidisation event that occurred between two diploids, resulting in the tetraploid cotton genome.

The sequences 5’ of base 505 in clone 2 and 5’ of base 385 in clone 5 did not align with the inserts of clones 6, 9 or one another (Appendix 5, Figure B). Upon further analysis, runs of poly T’s were identified near the 5’ end of clone 2 and 5 suggesting that these two clones may also be chimeric and composed of two different cDNAs (Appendix 5, Figure A). So, although one of the cDNAs making up each chimeric cDNA appeared to be an extension of the *Mutator* transposase-like partial cDNA, a full-length sequence was not obtained from these clones so the two sequences were not considered further.

Clones 6 and 9 were extremely similar other than four base pair differences along their sequence lengths. This gave approximately 1.4 kb of extra 5’ cDNA sequence compared to the partial *Mutator* transposase-like cDNA. To determine if this extra sequence gave the full-length *Mutator* transposase-like coding sequence, the full-length sequences were assembled and translated in the six possible reading frames. Results showed that each of the two sequences could be translated into long open reading frames (ORFs), but no start codon was identified (Figures 20 and 21, Frame A). The ORFs were approximately 1.75 kb and 1.4 kb when using clone 6 and 9 added to the partial cDNA sequence respectively. The shorter ORF, using clone 9, was due to a nucleotide difference between it and the other two clones in the region of overlap, encoding a stop codon. The size of these ORFs were much smaller than those reported for other *Mutator* transposase-like cDNAs. Therefore, it was likely that more coding seqeunce 5’ of the sequences obtained in clones 6 and 9 was needed to encode the full-length cDNA. To further attempt to isolate the full-length *Mutator* transposase-like cDNA another round of 5’ RACE was required.

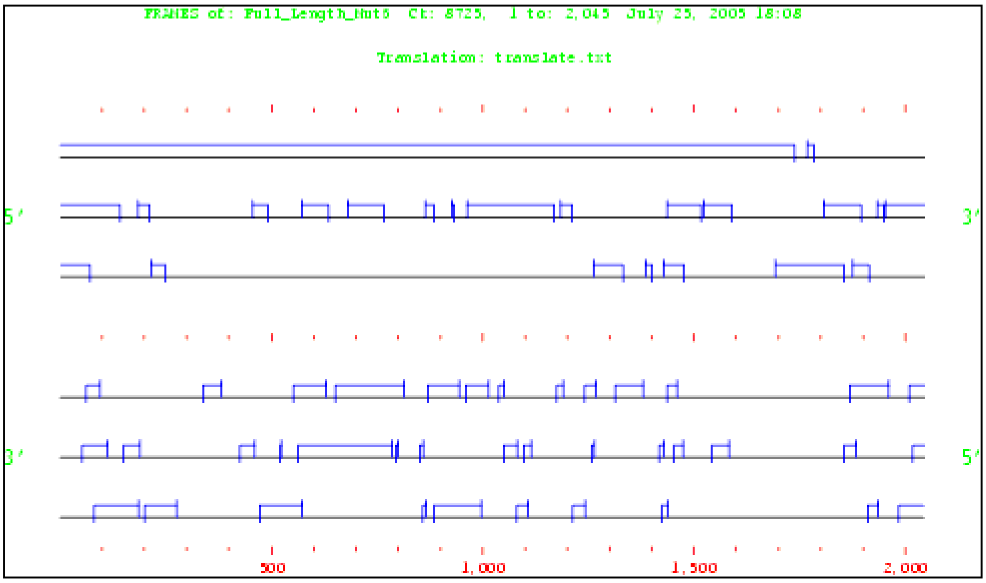


Figure 20: Translation of the potential full-length *Mutator* transposase-like cDNA sequence in all six possible open reading frames. Imput sequence consisted of the 5' RACE product encoded by clone 6 virtually added to the known partial cDNA sequence. Horizontal blue lines represent open reading frames, while vertical blue lines represent start and stop codons.

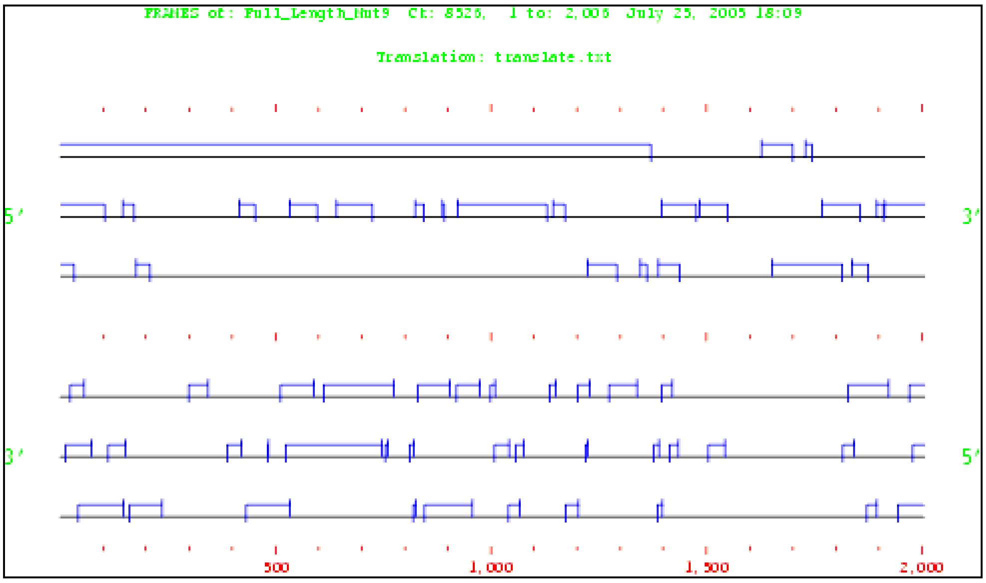


Figure 21: Translation of the potential full-length *Mutator* transposase-like cDNA sequence in all six possible open reading frames. Imput sequence consisted of the 5' RACE product encoded by clone 9 virtually added to the known partial cDNA sequence. Horizontal blue lines represent open reading frames, while vertical blue lines represent start and stop codons.

5.6 *Mutator* transposase-like (Round II)

5.6.1 Primary 5’ RACE-PCRs

To undergo another round of 5’ RACE on the *Mutator* transposase-like cDNA, new gene-specific and nested reverse primers (Mut GSP 2 and Mut Nested 2, section 3.1) were required. These were designed to bind near the 5’ end of a region common to clones 6 and 9, and to allow for a region of overlap to check that any products amplified were specific for this cDNA (Appendix 6).

Results from round II of 1° RACE showed bands amplified at approximately 1.5 kb, 1.0 kb and 300 bp (Figure 22, lanes 1 & 2), however bands at 1.5 kb and 300 bp were also amplified in the negative control lane (Figure 22, lanes 6 & 7). In light of this contamination only the 1.0 kb band was cut from the gel, DNA extracted and concentrated by vacuum centrifugation. A sample of this was run on a gel to check that DNA had been retained (Figure 23). A faint band could be seen so the remaining sample was concentrated further and used in a ligation reaction.

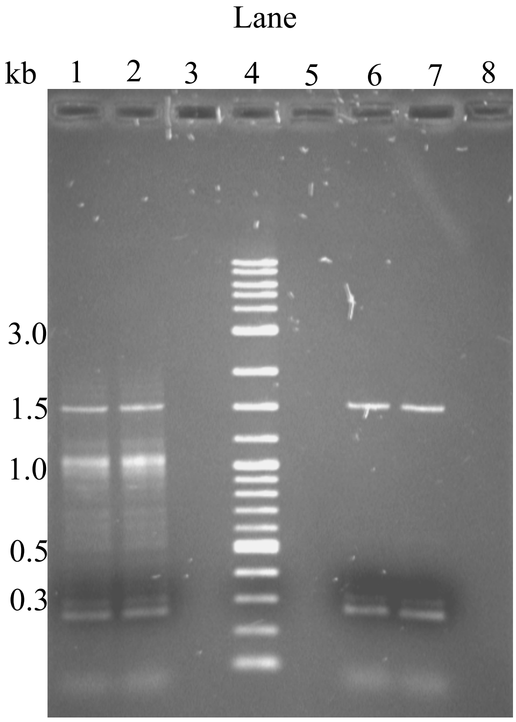


Figure 22: Agarose gel of primary 5' RACE reaction round II using the AP1 and Mut GSP 2 primers. Lane 1 & 2: 30μL of *Mutator* transposase-like 5' RACE reaction loaded across the two lanes. **Lane 3:** No sample loaded. **Lane 4:** 2-log DNA size markers in kb. **Lane 5:** No sample loaded. **Lanes 6 & 7:** 30μL of no template negative control loaded across the two lanes. **Lane 8:** No sample loaded.

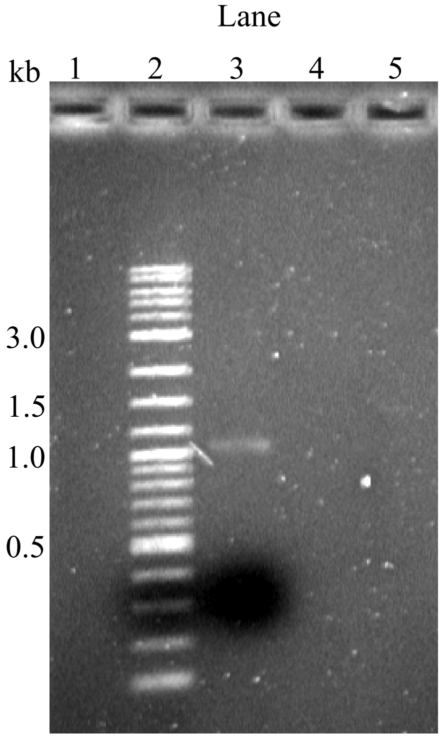


Figure 23: Agarose gel showing a sample of the *Mutator* transposase-like primary 5' RACE round II product after vacuum centrifugation. Lane 1: No sample loaded. **Lane 2:** 2-log DNA size markers in kb. **Lane 3:** 1.0 kb 5' RACE product. **Lane 4 & 5:** No samples loaded.

