In Planta and *In Silico* Analysis of Soybean Lectin Promoters

Hanaa Saeed

Department of Plant Science
McGill University, Montreal, Quebec, Canada

August, 2007

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of a Masters in Science.

Abstract

Soybean seed lectin, *Le1*, is specifically located in seeds of soybean, *Glycine max*, (L.) Merr., due to its promoter. Gene homologues of *Le1* were previously identified as possibly located in other parts of soybean. We cloned two novel promoters from these genes, and show that they drive reporter gene expression in transgenic *Arabidopsis*. A total of 1.3kb was isolated from each of the *Le2* and *Le3* 5' promoter regions and fused with the GUS reporter gene. A previously cloned *Le1* 5' promoter was used as a control and the constructs were introduced into *Arabidospis*. GUS expression in transformed plants reveals that GUS driven by *Le3* is found predominantly in vegetative tissues whereas GUS driven by *Le2* show low expression in all tissues examined. The expression patterns resulting from the three different lectin promoters are distinct and consistent with regulatory motifs computationally identified in the sequences.

Résumé

Chez le soja (*Glycine max*), le promoteur du gene lectine *Le1* dirige l'expression spécifique dans les graines. Des homologues de *Le1* existent dans le genome du soja et sont exprimées ailleurs dans la plante. Nous avons isolé deux promoteurs de ces homologues de lectine, et décrivons le patron d'expression qu'ils dirigent. Un total de 1.3 kilobase des regions 5' des promoteurs, en amont du gène, a été isolé pour chacune des copies *Le2* et *Le3*, et fusionné avec le gène rapporteur GUS. Le promoteur de *Le1* étant déjà connu, il sert de controle. L'*Arabidopsis* transformée avec ces constructions, montre que le promoteur de *Le3* dirige l'expression dans les tissues végétatifs, tandis que le promoteur de *Le2* procure un niveau minimal d'expression dans tous les tissus examinés. De plus, des analyses bioinformatiques identifient des motifs spécifiques dans les sequences de promoteurs qui confirment les patrons d'expression que nous avons démontrés.

Acknowledgements

I would like to take this opportunity to sincerely thank my supervisor, Dr. Martina Strömvik, for her guidance, patience and encouragement throughout my degree.

I would also like to give my thanks to my committee members Dr. Jabaji-Hare and Dr. Seguin for their guidance and support.

Special thanks to the members of Dr. Strömvik's lab, Annie, Kei-Chin, Francois, Catherine, Fred, Eve, and Julie for their friendship and help. Special thanks to Annie for all the help, encouragement and abstract translation, and Francois for the help with the bioinformatics analysis.

Finally, I would like to thank my parents, Amjad and Neelam, and my sisters, Sabina and Bushra, for their support.

This project would not have been possible without the financial support from NSERC.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ψ*BCH*	Tomato class I basic chitinase gene
ABA	Abscisic acid
ABRE	Abscisic acid responsive element
AsGlo1	Oat globulin
Atmyb2	*A. thaliana* MYB2
AUX/IAA	Auxin/ indole-3-acetic acid
CaMV 35S	Cauliflower mosaic virus 35S promoter
CARE	Cis-acting regulatory elements
cv	Cultivar
DB58	*Dolichos biflorus* lectin 58
DBSL	*D. biflorus* seed lectin
DcMYB1	*Daucus carota* (carrot) MYB1
DcPal1	*Daucus carota* (carrot) phenylalanine ammonialyase
DNA	Deoxyribonucleic acid
ELISA	Enzyme-Linked ImmunoSorbent Assay
ERD15	EARLY RESPONSIVE TO DEHYDRATION 15 gene
ERF-1	Ethylene response factor 1
EST	Expressed sequence tag
FGAM1/ FGAM2	Soybean phosphoribosylformyl-glycinamide synthase 1/2
GA	Gibberellic acid
GFP	Green fluorescent protein (*Aequorea victoria)*
GMO	Genetically modified organism
GmSBP	*Glycine max* sucrose binding protein
GNA	*Galanthus nivulis* agglutinin
GUS	Glucuronidase
gusA (uidA)	*E. coli* β-glucuronidase gene
hptII	Hygromycin resistance gene
hsp17.3-B	Soybean heat shock promoter 17.3-B
KTi	Kunitz trypsin inhibitors

LB	Luria-Bertani media
Le1	Soybean lectin 1
LE1	Soybeen seed lectin protein
Le2	Soybean lectin 2
LE2	Soybean lectin 2 protein
Le3	Soybean lectin 3
Le4	Soybean lectin 4
legA	Lectin legume alpha region/site
legB	Lectin legume beta region/site
maize *Ubi1*	Maize ubiquitin 1
MAP	Multiple (sequence) alignment program
mRNA	Messenger RNA
Msg	Soybean major latex protein gene
NMD	Nonsense-mediated mRNA decay
nos	*nopaline synthase*
nt	nucleotide
PAL	phenylalanine ammonialyase
PCR	Polymerase chain reaction
PLACE	Plant *cis*-acting regulatory DNA elements (database)
rd22	*A. thalina* drought responsive gene
RNA	Ribonucleic acid
RNAi	RNA interference
Rplec	*Robina pseudoacacia* lectin
SBL	Soybean lectin (LE1)
SVL/LE3	Soybean vegetative lectin
TBP	TATA-box binding protein
TMV	Tobacco mosaic virus
Tnos	Nopaline synthase terminator
TOGT	Tobacco glucosyltransferase
UTR	Untranslated region
VSPα/VSPβ	Vegetative soybean protein alpha/beta

Section 1:

Introduction

1.1 General Introduction

While the predicted number of genes in soybean is about 61 000 (Vodkin *et al.*, 2004), less than 20 soybean gene promoter sequences are well-characterized. Promoters are sequences upstream of a coding region for a gene and by interacting with transcription factors (proteins), promoters regulate gene transcription levels and patterns (profiles) (Wray *et al.*, 2003). The effect of a promoter is based on the combination of motifs found within the promoter and regulation of the transcription factors that bind to them (Singh, 1998), so that the promoter may drive gene expression in a certain tissue, organ or cell type (tissue-specific promoter), only during certain conditions, as a result of specific signals (induced promoter), or at all times and locations (constitutive promoter) (Potenza *et al.*, 2004). The exact motifs or combination of motifs to result in a certain expression profile is as of yet not well known, especially for plants.

Many genes are part of gene families within a genome, where the gene products are very conserved, but often expressed in different tissues. For example, the three Kunitz trypsin inhibitor genes in soybean have different expression profiles which are highly regulated (Jofuku and Goldberg, 1989). Similarily, the soybean sucrose binding protein gene family contains at least two non-allelic genes, *GmSBP1* and *GmSBP2*. The latter is seed-specific, while the former is expressed in seed, fruit, stem root and leaves (Contim *et al.*, 2003; Elmer *et al.*, 2003; Waclawovsky *et al.*, 2006). The soybean legume lectin family is also a small gene family (Strömvik *et al.*, 2004).

While the expression profile of the lectin *Le1* has been shown to be seed-specific in soybean (Goldberg *et al.*, 1983; Vodkin *et al.*, 1983) and its promoter has been shown to drive seed-specific gene expression in transgenic tobacco, soybean and *Arabidopsis* (Lindstrom *et al.*, 1990; Cho *et al.*, 1995; Philip *et al.*, 2001; Darnowski and Vodkin, 2002), the promoters of its homologous genes have not previously been isolated nor studied.

In this study, we have isolated the promoter and coding regions of two *Le1* lectin gene homologues from soybean. The promoter region was tested *in planta* by transforming *Arabidopsis* plants and determing the reporter gene activity. An *in silico* analysis was done on the promoter region to determine if the motifs found correlated with previously predicted *in silico* expression profiles (Strömvik *et al.*, 2004), as well as the *in planta* results here.

Research into non-constitutive promoters, including the soybean lectin promoters will increase the basic knowledge of promoters and plant gene regulation, as well as provide a greater diversity of promoters that can be useful tools in genetic engineering.

All parts of this work was carried out by myself except for the extraction of promoter motifs from the PLACE database, which was carried out by Francois Fauteux, but interpreted by myself.

1.2 Hypotheses

I. Putative transcription factor binding sites that determine tissue specificity can be predicted by analyzing promoter sequences with bioinformatic methods.

II. The soybean lectin genes *Le2* and *Le3* have specific developmental and/or tissue specific mRNA expression patterns, which are determined by their respective promoter sequences.

III. Soybean sequences from the 5' and 3' promoter region of the lectin genes *Le2* and *Le3* can be used to drive reporter gene expression in transformed *Arabidopsis* plants in tissue specific patterns.

1.3 Objectives

Objective 1: Isolation and Sequence Analysis of Lectin Promoters

Aim I. To isolate and sequence 1.3 kb from the 5' region, and the coding sequence of the soybean lectin genes *Le2* and *Le3* from genomic DNA.

Aim II. To analyze these 5' regions using online bioinformatics tool PLACE (http://www.dna.affrc.go.jp/PLACE/), to locate known putative transcription binding sites.

Objective 2: Promoter Profiling

Aim I. To create a series of promoter-GUS reporter gene constructs using the 5' regions isolated from the soybean lectin genes *Le2*, and *Le3*, as well as the 5' region from *Le1* (obtained from collaborator).

Aim II. To transform *Arabidopsis thaliana* ecotype Columbia (Col-0) using the constructs, as well as positive and negative controls, and to detect reporter gene expression in transformed plants.

Section 2:
Literature Review

2.1 Soybean (*Glycine max*)

The increasing development of non-transgenic and transgenic crops encourages research of the soybean (*Glycine max* (L.) Merr.) genome in order to produce better cultivars that express desirable genes more specifically and effectively. For example, transgenic plants could be developed that would express anti-fungal compounds only in the tissues that fungi attack, such as in the root and stem to fight Rhizoctonia root and stem rot (*Rhizoctonia solani*), and not in the tissues used for human or animal purposes, such as the seed.

In order to achieve this goal, the soybean genome has been, and continues to be, well researched, and updated versions of both genetic linkage (Song *et al.*, 2004) and physical maps (Wu *et al.*, 2004) of the soybean genome have been published. An online database, the Soybean Genomic Database (SoyGD, http://soybeangenome.siu.edu), has integrated the known soybean physical map, bacterial artificial chromosome fingerprint database and genetic map associated genomic data (Shultz *et al.*, 2006). In addition to this, 371, 817 ESTs (Expressed Sequence Tags) for soybean are published in GenBank (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html, as of July 10, 2007), most of which come from Shoemaker *et al.*, 2002 and Vodkin *et al.*, 2004. This data is also available at the Soybean Genome Initiative site (http://soybean.ccgb.umn.edu/) and the Legume Information System (http://www.comparative-legumes.org/) (Gonzales *et al.*, 2005). The soybean genome is 1.115 Mb (n=20) in size, (Arumuganathan and Earle,

1991) and like most other plant genomes, is highly repetitive. Based on genetic mapping technologies, (Hadley and Hymowitz, 1973; Fischer and Goldberg, 1982; Shoemaker *et al.*, 1996; Wendel, 2000), and supported by EST sequence analysis (Schlueter *et al.*, 2004; Nelson and Shoemaker, 2006) soybean is generally thought to be a diploidized tetraploid. Gene and genome duplications are beneficial not only for the extra genes available for protein synthesis, but these duplications also increase the opportunities for diversification of gene function (Sparvoli *et al.*, 2001). Many multigene families (having two or more well-defined subgroups of more or less closely related genes) are found in soybean such as the glycinin, actin and lectin gene families (Hightower and Meagher, 1985; Nielsen *et al.*, 1989; Shoemaker *et al.*, 1996; Strömvik *et al.*, 2004).

2.2 The lectin gene family

2.2.1 Types of lectins

The legume lectin gene family is one example of a multi-gene family. Extensive research has been carried out on lectins from a wide array of plants. As early as 1888, Hermann Stillmark described an extract, which contained a lectin known as ricin, from castor beans, that could agglutinate blood cells from different animals (Rüdiger and Gabius, 2001). Other types of lectins have also been found in bacteria, slime molds, sponges, invertebrates and vertebrates (Rüdiger, 1984).

Lectins are now defined by three characteristics: 1) a lectin must be a carbohydrate-binding (glyco)protein 2) lectins are of non-immune origin and 3) the carbohydrate bound to the lectin cannot be biochemically changed (Goldstein *et al.*, 1980; Rüdiger and Gabius, 2001). Grouped by structural similarity and evolutionary evidence,

lectins can be divided into eight families: legume lectins, chitin-binding lectins, type-2 ribosome-inactivating proteins, monocot mannose-binding lectins, amaranthins, cucurbitaceous phloem lectins and jacalin-related lectins (Van Damme *et al.*, 1998; Zhang *et al.*, 2000).

Although lectins as a group are relatively well defined, their biological functions are not (Rüdiger and Gabius, 2001). They are found in the seed, or storage organs, but also in bark, leaves, roots and nodules, stem, and leaves (Van Damme *et al.*, 1995; Bauchrowitz *et al.*, 1996; Zhu *et al.*, 1996; Etzler, 1998), although these lectins are not as well-characterized as those found in seeds (Spilatro *et al.*, 1996). The variability of lectin binding ability and structure suggest they have variable biological roles and that the same lectin may have multiple roles (Rüdiger, 1984).

Several lectins that are toxic to insects have been studied. The mannose-binding lectin GNA gene (*Galanthus nivulis* agglutinin) from *Galanthus nivulis* (snowdrop) has been shown to be insecticidal for a broad range of insects, but non-toxic to mammals. It has been transformed into various crops, including potato and rice, where it offered protection against insects (Wool *et al.*, 1992; Pusztai *et al.*, 1993; Down *et al.*, 1996; Nagadhara *et al.*, 2004). This defensive role of legume lectins, and a role for lectins in legume-bacteria symbiosis are both common suggestions for legume lectin function. It has been hypothesized that the latter function may have originated from the lectins agglutinating and immobilizing bacteria on the roots as a defensive measure, which eventually evolved into the current symbiotic relationship (Chrispeels and Raikhel, 1991; Wool *et al.*, 1992). As evidence of the symbiosis function, researchers found that when a

pea lectin gene was inserted into white clover, the expression of the pea lectin in the roots allowed the white clover to host pea-specific symbiotic bacteria (Diaz *et al.*, 1989).

2.2.2. Differentially expressed lectin gene homologues

Gene and genome duplications, found in all organisms, results in the production of gene families, which allows for a diversification of gene function (Sparvoli *et al.*, 2001). Several legumes have been found to contain more than one legume lectin, which may be expressed at different times and locations within the plant (Talbot and Etzler, 1978; Etzler, 1985). The legume lectin proteins are highly similar, despite their varied functions and differing localization in the plant, suggesting that they have important biological roles and that they are under selective pressure to stay conserved (Rüdiger, 1984; Spilatro *et al.*, 1996). For example, *Dolichos biflorus* contains two legume lectin genes, *DB58* and *DBSL*, the former expressed in the stem and leaf, and the latter exclusively in the seed (Harada *et al.*, 1990). These genes have over 90% nucleotide sequence identity in both the coding and 5' and 3' flanking regions of the genes. However, the *D. biflorus* seed lectin promoter contains a 116 bp region which is not present in the stem and leaf lectin promoter and which is thought to be the cause of the differential expression of these genes (Harada *et al.*, 1990). The genes are found within 3kb of each other and appear to be an example of where the duplication of a gene, including the 5' and 3' flanking regions, allowed for the evolution of a new gene function following the mutation of the promoter region of one of the genes (Harada *et al.*, 1990).

In soybean, four classical legume lectins have been found; *Le1* (SBL - SoyBean Lectin, also know as phytohaemagglutinin or agglutinin) (Goldberg *et al.*, 1983; Vodkin

et al., 1983), *Le2* (Goldberg *et al*., 1983; Vodkin *et al*., 1983), *Le3* (SVL/LE3) (Spilatro *et al*., 1996; Strömvik *et al*., 2004) and *Le4* (Strömvik *et al*., 2004). The soybean lectin *Le1* gene and promoter region has been sequenced and shown to be seed-specific in several studies (Lindstrom *et al*., 1990; Cho *et al*., 1995; Philip *et al*., 1998; Philip *et al*., 2001). The gene product is coded by a 857bp intron-less gene (*Le1*) and forms a 120 kDa tetrameric glycoprotein (Vodkin and Raikhel, 1986; Cho *et al*., 1995). *Le1* mRNA begins to accumulate during early maturation and begins to decrease at the late maturation stage during embryogenesis (Goldberg *et al*., 1981; Goldberg *et al*., 1989). Constructs using 1.7 kb of the *Le1* promoter, fused with the GUS reporter gene, and 325bp of the *Le1* 3' UTR (untranslated) sequence showed expression in the developing cotyledons of 30 day old tobacco embryos (Cho *et al*., 1995). A developmental series of transgenic tobacco seeds from 0 to 6 days post emergence from the seed coat showed strong *Le1* promoter driven GUS expression in the cotyledons, which had disappeared by day 6 (Philip *et al*., 1998). Using GFP in the place of the GUS gene, this seed-specific pattern was maintained in transgenic *Arabidopsis thaliana* plants, where the *Le1* promoter directed GFP expression in the protein storage vacuoles (Darnowski and Vodkin, 2002). Assays examining the concentration of lectin in soybean sprouts parallelled this timeline of lectin concentration in soybean seedlings. In addition, a quantitative ELISA assay found 0.05±0.002 mg/g of sprout dry weight of LE1 protein in 10-day old seedlings (which were 19.1 – 22 cm long) (Rizzi *et al*., 2003). A related homologue, *Le2*, was reported from soybean genomic DNA (Goldberg *et al*., 1983). However, because of a frameshift mutation, resulting in a premature termination, and the scarcity of *Le2* mRNA, it has been thought to be a non-expressed pseudogene (Goldberg *et al*., 1983; Vodkin *et al*., 1983). Recently however, a

single *Le2* transcript clone was identified in a cDNA library constructed from mRNA of etiolated seedling shoot tips (Strömvik *et al.*, 2004). The *Le3* soybean lectin sequence was found in EST (Expressed Sequence Tag) data and is 56% identical to *Le1* at the protein level (Strömvik *et al.*, 2004). Only 21 amino acids have been sequenced and published for the *s*oybean *v*egetative *l*ectin (SVL) but the corresponding region in the predicted peptide sequence for *Le3* is 100% identical, suggesting that *Le3* is indeed the gene for SVL (Strömvik *et al.*, 2004). Both immunoblot and electronic Northern assays have shown that LE3/SVL is located in vegetative bud, leaves, petioles, stems, cotyledons of seedlings and relatively highly in floral meristem tissue, but not in seeds (Spilatro *et al.*, 1996; Strömvik *et al.*, 2004). In addition to this, SVL/LE3 protein production has been induced by removing sink organs such as seed pods and by phloem girdling, and has been the first lectin shown to be induced by treatment with low levels ($4\mu L\ L^{-1}$) methyl jasmonate (Spilatro *et al.*, 1996). As with the other lectins, the physiological function of SVL/LE3 is not yet determined, however it has been suggested that lectins in vegetative tissues are involved in plant defense, carbohydrate metabolism, packaging of seed storage proteins and stress physiology (Rüdiger, 1984; Spilatro *et al.*, 1996). Sequences representing mRNA from a fourth, less well-studied vegetative lectin gene called *Le4*, have recently been found among sequences from cDNA libraries constructed from mRNA of stem and seedlings (Strömvik *et al.*, 2004). *Le4* is more similar to soybean lectin *Le3* in sequence than to either *Le1* or *Le2* and its physiological role has yet to be established. The soybean lectin genes *Le2* and *Le1* are more similar to each other than either is to *Le3* or *Le4* (Strömvik *et al.*, 2004).