5.6.2 Cloning of the 5' RACE-PCR Products

Ligations and transformations were performed as in section 5.4 with four white colonies and two colonies of uncertain colour obtained. These colonies were tested for correctly sized inserts by colony PCR and *EcoRI* restriction digestion as performed previously. A fragment of around 1.25 kb was expected for a correctly sized insert in the colony PCRs. Results showed that clones 2, 4, 6 and 7 may contain the correct sized insert (Figure 24, lanes 2, 4, 9 and 10). A culture could not be grown for clone 2 so *EcoRI* restriction digests were performed on clones 4, 6, 7 and blue clones 5, 8 and 9 as controls. Results showed that the blue clones contained the expected 3.0 kb insert (Figure 25, lanes 7, 8 and 9). The other clones may have contained the correct sized insert, but only one would most likely be the correct 5' cDNA extension as a different restriction pattern was observed for each of the three clones (Figure 25, lanes 1, 2 and 3). Clone 4 contained two fragments, one at 600 bp and the other at 500 bp indicating an insert of around 1.1 kb with an internal *EcoRI* restriction site (lane 1). Clone 6 contained an expected sized fragment of approximately 1.1 kb (lane 2). Clone 7 contained a fragment at 950 bp and it is possible another fragment between 50-100 bp was present, but could not be seen due to its low intensity (lane 3). This would also give an insert of around 1.0 kb with an internal *EcoRI* restriction site. Consequently clones 4, 6 and 7 were all sequenced in the forward and reverse directions using M13 primers.

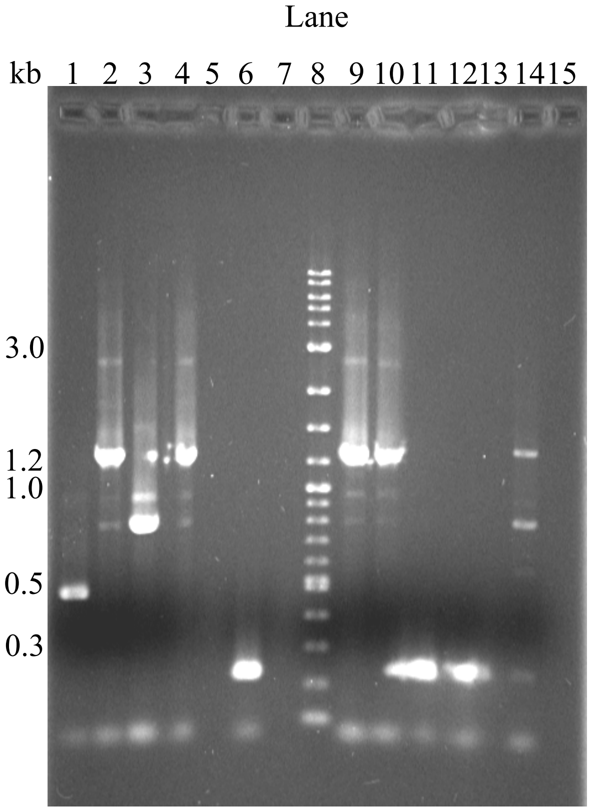


Figure 24: Agarose gel of colony PCRs using M13 primers. Colonies obtained from ligations between 1.0 kb gel extraction of *Mutator* transposase-like primary 5' RACE product (round II) and pGEM-T Easy vectors. Lanes 1-4: Clones 1-4. **Lane 5:** No sample loaded. **Lane 6:** Clone 5. **Lane 7:** No sample loaded. **Lane 8:** 2-log DNA size markers. **Lanes 9-12:** Clones 6-9. **Lane 13:** No sample loaded. **Lane 14:** Negative control containing no bacterial cells. **Lane 15:** No sample loaded.

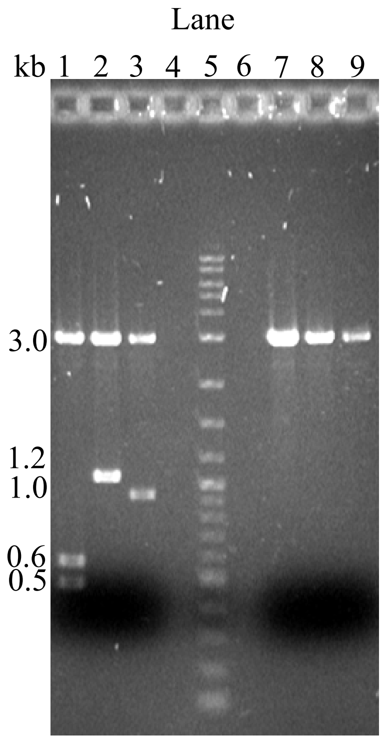


Figure 25: Agarose Gel of *EcoRI* restriction digests on clones obtained from ligations between 1.0 kb gel extraction of *Mutator* transposase-like primary 5' RACE product (round II) and pGEM-T Easy vectors **Lane 1-3:** Digested clones 4, 6 and 7. **Lane 4:** No sample loaded. **Lane 5:** 2-log DNA size markers in kb. **Lane 6:** No sample loaded. **Lanes 7-9:** Digested clones 5, 8 and 9.

5.6.3 Sequence Analysis of Cotton *Mutator* transposase-like cDNAs

A full-length insert sequence for each of the clones was obtained as previoiusly described (see section 5.5). Attempts to align the full-length inserts of clones 6 and 7 to the inserts of clones 6 and 9, isolated during the first round of 5' primary RACE-PCR, failed as no reverse Mut GSP 2 primer binding site could be identified. Upon further analysis, it was observed that the 3' end of the insert in clones 6 and 7 consisted of a sequence complimentary to the adaptor primer (data not shown). Therefore, during this second round of 1° RACE, two adaptor primers were able to amplify cDNA sequences that did not correspond to the *Mutator* transposase-like cDNAs. The insert of clone 4 however, aligned perfectly with the inserts of clones 6 and 9 in the approximately 250 bp of overlap between the sequences (Appendix 7). The size of clone 4 added approximately another 700 bp of sequence to the partial *Mutator* transposase-like sequence giving the full-length assembled cDNA an approximate total length of 2.7 kb. This was more consistent with other known *Mutator* transposase-like proteins that had reported cDNA lengths of around 3.0 kb.

5.7 Sequence Analysis of a full-length *Mutator* transposase-like cDNA from cotton

The 2.7 kb sequence was translated into the six possible reading frames, which showed that a full-length ORF was present in frame A of the translations and that a potential start codon had been identified (Figure 26). A 771 amino acid peptide sequence was encoded by the ORF (Appendix 8). No known nuclear localisation signal was identified from the amino acid sequence (see section 3.10).

The 2.7 kb sequence was analysed using Blastx to determine if the sequence still encoded a protein with homology to other plant *Mutator* transposases. The top four results from the program corresponded to *Arabidopsis* unknown and *Arabidopsis* *Mutator* transposase-like proteins with 50%-60% amino acid homology to the full-length coding sequence. The majority of the other results corresponded to *Mutator* transposase-like proteins from *Arabidopsis* and rice, but homology was only present to part of the coding region.

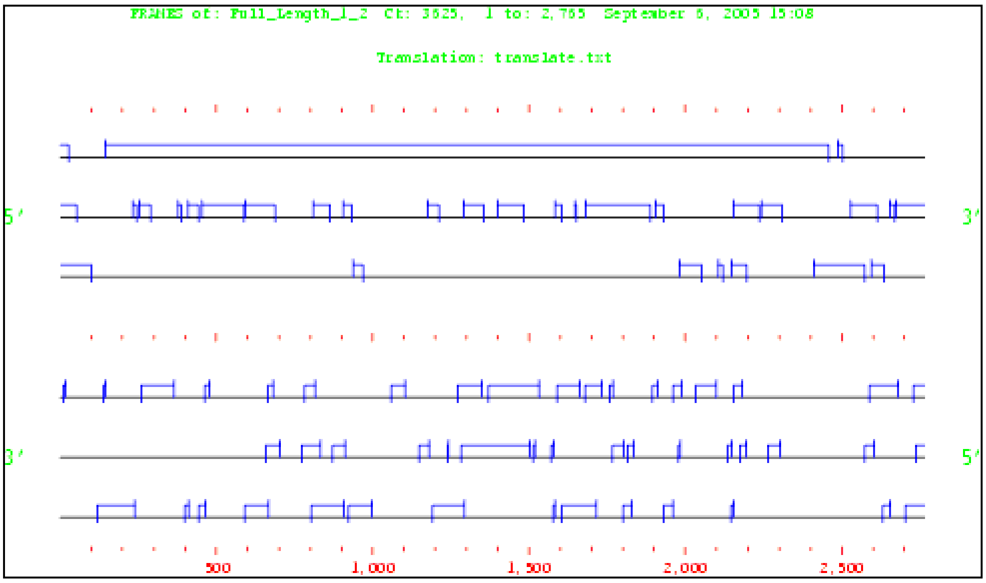


Figure 26: Translation of the potential full-length *Mutator* transposase-like cDNA sequence in all six possible open reading frames. Imput sequence consisted of the round II 5' RACE product, encoded by clone 4, virtually added to the known partial cDNA sequence (original partial cDNA plus round I 5' RACE product encoded by clone 6). Horizontal blue lines represent open reading frames, while vertical blue lines represent start and stop codons.

Now that a potential full-length coding sequence for the *Mutator* transposase-like cDNA had been isolated, its function as a putative transcription factor could be tested further. This would firstly involve isolation of the full-length sequence and cloning into a Gateway vector so that the sub-cellular localisation of the protein could be visualised.

6. Cloning of *Mutator* transposase-like cDNAs into Gateway™ Vectors

To generate a full-length *Mutator* transposase-like sequence that could be cloned into an appropriate Gateway entry vector, an approach was needed to piece together the three regions identified from the initial yeast-one-hybrid assay (Delaney 2005) and the two rounds of primary 5' RACE reactions ('A', 'B' and 'C') (Figure 27).

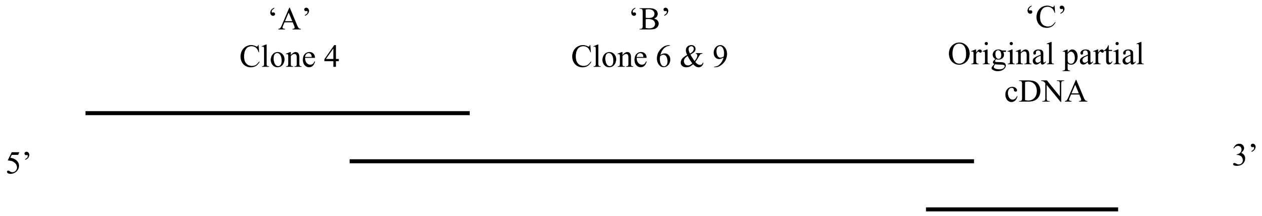


Figure 27: Full-length *Mutator* transposase-like cDNA coding sequence represented by three partial sequences 'A', 'B' and 'C'. Sequence 'A' isolated from primary 5' RACE, round II (see section 5.8). Sequence 'B' isolated from primary 5' RACE, round I (see section 5.5.2). Sequence 'C' isolated from yeast-one-hybrid assay (Delaney 2005).

6.1 Primer Design

Whilst designing primers, consideration was given to how the full-length insert would be placed into the Gateway entry vector (pENTR1a). pENTR1a contains a multiple cloning site, so restriction digestion would be used to ligate the full-length sequence into pENTR1a. Therefore, primers designed to bind to either the 5' end of sequence 'A' or the 3' end of sequence 'C' would also need to contain appropriate restriction sites (restriction sites that did not cut the full-length sequence) to enable this to occur.