Like the lectin genes of *D. biflorus*, the protein coding sequences for the soybean lectin genes are similar, while their expression profiles are distinct from each other. Of the soybean lectin, only the *Le1* promoter has been sequenced and is known to be specifically activated in the seed (Goldberg *et al.*, 1983), while the expression profiles of *Le2* and *Le3* have yet to be verified experimentally.

2.3 Plant Promoters

2.3.1 Promoter basics

Promoters are sequences upstream of a coding region for a gene and together with transcription factors (proteins) they regulate gene transcription levels and patterns (profiles). The promoter is as important to the function of the gene as is the coding region (Wray *et al.*, 2003). The structure of promoters is less strict than that of coding regions in the genome, however, it typically consists of a group of "control motifs" around a basal or core promoter, which is the initiation site of RNA polymerase II (which transcribes protein coding genes into mRNA) (Wray *et al.*, 2003; Potenza *et al.*, 2004). A simplified diagram of a promoter region can be seen in Figure 2.1. The length of a promoter region for a given gene can vary widely and how far upstream or downstream it extends cannot easily be delimited, but has been estimated to be approximately 1-4kb (Rombauts *et al.*, 2003; Shahmuradov *et al.*, 2005). Examples of characterized promoter sequences can extend to 4.0kb although this seems to be at the discretion of the researchers, and shorter sequences, around 2.5kb more often used in reporter gene analysis (Kluth *et al.*, 2002; Potenza *et al.*, 2004). In soybean, evidence shows that specific motifs can reside up to

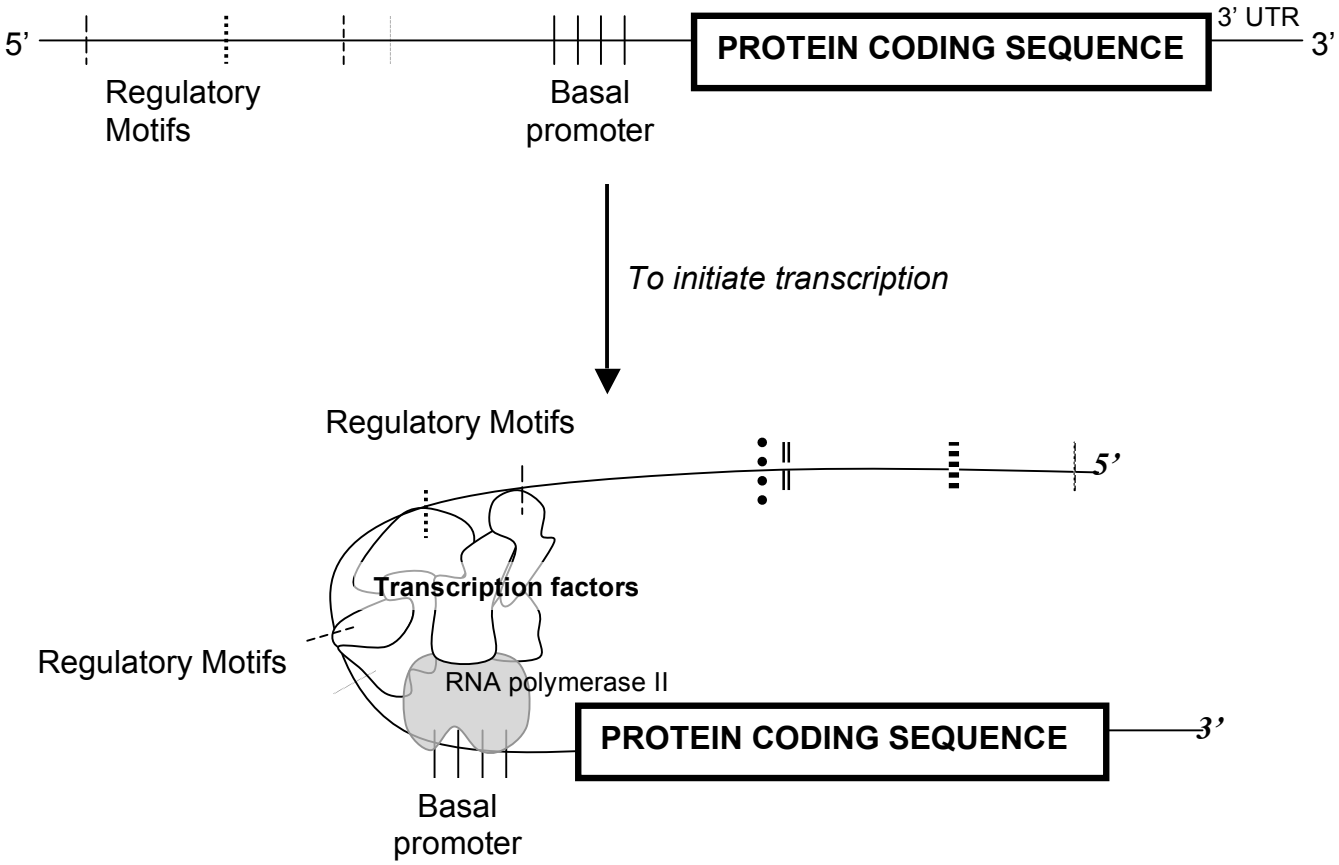


Figure 2.1: Model of transcription initiation.

A simplified model of transcription initiation, using the enhancer mechanism model is shown. Transcription factors coupled with RNA polymerase II bind to core promoter DNA elements, (e.g. TATA box) forming an initiation complex to begin transcription. Regulatory motifs further 5' in the DNA bind enhancer DNA binding proteins (transcription factors) forming an enhancer complex, which is brought close to the initiation complex by DNA looping. (Figure loosely based on (Potenza *et al.*, 2004))

2kb 5' in a promoter region (Strömvik *et al*., 1999), but also that a 190bp region can be enough for seed-specific expression (Lindstrom *et al*., 1990).

A common basal promoter element is the TATA box, which is found between -25 and -30bp upstream from the transcription start site. This is the binding site of the TATA-Box Binding Protein (TBP), which in turn recruits the RNA polymerase II complex. In dicots, the TATA-box is a 6 to 8bp motif of Ts and As, with a consensus sequence of TATAA/TA (Joshi, 1987; Grace *et al*., 2004). Some promoters do not have a TATA box (called TATA-less promoters) (Ohler and Niemann, 2001; Wray *et al*., 2003). As many as 50-70% of known promoters contain no TATA-box in the 45-25bp region upstream of the transcription start site (Shahmuradov *et al*., 2005). Although a basal promoter is essential, without additional control motifs or transcription factor binding sites, the gene expression would be insignificant or non-specific (Wray *et al*., 2003; Potenza *et al*., 2004). These transcription factor binding sites are most commonly called "enhancers" in literature, since they originally defined areas that raised transcription levels in a position and orientation independent fashion (Atchison, 1988). However, DNA regions in the genome may also repress transcription by methylation. The methylation of cytosine residues in DNA affects protein-DNA interactions which has an impact on the expression of the gene and the levels of DNA methylation in different stages during development vary (Finnegan *et al*., 1998). A DNA methylation map of the *Arabidopsis* genome revealed that pericentromeric heterochromatin, repetitive sequences, and regions producing small interfering RNAs were heavily methylated, but also that over a third of expressed genes were methylated in the transcribed region (Zhang *et al*., 2006).

Because of their various functions, areas that produce an effect on the transcription profile may also be called boosters, activators, insulators, repressors, locus control regions, upstream activating sequences or *cis*-elements (CAREs, *cis-a*cting *r*egulatory *e*lements), but in this thesis, they will be grouped under the term "motif".

These motifs are found in the promoter region, but may also be found farther upstream, downstream or within introns (Potenza *et al.*, 2004). The size, number and location of motifs in a promoter vary and they can still potentially influence transcription of at least one locus found hundreds or thousands of base pairs away (Wray *et al.*, 2003; Potenza *et al.*, 2004). In addition to this, mechanisms within the genome itself may also control the motif. For example, insulator sequences act as limits wherein the promoter can function (Nagaya *et al.*, 2001). The basal promoter may also selectively interact with certain types of enhancers over others. Lastly, transcription factor complexes found far from the basal promoter can be selectively recruited by 5' regions located before the basal promoter in a process known as selective tethering (Wray *et al.*, 2003). The different combinations of the transcription factors interacting with these motifs result in a variety of expression patterns. This is known as combinatorial control, meaning the expression of the gene is not only controlled by the types, quantities, locations and positions of motifs in the promoter region of the gene, but more by the interaction of these motifs with their associated transcription factors (Singh, 1998; Potenza *et al.*, 2004).

In this way, promoters are naturally highly tailored to the role of the gene they regulate, to give constitutive (expressed all times, every tissue), inducible (expressed in response to stress or signals not normally present in the plant), tissue-specific (expressed

in specific tissues), cell-specific (expressed in specific cell types only), and organelle-specific expression (expressed in specific organelles only) (Potenza *et al.*, 2004).

3.3.2 Plant and viral promoters in genetic engineering

Promoters used in plant transformation research may be viral, plant or synthetic in origin (Potenza *et al.*, 2004). Because of the lack of availability of specific promoters, transgenic plants are commonly produced with genes regulated by constitutive promoters. The cauliflower mosaic virus (CaMV) 35S promoter is the most frequently used promoter, as it can give high transgene expression in dicots as well as monocots (Odell *et al.*, 1985). Although CaMV 35S contains two important domains for specific gene expression in certain tissues, like most viral promoters, it is used for constitutive over-expression of a gene in all regions of the transformed plant (Odell *et al.*, 1985; Lam and Chua, 1989; Benfey and Chua, 1990). However, because of the viral origin of the promoter sequence, plant cells may be able to recognize the region as foreign and silence their activity (Elmayan and Vaucheret, 1996). This type of gene silencing may be avoided if promoters from plants are used instead (Potenza *et al.*, 2004). Often, constitutive plant promoters are taken from genes such as actin (McElroy *et al.*, 1990) or ubiquitin (Toki *et al.*, 1992), which are needed in all cell types.

Particular transgenes that are over-expressed in tissues or during developmental times at which they would not normally be present may give unexpected artificial results. For example, the over-expression of the ERF-1 (*e*thylene *r*esponse *f*actor 1) protein in *Arabidopsis* actually led to increased susceptibility to the *Pseudomonas syringae tomato* DC3000 pathogen, as compared to plants with normal protein expression levels, possibly

because of interference with the salicylic acid defense pathway (Berrocal-Lobo *et al.*, 2002).

A study of pineapple transformed with the constitutive promoters, maize Ubi1 (driving *bar* gene), OCS-35S CaMV-rice actin I (driving class-1 bean *chitinase* gene), and CaMV35S (driving tobacco *ap24* gene), showed decreased levels of aldehydes, and changes in the levels of total chlorophyll and phenolics (free and cell-wall linked) (Yabor *et al.*, 2006). Because of the relationship between aldehydes and stress tolerance, as well as chlorophyll and photosynthesis efficiency, there may be unintended, and undesirable effects in the plant. In this case, a tissue or developmentally-specific promoter may be more desirable than a constitutive promoter.

Typically, plants over-expressing pathogen-resistance genes show greater resistance when infected/inoculated with the pathogen, however this is not always the case. Disease resistance was shown in earlier studies to correlate with high constitutive levels of scopoletin and scopolin in hybrid *Nicotiana debnei x Nicotiana glutinosa*, but transgenic plants that had lower levels of these two compounds had higher sensitivity to tobacco mosaic virus (Goy *et al.*, 1993; Chong *et al.*, 2002). In another study, transgenic tobacco plants over-expressing TOGT (tobacco glucosyltransferase), using the CaMV35S promoter, contained higher levels of scopoletin and scopolin than control plants, however, they did not show increased resistance to the virus, and may have shown decreased resistance (Gachon *et al.*, 2004). If an inducible promoter could be used, defense genes would be expressed only when the plant is attacked, which is more representative of what actually occurs in plants, and would perhaps stop problems seen with the over-expression of defencse-related genes.

In some situations, complete gene silencing or over-expression both give undesirable results. In the case of the EARLY RESPONSIVE TO DEHYDRATION 15 (ERD15) gene, overexpression in *Arabidopsis* reduced sensitivity to abscisic acid, resulting in less drought tolerance, less freezing tolerance, but increased resistance to a bacterial necrotroph (Kariola *et al*., 2006). RNAi silencing of the ERD15 protein produced plants that were hypersensitive to abscisic acid, but more tolerant to drought and freezing. This protein was postulated to be a mediator of stress-related abscisic acid signaling in *Arabidopsis*, therefore any modifications in its expression would have to be more tightly controlled than a generalized "over-expression" or "no expression" (Kariola *et al*., 2006). When the expression is modified of important compounds such as ABA (abscisic acid), which is involved in many plant pathways, a promoter that simply and generally increases or decreases expression is not nearly as useful as one that can specifically control expression times and locations.

Results of this nature demonstrate a need for promoters that target specific areas in the plant. In addition to this, the use of inducible and/or organ-, tissue, or cell-specific promoters in transgenic plants would show a more controlled design of GMOs (genetically modified organisms), which would be more readily accepted by the general public (Potenza *et al*., 2004). The use of promoters that target specific areas in the plant requires research of the plant that is to be transformed. Unlike constitutive promoters, such as CaMV 35S that would give expression both in monocots and in dicots, a plant promoter from one plant with a specific pattern of expression may give a different pattern of expression when transformed into a different plant (Strömvik *et al*., 1999). This was the case for a study, which showed that there were differences in promoter requirements

between monocots and dicots for the expression of the same gene, *rbcS*, which is common in all plants (Schäffner and Sheen, 1991). Researchers concluded that there would be numerous molecular differences between monocots and dicots when it came to transcriptional regulation, RNA splicing and developmental patterns (Schäffner and Sheen, 1991), which are all important factors to consider when deciding on promoters to use in the making of a transgenic plant. In addition to this, when the promoter of a common seed storage protein in cotton, α-globulin B, was transformed into cotton, *Arabidopsis* and tobacco plants, the expression of the reporter gene was highly varied between the different species (Sunilkumar *et al*., 2002). The need for a diverse array of specific promoters required for use in transgenic plants is further shown by cases where the insertion of a transgene with homologous promoters in the plant could led to gene silencing of either the plant gene and/or the transgene (Vaucheret *et al*., 1998; Sunilkumar *et al*., 2002).

2.3.3 Synthetic promoters

To increase the control over the transgene expression, synthetic promoters have also been developed. To study how the promoter region worked, deletion series were made from the promoter regions of viral sequences, such as the CaMV35S (Odell *et al*., 1985; Pauli *et al*., 2004) and the mannopine synthase gene promoter in sunflower crown gall (DiRita and Gelvin, 1987) or plant gene sequences, such as root specific phosphate transporters in *Medicago truncatula* (Xiao *et al*., 2006). The expression profiles of the transformed plants gave clues as to how the promoter functioned. Naturally, this was followed by attempts to modify the promoter regions. The CaMV35S promoter, which

was already well-studied, was transformed into tobacco with either single, or double copies of the 250bp upstream promoter sequences, and was found to give higher expression when duplicated (Kay *et al*., 1987). Combinations of promoters from different organisms were also made. Synthetic promoters are made by combining a core promoter with repeats and or combinations of motifs that can function without their natural core promoter (Potenza *et al*., 2004; Venter, 2007). Often, the CaMV 35S basal promoter is used in conjunction with enhancer motifs such as the tobacco mosaic virus (TMV) omega sequence, which can increase gene transcription in eukaryotes as well as prokaryotes (Gallie and Walbot, 1992). However, sequences originating from viruses are not the only ones used. For example, in one of the first examples of a synthetic plant promoter, a 36bp upstream fragment of the soybean hsp17.3-B gene containing two partly-overlapping heat shock elements, gave heat-inducible reporter gene expression in transgenic tobacco, but expression was not organ-specific and was unaffected by light levels. However, once this 36bp region was inserted into the pea rbcS-3A 5'region, the expression of reporter gene became both light-inducible and organ-specific: creating a new and unique expression pattern as compared to the expression profiles of the two wild-type promoter regions used (Strittmatter and Chua, 1987).

Gradually, promoters have become much more targeted in their design to have more precise control over expression. Promoters have been found that are bidirectional (Li *et al*., 2004), inducible by pathogens (Hong *et al*., 2005; Yevtushenko *et al*., 2005), or by specific chemicals or compounds (Xu and Timko, 2004; Morikami *et al*., 2005).

Stress and defence motifs function to restrict the expression of certain proteins to only when and where they are needed, so as to conserve the energy of the plant. Specific

transcription factors are induced by signals, such as fragments of the cell wall, that are related to the stress, and bind to matching motifs in promoters for stress related genes. For example, plants use phenylpronoid compounds as a defense against pathogenic fungi. The carrot phenylalanine ammonialyase (PAL) gene *DcPal1* plays an important role in the allocation of energy between the primary and phenylpropanoid metabolism. The promoter region of the *DcPal1* gene contains several Box-L-*like* motifs (ACC(A/T)(A/T)CC) which have been shown to be critical in the activation of this gene, through site-directed mutagenesis of the Box-L-like motifs (Maeda *et al*., 2005). Using purified cell wall fragments from the pathogenic fungus *Chaetomium globosum* Kinze, a transciption factor found in carrot (*Daucus carota*), *DcMYB1*, was induced and found to bind to these Box-L-like motifs, which induced the expression of *DcPal1* (Maeda *et al*., 2005).

Similarly, environmental conditions can also induce the expression of certain proteins such as heat, cold, high salinity and drought. Drought stress has a strong impact on plant functions, therefore it is important to identify and react to the loss of water. In the case of *Arabidopsis*, when drought conditions are detected, abscisic acid (ASA) is synthesized, which then induces the *Atmyb2* gene (Abe *et al*., 1997). The *Atmyb2* protein is a transcription factor that binds the consensus I MYB DNA binding site (C(G/C)GTT(G/A)). This sequence has been found in the 5' promoter sequence of the *Arabidopsis* drought-responsive gene *rd22* and shown to be involved in its drought-induced transcription (Urao *et al*., 1993; Solano *et al*., 1995; Abe *et al*., 1997).