To allow for directional ligation of the full-length sequence into pENTR1a, different restriction sites would need to be present on either end of the full-length sequence. Consequently *SalI, NotI* and *XhoI* restriction sites were incorporated into primers MUTFLF, MUTFLRXhoI and MUTFLRNotI respectively (see section 3.1). The full-length sequence also needed to be inserted into pENTR1a in a particular frame so that the correct open reading frame of the sequence was maintained. To facilitate this MUTFLF was designed so that the *SalI* restriction site was immediately followed by the sequence complimentary to the start codon of the *Mutator* transposase-like coding sequence.

A variety of primers were designed so that combinations of the ‘A’, ‘B’ or ‘C’ sequences could be amplified from the 6 DPA cDNA. Reverse primers designed for the first and second rounds of 5’ RACE could also be used. Figure 28 illustrates the binding sites of all the primers that could be used (see also section 3.1).

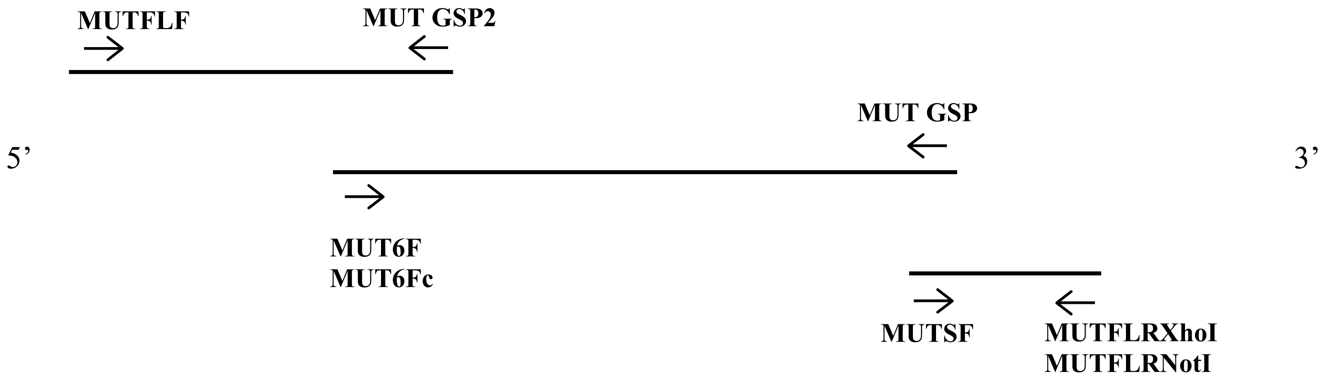


Figure 28: Binding site locations of primers designed to amplify a partial or full-length *Mutator* transposase-like coding sequence. Arrows indicate the 5’ to 3’ orientation of each primer.

6.2 Amplification of full-length *Mutator* transposase-like Coding Sequence

Attempts to amplify the full-length sequence from 6 DPA cDNA using primers MUTFLF with either MUTFLRXhoI or MUTFLRNotI were unsuccessful (Figure 29, lanes 2-3 and 5-6). A fragment of 2340 bp was expected. However, attempts to amplify the 'BC' fragment using primers Mut6 and either MUTFLRXhoI or MUTFLRNotI were both successful (Figure 29, lanes 8-9 and 11-12). The expected size fragment for 'BC' using both sets of primers was 1750 bp. Both of the 'BC' fragments were cut from the gel and DNA extracted.

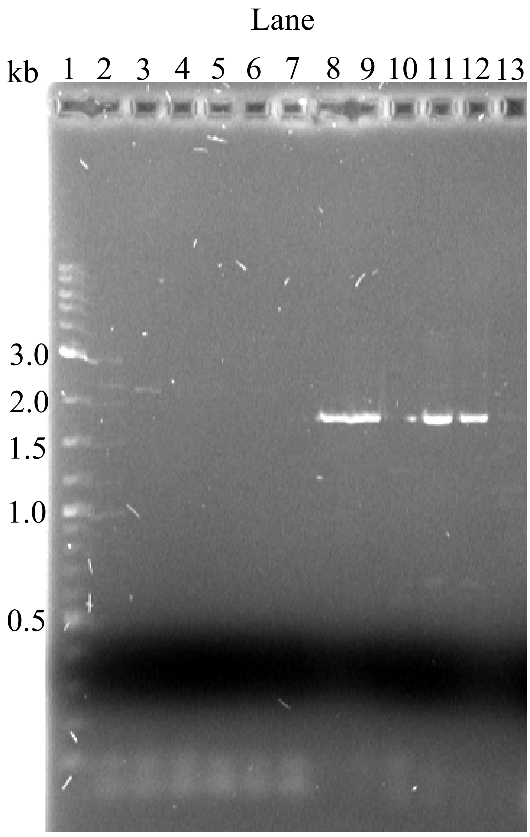


Figure 29: Agarose gel showing amplification of different regions of the *Mutator* transposase-like coding sequence by PCR from 6 DPA cDNA. **Lane 1:** 2-log DNA size markers in kb. **Lanes 2 & 3:** Full-length *Mutator* transposase-like coding sequence using MUTFLF and MUTFLRXhoI primers. **Lane 4:** No template negative control using MUTFLF and MUTFLRXhoI primers. **Lanes 5 & 6:** Full-length *Mutator* transposase-like coding sequence using MUTFLF and MUTFLRNotI primers. **Lane 7:** No template negative control using MUTFLF and MUTFLRNotI primers. **Lanes 8 & 9:** BC region of *Mutator* transposase-like coding sequence using MUT6F and MUTFLRXhoI primers. **Lane 10:** No template negative control using MUT6F and MUTFLRXhoI primers. **Lanes 11 & 12:** BC region of *Mutator* transposase-like coding sequence using MUT6F and MUTFLRNotI primers. **Lane 13:** No template negative control using MUT6F and MUTFLRNotI primers.

The next step involved amplification of the 'AB' fragment from 6 DPA cDNA using primers MUTFLF and MUTGSP. Results showed that this was also successful (Figure 30, lane 1) with the amplification of a fragment at around 2.0 kb (expected size was 2090 bp). The 'AB' fragment was cut from the gel and DNA extracted.

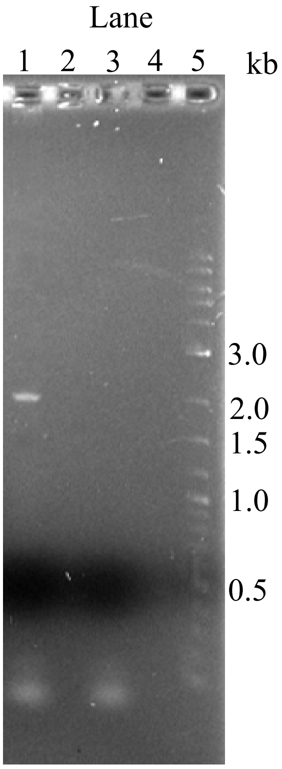


Figure 30: Agarose gel showing amplification of the AB region of the *Mutator* transposase-like coding sequence by PCR from 6 DPA cDNA using MUTFLF and MUTGSP primers. Lane 1: AB region of *Mutator* transposase-like coding sequence. **Lane 2:** No sample loaded. **Lane 3:** No template negative control. **Lane 4:** No sample loaded. **Lane 5:** 2-log DNA size markers in kb.

Since PCR products for the ‘AB’ and ‘BC’ regions had been obtained, a modified PCR could be attempted to obtain the full-length coding sequence. This involved using the ‘AB’ and ‘BC’ PCR products as template, with the MUTFLF and either MUTFLRXhoI or MUTFLRNotI primers, to amplify the full-length coding sequence instead of using cDNA as template. A full-length sequence of 2340 bp was expected (Figure 31). Results showed that a fragment had been amplified corresponding to this size. These bands were cut from the gel and DNA extracted.

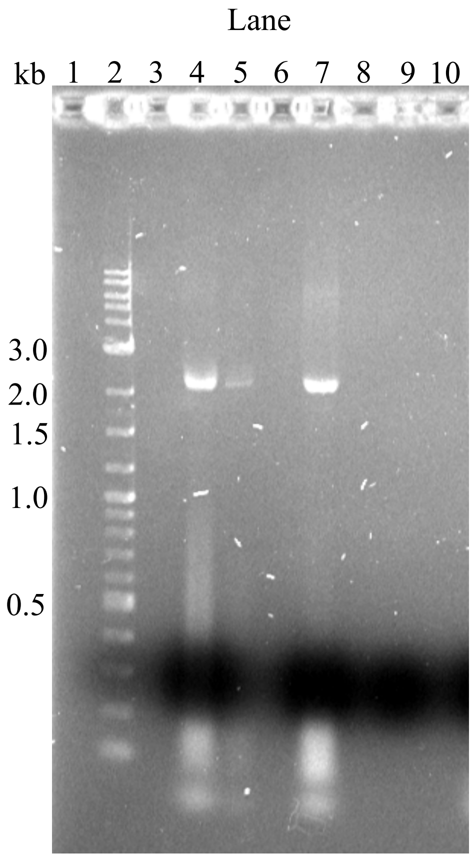


Figure 31: Agarose gel showing amplification of the full-length *Mutator* transposase-like coding sequence by PCR using MUTFLF and either MUTFLRXhoI or MUTFLRNotI primers. Lane 1: No sample loaded. **Lane 2:** 2-log DNA size markers in kb. **Lane 3:** No sample loaded. **Lane 4 & 5:** Full-length *Mutator* transposase-like coding sequence (MUTFLRXhoI). **Lane 6:** No sample loaded. **Lane 7:** Full-length *Mutator* transposase-like coding sequence (MUTFLRNotI). **Lane 8:** No sample loaded. **Lane 9:** No template negative control. **Lane 10:** No sample loaded.

Restriction digests of both the full-length sequence and the pENTR1a entry vector were performed (data not shown) using a *Sal / NotI* double digest. Ligations were performed between the two, but no transformants were obtained. This may have been due to low concentrations of both insert and vector DNA.

7. Discussion

In a previous study, an 84 bp region of the cotton *FSltp6* promoter was identified as the necessary component required to drive expression of the gene specifically in fibre cells (DELANEY 2005). The objective of this project was to further analyse *Mutator* transposase-like and HMGA-like partial cDNAs, identified as binding to the 84 bp region in a yeast-one-hybrid assay, to ascertain if these cDNAs are transcription factors that control the *FSltp6* promoter.

The initial aim of this project was to isolate a full-length sequence for each of the *Mutator* transposase-like and HMGA-like cDNAs using 5' rapid amplification of cDNA ends (5' RACE). Through two rounds of the 5' RACE procedure, a putative full-length *Mutator* transposase-like cDNA was isolated. This cDNA was predicted to encode a 771 amino acid peptide sequence. Due to time constraints attempts to isolate a full-length sequence for the HMGA-like partial cDNA were unsuccessful.

A further aim of this project was to perform functional analysis on the *Mutator* transposase-like and HMGA-like cDNAs. The first part of this analysis involved reverse transcription PCR (RT-PCR) using primers directed towards the partial cDNA sequences obtained in the yeast-one-hybrid assay. Results showed that both mRNAs were present in ovule, 6 DPA fibre, 12 DPA fibre, leaf, flower, stem and root cotton tissues. Both mRNAs appeared to be present at higher levels in ovule and 6 DPA fibres compared to 12 DPA fibres, suggesting a potential role for these proteins in early fibre development. However, this result would need to be confirmed using quantitative real time PCR (see section 4.2.2).

An interesting result from the RT-PCRs was the amplification of a HMGA-like fragment (Figure 3) larger than the expected size determined from the partial cDNA sequence (Figure 1). Upon alignment of the RT-PCR and partial cDNA sequences, an extra 54 bp sequence was identified in the RT-PCR sequence (Appendix 3). As the cDNA used in the RT-PCR was derived from the same cotton cultivar as the cDNA used in the yeast-one-hybrid assay, the additional sequence was not due to sequence differences between different cotton cultivars. As only one fragment was isolated in the RT-PCRs the two sequences were unlikely to be alloalleles of the same gene. One hypothesis for the additional 54 bp in the RT-PCR sequence may be that a partial deletion of the cDNA occurred during the yeast-one-hybrid assay. An AAGAAG sequence is present immediately 5' of the additional sequence and a TTCTTT sequence at the 3' end of the additional sequence (Appendix 3). These bases may have paired to form a loop that was recombined out of the cDNA sequence during the yeast-one-hybrid assay. This may have occurred during the recombination phase between the cDNA library and yeast vectors.

The second part of the functional analysis was to determine the sub-cellular localisation of each of the proteins in onion epidermal cells. Onion cells are used for this purpose as they have large nuclei that can be easily visualised and emit less autofluorescence compared to cotton fibre cells. This required a construct to be made encoding a fusion protein, between green fluorescent protein (GFP) and either the *Mutator* transposase-like or the HMGA-like protein. Microprojectile bombardment of the construct into onion epidermal cells would allow detection of protein localisation through GFP visualisation. To do this, the full-length cDNA sequence for each protein was required, thus only the *Mutator* transposase-like cDNA was investigated. A PCR strategy was used to isolate the full-length *Mutator* transposase-like cDNA that encoded the predicted 771 amino acid peptide sequence. Results showed that an expected 2.3 kb fragment was amplified from 6 DPA fibre cDNA (Figure 31). Due to time constraints, it was not possible to clone this fragment into the required construct, so the sub-cellular localisation of the *Mutator* transposase-like protein could not be determined.

Interestingly, no nuclear localisation signal was identified in the predicted *Mutator* transposase-like protein sequence (see sections 5.7 & 3.10). However, the absence of this signal does not rule out the possibility that the protein is a transcription factor. The protein sequence may contain a nuclear localisation signal not recognised by the computer algorithm used or the protein may need to bind to another protein with a nuclear localisation signal to gain entry into the nucleus. If the latter is true, this needs to be considered when performing the sub-cellular localisation experiments in the onion cells. If the protein that is required to aid *Mutator* transposase-like entry into the nucleus is a cotton specific protein, then localisation of GFP to the nucleus may not occur.

Future work on this project would firstly involve continuing construction of the vector encoding the fusion protein between GFP and the putative *Mutator* transposase-like protein. Following sequencing of the construct to confirm that the correct *Mutator* transposase-like sequence was present, the sub-cellular localisation experiment could be performed. Secondly, the full-length HMGA-like coding sequence should be identified and the same sub-cellular localisation experiment performed. If they showed localisation to the onion nuclei this would provide further information to confirm that these proteins are transcription factors. It would also be interesting to perform the RT-PCR experiments directed to the full-length or a larger region of the *Mutator* transposase-like and HMGA-like cDNA sequences. This would ensure that the results were specific for these cDNAs and not others of the same family with similar sequences.

Longer-term work could involve use of a model system, such as tobacco, to determine if the *Mutator* transposase-like and HMGA-like proteins do interact with the *FSltp6* promoter. It would also be interesting to see if these proteins interact with other fibre-specific promoters that drive expression at the same developmental stage, to see if a conserved mechanism exists for fibre-specific expression. This would involve a comparison of sequence motifs identified in the 84 bp region of the *FSltp6* promoter with other fibre-specific promoters. If similar sequences existed, experiments could be performed to determine if these proteins are able to bind to the other fibre-specific promoters.

Information about cotton fibre-specific promoters and the transcription factors that regulate these promoters, such as the ones investigated in this study, could be used in the development of transgenic cotton with improved fibre quality and yield. For example, various fibre-specific promoters could be available to drive expression of appropriate genes in fibre cells at particular developmental stages, to increase fibre length and/or strength. Alternatively, these promoters could be used to drive expression of appropriate transcription factors, at particular developmental stages, to increase expression of particular transgenes or endogenous genes, again resulting in improved fibre properties.

Use of fibre-specific promoters and their transcription factors in transgenic plants may also be beneficial in increasing fibre yield. To date, transgenic cotton crops have had transgenes expressed constitutively throughout the plant even when expression was only required in a single tissue type, such as fibre cells. The metabolic energy expended by a plant to express a transgene constitutively could instead be more efficiently utilised. For example, if the transgene was driven by a fibre-specific promoter then the energy saved could be used for other purposes, such as for growth and development.

Fibre-specific promoters and their associated transcription factors will play a significant role in cotton biotechnology and the generation of new cotton varieties with improved fibre characteristics. These new varieties have the potential to contribute significant economic benefits to Australian cotton growers and the Australian economy.

8. References

ABDALLA, A. M., O. U. K. REDDY, K. M. EL-ZIK and A. E. PEPPER, 2001 Genetic diversity and relationships of diploid and tetraploid cottons revealed using AFLP. Theoretical and applied genetics **102:** 222-229.

BASRA, A. S., and C. P. MALIK, 1984 Development of the Cotton Fiber. International Review of Cytology **89:** 65-113.

BASRA, A. S., and S. SAHA, 1999 Chapter 2. Growth Regulation of Cotton Fibers in *Cotton Fibers: Developmental Biology, Quality Improvement, and Textile Processing*, edited by A. S. BASRA. Haworth Press Inc.

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 Fibers of Cottons which Differ in their Lint Lengths and Extent of Fuzz. Physiologia plantarum **45:** 77-82.

DELANEY, S., 2005 Molecular Regulation of Fibre-specific Gene Expression in Cotton (*Gossypium hirsutum*), *School of Molecular and Biomedical Sciences, Department of Genetics*. University of Adelaide, Adelaide, South Australia.

EDWARDS, G. A., and J. E. ENDRIZZI, 1975 Cell Size, Nuclear size and DNA content relationships in Gossypium. Canadian journal of Genetics and Cytology **17:** 181-186.

ENDRIZZI, J. E., E. L. TURCOTTE and R. J. KOHEL, 1985 Genetics, Cytology and Evolution of Gossypium. Advances in Genetics **23:** 271-375.

GORDON, D. S., 2001 An Odyssey in Fibres and Space, pp. in *Textile Institute 81st World Conference*, edited by D. S. GORDON. CSIRO Textile and Fibre Technology, Melbourne, Australia.

GRASSER, K. D., 2003 Chromatin-associated HMGA and HMGB proteins: versatile co-regulators of DNA-dependent processes. Plant Molecular Biology **53:** 281-295.

GRAVES, D. A., and J. M. STEWART, 1988 Chronology of the differentiation of cotton (Gossypium hirsutum L.) fiber cells. Planta **175:** 254-258.

HSU, C.-Y., R. G. CREECH, J. N. JENKINS and D.-P. MA, 1999 Analysis of promoter activity of cotton lipid transfer protein gene LTP6 in transgenic tobacco plants. Plant Science **143:** 63-70.

HUMPHRIES, J. A., A. R. WALKER, J. N. TIMMIS and S. J. ORFORD, 2005 Two WD-repeat genes from cotton are functional homologues of the Arabidopsis thaliana TRANSPARENT TESTA GLABRA1 (TTG1) gene. Plant Mol Biol **57:** 67-81.

INOUE, H., H. NOJIMA and H. OKAYAMA, 1990 High Efficiency Transformation of *Escherichia coli* with plasmids. Gene **96:** 23-28.

JIANG, N., Z. BAO, X. ZHANG, S. R. EDDY and S. R. WESSLER, 2004 Pack-MULE transposable elements mediate gene evolution in plants. **431:** 569-573.

JOHN, M. E., and J. M. STEWART, 1992 Genes for Jeans: biotechnological Advances in Cotton. Trends in biotechnology **10:** 165-170.

KIM, H. J., and B. A. TRIPLETT, 2001 Cotton fiber growth in planta and in vitro. Models for plant cell elongation and cell wall biogenesis. Plant Physiol **127:** 1361-1366.

KLOSTERMAN, S. J., and L. A. HADWIGER, 2002 Plant HMG proteins bearing the AT-hook motif. Plant Science **162:** 855-866.

LISCH, D., 2002 Mutator transposons. Trends in Plant Science **7:** 498-504.

LIU, H.-C., R. G. CREECH, J. N. JENKINS and D.-P. MA, 2000 Cloning and promoter analysis of the cotton lipid transfer protein gene Ltp3. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids **1487:** 106-111.

MA, D. P., H. C. LIU, H. TAN, R. G. CREECH, J. N. JENKINS *et al.*, 1997 Cloning and characterization of a cotton lipid transfer protein gene specifically expressed in fiber cells. Biochim Biophys Acta **1344:** 111-114.

MA, D. P., H. TAN, Y. SI, R. G. CREECH and J. N. JENKINS, 1995 Differential expression of a lipid transfer protein gene in cotton fiber. Biochim Biophys Acta **1257:** 81-84.

MEINERT, M. C., and D. P. DELMER, 1977 Changes in Biochemical Composition of the Cell Wall of the Cotton Fiber During Development. Plant Physiology **59:** 1088-1097.

ORFORD, S. J., and J. N. TIMMIS, 1997 Abundant mRNAs specific to the developing cotton fibre. TAG Theoretical and Applied Genetics **94:** 909-918.

PAYNE, T., J. CLEMENT, D. ARNOLD and A. LLOYD, 1999 Heterologous myb genes distinct from GL1 enhance trichome production when overexpressed in Nicotiana tabacum. Development **126:** 671-682.