Because of the nature of promoters to contain multiple motifs that may or may not be functional in that promoter, as well as the effect of the motifs' position and the

interdependent nature of motifs regarding gene expression, it is difficult to predict gene expression using the promoter sequence without experimental confirmation. Although there are better promoter prediction tools available for non-plant sequences, work is being done to better analyze the structure and sequence of plant promoters using bioinformatics tools (Shahmuradov *et al.*, 2005). To increase the understanding of promoter regulatory sequences, the fully sequenced *Arabidopsis* genome is often used as a model. In one study, microarray experiments were used to find motifs that showed correlation with at least two related abiotic stresses such as heat and hydrogen peroxide (Geisler *et al.*, 2006). After using the *in silico* analysis to predict motifs, three GUS reporter constructs were made using short promoter fragments from the native promoter and stably transformed into *Arabidopsis* cell suspension cultures. Promoter induction was indicated by the blue stain indicative of GUS activity. In each construct, researchers were able to confirm an *in silico* prediction of an inducer for the respective promoter, although false predictions were also made (Geisler *et al.*, 2006). This shows that the use of bioinformatics tools can greatly increase the speed and efficiency of motif prediction.

In general, even when using well-known and well-studied promoters in plant transformation, the level of transcription can vary because of different synergistic promoter effects, the position effects by the gene insert location, and 3' UTR (untranslated) sequences (mRNA stability). Differences in transcription factors between the plants or tissues is of immense importance, although many transcription factors and their expression patterns are highly conserved because of their importance and effect on other genes. Mutating one transcription factor affects transcription of all the associated genes, which is far more likely to be harmful to the organism than the effect of mutating

individual gene promoter sequences, which would most likely affect the expression of that gene only (Doebley and Lukens, 1998; Wray *et al.*, 2003).

2.3.4 Promoter evolution

There are several examples in both the plant and animal world of protein and promoter function conservation over millions of years, as well as conservation of protein function and changes in promoter function (Doebley and Lukens, 1998). In a recent study, researchers studying the genomes of *Caenorhabditis elegans* and *C. briggsae* found that the evolution of protein and regulatory sequences was only somewhat linked in orthologous sequences (homologs derived by speciation and with conserved function), and not linked in paralogous (homologs derived by duplication and without preserved function) sequences (Castillo-Davis *et al.*, 2004). This showed that promoter sequences duplicated because of speciation were more likely to remain conserved that those duplicated within the same genome (Castillo-Davis *et al.*, 2004).

As early as 1969, an evolutionary model of change was proposed that suggested areas of the genome that regulated structural genes were more likely to evolve than the structural genes themselves (Britten and Davidson, 1969; Doebley and Lukens, 1998). This seems to be more often the case with plant genes. As mentioned earlier, plant genes are more likely to be duplicated within the genome than animal genes, which are more likely to undergo alternative splicing (Hofer and Ellis, 2002; Kazan, 2003). In addition to the less stringent conservation constrains on promoter sequences, their alterations would be less likely to seriously disrupt the development of the plant than alterations in coding regions (Doebley and Lukens, 1998).

The phosphoribosylformyl-glycinamide (FGAM) synthase gene in soybean, *FGAM1*, was studied based on its location in a genomic interval that is up-regulated during nematode feeding (Vaghchhipawala *et al.*, 2004). A homologue, also in soybean, *FGAM2*, was found to have 95.5% sequence identity between the open reading frames, and 85% similarity in approximately 2.5kb of the promoter sequence, which suggested a relatively recent gene duplication. Promoter analysis was carried out using transgenic *Arabidopsis thaliana* inoculated with the soybean cyst nematode (*Heterodera glycines*). Based on their results, they determined that the proteins were differentially expressed, although they perform duplicate functions. *FGAM1* has a function as a housekeeping protein, while *FGAM2* was induced by environmental stimuli (nematode) (Vaghchhipawala *et al.*, 2004).

2.4 Methods for promoter isolation and dissection

There are different experimental techniques to identify promoters including promoter trapping, genome walking, and deletion studies.

Random promoters can be found using a technique known as promoter trapping where a promoterless reporter gene, such as luciferase or *gusA*, is randomly inserted (using T-DNA or transposons) into the genome followed by detection of reporter gene activity (Springer, 2000; Alvarado *et al.*, 2004). The drawback of promoter trapping is that cryptic promoters can be found in intergenic regions. Cryptic promoters are DNA sequences that can promote gene expression but that are not adjacent to a coding gene in the genome (Alvarado *et al.*, 2004).

Promoters from genes with known coding sequence can be identified and sequenced by "walking" along the genome using kits such as the commercially available Universal Genome Walker kit (Clontech Inc.). The procedure involves digesting genomic DNA with a restriction enzyme and ligating the fragments to adaptors, provided with the kit. A primary PCR reaction is performed with a primer for the adaptor and a primer from within the coding region of the gene. A secondary PCR reaction with nested primers is performed using the primary PCR as a template. PCR fragments that are larger than the distance from the start codon to the secondary primer location are ones that contain the 5' region of the gene and are cloned and sequenced. A flowchart of the GenomeWalker procedure can be seen in Figure 2.2.

Once a promoter region has been sequenced, there are several methods for further analysis. Because motifs within the promoter can be found thousands of base pairs from the start of transcription, as well as just upstream, downstream, or inside exons or introns (Potenza *et al.*, 2004), it can be difficult to accurately determine the boundaries of promoters. The effect of the promoter region sequenced must be tested in several ways, such as with promoter fusion analysis with a reporter gene, mRNA expression analysis, and bioinformatic analysis.

Deletion studies are the most classical methods of promoter analysis and are useful in plant studies where transgenic plants (for example, *Arabidopsis*, tobacco) can be made easily and most cells can be studied throughout the life cycle of the plant (Benfey and Chua, 1990). Progressive deletions of the promoter are cloned together with reporter genes and tested in transgenic plants (Benfey and Chua, 1990). Deletion studies done on the soybean *Le1* promoter showed that sequence motifs found in the soybean *Le1*

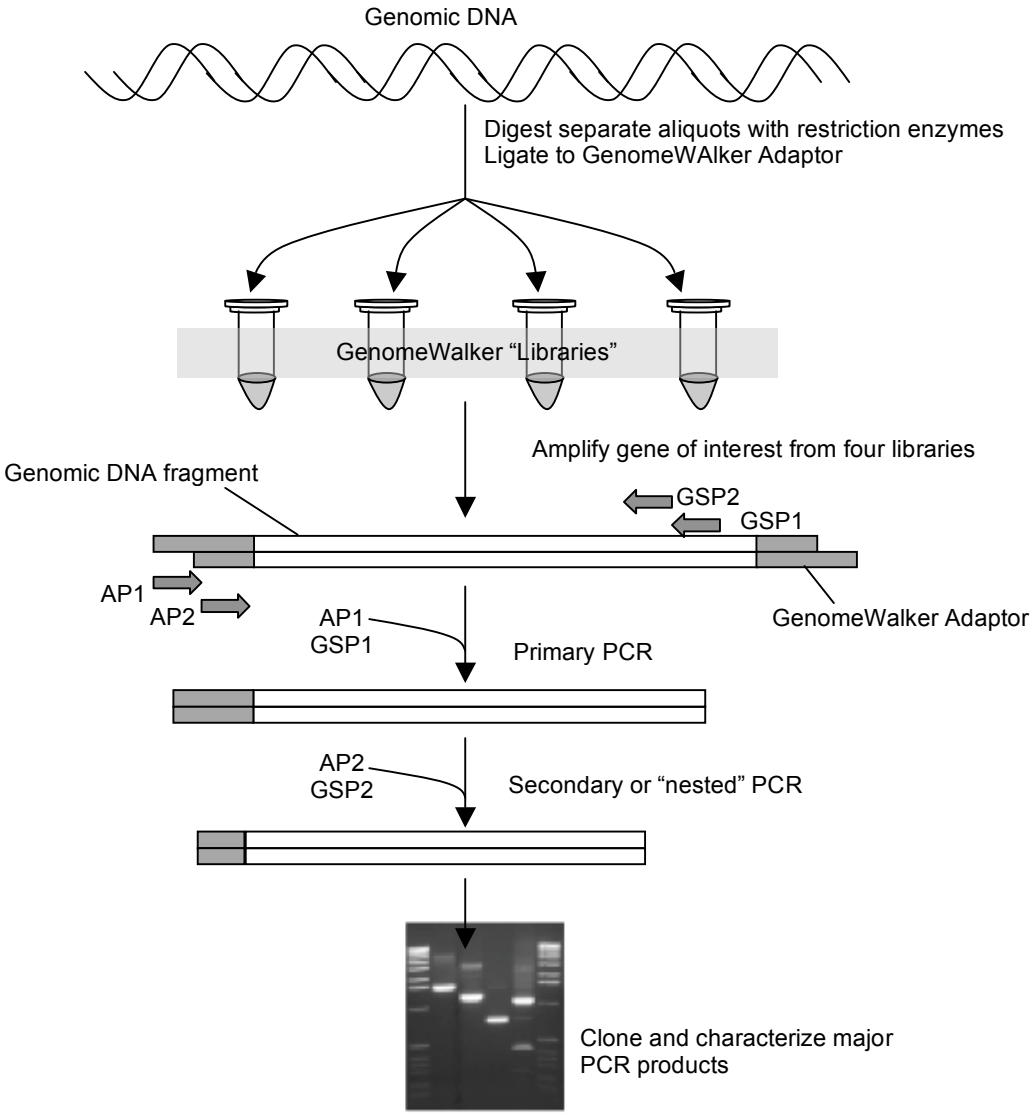


Figure 2.2: Flowchart of BD Biosciences GenomeWalker kit protocol.

Four libraries are made by digesting genomic DNA with four different restriction enzymes to create blunt ended fragments, which are ligated to GenomeWalker adaptors. A primary PCR is carried out using an adaptor primer (AP1) and a gene-specific primer (GSP1). A secondary PCR is done on the primary PCT using a nested adaptor primer (AP2) and nested gene-specific primer (GSP2). A single strong band should be seen in at least one library after the secondary PCR (GenomeWalker protocol, August 2004).

promoter remained active when transformed into another plant species (tobacco) (Lindstrom *et al.*, 1990). A series of constructs containing varying sizes of 5' and 3' sequence showed that a small amount of the 5' and 3' sequence (-190 bp and +194 bp respectively) were required for expression, but optimal expression was achieved between -338 bp and –700bp of the 5' region, and the largest construct, with 3000 bp of 5' and 1500 bp of 3' actually showed lower expression. The large construct may have shown lower expression due to the presence of suppressor elements further upstream that were cut out in the shorter constructs (Lindstrom *et al*., 1990).

A deletion study of the CaMV35S promoter identified the minimal sequence required for sufficient transcription of a reporter gene (Odell *et al.*, 1985). As a whole, this sequence showed constitutive expression (Odell *et al*., 1985; Benfey and Chua, 1990), but an additional, more extensive deletion study done on the CaMV promoter showed that the promoter could be split into subdomains, which contained specifically organized motifs (Benfey and Chua, 1990). These motifs led to different expression patterns when combined as opposed to when they drove expression alone, thereby demonstrating the existence of a synergy between different motifs, as well as the fact that different combinations of motifs showed different patterns of expression in two different plants tested (tobacco and petunia) (Benfey and Chua, 1990).

Although studies are usually done by deleting the promoter from the 5' end, a study on the soybean gene *Msg* showed that a complex developmental pattern could also be achieved by a promoter with or without the proximal 650bp, including the TATA box, before the start of transcription (Strömvik *et al.*, 1999). Researchers made a conventional deletion series, with the promoter region successively shortened from the 5' end, as well

as a more unconventional deletion series that eliminated 650bp of the 3' end. This study showed that the 650bp fragment did not affect the pattern of expression of the reporter gene, but did help to increase the level of expression and that the tissue specific motifs resided far 5' (Strömvik *et al.*, 1999).

2.5 Methods for Plant Transformation

Studies performed on plant promoters invariably involve plant transformation to test the expression of the modified promoter in either the plant from which the promoter was isolated, or in a different species. While some plants can be transformed relatively easily, others require more effort and optimized protocols are still lacking.

Arabidopsis is easily transformed using *Agrobacterium tumefaciens* transformed with a binary T_i vector containing the gene and/or promoter of interest. The "floral dip" method of transformation is used, whereby developing floral tissues are submersed in a solution containing the *A. tumefaciens*, and the seeds from those tissues are grown on selective medium to isolate the transformed plants (Clough and Bent, 1998).

The transformation of legumes has been done using direct DNA transfer methods such as microinjection (Reich *et al.*, 1986), electroporation (Akella and Lurquin, 1993) and microprojectile bombardment (Klein *et al.*, 1987; Christou *et al.*, 1990). Despite a great deal of research done on soybean, like many of the legumes, it remains difficult to transform (Somers *et al.*, 2003). Although non-tissue culture transformations (i.e. floral dip) of soybean have not been reported, successful transformations have been achieved using microprojectile bombardment (Christou *et al.*, 1988), and from regenerated shoots

from the cotyledonary node or other meristematic explants, following *Agrobacterium* infection (Olhoft *et al*., 2001; Somers *et al*., 2003).

2.6 Reporter Gene Expression

The term "reporter gene" refers to a variety of genes which are used to test transformation since they can easily be detected in the plant. A variety of these types of genes are available, but the most useful are the GFP or GUS proteins (Springer, 2000).

GFP (*g*reen *f*luorescent *p*rotein) is a fluorescent protein from the jellyfish *Aequorea victoria* that is detected by fluorescence (λ_{max}=508-515nm) following illumination (λ_{max}=470nm), without the use of a substrate. Because of this, it can be detected in live tissue over time without killing the cells (Springer, 2000). However, an appropriate light source is needed to detect the protein (Springer, 2000). In 1997, it was modified so that it could also be expressed in plant tissues (Haseloff *et al*., 1997).

The GUS reporter protein comes from the bacterial (*E. coli*) *gusA (uidA)* gene and encodes β-glucuronidase (Springer, 2000). Like GFP, it remains stable when it is fused to other proteins (Jefferson *et al*., 1987). GUS activity is detected by histochemical staining of transformed tissues submersed in substrate buffer (Alvarado *et al*., 2004). The assay is destructive (kills the tissue), however, the high level of sensitivity in detecting the GUS activity allows for the analysis down to single cells (Jefferson *et al*., 1987; Springer, 2000).

2.7 Bioinformatic Analysis of Promoter Regions

Once the promoter region has been sequenced, to find the actual promoter motif, computer programs/algorithms are used. This involves several computer programs and databases available, such as PlantCARE (Lescot *et al.*, 2002) and PlantProm (Shahmuradov *et al.*, 2003), to look for known motifs in the sequence. Motifs in these databases have been known to direct specific types of expression or are otherwise common in promoters (for example, the TATA box). These bioinformatics tools save time by automatically looking for a large number of specific sequences. More specialized computer algorithms can also help to predict novel conserved motifs. This is difficult to do manually (by eye) since the promoter motifs can vary in size (5-15bp), may occur randomly, in different combinations, duplications and orientations. They may also be structural, detected only by looking at the secondary or tertiary levels. Furthermore, the motif sequences may differ from the consensus of the motif. Many factors make motif identifications a challenge, and often false positives are found and some motifs can be missed (Rombauts *et al.*, 2003).

A commonly used database for locating plant *cis*-acting regulatory elements, enhancers and repressors, is PlantCARE (Rombauts *et al.*, 1999; Lescot *et al.*, 2002). At present, their website (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/) states that it contains 435 different names of plant transcription sites describing over 159 plant promoters (as of March 23, 2007). After entering a promoter sequence, the description for specific transcription factor sites as well as confidence level for the experimental evidence are given (Lescot *et al.*, 2002).

Another large online database of plant promoter sequences is the PlantProm database (Shahmuradov *et al.*, 2003). This is an annotated, non-redundant collection of experimentally determined promoter sequences located several hundred nucleotides from the transcription start site. The first release (2002.01) contained 305 entries, but currently contains 1211 regulatory elements (http://mendel.cs.rhul.ac.uk/, http://softberry.com/berry.phtml?topic=plantprom&group=data&subgroup=plantprom) (Shahmuradov *et al.*, 2003).

A third database of nucleotide sequence motifs, which searches for plant promoter motifs, is the PLACE database (*pla*nt *c*is-acting regulatory DNA *e*lements) (Higo *et al.*, 1999), which contains 451 entries. Like the databases previously mentioned, motifs were collected from reports and article reviews on regulatory regions that had been published earlier. Motif variations in other genes or plant species were also added to the database (Higo *et al.*, 1999). Motifs are given unique identifiers and accession numbers and results from a query include the accession number of the motifs found, access to PubMed to find an abstract of the literature and access to the GenBank annotation (Higo *et al.*, 1999).

Although it is not a database of plant promoters, the TIGR plant repeat database can also be useful. Repetitive sequences from 12 plant genera, one of which is *Glycine*, were collected from GenBank and coded into 5 classes: transposable elements, centromere related, telomere related, rDNA and unclassified repetitive sequences (Ouyang and Buell, 2004). Although the function of the sequences are not provided, a hit may still provide useful information if the location and expression of the other proteins with similar repeats is known.

The use of many databases is important, as each one may contain different pieces of information regarding the promoter and the motifs contained therein. In addition to this, while some newly discovered motifs may be entered onto one database and not the others so results from one database should be confirmed in the others.

2.8 Summary

Soybean is an important and widely used crop for which 61 000 genes have been predicted (Shoemaker *et al.*, 2002; Vodkin *et al.*, 2004). However less than 20 soybean gene promoter sequences are well-characterized. One of the better characterized is that of the soybean lectin *Le1* (Goldberg *et al.*, 1983; Vodkin *et al.*, 1983). Three *Le1* homologues have been found in soybean: *Le2* (Goldberg *et al.*, 1983; Vodkin *et al.*, 1983), *Le3* (SVL/LE3) (Spilatro *et al.*, 1996; Strömvik *et al.*, 2004) and *Le4* (Strömvik *et al.*, 2004). The *Le1* promoter has been shown to be seed-specific, while the expression profiles of its' homologues are unknown. However, *Le2* is thought to be a pseudogene and *Le3* is thought to be the gene coding for the soybean vegetative lectin. Unknown promoter sequences, such as those for *Le2* and *Le3*, can be isolated through various methods. The Clontech GenomeWalker kit uses the known coding region sequence to quickly isolate the promoter region of the gene, which can then be fused to a reporter gene and stably transformed into *Arabidopsis* using the floral dip method to determine their expression profile. The detection of the reporter gene in different tissues and at different developmental stages reveals effect of the promoter isolated. Using the experimental results and computer analysis, specific motifs controlling the expression profile can be found in the promoter region. The motifs can be used to predict the profile of other promoters. This would be useful in the design of promoters for genetic

engineering to avoid problems seen with the current plant, viral and synthetic promoters used.