STEWART, J. M., 1975 Fiber Initiation on the Cotton Ovule (Gossypium Hirsutum). American Journal of Botany **62:** 723-730.

SUNILKUMAR, G., J. P. CONNELL, C. W. SMITH, A. S. REDDY and K. S. RATHORE, 2002 Cotton α-Globulin Promoter: Isolation and Functional Characterization in Transgenic Cotton, *Arabidopsis*, and Tobacco. Transgenic Research **11:** 347-359.

TIWARI, S. C., and T. A. WILKINS, 1995 Cotton (Gossypium hirsutum) seed trichomes expand via diffuse growing mechanism. Canadian journal of botany **73:** 746-757.

WENDEL, J. F., 1989 New World tetraploid cottons contain Old World cytoplasm. Proceedings of the National Academy of Sciences of the United States of America **86:** 4132-4136.

WENDEL, J. F., C. L. BRUBAKER and A. E. PERCIVAL, 1992 Genetic Diversity in Gossypium Hirsutum and the Origin of Upland Cotton. American Journal of Botany **79:** 1291-1310.

WILKINS, T. A., K. RAJASEKARAN and D. M. ANDERSON, 2000 Cotton Biotechnology. Critical Review in Plant Sciences **196:** 511-550.

9. Appendices

Appendix 1: HMGA-like Partial cDNA Sequence

```
   1  GGGCATGGTC AAGAAATCTT ATAAGCTTGC TTCTACTGCT AGATCTGAGG
                 ←---------            ---------→
  51  TTCCTATCCT CGATTCTACT CCTTCCAACC CTCCCGATGT TTCTTCTCCT
      ←---------
 101  CCTGGATTCA AACGAAGCCG TGGTCGTCCT CCTAAACCCA AACCCACAAT
 151  TTCTGCTCCT GCCGACCCTA TTCCTCAGCA GCAGCAGCAG CAACAGCAAC
 201  CGCTGCCTGC TCCGATTCCC GATGCCACTA AAAGGTCTCC CGGTAGGCCC
 251  AGGAAGAATG GTCCCGTTGC TCCACTTGGG GTTAGGAAGG GCCGGGGTCG
 301  TCCACCCAAA ACTGGCCCTA AGAAGAGTCC CGGTCGTCCC AGGAAGCCCA
 351  AAACGGTCAG GTCGGTGGTG GGGGCTAATG CGATGAAAAG GGGTCGTGGG
 401  CGGCCGCCTA AGGTCTTGAA CCAGATGCCT CAACCGGCTG TTATGCCGAT
 451  CCAGGGTCAG CCCATGGCTG TTCCTTATGC TGATACCGCT GCTGCTGTTC
 501  CTACTACTAC TGCCGTTGCT GCTGGTCCGA GGCCTAGAGG AAGGCCAAAA
 551  GGCACTTCCG TGGCTCCTGC TGGCCTTGCG GTCCCTGGAA AGGGGAGAGG
 601  TCGACCACCC AAGAGTAGTG GGGTCGCCGC GAAGCCCATC AAGCCCAAGA
 651  AATATACTGG AAAGCCTGTT GGCCGCCCCA AGAAGCAATC AAAAGTGAAG
 701  CAAGCAGTTG GAGTACTAAA GTCTCAGTTC AGCAGTGAAA GCAACATCAG
 751  TGCAATCGAT GCTATCCAAG AGTTAGAAGT ACTTGCAGCT ATGGACATTA
 801  ACAAACCATT TAAAGATGAT GCTCAGCCAC CGCCCCCACC ACCCCCAGCA
 851  CCAGAACCGA CACAGCCAGC ACCGATGCCC CAAAATATGG AAGGACAAGT
 901  GTATTAATAA CGAAGCCCTA TCTCTCTGTA TGATTATTAA GAAGGCAAGG
 951  CGTGGAAGAA AATGTTCTGA CCCTAGACAT ATGATATTTC TGTAAAGGGA
                                                  ←---------
1001  AAAGTAGGAA GTTGCAGATC ATCTAATTTG TGTTGTTATT GTTGTTGAGG
1051  TGTGTGTTCG TTTGTGAACC CTGTTCATCG ACTTTTAATT ATTCCCCCTC
1101  ATTTAGTGCA ACATTAGTCA AGCATCCTTG TTCGCTTGGT TTTTGCTTGA
1151  GTGTTATTAT TATTATTATG AATTTGATAT TAGTAGTAGT GTATTGGAAA
1201  AAAAAAAAAA AAAAAAAAAA AAAAA
```

→ Binding sites for HMGA-like RT-PCR primers (HMGF1 & HMGR1) highlighted in yellow.

→ Binding site for HMGA-like gene-specific primer (HMG GSP) in blue.

→ Binding site for HMGA-like nested primer (HMG NESTED) in pink.

Appendix 2:
Mutator transposase-like Partial cDNA Sequence

```
  1   ATTATGGCCG GGGGGCCGAT GGGATTGCAG TTGTAAAGGT TGGCAACTTA
                                     <-------
 51   CAGCTTTACC ATGTTCTCAT GCTATTGCCG TTATCAGTTG CATTGGACGA
    --->                  <-------
101   AGCCCATATG ACTTTTGCTC CAGATACTTT ACAACCGAGA GCTACAGGTT

151   AACCTATGCA GAGTCTATAC AACCTATTCC AGATGTGGAG AGACCTCTGC

201   AGAAGGATAC TTCATCAGCT CTAGTGACTG TAACTCCTCC TCCAACACGC

251   CGTCCACCAG GTAGGCCCAC CACAAAAAAA GTTGGAACAC CAGAGGTAAT

301   GAAGCGTCAA CTACAGTGCA GTAGATGCAA GGGTTTAGGG CACAACAAGT

351   CCACTTGCAA AGAGCTCTTG TAGAGTGTTG GCCACTGGAT CGAGTTAAGG

401   ATGACACAGG TTGGCTGATG GATCTAATGG TAGTTCTTCA ATGTCAAATA

451   GTTTTATTAG GTGTAATGCT AGTTTTCCTT GGTCTCTAGT TAAAAACTTT

501   ATGATAAATG AACATGGCAG TGACTCGTAG GGAAGGATTG GTTATTTAAA

551   GTAGCTATTT GTATCCATGT TGATGGAAAG CTAAATGTTA CCATCGACTT
                   <-------
601   GGATCAAATT TTGTTACTGT ATTATCAGAA AGAGATTGGT TAGGTTTTCA

651   AAAAAAAAAA AAAAAAAAAA AAAAAAAA
```

→ Binding sites for *Mutator* transposase-like RT-PCR primers (MutF1 & MutR1) highlighted in yellow.

→ Binding site for *Mutator* transposase-like gene-specific primer (MUTGSP) for use in round one of 5’ primary RACE-PCR in blue.

→ Binding site for *Mutator* transposase-like nested primer (MUT NESTED) for use in round one of 5’ nested RACE-PCR in pink.

Appendix 3:
Alignment of HMGA-like Partial cDNA & HMGA-like RT-PCR Sequences

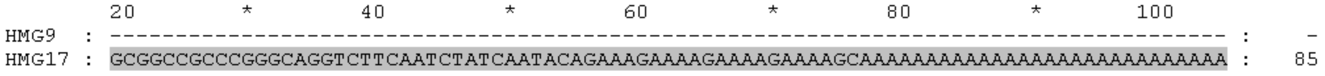
The sequence below shows part of the alignment between the HMGA-like partial cDNA (HMG) and the HMGA-like RT-PCR (HMG10) sequences.

```
                  *       440         *       460         *       480         *       500
HMGA10 : CGCTGCTGCTGTTCCTACTACTACTGCCGTTGCTGCTGGTCCGAGGCCTAGAGGAAGGCCAAAAGGCACTTCCGTGGCTCCTGC :  504
HMGA   : CGCTGCTGCTGTTCCTACTACTACTGCCGTTGCTGCTGGTCCGAGGCCTAGAGGAAGGCCAAAAGGCACTTCCGTGGCTCCTGC :  504
         CGCTGCTGCTGTTCCTACTACTACTGCCGTTGCTGCTGGTCCGAGGCCTAGAGGAAGGCCAAAAGGCACTTCCGTGGCTCCTGC

              *       520         *       540         *       560         *       580
HMGA10 : TGGCCTTGCGGTCCCTGGAAAGGGGAGAGGTCGACCACCCAAGAGTAGTGGGGTCGCCGCGAAGCCCATCAAGCCCAAGAAATC :  588
HMGA   : TGGCCTTGCGGTCCCTGGAAAGGGGAGAGGTCGACCACCCAAGAGTAGTGGGGTCGCCGCGAAGCCCATCAAGCCCAAGAAATA :  588
         TGGCCTTGCGGTCCCTGGAAAGGGGAGAGGTCGACCACCCAAGAGTAGTGGGGTCGCCGCGAAGCCCATCAAGCCCAAGAAAT

          *       600         *       620         *       640         *       660         *
HMGA10 : TACTGGAAAGCCTGTTGGCCGCCCCAAGAAGACAACGGATGGAGCAGCATCATATGGCGATCTGAAGAGAAAACTTGAATTCTT :  672
HMGA   : TACTGGAAAGCCTGTTGGCCGCCCCAAGAAG----------------------------------------------------- :  619
         TACTGGAAAGCCTGTTGGCCGCCCCAAGAAG

              680         *       700         *       720         *       740         *
HMGA10 : TCAATCAAAAGTGAAGCAAGCAGTTGGAGTACTAAAGTCTCAGTTCAGCAGTGAAAGCAACATCAGTGCAATCGATGCTATCCA :  756
HMGA   : -CAATCAAAAGTGAAGCAAGCAGTTGGAGTACTAAAGTCTCAGTTCAGCAGTGAAAGCAACATCAGTGCAATCGATGCTATCCA :  702
          CAATCAAAAGTGAAGCAAGCAGTTGGAGTACTAAAGTCTCAGTTCAGCAGTGAAAGCAACATCAGTGCAATCGATGCTATCCA

          760         *       780         *       800         *       820         *       840
HMGA10 : AGAGTTAGAAGTACTTGCAGCTATGGACATTAACAAACCATTTAAAGATGATGCTCAGCCACCGCCCCCACCACCCCCAGCACC :  840
HMGA   : AGAGTTAGAAGTACTTGCAGCTATGGACATTAACAAACCATTTAAAGATGATGCTCAGCCACCGCCCCCACCACCCCCAGCACC :  786
         AGAGTTAGAAGTACTTGCAGCTATGGACATTAACAAACCATTTAAAGATGATGCTCAGCCACCGCCCCCACCACCCCCAGCACC
```

Appendix 4:
Alignment of HMGA-like Partial cDNA Sequence and two HMGA-like 5' Nested RACE-PCR Products

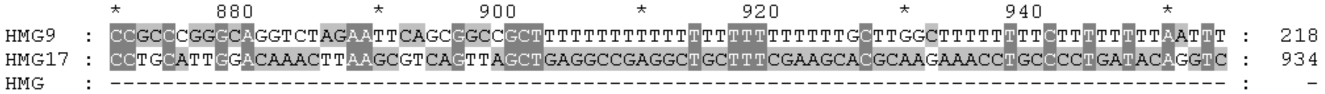
The figures below show parts of the alignment between the HMGA-like partial cDNA sequence (HMG) and sequences isolated from 5' Nested RACE-PCR (HMG9 & HMG17).

Figure A: Poly A region identified in HMG17.



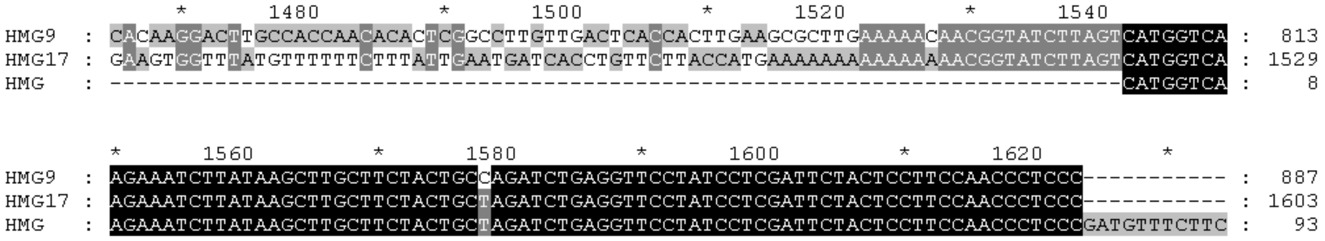
```
        20        *         40        *         60        *         80        *         100
HMG9  : ------------------------------------------------------------------------------------- :    -
HMG17 : GCGGCCGCCCGGGCAGGTCTTCAATCTATCAATACAGAAAGAAAAGAAAAGAAAAGCAAAAAAAAAAAAAAAAAAAAAAAAAAAA :   85
HMG   : ------------------------------------------------------------------------------------- :    -
```

Figure B: Poly T region identified in HMG9.



```
        *         880       *         900       *         920       *         940       *
HMG9  : CCGCCCGGGCAGGTCTAGAATTCAGCGGCCGCTTTTTTTTTTTTTTTTTTTTTTTGCTTGGCTTTTTTTTCTTTTTTTTAATTT :  218
HMG17 : CCTGCATTGGACAAACTTAAGCGTCAGTTAGCTGAGGCCGAGGCTGCTTTCGAAGCACGCAAGAAACCTGCCCCTGATACAGGTC :  934
HMG   : ------------------------------------------------------------------------------------- :    -
```

Figure C: Region of overlap between the partial cDNA sequence and sequences isolated from 5' nested RACE-PCR.

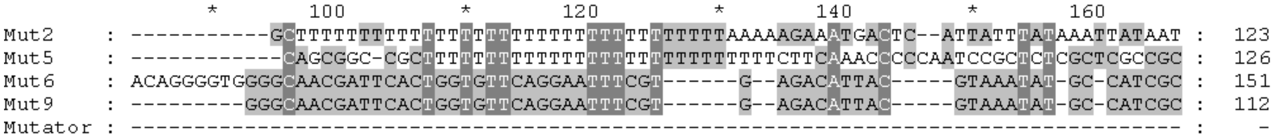


```
             *         1480      *         1500      *         1520      *         1540
HMG9  : CACAAGGACTTGCCACCAACACACTCGGCCTTGTTGACTCACCACTTGAAGCGCTTGAAAAACAACGGTATCTTAGTCATGGTCA :  813
HMG17 : GAAGTGGTTTATGTTTTTTCTTTATTGAATGATCACCTGTTCTTACCATGAAAAAAAAAAAAAACGGTATCTTAGTCATGGTCA : 1529
HMG   : -----------------------------------------------------------------------------CATGGTCA :    8

        *         1560      *         1580      *         1600      *         1620      *
HMG9  : AGAAATCTTATAAGCTTGCTTCTACTGCCAGATCTGAGGTTCCTATCCTCGATTCTACTCCTTCCAACCCTCCC----------- :  887
HMG17 : AGAAATCTTATAAGCTTGCTTCTACTGCTAGATCTGAGGTTCCTATCCTCGATTCTACTCCTTCCAACCCTCCC----------- : 1603
HMG   : AGAAATCTTATAAGCTTGCTTCTACTGCTAGATCTGAGGTTCCTATCCTCGATTCTACTCCTTCCAACCCTCCCGATGTTTCTTC :   93
```

Appendix 5:
Alignment of *Mutator* transposase-like Partial cDNA Sequence and four *Mutator* transposase-like 5' Primary RACE-PCR Products

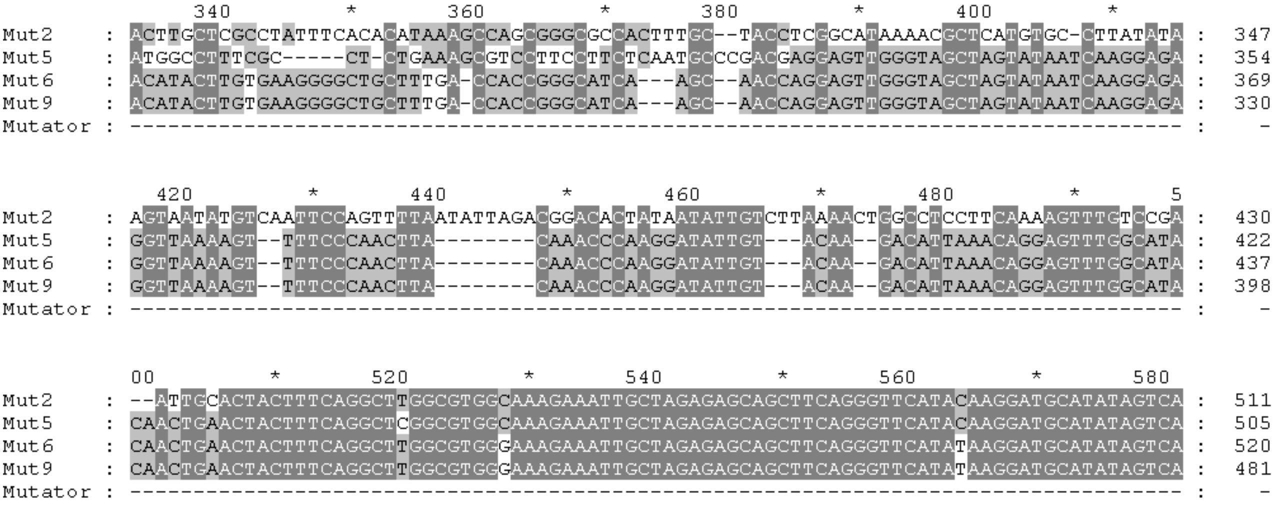
The figures below show parts of the alignment between the *Mutator* transposase-like partial cDNA (Mutator) and sequences isolated from 5' Primary RACE-PCR (Mut 2, Mut5, Mut 6 & Mut9).

Figure A: Poly T regions identified in Mut2 & Mut5.



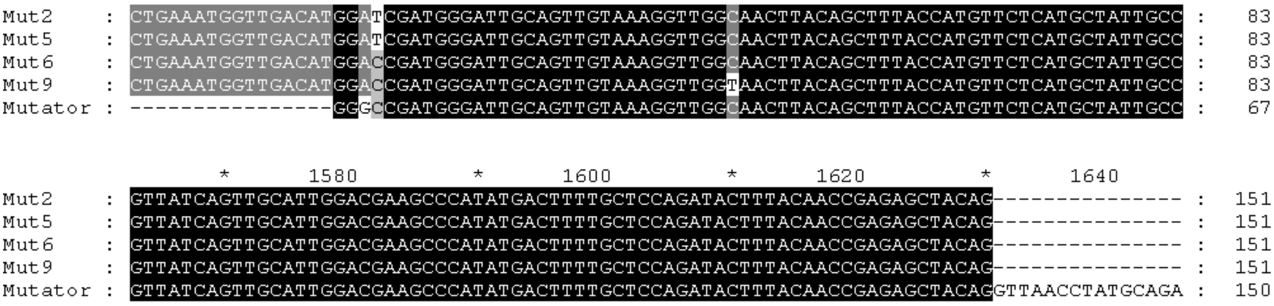
```
                *       100         *       120         *       140         *       160
Mut2    : -----------GCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTAAAAAGAAATGACTC--ATTATTTATAAATTATAAT :  123
Mut5    : ------------CAGCGGC-CGCTTTTTTTTTTTTTTTTTTTTTTTTTTTTCTTCAAACCCCCAATCCGCTCTCGCTCGCCGC :  126
Mut6    : ACAGGGGTGGGGCAACGATTCACTGGTGTTCAGGAATTTCGT------G--AGACATTAC-----GTAAATAT-GC-CATCGC :  151
Mut9    : ---------GGGCAACGATTCACTGGTGTTCAGGAATTTCGT------G--AGACATTAC-----GTAAATAT-GC-CATCGC :  112
Mutator : ----------------------------------------------------------------------------------- :    -
```

Figure B: Mut2 & Mut5 no longer align with Mut6 & Mut9.



```
               340         *       360         *       380         *       400         *
Mut2    : ACTTGCTCGCCTATTTCACACATAAAGCCAGCGGGCGCCACTTTGC--TACCTCGGCATAAAACGCTCATGTGC-CTTATATA :  347
Mut5    : ATGGCCTTTCGC-----CT-CTGAAAGCGTCCTTCCTTCTCAATGCCCGACGAGGAGTTGGGTAGCTAGTATAATCAAGGAGA :  354
Mut6    : ACATACTTGTGAAGGGGCTGCTTTGA-CCACCGGGCATCA---AGC--AACCAGGAGTTGGGTAGCTAGTATAATCAAGGAGA :  369
Mut9    : ACATACTTGTGAAGGGGCTGCTTTGA-CCACCGGGCATCA---AGC--AACCAGGAGTTGGGTAGCTAGTATAATCAAGGAGA :  330
Mutator : ----------------------------------------------------------------------------------- :    -

            420         *       440         *       460         *       480         *       5
Mut2    : AGTAATATGTCAATTCCAGTTTTAATATTAGACGGACACTATAATATTGTCTTAAAACTGGCCTCCTTCAAAAGTTTGTCCGA :  430
Mut5    : GGTTAAAAGT--TTTCCCAACTTA--------CAAACCCAAGGATATTGT---ACAA--GACATTAAACAGGAGTTTGGCATA :  422
Mut6    : GGTTAAAAGT--TTTCCCAACTTA--------CAAACCCAAGGATATTGT---ACAA--GACATTAAACAGGAGTTTGGCATA :  437
Mut9    : GGTTAAAAGT--TTTCCCAACTTA--------CAAACCCAAGGATATTGT---ACAA--GACATTAAACAGGAGTTTGGCATA :  398
Mutator : ----------------------------------------------------------------------------------- :    -

          00         *       520         *       540         *       560         *       580
Mut2    : --ATTGCACTACTTTCAGGCTTGGCGTGGCAAAGAAATTGCTAGAGAGCAGCTTCAGGGTTCATACAAGGATGCATATAGTCA :  511
Mut5    : CAACTGAACTACTTTCAGGCTCGGCGTGGCAAAGAAATTGCTAGAGAGCAGCTTCAGGGTTCATACAAGGATGCATATAGTCA :  505
Mut6    : CAACTGAACTACTTTCAGGCTTGGCGTGGGAAAGAAATTGCTAGAGAGCAGCTTCAGGGTTCATATAAGGATGCATATAGTCA :  520
Mut9    : CAACTGAACTACTTTCAGGCTTGGCGTGGGAAAGAAATTGCTAGAGAGCAGCTTCAGGGTTCATATAAGGATGCATATAGTCA :  481
Mutator : ----------------------------------------------------------------------------------- :    -
```