Section 3:
Materials & methods

3.1 Isolation of soybean lectin promoter genomic clones

Soybean (*Glycine max* (L.) Merr. cv Williams 82) seedlings were germinated in autoclaved glass Petri dishes containing several sterile, wet Whatman filter papers. Genomic DNA was extracted from the seedling roots using commercially available kits DNeasy Plant Mini (Cat. No 69104) and DNeasy Plant Maxi Kit (Cat. No 69104)) (Qiagen Inc. (Canada), Missisagua, Ontario). To construct the GenomeWalker DNA Libraries (Clontech Universal GenomeWalker Kit, (Cat. No. 638904), Mountainview, California), genomic DNA was digested in separate tubes with *Dra I, Stu I, Eco RV, Sma I* and *Pvu II* and ligated to the adaptor provided with the kit. A primary PCR was performed on each library using the adaptor-specific primer from the kit, and the following lectin gene-specific primers, (Table 3.1a and 3.1b): Le2_7.rev for *Le2* 5' region, Le2_17.for for *Le2* coding and 3' region, Le3_3.rev rev for *Le3* 5' region, and Le3_11.for for *Le3* coding and 3' region. A secondary PCR was performed using the PCR product from selected libraries and using the second (nested) adaptor-specific primer from the kit, together with the following gene-specific nested primers: Le2_8.rev for *Le2* 5' region, Le2_18.for for *Le2* coding and 3' region, Le3_4.rev for *Le3* 5' region, and Le3_12.for for *Le3* coding and 3' region. The program for the primary PCR was: 7 cycles: 94°C for 2 sec, 72°C for 3min, 32 cycles: 94°C for 2sec, 67°C for 3min, 67°C for 4 min, and for the secondary PCR was: 5 cycles: 94°C for 2 sec, 72°C for 3min, 24 cycles: 94°C for 2sec, 67°C for 3min, 67°C for 4 min. To facilitate cloning, the primers

were designed with restriction enzyme site overhangs, and if needed, secondary PCR products were digested and subcloned into a pUC19 vector (Invitrogen, Cat. No. 15364011, Carlsbad, California) to obtain the following plasmids: HS3_Le2.4 (*Le2* 5'-upstream region), HS2_Le3.6 (*Le3* 5'-upstream region), HS27_Le2.4 (*Le2* 3'-downstream region) and HS136_Le3.1 (*Le3* coding region). The plasmids were transformed into *E. coli* (Top10, Invitrogen, Cat. No.C4040-10), on LB-ampicillin (100μg/ml) selection medium and positive colonies were selected using blue-white lac screening. All plasmids used for sequencing and further experiments were purified using the Qiagen plasmid mini kit (Qiagen, QIAprep Spin Miniprep kit (Cat. No. 27104)).

3.2 Sequencing and sequence analysis

Sequencing was performed at the McGill University and Genome Quebec Innovation Center, (http://www.genomequebec.mcgill.ca/), using the primers listed in Table 3.1. The sequences were edited using 4Peaks (by A. Griekspoor and Tom Groothuis, mekentosj.com. Version 1.7.2). Signal peptides were predicted using SignalP 3.0 (Nielsen *et al.*, 1997; Bendtsen *et al.*, 2004). Primers were designed by eye. Sequences were assembled on the BCM Multiple Sequence Alignments webpage, using the MAP alignment method (http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html) (Huang, 1994), and compared to lectin EST contig sequences (Strömvik *et al.*, 2004) to ensure that the right lectin gene was cloned. The *Le2* sequence will appear in GenBank under the accession no. EU070414 and *Le3* under accession no. EU070415.

Sequences were entered into the PROSITE database (http://ca.expasy.org/prosite/) to locate protein domains and functional sites (de Castro *et al.*, 2006).

Table 3.1: A. Cloning primers made to isolate promoter and coding region of *Le2* and *Le3*. B. Sequencing primers made to determine sequence of promoter and coding regions.

Primer Name	Sequence (5'-3')	# of nt*	Location (relative to start of sequence)	Location (relative to start codon)	Region	GC %	Comments
A. Cloning Primers							
Le2_7.rev	5'GTACTGACACACAACAACGTTATCTCTG3'	28	1707 <--- 1736	414 <--- 443	coding	57%	- primary primer used for cloning 5' fragment from GenomeWalker kit
Le2_8.rev	5'CTGTTTATGCGTCCGACATAGCAAATC3'	27	1598 <--- 1626	306 <--- 333	coding	55%	- secondary primer used for cloning fragment in the GenomeWalker kit - used for sequencing
Le2_9.for	5'[GGAATTCC]TCGTCACATACACTGCAATT3'	28	1273 ---> 1293	-20 ---> -1	5'	57%	- used to clone Le2 coding region - added EcoR I cut site to 5' end for cloning (bracketed)
Le2_10.rev	5'GCATCTCGCTTCTTCCTAGTG[GGTCGACC]3'	29	2148 <--- 2168	855 <--- 875	3'	41%	- used to clone Le2 coding region - added Sal I cut site to 3' end for cloning (bracketed)
Le2_13.rev	5'CTGTTTATGCCTCCGACATAGCAAATCT[CCCAAGCTTGGG]3'	39	1598 <--- 1626	306 <--- 333	coding	51%	- used to clone Le2 5' region - added HindIII cut site to the 3' end for cloning (bracketed)
Le2_14.rev	5'TCACATACACTGCAATTATG[CCAGATCTGGC]3'	31	1276 <--- 1296	-17 <--- 3	5'coding	40%	- used for cloning Le2 promoter without Le2 signal peptide - added Bgl II cut site to the 3' end for cloning (bracketed)
Le2_15.rev	5'CTCACCATGACCAAGGTAAACTCA[CCAGATCTGG]3'	34	2148 <--- 2168	66 <--- 90	coding	50%	- used for cloning Le2 promoter with Le2 signal peptide - added Bgl II cut site to the 3' end for cloning (bracketed)
Le2_16.for	5'[GGGAAGCTTCCC]AAAGTTTTATAATAATATTTAATAAA3'	38	1 ---> 26	-1293 ---> 1268	5'	38%	- used for cloning Le2 5' promoter - added Hind III cut site on 5' end for cloning (bracketed)
Le2_17.for	5'AGAGGGTGAAGGTGAGTGTTAGCTAGTA3'	28	1221 ---> 1248	-73 ---> -46	5' UTR	46%	- primary primer used for cloning coding region and 3' fragment from GenomeWalker kit
Le2_18.for	5'CCCATGCATCGTCACATACACTGCAATT3'	28	1266 ---> 1293	-28 ---> -1	5' UTR	46%	- secondary primer used for cloning coding region and 3' fragment from GenomeWalker kit - used for sequencing
Le2_21.for	5'[GGGAAGCTT]CCCATGCATCGTCACATACACTGCAATT3'	37	1266 ---> 1293	-28 ---> -1	5' UTR	46%	- used to clone Le2 coding and 3' region - added HindIII cut site to the 3' end for cloning (bracketed)
Le2_26.for	5'[CGCGGATCCGCG]GAGTGATGCCACGAGAGGAATTGAGTGG3'	40	1022 ---> 1049	-272 ---> -245	5'UTR	54%	- used to clone Le2 5' region from genomic DNA - added BamHI cut site to the 5' end for cloning (bracketed)
Le2_27.rev	5'GACTCTCGCTTCTTCCTAGTGACTG[CCCAAGCTTGGG]3'	37	2148 <--- 2172	855 <--- 879	3'UTR	52%	- used to clone Le2 3' region from genomic DNA - added HindIII cut site to the 5' end for cloning (bracketed)
Le3_3.rev	5'TCCAATGGACAAGTGGCGGAGATTCTCGT3'	29	1835<--- 1863	556 <--- 584	Coding	48%	- primary primer used for cloning 5' fragment from GenomeWalker kit
Le3_4.rev	5'ATTTTCGCACCCAACAAATCAAACTCAGCT3'	30	1589 <--- 1618	310 <--- 339	Coding	60%	- secondary primer used for cloning fragment in the GenomeWalker kit - used for sequencing

Table 3.1 continued

Le3_6.rev	5'ATTTTCGCACCCAACAAATCAAACTCAGCT[CCCAAGCTTGGG]3'	42	1589 <--- 1618	310 <--- 339	Coding	52%	- used to clone Le3 coding region - added Sal I cut site to 5' end for cloning (bracketed)
Le3_8.rev	GAGTTCAAACAAATCAAAGCCATG[CCAGATCTGGC]	35	1259 <--- 1282	-21 <--- +3	5'coding	38%	- used for cloning Le3 promoter without Le2 signal peptide - added Bgl II cut site to the 3' end for cloning (bracketed)
Le3_9.rev	5'TGCTACTTACCAAGGCACACTCG[CCAGATCTGGC]3'	34	1335 <--- 1357	56 <--- 78	coding	56%	- used for cloning Le3 promoter with Le2 signal peptide - added Bgl II cut site to the 3' end for cloning (bracketed)
Le3_10.for	5'[GGGAAGCTTCCC]ATCCCACGTGTTGAACGTGG3'	32	1 ---> 19	-1279 ---> -1261	5' UTR	55%	- used for cloning Le3 5' promoter - added Hind III cut site on 5' end for cloning (bracketed)
Le3_11.for	5'GACACAGTCATAGTCCTATCCTTGCACTA3'	29	1191 ---> 1219	-89 ---> -61	5' UTR	45%	- primary primer used for cloning coding region and 3' fragment from GenomeWalker kit
Le3_12.for	5'CACAACCCGATGAAAGTCCTATGCAT3'	26	1226 ---> 1251	-54 ---> -29	5' UTR	46%	- secondary primer used for cloning coding region and 3' fragment from GenomeWalker kit - used for sequencing
Le3_13.for	5'[GGGAAGCTTCCC]ACGTGTTGAACGTGG3'	27	6 ---> 19	-1273 ---> -1261	5' UTR	61%	- used for cloning Le3 5' promoter - added Hind III cut site on 5' end for cloning (bracketed)
Le3_14.for	5'[CGCAACGTTGCG]CACAACCCGATGAAAGTCCTATGCAT3'	38	1226 ---> 1251	-54 ---> -29	5' UTR	46%	- secondary primer used for cloning coding region and 3' fragment from GenomeWalker kit - added HindIII cut site to the 3' end for cloning (bracketed)
Le3_20.for	5'[GGAATTCC]TACGTTGTCACATTACTAGAGG3'	30	1061 ---> 1082	-219 ---> -198	5' UTR	59%	- used to clone Le3 5' region from genomic DNA - added EcoRI cut site to the 5' end for cloning (bracketed)
Le3_21.rev	5'CCTAGCTTGCTAGAGGCTGGTGCTACTA[CCCGTCGACGGG]3'	40	2176 <--- 2203	897 <--- 924	3' UTR	54%	- used to clone Le3 3' region from genomic DNA - added SalI cut site to the 5' end for cloning (bracketed)
Le3_22.for	5'[CCCGTCGACGGG]CACAACCCGATGAAAGTCCTATGCAT3'	38	1226 ---> 1251	-54 ---> -29	5' UTR	46%	- secondary primer used for cloning coding region and 3' fragment from GenomeWalker kit - added SalI cut site to the 3' end for cloning (bracketed)
B. Sequencing Primers							
Le2_8.rev	5'CTGTTTATGCGTCCGACATAGCAAATC3'	27	1598 <--- 1626	306 <--- 333	coding	55%	- secondary primer used for cloning fragment in the GenomeWalker kit - used for sequencing
Le2_11.rev	5'CCATTCTGTTACATTCTTGTCC3'	22	825 <--- 847	-470 <--- -449	5'	59%	- used to sequence Le2 5' region
Le2_12.rev	5'GTTCACAATCACAGTCTCTCGCTT3'	24	256 <--- 279	-1038 <--- -1015	5'	54%	- used to sequence Le2 5' region
Le2_18.for	5'CCCATGCATCGTCACATACACTGCAATT3'	28	1266 ---> 1293	-28 ---> -1	5' UTR	46%	- primary primer used for cloning coding region and 3' fragment from GenomeWalker kit - used for sequencing
Le2_19.for	5'TTGACACTCAGCCTCAGAC3'	19	1659 ---> 1677	366 ---> 384	coding	53%	- used to sequence Le2 coding and 3' region
Le2_23.for	5'CTCATTACCTATGATGCCTCC3'	21	1846 ---> 1866	553 ---> 573	coding	43%	- used to sequence Le2 coding and 3' region
Le2_24.rev	5'GTCAATTTTGCCAACCATGG3'	20	3852 <--- 3871	2559 <--- 2578	3'UTR	45%	- used to sequence Le2 coding and 3' region

Le2_25.for	5'CTTTGCATACCGACCC3'	16	1105 ---> 1120	-189 ---> -174	5'UTR	56%	- used to check GUS was in frame in pCAMBIA vector
Le2_28.rev	5'GAGGGTGAAGGTGA3'	14	1222 <--- 1235	-72 <--- -59	5' UTR	57%	- used to sequence Le2 5' region
Le2_29.for	5'GACGGACTTCAACTGCATG3'	19	2483 ---> 2501	1190 ---> 1208	3'UTR	47%	- used to sequence Le2 coding and 3' region
Le2_30.rev	5'GGAAGAGTCAAGAGCAGGGGT3'	21	3200 <--- 3220	1907 <--- 1927	3'UTR	57%	- used to sequence Le2 coding and 3' region
Le2_31.rev	5'GACGGACTTCAACTGCATG3'	19	2483 <--- 2501	1190 <--- 1208	3'UTR	47%	- used to sequence Le2 coding and 3' region
Le2_32.for	5'GGAAGAGTCAAGAGCAGGGGT3'	21	3200 ---> 3220	1907 ---> 1927	3'UTR	57%	- used to sequence Le2 coding and 3' region
Le2_33.for	5'GGGTATTGATGAAAATGGTG3'	20	3716 ---> 3735	2423 ---> 2442	3'UTR	40%	- used to sequence Le2 coding and 3' region
Le3_4.rev	5'ATTTTCGCACCCAACAAATCAAACTCAGCT3'	30	1589 <--- 1618	310 <--- 339	Coding	60%	- secondary primer used for cloning fragment in the GenomeWalker kit - used for sequencing
Le3_5.rev	5'TACGTTGTCACATTACTAGAGG3'	22	1061 <--- 1082	-219 <--- -198	5' UTR	59%	- used to sequence Le3 5' region
Le3_7.rev	5'CGCTATATGCTAGCTGGCATTC3'	22	179 <--- 200	-1101 <--- -1078	5' UTR	50%	- used to sequence Le3 5' region
Le3_12.for	5'CACAACCCGATGAAAGTCCTATGCAT3'	26	1226 ---> 1251	-54 ---> -29	5' UTR	46%	- secondary primer used for cloning coding region and 3' fragment from GenomeWalker kit - used for sequencing
Le3_16.for	5'CGTTGTCACATTACTAGAGGTTT3'	22	1061 ---> 1082	-219 ---> -198	5' UTR	59%	- used to check GUS was in frame in pCAMBIA vector
Le3_17.for	5'GGGTGTTAGCAATTAG3'	16	445 ---> 460	-835 ---> -820	5' UTR	44%	- used to sequence Le3 5' region
Le3_18.rev	5'GGGTGTTAGCAATTAG3'	16	445 <--- 460	-835 <--- -820	5' UTR	44%	- used to sequence Le3 5' region
Le3_19.for	5'GTGTAATGGGTATCG3'	15	892 ---> 906	-388 ---> -374	5' UTR	47%	- used to sequence Le3 5' region

* nt = nucleotide

3.3 Promoter sequence analysis

The 5' upstream regions of *Le1*, *Le2* and *Le3* were scanned for known motif sequences by analysis against the PLACE online database (http://www.dna.affrc.go.jp/PLACE/) (Higo *et al.*, 1999). The PLACE sequence file was downloaded in ascii format (http://ftp.dna.affrc.go.jp/pub/dna_place/place.seq) and unix and perl scripts were used to index the place sequence file into a mySQL database, with the following fields indexed: accession number, description, keyword, sequence, and reference in Pubmed. A perl script was written that translated each motif sequence in PLACE into regular expressions using a degenerate IUPAC code (Cornish-Bowden, 1985). Bioperl modules were used to load the lectin sequence files into sequence objects (Stajich *et al.*, 2002). Sequences were scanned with overlap and if a match to the regular expression was found, a score of one was added to the total score. For each sequence, the total score (number of overlapping motifs) was repeated in the database for each sequence separately. The total number of occurrences for each motif in each lectin promoter sequence were added to the mySQL database.

Once results were obtained, the function of the motif was determined by looking at the original and any subsequent papers studying the effect of the motif on gene expression. Expression was grouped into the following nine general categories: seed (included developing and mature), vegetative (germinating seed, vegetative tissue uninduced by anything but light), defense (jasmonic acid, ethylene, wounding, stress (biotic, drought, flooding, high salt, cold, heat), root (nodule, root), pollen/flower/fruit, etiolated, ubiquitously expressed (cell cycle), hormone induced (AUX/IAA, GA).

The number of motifs, as well as the number of times the motifs occurred were counted for each category. To account for the extra length of the *Le1* promoter, which, based on its size, had more different motifs and also motif occurrences (i.e. types of motif and how many motif sites), the number of motifs for each category was divided by the total number of motifs in that promoter. Similarly, the total number of times the motifs in one category occurred was divided by the total number of motif occurrences in that promoter. The different motifs and also motif occurrences were also divided by the total length of the promoter sequence to determine the number of each per base pair. Results were calculated using all the motifs and motif occurrences found in the promoters, as well as after excluding any motifs found in all three promoters.

The following published promoters were analyzed through the PLACE webinterface and manually inspected for motifs. The seed-specific promoters: β-phaseolin from *Phaseolus vulgaris* (accession no. J01263.1) (Slightom *et al.*, 1983), beta-conglycinin alpha subunit from soybean (accession no. AB237643.1)(Yoshino *et al.*, 2001), and globulin (*AsGlo1*) from *Avena sativa* (oat) (accession no. AY795082.1) (Vickers *et al.*, 2006). The pseudogene class I basic chitinase gene, ψ*BCH* (tomato) (accession no. AY185815.1) (Baykal *et al.*, 2006) was compared to *Le2*. The vegetative promoters analyzed were the soybean VSPα and VSPβ (accession no.'s M76981.1 and M76980.1, respectively) (Wittenbach, 1983; Staswick, 1988) and DB58 in *D. biflorus* (accession no. M34271.1) (Harada *et al.*, 1990). The number of occurences for certain selected seed-specific (G box, E box, RY motifs) vegetative (MybST1 core motifs, EPB-1 pyrimidine box), defence and stress motifs (box-L-like motif, MYB DNA binding site,

core DRE and CBF/CRT/DRE motifs) were recorded for all the promoters, including the soybean lectins.

3.4 Construction of gene fusions of lectin 5' upstream regions with the *gusA* reporter gene

All promoter constructs were made in the binary vector pCAMBIA 1391Xa (Hajdukiewicz *et al*., 1994), which contains the *gusA* reporter gene sequence (Jefferson *et al*., 1987) without the start codon, and the *nos (nopaline synthase)* 3' terminator region. pCAMBIA1391Xa also contains a hygromycin resistance (*hptII*) cassette driven by the CaMV 35S promoter in the opposite orientation as the cloned soybean promoter for plant selection. Diagrams of constructs are presented in the appendix.

The 5' region of *Le1*, with and without the signal peptide was amplified from the pGLeGUS-7 vector (Cho *et al*., 1995) (kindly provided by Dr. Lila Vodkin, University of Illinois at Urbana-Champaign) to introduce restriction sites (*Eco RI* and *Bgl II*) and then ligated to pCAMBIA 1391Xa. Plasmid pHS130_Le1.1 contains the 5' promoter region as well as the predicted signal peptide from the soybean *Le1* gene (-968 to +96bp) and plasmid pHS131_Le1.1 consists of the 5' promoter region (-968 to +3bp) from the soybean *Le1* gene from the soybean *Le1* gene.

The 5'-upstream region of the soybean *Le2* gene, with and without the signal peptide (-1293 to +90bp, and –1293 to +3bp, respectively), and for the *Le3* gene, with and without the signal peptide (–1280 to +78, and –1280 to +3) were amplified from the pHS3_Le2.4, and pHS2_Le3.6 plasmids, respectively, to introduce restriction sites (*Hind III* and *Bgl II*). After double digestion with *Hind III* and *Bgl II*, the fragments were cloned

into the pCAMBIA 1391Xa vector, digested with the same restriction enzymes, to obtain the pHS17_Le2.1 and pHS18_Le2.1 vectors (with and without the predicted *Le2* signal peptide, respectively) and the pHS19_Le3.1 and pHS20_Le3.1 vectors (with and without the predicted *Le3* signal peptide, respectively).

All six promoter-GUS fusion constructs were sequenced across the ligation site at the translational start to ensure that they were in frame, using primers shown in Table 3.1a and b.

3.5 *Arabidopsis* plant transformation using *Agrobacterium*

The unchanged pCAMBIA 1391 Xa vector was used as a negative control and the pCAMBIA1301 vector was used as a positive control. pCAMBIA1301 contains the same TDNA region as pCAMBIA 1931 Xa, but has the *gusA* gene driven by the CaMV35S promoter. The lectin promoter constructs, pCAMBIA 1391 Xa (promoterless *gusA*, negative control) and pCAMBIA 1301 (35S:*gusA*, positive control), were introduced into *Agrobacterium tumefaciens*, GV3101 strain (Koncz and Schell, 1986), by the freeze-thaw method (Holsters *et al.*, 1978). Agrobacteria were grown on 100μg/ml rifampicin/25μg/ml kanamycin LB medium. Cultures were tested by PCR to confirm presence of the desired insert using a reverse primer in the *gusA* gene (GUS_4.rev) and a promoter-specific primer (Le1_4.for, Le2_25.for, Le3_16.for, described in Table 3.1) before being used for plant transformation.

Agrobacteria containing the constructs were used to transform *Arabidopsis thaliana* by the floral dip method (Bechtold *et al.*, 1993; Clough and Bent, 1998). *Arabidopsis thaliana* (ecotype Columbia) seeds, sown on meshed pots, were incubated

for 48 hours at 4°C in the dark, and then grown in a growth chamber with a light intensity of 85-110μmoles/m^2/s for 16 hours of illumination at 22°C and 8 hours of darkness at 18°C. Humidity and carbon dioxide levels were at ambient levels. Bolting plants were cut down to stimulate more bud formation. One *Agrobacterium* colony was used to make a 10 ml LB-(100μg/ml rifampicin/25μg/ml kanamycin) overnight culture, which was transferred to 150 ml LB- (100μg/ml rifampicin/25μg/ml) kanamycin. This culture was grown 24 hours in 28°C shaking at 220 rpm. The culture was spun down at 5500rpm in a JA-14 rotor, in a Sorvall RC5B PLUSS centrifuge at 20°C for 15min and the bacterial pellet was resuspended in 250ml of transformation buffer (5% sucrose solution, 0.05% Silwet L-77). The *Arabidopsis* inflorescences were completely immersed in the bacterial suspension for 15-30 seconds. Treated plants were returned to the growing shelves but lain on their side and kept under cover to maintain humidity for one night, after which they were uncovered and allowed to grow to maturity normally. Seeds were harvested and selected on 1/2X Murashige and Skoog basal medium plates with Gamborg's vitamins (Sigma Chemicals, St. Louis, MO, USA, Cat. No. M0404), 0.8% agar (Sigma Chemicals, Cat. No. A1296), containing 50μg/ml hygromycin under the growing conditions described above. Transformed plantlets were transferred to soil in individual pots after the first set of rosette leaves emerged (approximately 8-10 days).

3.6 Detection of reporter gene expression: Histochemical GUS assay

Histochemical GUS assay was performed on the T_1 generation of flowers, cauline leaves, rosette leaves, uncut siliques and cut siliques following the method of Jefferson *et al.*, 1987, with the adjustment outlined in Stomp (Stomp, 1992). Plant tissues were

incubated at 37°C for 24 hours in the GUS assay buffer, which contained 5-bromo-4-chloro-3-indoyl-β-D-glucuronic acid (X-gluc) (Sigma Chemicals, St. Louis, MO, USA, Cat. No. B5285).

3.7 Detection of promoter impact on developmental gene expression

Histochemical GUS assay was performed on the T_2 generation to monitor GUS expression during development. T_2 seeds from three lines of each construct were plated on selection medium as described above and kept at 4°C for two days. Seeds were tested for GUS activity before being plated ("seed"), and seedlings tested at 14:00 hrs every day from the day before being placed in the growth chamber (Day 0) to 13 days after (Day 1 to 13, second set of rosette leaves emerging). At Day 13, five seedlings were transplanted to soil and, once mature, histochemical GUS assays on the T_2 generation were performed on flowers, cauline leaves, rosette leaves, uncut siliques and cut siliques following the method outlined above. The developmental series was repeated three times.

Section 4:
Results

4.1 Isolation and sequencing of the *Le2* and *Le3* genes

To clone homologues of *Le1*, gene primers specific to the *Le2* and *Le3* genes were designed and used with GenomeWalker libraries. A 1662bp band containing the *Le2* promoter region was isolated from the secondary PCR carried out on the GenomeWalker library made using *DraI*. This was cloned into a pUC19 vector and found to contain 1293bp of the *Le2* promoter region and 333bp of the coding region. A 2679bp band containing the *Le2* coding and 3'UTR regions was isolated from the secondary PCR done on the GenomeWalker library made using *PvuII*. This was cloned into a pUC19 vector and found to contain 28bp of the *Le2* promoter region and the complete coding region (804bp) and 1829bp of sequence after the stop codon that aligns with that in *Le1* and *Le3*. Early termination codons occurred at 366bp and 417bp into the coding sequence. Figure 4.2 shows the sequence of *Le2*.

A 1655bp band containing the *Le3* promoter region was isolated from the secondary PCR done on the GenomeWalker library made using *StuII*. This was cloned into a pUC19 vector and cotained 1279bp of the *Le3* promoter region and 340bp of the coding region. A 978bp region of the *Le3* coding region was isolated from genomic soybean DNA using gene-specific cloning primers. This was cloned into a pCR®2.1-TOPO® vector and contained the entire 849bp coding region and 75bp of sequence after the stop codon. The sequence of *Le3* is shown in Figure 4.3.

No introns are present in *Le1*, *Le2* or *Le3*. Since a signal peptide for *Le1* has been confirmed (shown to be from +1bp to +96bp), we predicted signal peptides for *Le2* and

Le3 are from 1bp to 90bp and 1bp to 78bp, respectively. The *Le1* protein has been shown to be 286 amino acids long (Vodkin *et al.*, 1983). The truncated *Le2* protein is predicted to be 122 amino acids long, and 268 amino acids long with the two putative read-through codons at +366bp and +417bp. The predicted full length amino acid sequence length of *Le3* is 282aa. Figure 4.4 shows an alignment of the genes and gene products.

When the LE1 sequence was entered into PROSITE, the lectin legume beta (legB) site was found at 153-159aa, and the lectin legume alpha (legA) site was found at 232-241aa. The predicted sequences for LE2 (read though codons removed) and LE3 both contained only one lectin legume alpha (legA) site each, at 218-227aa and 229-238aa, respectively.

4.2 Analysis of promoter sequence motifs

The *Le1, Le2* and *Le3* sequences were analyzed against the PLACE database (Higo *et al.*, 1999) to locate known motifs within the promoter sequence. Of the three promoters, *Le2* was found to have a slightly lower variety of motifs, but a higher frequency of each motif (1motif/20.5bp, 1 occurrence/9.6bp) as compared to *Le1* and *Le3* (1 motif/23.7bp, 1 occurrence/7.4bp and 1 motif/23.7bp, 1 occurrence/7.4bp respectively), even though no significant expression of *Le2* has been seen in soybean.

Although the *Le1* promoter contained a variety of motifs, it contained more seed-related motifs than the other two promoters (Figure 4.5). The motifs were divided according to function, and after filtering out motifs that were common to all three promoters the largest grouping of motifs in the *Le1* promoter were the seed-specific ones. No other motif grouping in *Le1*, such as the stress and defense, root, or hormone-related

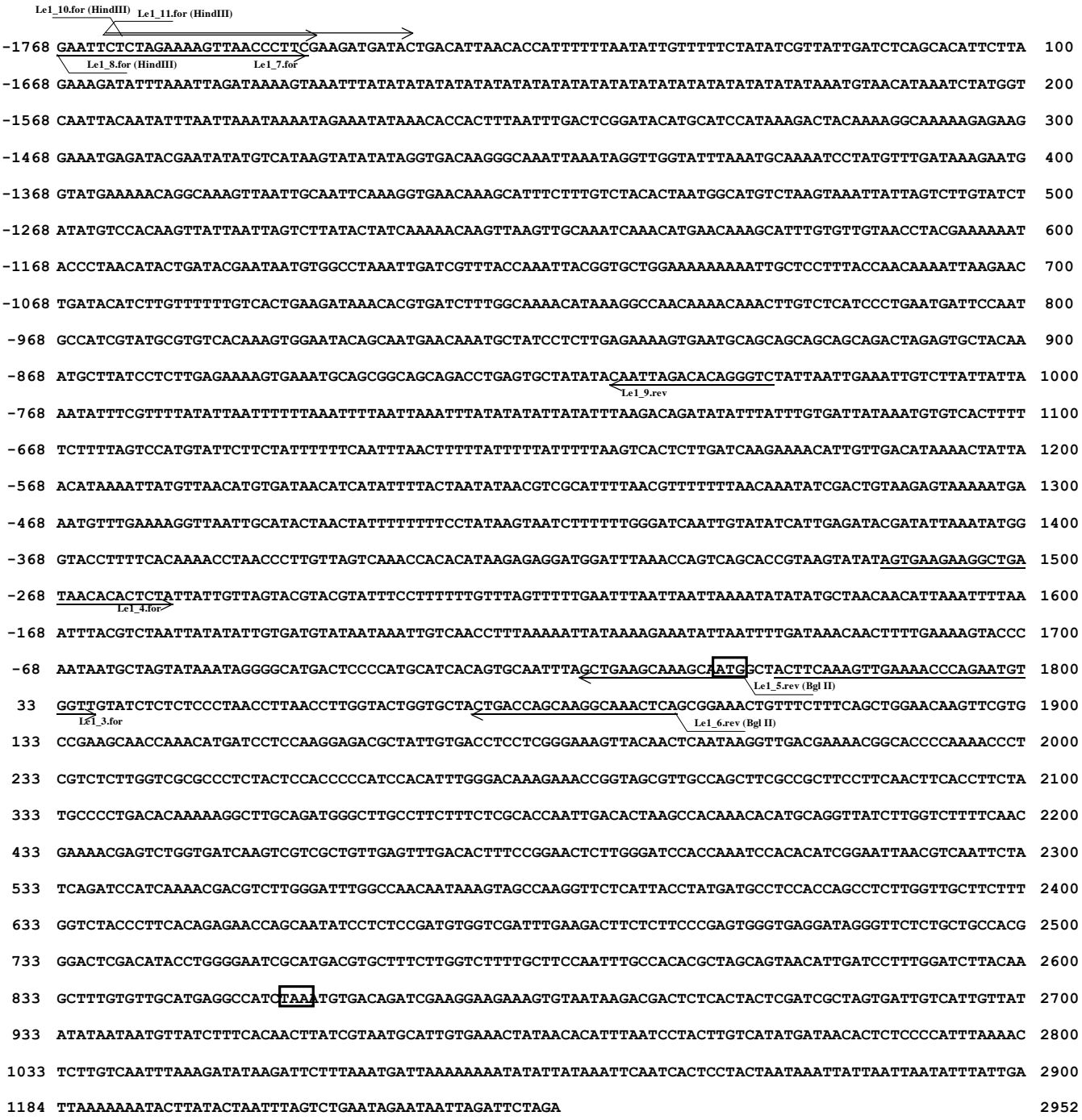


Figure 4.1: The *Le1* complete gene map.

Le1 5', coding region and 3' assembled from sequencing vector, or Genbank (Vodkin pGLGUS-21 and Accession K00821 M30884 for coding and 3' region). Stop and start codons are boxed. Total sequence length 2952bp (5' region =1768bp, coding region = 858bp, 3' region = 326bp)

Figure 3.2: The *Le2* complete gene map. (following page)

Le2 5' total sequence length 3929bp. Stop and start codons boxed. (5' region = 1293bp, coding region = 804bp, 3' region =1829bp) pHS3_Le2.4 region sequence is from -1293bp to 333bp, and pHS27_Le2.4 coding region is from -28bp to 2636bp.