Figure C: Region of overlap between the partial cDNA sequence and sequences isolated from 5' primary RACE-PCR.



```
          1480      *      1500         *      1520         *      1540         *      1560
Mut2    : CTGAAATGGTTGACATGGATCGATGGGATTGCAGTTGTAAAGGTTGGCAACTTACAGCTTTACCATGTTCTCATGCTATTGCC :   83
Mut5    : CTGAAATGGTTGACATGGATCGATGGGATTGCAGTTGTAAAGGTTGGCAACTTACAGCTTTACCATGTTCTCATGCTATTGCC :   83
Mut6    : CTGAAATGGTTGACATGGACCGATGGGATTGCAGTTGTAAAGGTTGGCAACTTACAGCTTTACCATGTTCTCATGCTATTGCC :   83
Mut9    : CTGAAATGGTTGACATGGACCGATGGGATTGCAGTTGTAAAGGTTGGTAACTTACAGCTTTACCATGTTCTCATGCTATTGCC :   83
Mutator : ----------------GGGCCGATGGGATTGCAGTTGTAAAGGTTGGCAACTTACAGCTTTACCATGTTCTCATGCTATTGCC :   67

                 *      1580         *      1600         *      1620         *      1640
Mut2    : GTTATCAGTTGCATTGGACGAAGCCCATATGACTTTTGCTCCAGATACTTTACAACCGAGAGCTACAG--------------- :  151
Mut5    : GTTATCAGTTGCATTGGACGAAGCCCATATGACTTTTGCTCCAGATACTTTACAACCGAGAGCTACAG--------------- :  151
Mut6    : GTTATCAGTTGCATTGGACGAAGCCCATATGACTTTTGCTCCAGATACTTTACAACCGAGAGCTACAG--------------- :  151
Mut9    : GTTATCAGTTGCATTGGACGAAGCCCATATGACTTTTGCTCCAGATACTTTACAACCGAGAGCTACAG--------------- :  151
Mutator : GTTATCAGTTGCATTGGACGAAGCCCATATGACTTTTGCTCCAGATACTTTACAACCGAGAGCTACAGGTTAACCTATGCAGA :  150
```

Appendix 6:
5’ Region of Consensus Sequence of *Mutator* Transposase-like cDNA after 5’ Primary RACE-PCR, Round I

The 5’ consensus sequence of the *Mutator* transposase-like cDNA was obtained from the alignment between Mut6 and Mut9 isolated in the first round of 5’ primary RACE-PCR (Appendix 5). Only a portion of the *Mutator* transposase-like cDNA is shown.

```
  1  AAAACCACTC AGCAGTGGCA GAATACAATT ACAGGGGTGG GGCAACGATT

 51  CACTGGTGTT CAGGAATTTC GTGAGACATT ACGTAAATAT GCCATCGCAC
                                                 ◄───────
101  ATCAGTTCGC ATTCAGGTAT AAAAAGAACG ATAGTCACCG CGTGACTGTT

151  AAGTGCAGAG CAGAAGGCTG CCCTTGGAGA ATCCATGCAT CAAGGTTGTC
                                          ◄──────
201  AACCACTCAG CTGATATGTA TCAAGAAGAT GAATCCAACA CATACTTGTG

251  AAGGGGCTGC TTTGACCACC GGGCATCAAG CAACCAGGAG TTGGGTAGCT

301  AGTATAATCA AGGAGAGGTT AAAAGTTTTC CCAACTTACA AACCCAAGGA

351  TATTGTACAA GACATTAAAC AGGAGTTTGG CATACAACTG AACTACTTTC

401  AGGCTTGGCG TGGGAAAGAA ATTGCTAGAG AGCAGCTTCA GGGTTCATAT

451  AAGGATGCAT ATAGTCAGTT ACCACTTTTC TGCGAGAGGA TAATGGAGAC
```

──► Binding site for *Mutator* transposase-like gene-specific primer (Mut GSP 2) for use in round two of 5’ primary RACE-PCR in blue.

──► Binding site for *Mutator* transposase-like nested primer (Mut Nested 2) for use in round two of 5’ nested RACE-PCR in pink.

Appendix 7:

Alignment of *Mutator* Transposase-like Sequences from First and Second Rounds of 5’ Primary RACE-PCR

The sequence below shows the region of overlap between the consensus sequence obtained from round one of 5’ primary RACE-PCR (consensus of Mut 6 & Mut9) and the sequence isolated from round two of 5’ primary RACE-PCR (Mut4).

```
                      *       740         *       760         *       780         *       800
Mut4       : TTCATCTATTCTTCCTCTTGCTAAAACCACTCAGCAGTGGCAGAATACAATTACAGGGGTGGGGCAACGATTCACTGGTG :  800
Mut6/9cons : ---------------------TAAAACCACTCAGCAGTGGCAGAATACAATTACAGGGGTGGGGCAACGATTCACTGGTG :   59
                                  TAAAACCACTCAGCAGTGGCAGAATACAATTACAGGGGTGGGGCAACGATTCACTGGTG

                      *       820         *       840         *       860         *       880
Mut4       : TTCAGGAATTTCGTGAGACATTACGTAAATATGCCATCGCACATCAGTTCGCATTCAGGTATAAAAAGAACGATAGTCAC :  880
Mut6/9cons : TTCAGGAATTTCGTGAGACATTACGTAAATATGCCATCGCACATCAGTTCGCATTCAGGTATAAAAAGAACGATAGTCAC :  139
             TTCAGGAATTTCGTGAGACATTACGTAAATATGCCATCGCACATCAGTTCGCATTCAGGTATAAAAAGAACGATAGTCAC

                      *       900         *       920         *       940         *       960
Mut4       : CGCGTGACTGTTAAGTGCAGAGCAGAAGGCTGCCCTTGGAGAATCCATGCATCAAGGTTGTCAACCACTCAGCTGATATG :  960
Mut6/9cons : CGCGTGACTGTTAAGTGCAGAGCAGAAGGCTGCCCTTGGAGAATCCATGCATCAAGGTTGTCAACCACTCAGCTGATATG :  219
             CGCGTGACTGTTAAGTGCAGAGCAGAAGGCTGCCCTTGGAGAATCCATGCATCAAGGTTGTCAACCACTCAGCTGATATG

                      *       980         *      1000         *      1020         *      1040
Mut4       : TATCAAGAAGATGAATCCAACACATACTTGTGAAGGGGCTGCT------------------------------------- : 1003
Mut6/9cons : TATCAAGAAGATGAATCCAACACATACTTGTGAAGGGGCTGCTTTGACCACCGGGCATCAAGCAACCAGGAGTTGGGTAG :  299
             TATCAAGAAGATGAATCCAACACATACTTGTGAAGGGGCTGCT
```

Appendix 8:
Full-length *Mutator* Transposase-like cDNA Consensus Sequence Identified by 5’ RACE-PCR