Figure 4.2

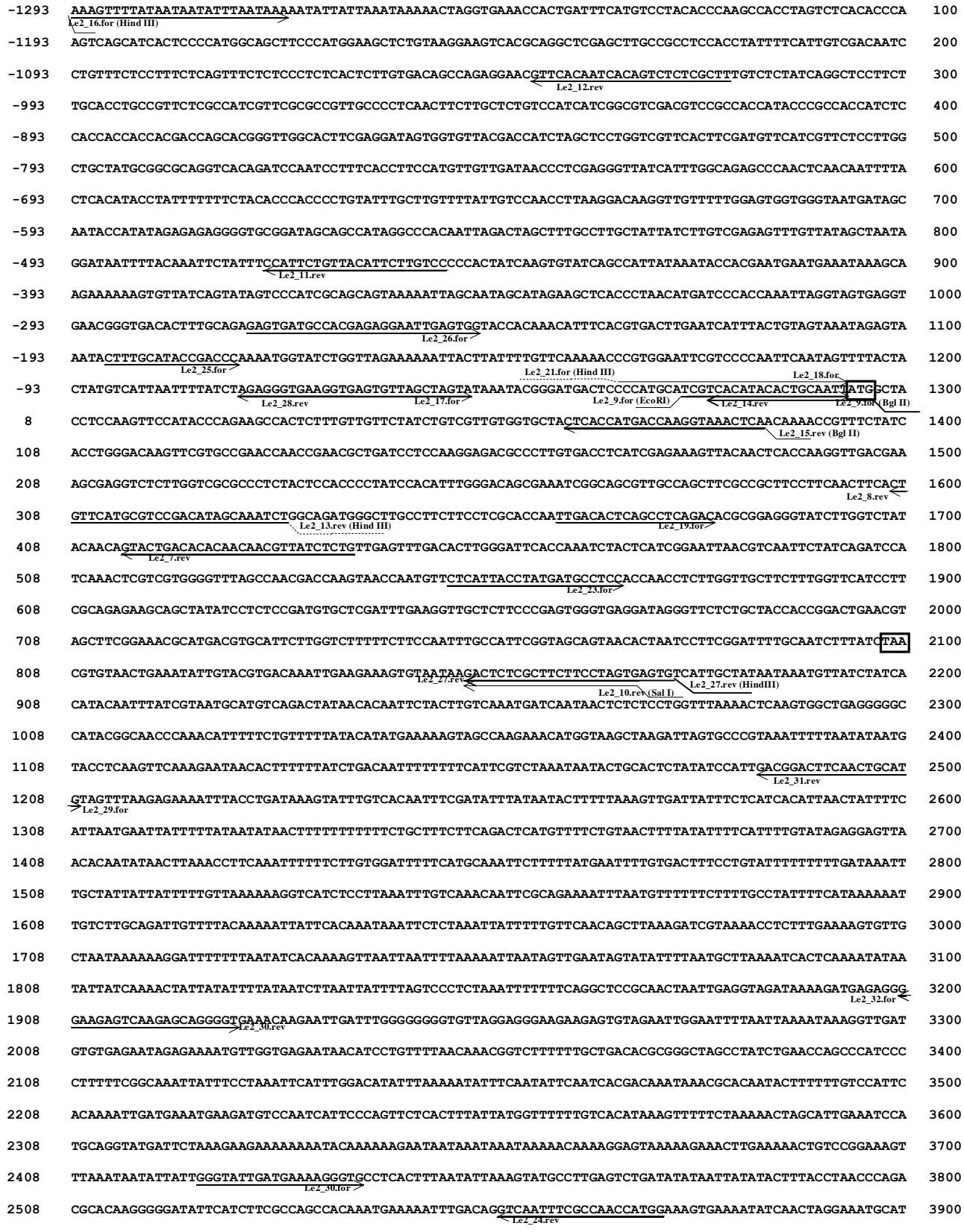

-1293 AAAGTTTTATAATAATATTTAATAAAAATATTATTAAATAAAAACTAGGTGAAACCACTGATTTCATGTCCTACACCCAAGCCACCTAGTCTCACACCCA 100
Le2_16.for (Hind III)
-1193 AGTCAGCATCACTCCCCATGGCAGCTTCCCATGGAAGCTCTGTAAGGAAGTCACGCAGGCTCGAGCTTGCCGCCTCCACCTATTTTCATTGTCGACAATC 200
-1093 CTGTTTCTCCTTTCTCAGTTTCTCTCCCTCTCACTCTTGTGACAGCCAGAGGAACGTTCACAATCACAGTCTCTCGCTTTGTCTCTATCAGGCTCCTTCT 300
Le2_12.rev
-993 TGCACCTGCCGTTCTCGCCATCGTTCGCGCCGTTGCCCCTCAACTTCTTGCTCTGTCCATCATCGGCGTCGACGTCCGCCACCATACCCGCCACCATCTC 400
-893 CACCACCACCACGACCAGCACGGGTTGGCACTTCGAGGATAGTGGTGTTACGACCATCTAGCTCCTGGTCGTTCACTTCGATGTTCATCGTTCTCCTTGG 500
-793 CTGCTATGCGGCGCAGGTCACAGATCCAATCCTTTCACCTTCCATGTTGTTGATAACCCTCGAGGGTTATCATTTGGCAGAGCCCAACTCAACAATTTTA 600
-693 CTCACATACCTATTTTTTTCTACACCCACCCCTGTATTTGCTTGTTTTATTGTCCAACCTTAAGGACAAGGTTGTTTTTGGAGTGGTGGGTAATGATAGC 700
-593 AATACCATATAGAGAGAGGGGTGCGGATAGCAGCCATAGGCCCACAATTAGACTAGCTTTGCCTTGCTATTATCTTGTCGAGAGTTTGTTATAGCTAATA 800
-493 GGATAATTTTACAAATTCTATTTCCATTCTGTTACATTCTTGTCCCCCACTATCAAGTGTATCAGCCATTATAAATACCACGAATGAATGAAATAAAGCA 900
Le2_11.rev
-393 AGAAAAAAGTGTTATCAGTATAGTCCCATCGCAGCAGTAAAAATTAGCAATAGCATAGAAGCTCACCCTAACATGATCCCACCAAATTAGGTAGTGAGGT 1000
-293 GAACGGGTGACACTTTGCAGAGAGTGATGCCACGAGAGGAATTGAGTGGTACCACAAACATTTCACGTGACTTGAATCATTTACTGTAGTAAATAGAGTA 1100
Le2_26.for
-193 AATACTTTGCATACCGACCCAAAATGGTATCTGGTTAGAAAAAATTACTTATTTTGTTCAAAAACCCGTGGAATTCGTCCCCAATTCAATAGTTTTACTA 1200
Le2_25.for
Le2_21.for (Hind III)
Le2_18.for
-93 CTATGTCATTAATTTTATCTAGAGGGTGAAGGTGAGTGTTAGCTAGTATAAATACGGGATGACTCCCCATGCATCGTCACATACACTGCAATTTATGGCTA 1300
Le2_28.rev
Le2_17.for
Le2_9.for (EcoRI)
Le2_14.rev
Le2_9.rev (Bgl II)
8 CCTCCAAGTTCCATACCCAGAAGCCACTCTTTGTTGTTCTATCTGTCGTTGTGGTGCTACTCACCATGACCAAGGTAAACTCAACAAAACCGTTTCTATC 1400
Le2_15.rev (Bgl II)
108 ACCTGGGACAAGTTCGTGCCGAACCAACCGAACGCTGATCCTCCAAGGAGACGCCCTTGTGACCTCATCGAGAAAGTTACAACTCACCAAGGTTGACGAA 1500
208 AGCGAGGTCTCTTGGTCGCGCCCTCTACTCCACCCCTATCCACATTTGGGACAGCGAAATCGGCAGCGTTGCCAGCTTCGCCGCTTCCTTCAACTTCACT 1600
Le2_8.rev
308 GTTCATGCGTCCGACATAGCAAATCTGGCAGATGGGCTTGCCTTCTTCCTCGCACCAATTGACACTCAGCCTCAGACACGCGGAGGGTATCTTGGTCTAT 1700
Le2_13.rev (Hind III)
Le2_19.for
408 ACAACAGTACTGACACACAACAACGTTATCTCTGTTGAGTTTGACACTTGGGATTCACCAAATCTACTCATCGGAATTAACGTCAATTCTATCAGATCCA 1800
Le2_7.rev
508 TCAAACTCGTCGTGGGGTTTAGCCAACGACCAAGTAACCAATGTTCTCATTACCTATGATGCCTCCACCAACCTCTTGGTTGCTTCTTTGGTTCATCCTT 1900
Le2_23.for
608 CGCAGAGAAGCAGCTATATCCTCTCCGATGTGCTCGATTTGAAGGTTGCTCTTCCCGAGTGGGTGAGGATAGGGTTCTCTGCTACCACCGGACTGAACGT 2000
708 AGCTTCGGAAACGCATGACGTGCATTCTTGGTCTTTTTCTTCCAATTTGCCATTCGGTAGCAGTAACACTAATCCTTCGGATTTTGCAATCTTTATCTAA 2100
808 CGTGTAACTGAAATATTGTACGTGACAAATTGAAGAAAGTGTAATAAGACTCTCGCTTCTTCCTAGTGAGTGTCATTGCTATAATAAATGTTATCTATCA 2200
Le2_27.rev
Le2_27.rev (HindIII)
Le2_10.rev (Sal I)
908 CATACAATTTATCGTAATGCATGTCAGACTATAACACAATTCTACTTGTCAAATGATCAATAACTCTCTCCTGGTTTAAAACTCAAGTGGCTGAGGGGGC 2300
1008 CATACGGCAACCCAAACATTTTTCTGTTTTTATACATATGAAAAAGTAGCCAAGAAACATGGTAAGCTAAGATTAGTGCCCGTAAATTTTTAATATAATG 2400
1108 TACCTCAAGTTCAAAGAATAACACTTTTTTATCTGACAATTTTTTTTCATTCGTCTAAATAATACTGCACTCTATATCCATTGACGGACTTCAACTGCAT 2500
Le2_31.rev
1208 GTAGTTTAAGAGAAAATTTACCTGATAAAGTATTTGTCACAATTTCGATATTTATAATACTTTTTAAAGTTGATTATTTCTCATCACATTAACTATTTTC 2600
Le2_29.for
1308 ATTAATGAATTATTTTTATAATATAACTTTTTTTTTTTCTGCTTTCTTCAGACTCATGTTTTCTGTAACTTTTATATTTTCATTTTGTATAGAGGAGTTA 2700
1408 ACACAATATAACTTAAACCTTCAAATTTTTTCTTGTGGATTTTTCATGCAAATTCTTTTTATGAATTTTGTGACTTTCCTGTATTTTTTTTTGATAAATT 2800
1508 TGCTATTATTATTTTTGTTAAAAAAGGTCATCTCCTTAAATTTGTCAAACAATTCGCAGAAAATTTAATGTTTTTTCTTTTGCCTATTTTCATAAAAAAT 2900
1608 TGTCTTGCAGATTGTTTTACAAAAATTATTCACAAATAAATTCTCTAAATTATTTTTGTTCAACAGCTTAAAGATCGTAAAACCTCTTTGAAAAGTGTTG 3000
1708 CTAATAAAAAAGGATTTTTTTAATATCACAAAAGTTAATTAATTTTAAAAATTAATAGTTGAATAGTATATTTTAATGCTTAAAATCACTCAAAATATAA 3100
1808 TATTATCAAAACTATTATATTTTATAATCTTAATTATTTTAGTCCCTCTAAATTTTTTTCAGGCTCCGCAACTAATTGAGGTAGATAAAAGATGAGAGGG 3200
Le2_32.for
1908 GAAGAGTCAAGAGCAGGGGTGAAACAAGAATTGATTTGGGGGGGGTGTTAGGAGGGAAGAAGAGTGTAGAATTGGAATTTTAATTAAAATAAAGGTTGAT 3300
Le2_30.rev
2008 GTGTGAGAATAGAGAAAATGTTGGTGAGAATAACATCCTGTTTTAACAAACGGTCTTTTTTGCTGACACGCGGGCTAGCCTATCTGAACCAGCCCATCCC 3400
2108 CTTTTTCGGCAAATTATTTCCTAAATTCATTTGGACATATTTAAAAATATTTCAATATTCAATCACGACAAATAAACGCACAATACTTTTTTGTCCATTC 3500
2208 ACAAAATTGATGAAATGAAGATGTCCAATCATTCCCAGTTCTCACTTTATTATGGTTTTTTGTCACATAAAGTTTTTCTAAAAACTAGCATTGAAATCCA 3600
2308 TGCAGGTATGATTCTAAAGAAGAAAAAAAATACAAAAAAGAATAATAAATAAATAAAAACAAAAGGAGTAAAAAGAAACTTGAAAAACTGTCCGGAAAGT 3700
2408 TTAAATAATATTATTGGGTATTGATGAAAAGGGTGCCTCACTTTAATATTAAAGTATGCCTTGAGTCTGATATATAATTATATACTTTACCTAACCCAGA 3800
Le2_30.for
2508 CGCACAAGGGGGATATTCATCTTCGCCAGCCACAAATGAAAAATTTGACAGGTCAATTTCGCCAACCATGGAAAGTGAAAATATCAACTAGGAAATGCAT 3900
Le2_24.rev
2608 AGTGGTTTGGTAATGGTGGCTCCCTTCAA 3929

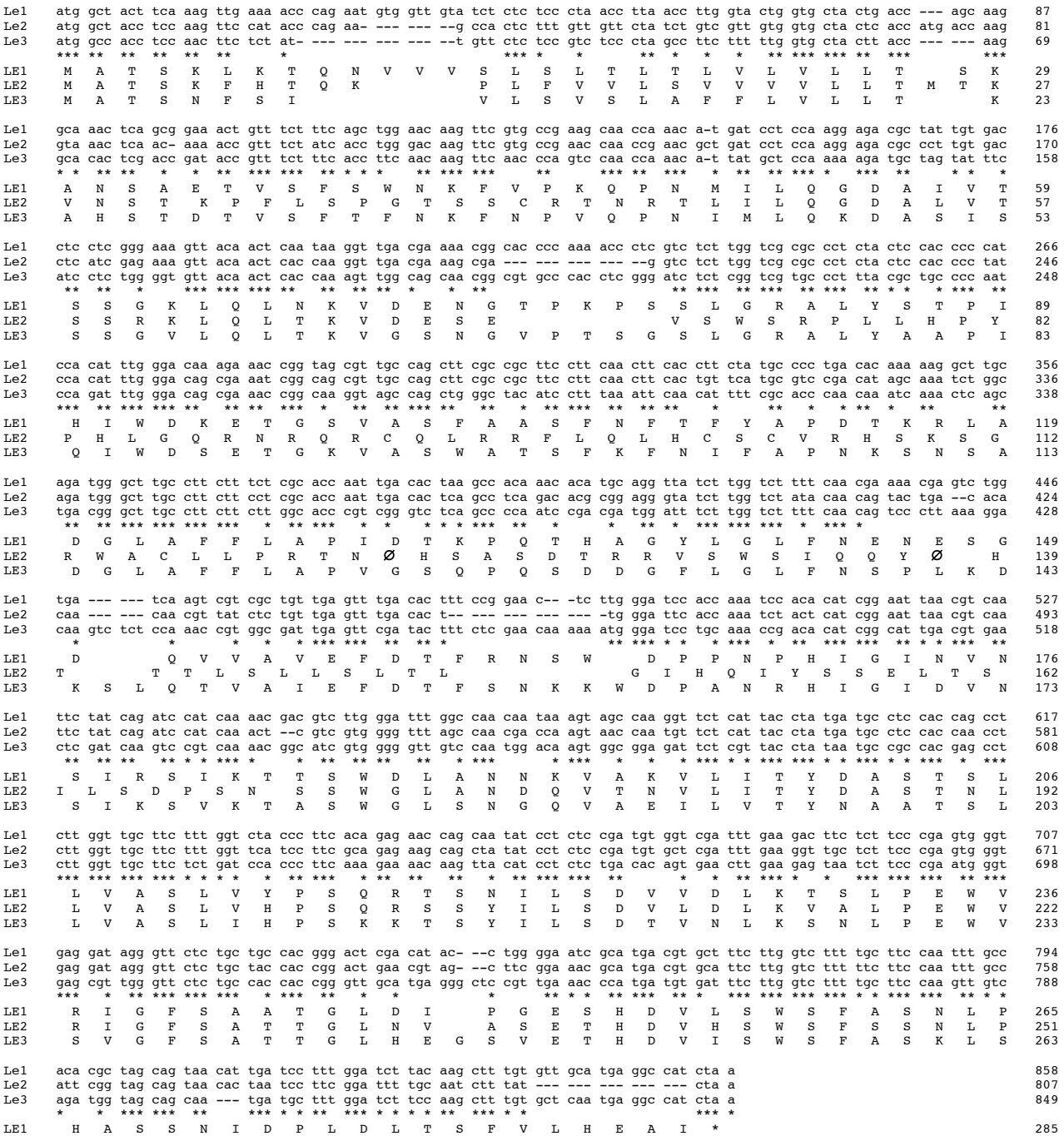
Figure 4.3: The *Le3* complete gene map.

Le3 total sequence length 2252bp. *Le3* 5' region sequence length 1279bp. Stop and start codons boxed. (5' region = 1279bp, coding region = 849bp, 3' region = 75bp) pHS2_Le3.6 region sequence is from -1279bp to 340bp and pHS136_Le3.1 contains the sequence region from -54bp to 924bp.

Figure 4.4: Nucleotide and amino acid alignment of the soybean lectins *Le1, Le2* and *Le3*.

Protein translation written below nucleotide sequence. Stars indicate perfect nucleotide alignment. Premature stop codins in *Le2* sequences marked by " ∅ ", gaps in nucleotide sequence marked by " - ", and final stop codons for all three sequences marked by "*" in amino acid sequence.

Figure 4.4:



motifs, showed any grouping of motif that had a higher number of motif occurrences in *Le1* as compared to the other lectin promoters.

Although the *Le1* and *Le3* promoter regions had similar proportions of seed motifs, *Le2* had less than the others, and when motifs that were present in all three promoter were excluded, nearly 40% of the motifs that were present in the 5' promoter region of *Le1* were seed-specific (Figure 4.5).

All three promoters had many examples of vegetative motifs which occurred at similar amounts relative to the total number of motifs and motif occurrences in their 5' promoter sequences. Despite having different expression profiles in vegetative tissue, no pattern could be seen from the PLACE database analysis. Motifs related to flower/pollen/fruit, etiolation, hormone-induced or ubiquitous expression profiles were found less often in the promoter sequences and their relative frequency were about equal between the promoters (Figure 4.5).

Motifs related to defence or stress were found much more often in the *Le2* 5' region than in either *Le1* or *Le3*. If motifs that occurred in all three promoters were removed, there were proportionally nearly twice as many stress motifs in *Le2* than *Le1*, and three times as many in *Le2* than *Le3*.

After filtering out motifs that were shared between all three promoters, the *Le3* promoter was found to have a higher proportion of motifs related to root expression, than did the *Le1* or *Le2* promoters. Unlike the *Le2* promoter, motifs related to stress and defense were particularly under-represented in the *Le3* promoter.

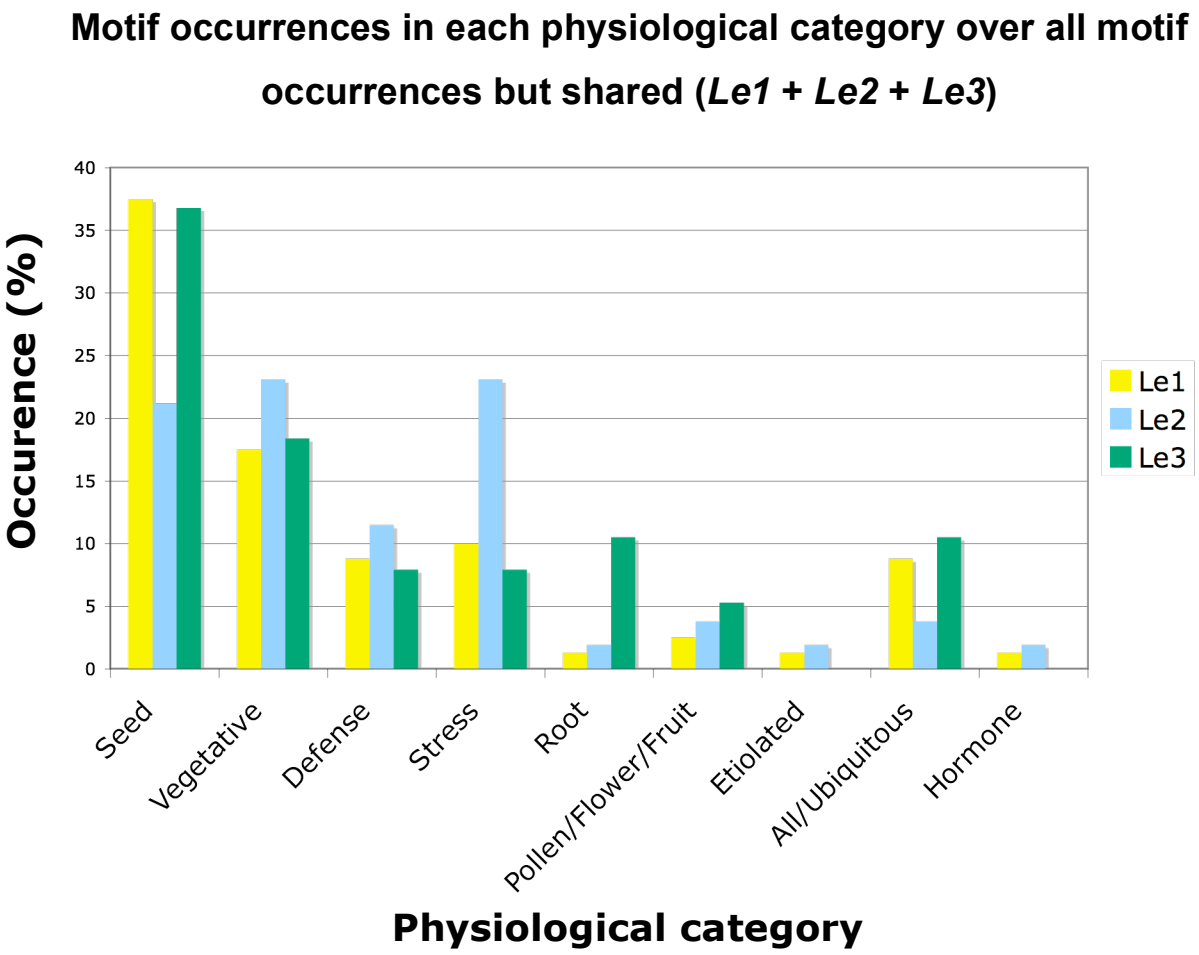


Figure 4.5: Portion of motifs related to tissue types in *Le1*, *Le2* and *Le3* promoter sequences.

Number of motif occurrences in each category divided by the total number of motif occurrences in promoter sequence, excluding those shared between all three soybean lectin promoters *Le1, Le2* and *Le3*.

4.3 Comparison of promoter sequence motifs to non-lectin soybean promoters

The seed-specific promoters β-phaseolin (*Phaseolus vulgaris*), β-conglycinin α subunit (soybean) and the oat globulin promoter *AsGlo1*, vegetative protein promoters VSPα (soybean), VSPβ (soybean) and DB58 (*D. biflorus),* and the pseudogene class I basic chitinase gene, ψ*BCH* (tomato) were analysed against the PLACE database (Higo *et al.*, 1999) to locate the known motifs within the promoter sequences. Results were compared to the PLACE motif results for the three soybean lectins and the presence of certain selected seed-specific, vegetative, defence and stress motifs were recorded, as shown in Table 4.1.

Several motifs for vegetative expression found in the *Le1* sequence were also found in other seed-specific promoters. The promoter sequences of β-phaseolin promoter from *Phaseolus vulgaris*, β-conglycinin α subunit gene from *Glycine max* and the oat globulin promoter *AsGlo1* each contained several motifs which had been designated as vegetative motifs (based on a literature review of the motifs), that were also present in the *Le1* promoter sequence. For example, the core motif from a potato MYB homologue gene (*MybSt1*) which was found to be a transcriptional activator in vegetative tissue, was found in all four seed-specific promoters (Baranowskij *et al.*, 1994). All but two of the fifteen vegetative motifs found in the *Le1* promoter sequence were also found in at least one of these three other seed-specific promoters. One of these two was the pyrimidine box found in the barley *EPB-1* promoter, which is related to the expression of a cystein protease in germinating seeds (Cercós *et al.*, 1999).

Table 4.1: Motifs present in promoter regions of selected genes.

		PROMOTER MOTIFS								
		Seed-specific motifs			Vegetative motifs		Stress and or defense motifs			
Gene Name	**Plant of origin**	**G box**	**E box**	**RY motifs**	**Core motif *MybSt1* (potato)**	**Pyrimidine box *EPB-1* (barley)**	**Box-L-like**	**Consensus I MYB DNA binding site (*Arabidopsis*)**	**Core DRE motif (maize)**	**CBF CRT/DRE motif (barley)**
Seed-specific promoters										
***Le1* lectin**	Soybean	1	6	3	1	1	0	0	0	1
β-phaseolin	*P. vulgaris*	2	10	6	1	0	0	1	0	0
β-conglycinin α subunit	Soybean	2	24	10	7	0	0	3	1	2
globulin promoter *AsGlo1*	Oat	0	6	0	2	0	0	3	1	0
Potential pseudogenes										
***Le2* lectin**	Soybean	1	4	1	3	0	1	2	1	3
class I basic chitinase (ψ*BCH)*	Tomato	0	18	0	3	0	2	8	1	1
Vegetative										
***Le3* lectin**	Soybean	1	3	1	0	0	0	0	0	0
VSPα	Soybean	2	10	2	1	0	1	3	0	0
VSPβ	Soybean	2	4	0	3	0	0	5	1	1
DB58	*D. biflorus*	2	6	0	0	0	0	0	0	0

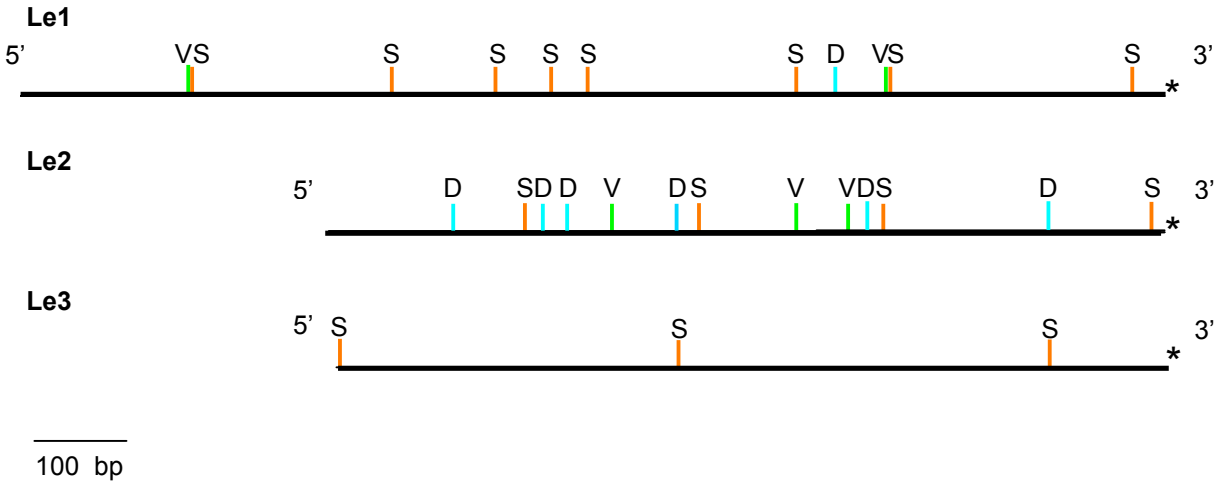


Figure 4.6: Selected promoter motifs in soybean lectin promoters.

Start codon indicated by star. Bars represent different motif categories marked by V for vegetative, S for seed and D for stress and defense. Bar represents 100bp of sequence.

Like *Le2*, the coding region of a tomato class I basic chitinase gene (ψ*BCH*) contains a frameshift mutation in the open reading frame causing it to code for a truncated protein (Baykal *et al*., 2006; Goldberg *et al*., 1983). Although thought to be a pseudogene, ψ*BCH* was shown to have a functional promoter (Baykal *et al*., 2006). Promoter analysis using the PLACE database revealed several stress and defense-related motifs that were not found in *Le1* or *Le3*, but were present in *Le2* and ψ*BCH,* including a MYB consensus I binding site (Abe *et al*., 1997; Solano *et al*., 1995; Urao *et al*., 1993) related to drought conditions, described above, as well as the core DRE motif site (Xue, 2002), related to cold and dehydration. The box-L-like (Maeda *et al*., 2005) sequence and GCC-box core (Brown *et al*., 2003) motifs were also found in *Le2* and ψ*BCH* and not *Le1* or *Le3*. Both were classified as defensive motifs, based on the literature.

4.4 Construction of lectin promoter::*gusA* fusions

In order to determine the tissue-specificity of the promoters of the *Le1* gene homologues *Le2* and *Le3*, promoter::*gusA* reporter gene constructs were made, as shown in Figure 4.7. The 5' region of the *Le2* and *Le3* lectin genes were amplified in two versions, either with or without their respective predicted signal peptide using the primers listed in Table 3.1, following a modified ligation-mediated PCR method (Scharf *et al*., 1986) in which restriction enzyme cut sites are added to the 5' end of the primers. The constructs contain the 5'-upstream region of the *Le1, Le2* and *Le3* genes, either with or without the respective signal peptides (of which *Le1* has been proven, and *Le2* and *Le3* have been predicted), and a start codon, ligated to a bacterial GUS-coding sequence

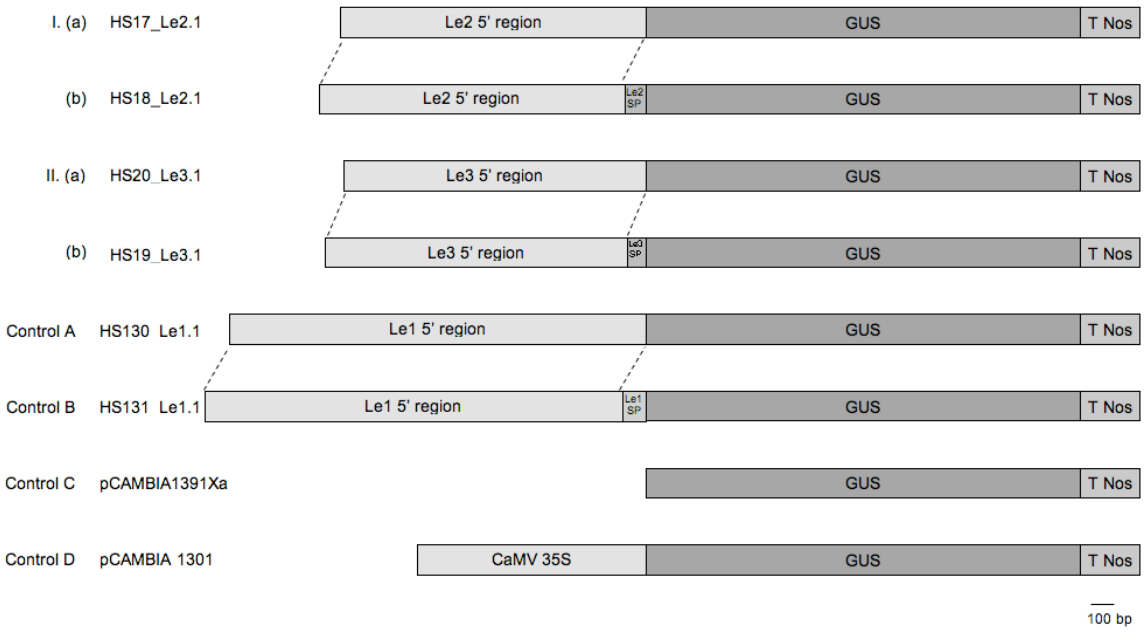


Figure 4.7: 5' Le construct series for plant transformation.

A series of constructs made from 1.3kb of the 5'-upstream regions of lectin genes *Le2* and *Le3* constructing promoter-GUS reporter series made of 5'-upstream region::*gusA*:: Tnos terminator, and 5'-upstream region+signal peptide::*gusA*::Tnos terminator. In addition to this, two positive controls that have known expression patterns were made: the *Le1* 5'-upstream region::*gusA*::Tnos, and *Le1* 5'-upstream region+*Le1* signal peptide::*gusA*::Tnos. An unmodified pCAMBIA 1391Xa vector (contains no start codon for GUS reported gene) was also tested.

(*gusA*) in the T-DNA of a binary vector pCAMBIA1391Xa. The signal peptides are not expected to change the tissue specificity, but were included because signal peptides have been shown to enhance the stability of the transgene product (Wandelt *et al.*, 1992; Sojikul *et al*., 2003).

Plasmid pHS130_Le1.1 contains the *Le1* 5' promoter region as well as the predicted signal peptide from the soybean *Le1* gene (-968 to +96bp) and plasmid pHS131_Le1.1 consists of the 5' promoter region (-968 to +3bp) from the soybean *Le1* gene from the soybean *Le1* gene (Figure 4.1, Appendix IV and III, respectively). The 5'-upstream regions of the soybean *Le2* gene, and of the soybean *Le3* genes, with and without their signal peptides (for *Le2*, -1293 to +90bp, and –1293 to +3bp, respectively, and for *Le3,* –1280 to +78, and –1280 to +3, with and without the signal peptide, respectively) were amplified by PCR and the fragments cloned into the pCAMBIA 1391Xa vector to obtain the pHS17_Le2.1 and pHS18_Le2.1 vectors (with and without the *Le2* signal peptide, respectively) and the pHS19_Le3.1 and pHS20_Le3.1 vectors (with and without the *Le3* signal peptide, respectively). (Figures 4.2, 4.3, Appendix V, VI, VII and VIII).

4.5 Transformation of *Arabidopsis* using the lectin promoters::gusA constructs

To study the promoter activities *in planta*, the constructs were transformed into *Agrobacterium,* which was used to transform *Arabidopsis thaliana* with the floral dip technique (Bechtold *et al.*, 1993; Clough and Bent, 1998). Transformation efficiency rates varied between 0.18-0.74% seeds transformed, as seen in Table 4.2.

Table 4.2: Transformation efficiency of *Arabidopsis* floral dip transformation.

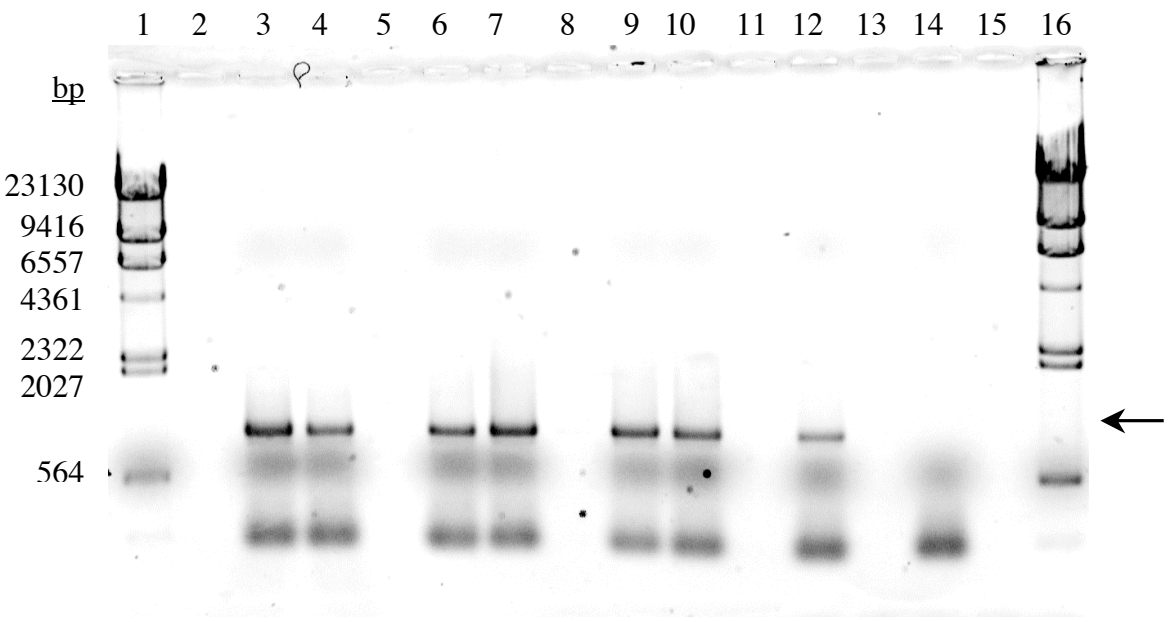
Vector	Contents	Transformation efficiency
HS130_Le1.1	Le1 5' region with signal peptide	0.56%
HS131_Le1.1	Le1 5' region without signal peptide	0.74%
HS17_Le2.1	Le2 5' region with signal peptide	0.30%
HS18_Le2.1	Le2 5' region without signal peptide	0.50%
HS19_Le3.1	Le3 5' region with signal peptide	0.35%
HS20_Le3.1	Le3 5' region without signal peptide	0.35%
pCAMBIA 1391Xa	No promoter, non-functional *gusA*	0.18%
pCAMBIA 1301	*gusA* driven by CaMV35s	0.38%

In total, 59 independent transgenic T_1 lines containing the *Le2* 5'-upstream region, (37 without and 22 with signal peptide) were obtained and characterized. Forty eight independent transgenic T_1 lines containing the *Le3* 5'-upstream region, (22 without and 16 with signal peptide) were obtained, and 60 independent transgenic T_1 lines containing the positive control *Le1* 5'-upstream region, (42 without and 18 with signal peptide) were obtained and characterized. Twenty one independent transgenic T_1 lines containing the promoterless::*gusA* pCAMBIA 1391 Xa vector (negative control), and 23 T_1 lines containing the pCAMBIA 1301 (positive control) vector were obtained and characterized. Genomic DNA from one T_2 line of each construct was tested by PCR for the GUS gene to confirm the plant transformation (Figure 4.8).

4.6 Tissue-specific expression patterns of soybean lectin promoter::gusA gene fusion constructs in *Arabidopsis*

Between 2-4 weeks after transplanting to soil, all lines described in Section 4.5 were tested for GUS activity in flowers, cauline leaves, rosette leaves, uncut siliques and cut siliques.

Plants transformed with the *Le1* promoter construct not containing the signal peptide showed seed-specific expression (Figure 4.9, pHS15_Le1.1). In contrast, plants transformed with the *Le3* promoter construct, without the predicted signal peptide, showed a high level of expression in all vegetative tissue tested, and no noticeable expression in the developing seed (Figure 4.9, pHS19_Le3.1). The *Le2* promoter construct, made without the predicted signal peptide, showed very low levels of expression in all tissue tested including developing seeds (Figure 4.9, pHS17_Le2.1).



Lanes: 1 – λ DNA *Hind III* ladder
2 – (empty)
3 – HS15_Le1.1.15.1 (with *Le1* signal peptide)
4 – HS16_Le1.1.52.5 (without *Le1* signal peptide)
5 – (empty)
6 – HS17_Le2.1.6.2 (with *Le2* signal peptide)
7 – HS18_Le2.1.38.2 (without *Le2* signal peptide)
8 – (empty)
9 – HS19_Le3.1.13.1 (with *Le3* signal peptide)
10 – HS20_Le3.1.24.3 (without *Le3* signal peptide)
11 – (empty)
12 – pCAMBIA1301.14.3 (*Arabidopsis* positive control)
13 – (empty)
14 – Untransformed *Arabidopsis* (negative control)
15 – (empty)
16 – λ DNA *Hind III* ladder

Figure 4.8: Confirmation of transformed *Arabidopsis* plants.

A PCR reaction was performed using genomic DNA of T_2 generation of transformed plants as template and *gusA* specific primers. Arrow points to bands of interest.

Figure 4.10 shows the results of the constructs with the signal peptide included. As can be seen, the results are similar to the constructs with the signal peptide. The signal peptide for *Le1* and predicted signal peptides for *Le2* and *Le3* were included to help make a more stable gene product, but did not seem to have any significant effect on the level of expression or the tissues where expression was observed.

Plants transformed with the pCAMBIA 1391Xa negative control vector showed no reporter gene expression, while plants transformed with the pCAMBIA 1301 positive control vector, containing the *gusA* gene driven by the CaMV 35S promoter, showed high expression in all tissues including developing seeds (Figure 4.9 and Figure 4.10).

Seeds from three T_1 lines showing representative expression profiles were selected for each of the six constructs, and vector controls, were grown on selection medium. Five T_2 plants from each of the three selected T_1 lines were tested for the same tissues, and showed expression profiles consistent with those seen in T_1 plants (data not shown).

4.7 Different lectin promoters drive differential reporter gene expression in developing *Arabidopsis* seedlings

To investigate whether there is a change in the lectin promoter activities over time, T_2 plants from each of the three selected T_1 lines were assayed for GUS activity in a developmental series. Ungerminated seeds, and germinating seeds/seedlings were assayed over 14 days.

Figures 4.11 shows the GUS activity as a result of the different promoters. The three T_2 lines containing the *Le1* promoter region (lines 11, 15 and 10 in pHS15_Le1.1 and lines 48, 52 and 53 in pHS16_Le1.1 plants) showed strong GUS activity in the seed

Figure 4.9: GUS assays on T_1 *Arabidopsis* plants transformed with soybean lectin promoters without signal peptide. (following page)

HS130_Le1.1, HS18_Le2.1, HS20_Le3.1. pCAMBIA1391Xa (negative control, with no promoter, *gusA* without start codon), or pCAMBIA1301 (positive control, *gusA* driven by CaMV 35S promoter) vectors are included. Negative represents an untransformed plant. Bar = 1mm.