Translation of the full-length *Mutator* transposase-like cDNA is shown below the nucleotide sequence. The start codon is boxed in red and the stop codon is boxed in green.

```
    CCGAAATTTTACAACATCAACCAAAAATAAAAAAAAAGGGAAATTTTTTATTTTGAGAATC
  1 ---------+---------+---------+---------+---------+---------+ 60


    TAATCCTCAGAAACTCACAATTTCAAGTATCGACGCTCTGATTGGTTGATTTTCAATTCA
 61 ---------+---------+---------+---------+---------+---------+ 120


    GAGAACGCCGCCTCCGGTCAGGTCATGACTACCAAAAGGGTGATAGCGGTTTGTCAATCT
121 ---------+---------+---------+---------+---------+---------+ 180
                            M  T  T  K  R  V  I  A  V  C  Q  S

    GGTGGTGAGTTTGAGACTGATAAGAACGGGGTGATGTTGTATAAAGGCGGAGATGCTCAT
181 ---------+---------+---------+---------+---------+---------+ 240
    G  G  E  F  E  T  D  K  N  G  V  M  L  Y  K  G  G  D  A  H

    GCGATAGACCTTGATGATCAAACGAAGTTCGACGACTTCAAGAAGGAGCTAGCAGAAATG
241 ---------+---------+---------+---------+---------+---------+ 300
    A  I  D  L  D  D  Q  T  K  F  D  D  F  K  K  E  L  A  E  M

    TTTAATTTGAACATGAGTACTATGTCTGTTAAATACTTCCTCCCTGGCAATAAAAAGACT
301 ---------+---------+---------+---------+---------+---------+ 360
    F  N  L  N  M  S  T  M  S  V  K  Y  F  L  P  G  N  K  K  T

    CTCATCAGTATCTCCAATGACAAGGATTTAAAGCGGTTAATAAAATTTCATGGGAATTCC
361 ---------+---------+---------+---------+---------+---------+ 420
    L  I  S  I  S  N  D  K  D  L  K  R  L  I  K  F  H  G  N  S

    ACCACTGTCGATGTGTTTATCCTAATGGAGGAAAATGTTCCTCATGATGTTTCCAACATG
421 ---------+---------+---------+---------+---------+---------+ 480
    T  T  V  D  V  F  I  L  M  E  E  N  V  P  H  D  V  S  N  M

    CCAGCCAGTAGGTCTAGCAGGACAACTTTGTCTGAAGCAGTGGCTCCAGCTCCGGATCCG
481 ---------+---------+---------+---------+---------+---------+ 540
    P  A  S  R  S  S  R  T  T  L  S  E  A  V  A  P  A  P  D  P

    GCTCCGGCTCCAGCTCCAGCTCTGGTTCTGGCTTCGCTTGATCCACCTCTAGATGTTATG
541 ---------+---------+---------+---------+---------+---------+ 600
    A  P  A  P  A  P  A  L  V  L  A  S  L  D  P  P  L  D  V  M

    GATGATACCACTCGATCGAATATTCCTCTTGCGGCCTCTCTTGATATTGTTGATGCTACC
601 ---------+---------+---------+---------+---------+---------+ 660
    D  D  T  T  R  S  N  I  P  L  A  A  S  L  D  I  V  D  A  T

    AATCATATTGATCCACACATTGATCTAGCACCTGAGTTTTCATCTATTCTTCCTCTTGCT
661 ---------+---------+---------+---------+---------+---------+ 720
    N  H  I  D  P  H  I  D  L  A  P  E  F  S  S  I  L  P  L  A

    AAAACCACTCAGCAGTGGCAGAATACAATTACAGGGGTGGGGCAACGATTCACTGGTGTT
721 ---------+---------+---------+---------+---------+---------+ 780
    K  T  T  Q  Q  W  Q  N  T  I  T  G  V  G  Q  R  F  T  G  V
```

```
     CAGGAATTTCGTGAGACATTACGTAAATATGCCATCGCACATCAGTTCGCATTCAGGTAT
 781 ---------+---------+---------+---------+---------+---------+ 840
     Q  E  F  R  E  T  L  R  K  Y  A  I  A  H  Q  F  A  F  R  Y

     AAAAAGAACGATAGTCACCGCGTGACTGTTAAGTGCAGAGCAGAAGGCTGCCCTTGGAGA
 841 ---------+---------+---------+---------+---------+---------+ 900
     K  K  N  D  S  H  R  V  T  V  K  C  R  A  E  G  C  P  W  R

     ATCCATGCATCAAGGTTGTCAACCACTCAGCTGATATGTATCAAGAAGATGAATCCAACA
 901 ---------+---------+---------+---------+---------+---------+ 960
     I  H  A  S  R  L  S  T  T  Q  L  I  C  I  K  K  M  N  P  T

     CATACTTGTGAAGGGGCTGCTTTGACCACCGGGCATCAAGCAACCAGGAGTTGGGTAGCT
 961 ---------+---------+---------+---------+---------+---------+ 1020
     H  T  C  E  G  A  A  L  T  T  G  H  Q  A  T  R  S  W  V  A

     AGTATAATCAAGGAGAGGTTAAAAGTTTTCCCAACTTACAAACCCAAGGATATTGTACAA
1021 ---------+---------+---------+---------+---------+---------+ 1080
     S  I  I  K  E  R  L  K  V  F  P  T  Y  K  P  K  D  I  V  Q

     GACATTAAACAGGAGTTTGGCATACAACTGAACTACTTTCAGGCTTGGCGTGGGAAAGAA
1081 ---------+---------+---------+---------+---------+---------+ 1140
     D  I  K  Q  E  F  G  I  Q  L  N  Y  F  Q  A  W  R  G  K  E

     ATTGCTAGAGAGCAGCTTCAGGGTTCATATAAGGATGCATATAGTCAGTTACCACTTTTC
1141 ---------+---------+---------+---------+---------+---------+ 1200
     I  A  R  E  Q  L  Q  G  S  Y  K  D  A  Y  S  Q  L  P  L  F

     TGCGAGAGGATAATGGAGACCAATCCAGGCAGTCTTGCTACATTCTCCACTAAGGAAGAC
1201 ---------+---------+---------+---------+---------+---------+ 1260
     C  E  R  I  M  E  T  N  P  G  S  L  A  T  F  S  T  K  E  D

     TCTAGTTTTCATCGCCTCTTTATCGCATTCCATGCTTCTCTGTCCGGCTTTCTACAAGGT
1261 ---------+---------+---------+---------+---------+---------+ 1320
     S  S  F  H  R  L  F  I  A  F  H  A  S  L  S  G  F  L  Q  G

     TGCAGACCTCTCCTTTTTCTTGATAGCATACCCTTGAATTCCAAATATCAAGGTACCTTG
1321 ---------+---------+---------+---------+---------+---------+ 1380
     C  R  P  L  L  F  L  D  S  I  P  L  N  S  K  Y  Q  G  T  L

     TTGGCAGCAACCGCTGCAGATGGGAATGATGATGTATTTCCTGTTGCCTTTTCTGTCGTT
1381 ---------+---------+---------+---------+---------+---------+ 1440
     L  A  A  T  A  A  D  G  N  D  D  V  F  P  V  A  F  S  V  V

     GATGCTGAAACTGATGAGAACTGGTATTGGTTTTTGTTACAACTGAAATCTGCCCTGTCA
1441 ---------+---------+---------+---------+---------+---------+ 1500
     D  A  E  T  D  E  N  W  Y  W  F  L  L  Q  L  K  S  A  L  S

     ACATCTTGTTCTATAACATTTGTTGCAGACATACAGAAGGGTTTACGGGAGTCAATCTCC
1501 ---------+---------+---------+---------+---------+---------+ 1560
     T  S  C  S  I  T  F  V  A  D  I  Q  K  G  L  R  E  S  I  S

     GAAATTTTCAAGGGATCATTTCATGGTTATTGTCTCCGATACTTAACTGAACAGCTCATT
1561 ---------+---------+---------+---------+---------+---------+ 1620
     E  I  F  K  G  S  F  H  G  Y  C  L  R  Y  L  T  E  Q  L  I

     AGAGACTTGAAAGGGCAGTTTTCCCATGAGATGAAACGGCTCATGATTGACGACCTTTAT
1621 ---------+---------+---------+---------+---------+---------+ 1680
     R  D  L  K  G  Q  F  S  H  E  M  K  R  L  M  I  D  D  L  Y

     AATGCAGCTTTGGCACCTAGACCTGAAGCCTTCCATTCGTATATCAACAGCATCAAAAAA
1681 ---------+---------+---------+---------+---------+---------+ 1740
     N  A  A  L  A  P  R  P  E  A  F  H  S  Y  I  N  S  I  K  K

     ATTTCACTGGAAGCTTACAATTGGATCATTCAGAGTGAGCCCCAATACTGGGCAAATTCA
1741 ---------+---------+---------+---------+---------+---------+ 1800
     I  S  L  E  A  Y  N  W  I  I  Q  S  E  P  Q  Y  W  A  N  S
```

```
     TATTTCCAGGGTGCTAGATATAACCATATGGCATCAAACTTTGGGGAACTCTTTTATAGT
1801 ---------+---------+---------+---------+---------+---------+ 1860
     Y  F  Q  G  A  R  Y  N  H  M  A  S  N  F  G  E  L  F  Y  S

     TGGGTGTCAGATGCACACGAATTACCTATAACTCAGATGGTTGATGTCATACGGCAAAAG
1861 ---------+---------+---------+---------+---------+---------+ 1920
     W  V  S  D  A  H  E  L  P  I  T  Q  M  V  D  V  I  R  Q  K

     GTTATGGAGTTAATTTATTCCCGCAGGGCTGAGTCTGATCATTGGTTAACAAGGCTTACT
1921 ---------+---------+---------+---------+---------+---------+ 1980
     V  M  E  L  I  Y  S  R  R  A  E  S  D  H  W  L  T  R  L  T

     CCATGCATGGAGGAAAAATTGGAAAAGGAGGGCTTGAAAGTCCGTACTCTTCAAGTGCTT
1981 ---------+---------+---------+---------+---------+---------+ 2040
     P  C  M  E  E  K  L  E  K  E  G  L  K  V  R  T  L  Q  V  L

     CTGACTTCTGGTAGCATATTTGAGGTCCGTGGTGATTCCACCACTGAAATGGTTGACATG
2041 ---------+---------+---------+---------+---------+---------+ 2100
     L  T  S  G  S  I  F  E  V  R  G  D  S  T  T  E  M  V  D  M

     GACCGATGGGATTGCAGTTGTAAAGGTTGGCAACTTACAGCTTTACCATGTTCTCATGCT
2101 ---------+---------+---------+---------+---------+---------+ 2160
     D  R  W  D  C  S  C  K  G  W  Q  L  T  A  L  P  C  S  H  A

     ATTGCCGTTATCAGTTGCATTGGACGAAGCCCATATGACTTTTGCTCCAGATACTTTACA
2161 ---------+---------+---------+---------+---------+---------+ 2220
     I  A  V  I  S  C  I  G  R  S  P  Y  D  F  C  S  R  Y  F  T

     ACCGAGAGCTACAGGTTAACCTATGCAGAGTCTATACAACCTATTCCAGATGTGGAGAGA
2221 ---------+---------+---------+---------+---------+---------+ 2280
     T  E  S  Y  R  L  T  Y  A  E  S  I  Q  P  I  P  D  V  E  R

     CCTCTGCAGAAGGATACTTCATCAGCTCTAGTGACTGTAACTCCTCCTCCAACACGCCGT
2281 ---------+---------+---------+---------+---------+---------+ 2340
     P  L  Q  K  D  T  S  S  A  L  V  T  V  T  P  P  P  T  R  R

     CCACCAGGTAGGCCCACCACAAAAAAAGTTGGAACACCAGAGGTAATGAAGCGTCAACTA
2341 ---------+---------+---------+---------+---------+---------+ 2400
     P  P  G  R  P  T  T  K  K  V  G  T  P  E  V  M  K  R  Q  L

     CAGTGCAGTAGATGCAAGGGTTTAGGGCACAACAAGTCCACTTGCAAAGAGCTCTTGTAG
2401 ---------+---------+---------+---------+---------+---------+ 2460
     Q  C  S  R  C  K  G  L  G  H  N  K  S  T  C  K  E  L  L  *


     AGTGTTGGCCACTGGATCGAGTTAAGGATGACACAGGTTGGCTGATGGATCTAATGGTAG
2461 ---------+---------+---------+---------+---------+---------+ 2520


     TTCTTCAATGTCAAATAGTTTTATTAGGTGTAATGCTAGTTTTCCTTGGTCTCTAGTTAA
2521 ---------+---------+---------+---------+---------+---------+ 2580


     AAACTTTATGATAAATGAACATGGCAGTGACTCGTAGGGAAGGATTGGTTATTTAAAGTA
2581 ---------+---------+---------+---------+---------+---------+ 2640


     GCTATTTGTATCCATGTTGATGGAAAGCTAAATGTTACCATCGACTTGGATCAAATTTTG
2641 ---------+---------+---------+---------+---------+---------+ 2700


     TTACTGTATTATCAGAAAGAGATTGGTTAGGTTTTCAAAAAAAAAAAAAAAAAAAAAAAA
2701 ---------+---------+---------+---------+---------+---------+ 2760


     AAAAA
2761 ----- 2765
```