Figure 4.9

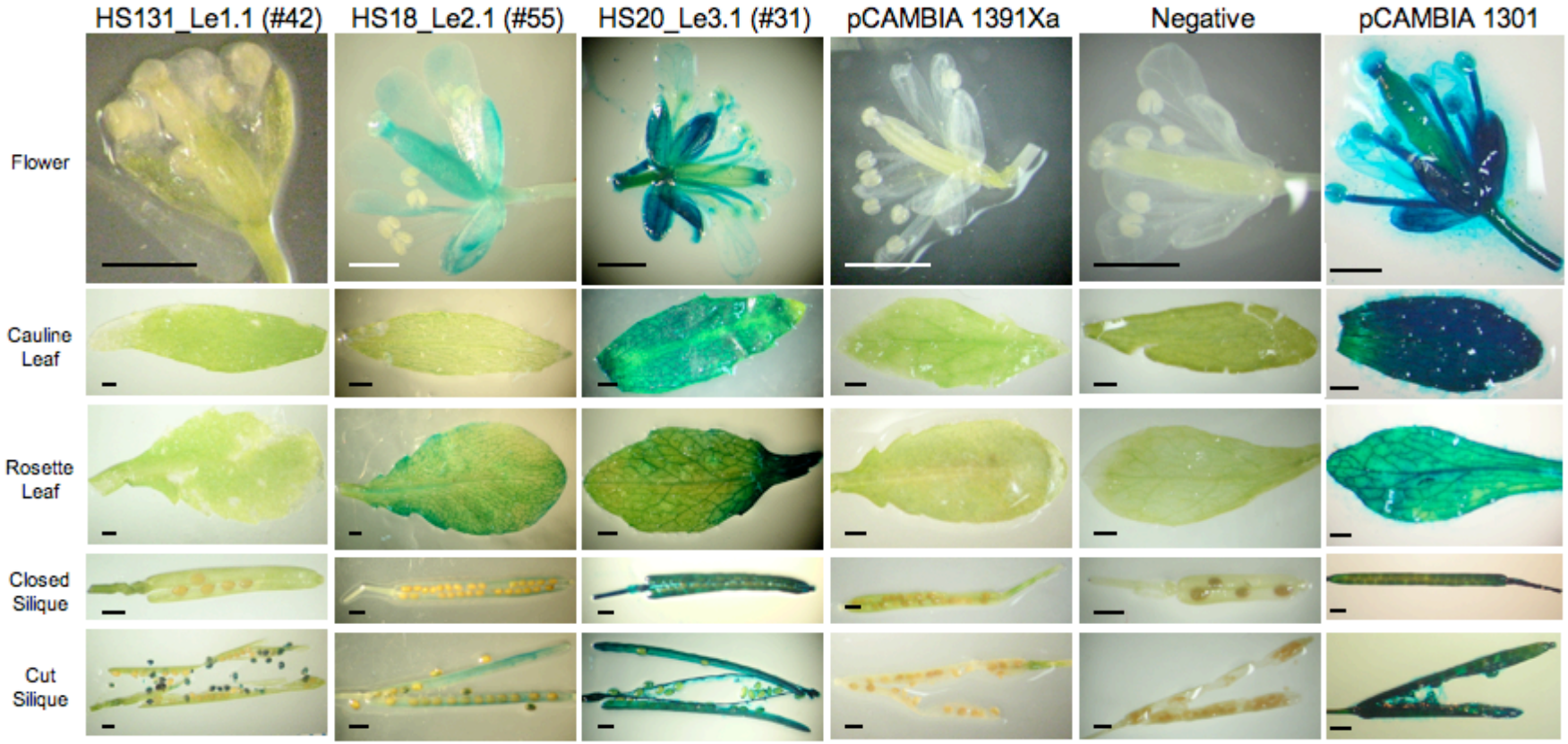
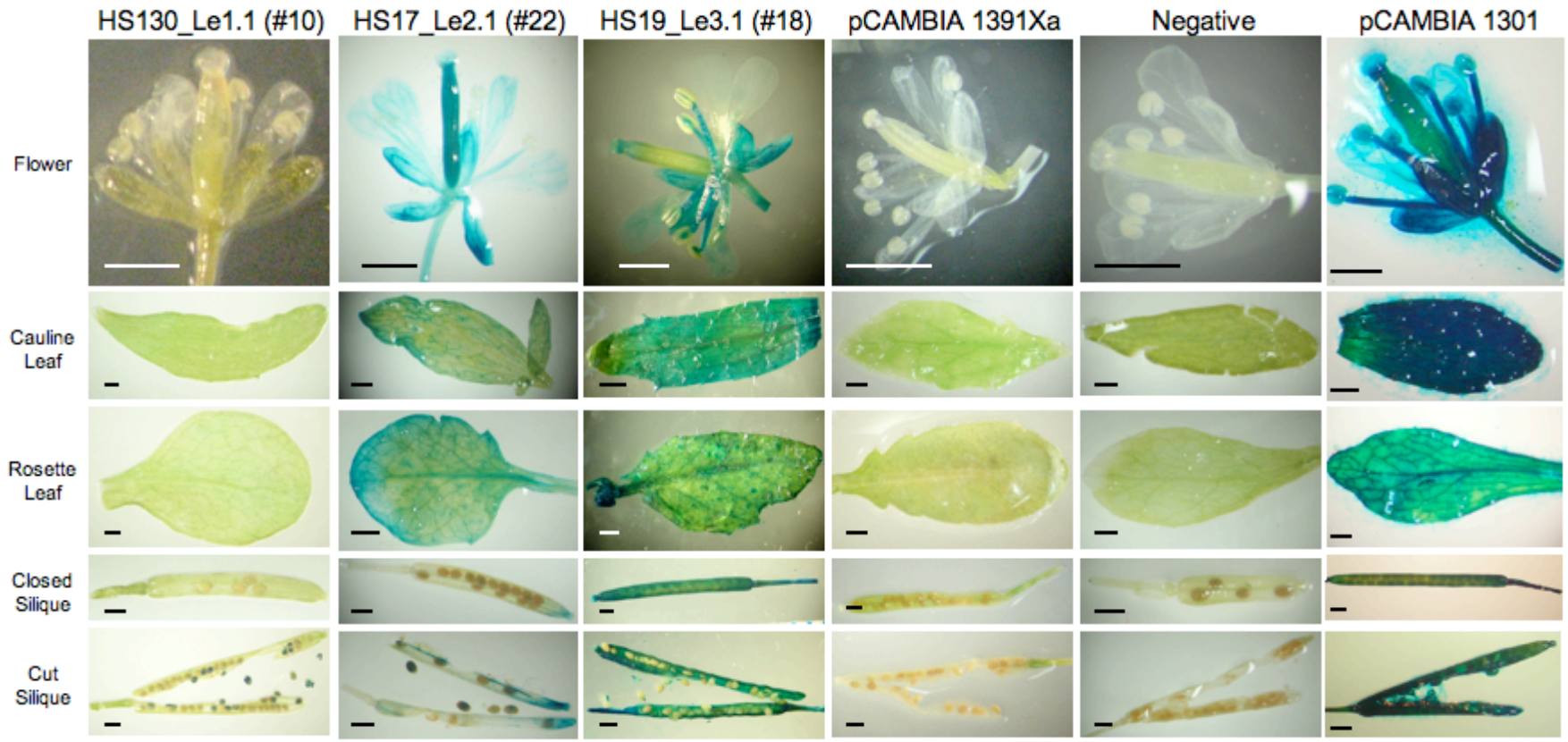

HS131_Le1.1 (#42)
HS18_Le2.1 (#55)
HS20_Le3.1 (#31)
pCAMBIA 1391Xa
Negative
pCAMBIA 1301
Flower
Cauline Leaf
Rosette Leaf
Closed Silique
Cut Silique

Figure 4.10: GUS assays on T_1 *Arabidopsis* plants transformed with lectin promoters and respective signal peptide or predicted signal peptide. (following page)

HS131_Le1.1, HS17_Le2.1, HS19_Le3.1. pCAMBIA1391Xa (negative control, with no promoter, *gusA* without start codon), or pCAMBIA1301 (positive control, *gusA* driven by CaMV 35S promoter) vectors are included. Negative represents an untransformed plant. Bar = 1mm.

Figure 4.10


HS130_Le1.1 (#10)
HS17_Le2.1 (#22)
HS19_Le3.1 (#18)
pCAMBIA 1391Xa
Negative
pCAMBIA 1301
Flower
Cauline Leaf
Rosette Leaf
Closed Silique
Cut Silique

before germination and the day prior to being placed in the growth chamber. This was quickly reduced by Day 3 to faint GUS activity in the cotyledons and hypocotyl, and after day 5/6, it was only seen in the apical meristem region, and in some plants, also faintly in rosette leaves, although the majority did not have it in these leaves. (Figure 4.11 pHS15_Le1.1 and pHS16_Le1.1)

Although plants containing the *Le2* promoter region had some GUS activity in the developing seeds, no seeds prior to selection or during early germination showed GUS activity. Lines 20, 6 and 22 were tested in pHS17_Le2.1 plants and lines 52, 38 and 55 were tested in pHS18_Le2.1 plants. No activity was seen in the *Le2* promoter-containing seedlings for at least 5 days, after which point faint activity was seen in the cotyledons. Rosette leaves also showed GUS activity, especially in vascular tissue, which was stronger than expression seen in the cotyledons of those plants (Figure 4.11 – pHS17_Le2.1 and pHS18_Le2.1).

The assays on T_1 plants transformed with the *Le3* constructs showed strong GUS activity in vegetative tissues, but not in developing seeds. Lines 20, 13 and 18 were tested in pHS19_Le3.1 plants and lines 14, 24 and 31 were tested in pHS20_Le3.1 plants. Interestingly, in the T_2 seedlings, faint GUS activity was seen in the cotyledons still within a small number of seeds tested before germination, although this was far less intense than GUS activity seen with the *Le1* promoter. In contrast to the plants transformed with the *Le1* constructs, in which the GUS activity was only seen until Day 2, plants with the *Le3* promoter very quickly intensified GUS activity in the cotyledons. By the second day after moving to the growth chamber (Day 2), GUS activity had extended to the root tissue, with the exception of the root tip. GUS activity in root tissues

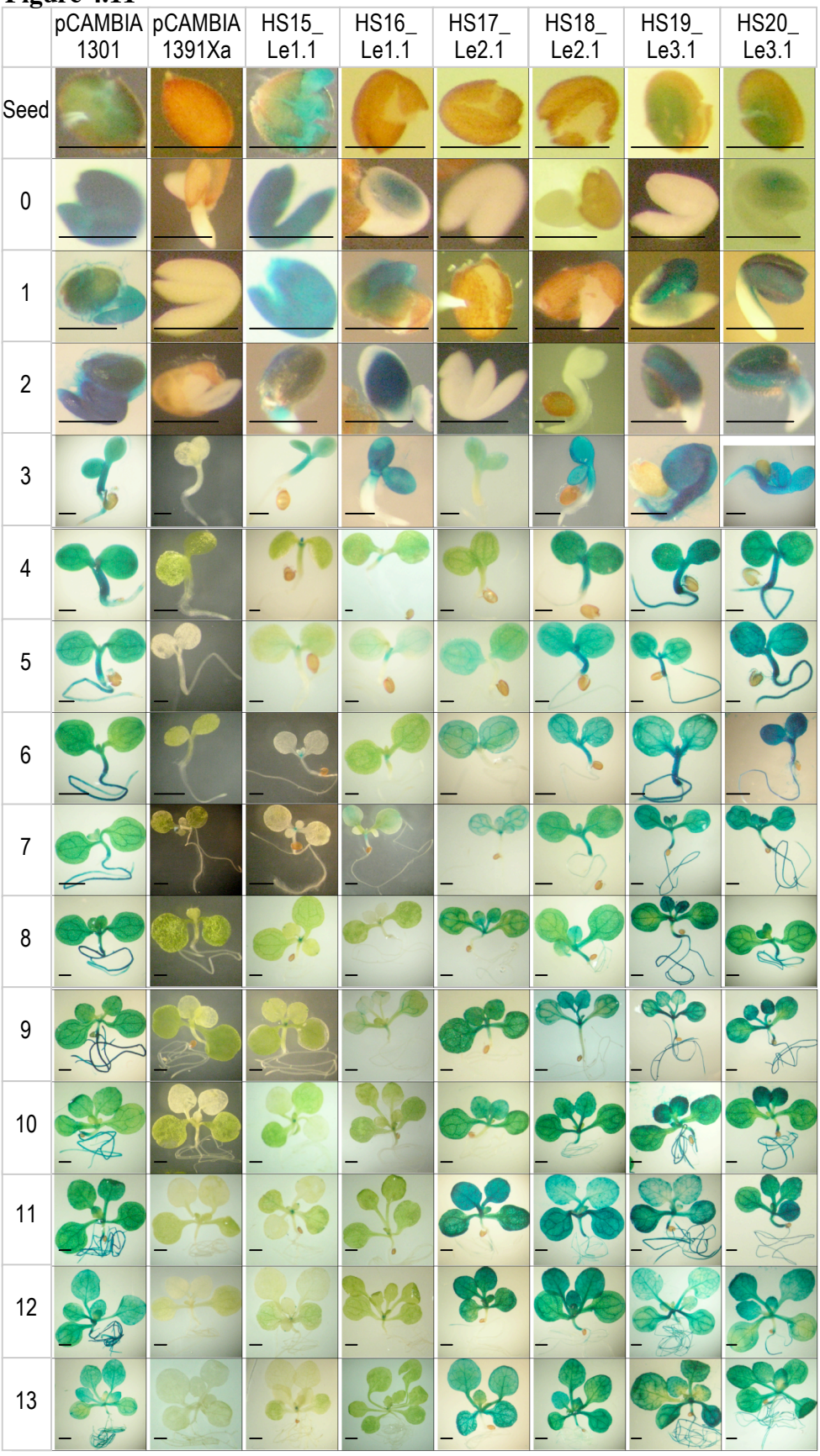
continued to be strong in the *Le3* plants throughout the series, while activity in the cotyledons seemed to fade. Rosette leaves showed strong GUS activity when young, but this became less intense and patchy as they matured, and by approximately Day 4, had similar patterns of GUS activity to the older set of rosette leaves.

As expected, the positive control CaMV35S::*gusA*-containing lines showed GUS activity in all tissues at all times, while the negative control pCAMBIA1391Xa vector with a promoter-less *gusA* gene showed no GUS activity.

Figure 4.11: Developmental series of GUS assay on T_2 seeds from *Arabidopsis* plants. (following page)

Plants transformed with contructs containing the *Le1*, *Le2* or *Le3* 5' promoter region, with or without the predicted signal peptide, or a positive control (CaMV 35S promoter) or negative control (pCAMBIA 1391Xa vector only). Bar from "seed" to Day 5 = 0.5mm, bar from Day 6 to 13 = 1mm.

Figure 4.11


pCAMBIA 1301
pCAMBIA 1391Xa
HS15_ Le1.1
HS16_ Le1.1
HS17_ Le2.1
HS18_ Le2.1
HS19_ Le3.1
HS20_ Le3.1
Seed
0
1
2
3
4
5
6
7
8
9
10
11
12
13

Section 5:
Discussion

We have cloned two soybean *Le1* gene homologues, *Le2* and *Le3*, and analyzed their promoters for tissue preference *in silico* and *in planta* in transgenic *Arabidopsis*. Based on similarity analyses (Strömvik *et al*., 2004), the *Le1*, *Le2* and *Le3* genes belong to a small gene family with conserved coding sequences, where the promoters for the individual genes have evolved differently, causing the individual genes to be differentially expressed. Gene duplication has long been thought to be a mechanism by which proteins can evolve new functions and the promoters of several small gene families, similar to the soybean lectins have previously been studied (Harada *et al*., 1990; Van Damme *et al*., 1995). While one protein remains to perform the original function, mutations in the coding sequence or in the promoter will over time lead to new protein characteristics or to expression in a novel tissue (Sparvoli *et al*., 2001). In a study of four phenylalanine ammonia-lyase (PAL) genes in parsley, the four gene products were shown to possess the same enzyme kinetic activity, however, while three had similar expression profiles and promoters, one (*PAL4*) was differentially expressed and differed in promoter structure (Logemann *et al*., 1995).

The soybean Kunitz trypsin inhibitors (KTi) are differentially expressed as well. The KTi family contains at least ten members, of which many are linked in tandem pairs (e.g. KTi1/Kti2, and Kti3/Kti4) (Jofuku and Goldberg, 1989). The KTi1 and KTi2 coding regions shared 97% similarity and 80% to the KTi3 coding region (Jofuku and Goldberg, 1989). The 5' regions of KTi1 and KTi2 were approximately 900bp apart, but only had 1bp difference in -335bp of the KTi2 start codon. The KTi3 5' region however, is about

80% similar to the KTi1/KTi2 5' regions. Despite such high levels of similarity between the three genes and promoter regions, all three were expressed at different levels during embryogenesis, with KTi3 expressed at a higher level than the others. In addition to this, KTi1/KTi2 were expressed in soybean stem, leaf and root, but KTi3 was found only in the stem and leaf, all at approximately 1000x less than the levels found during embryogenesis. Transgenic tobacco plants transformed with DNA fragments consisting of the coding regions, as well as the 5' and 3' flanking regions, maintained most expression profiles, however KTi1/ KTi2 were not found in root in tobacco, although it was present in soybean roots (Jofuku and Goldberg, 1989).

Legumes lectins are a large family of proteins with a diverse array of functions. One of the defining characteristics of lectins is the carbohydrate-binding property (Goldstein *et al.*, 1980; Rüdiger and Gabius, 2001). While the protein sequences for soybean lectins *Le1* and *Le3* are very similar, the different carbohydrate binding properties (Spilatro and Anderson, 1989) suggest they have different functions in the plant, which is reflected also by their expression profiles. The way in which they are expressed in soybean is controlled by different motifs in the promoter sequences.

The soybean sucrose binding proteins families have a differential expression profile similar to that seen with the soybean lectins. Two sucrose binding proteins share a structural homology with globulin-like seed storage proteins (Elmer *et al.*, 2003). The *GmSBP1* (*Glycine max s*ucrose *b*inding *p*rotein 1) is a seed storage protein that is found in the prevacuolar compartment and the mature protein storage vacuole, however, it has been detected in young sink leaves as well (Elmer *et al.*, 2003). Like the seed storage protein *β-conglycinin*, mRNA levels for both first appear in 5.5mm cotyledons, increase

during seed filling and decrease as seed approached maturity. The proteins however, are detected after imbibition and decrease beginning from 24 hours after imbibition (Elmer *et al.*, 2003). A -2000bp fragment of the promoter of the second sucrose binding protein found in soybean, *GmSBP2*, directs seed and fruit-specific expression, as well as phloem-specific expression in the roots, stem and leaves of reporter genes in tobacco (Contim *et al.*, 2003; Waclawovsky *et al.*, 2006). The GmSBP2 -2000bp promoter region has been shown to act in a combinatorial manner with silencing and activating regions. The two genes are 92% identical in the amino acid sequences and the first 200bp upstream regions of both from the translational start sites display a high level of conservation as compared to sequences further upstream (Elmer *et al.*, 2003). However, there is evidence the two genes are not functional analogues and *GmSBP2* may be involved in the sucrose uptake system with the long distance sugar translocation pathway (Contim *et al.*, 2003). Because the GmSBP proteins are differentially expressed in reproductive and vegetative tissues, they may have different functions (Waclawovsky *et al.*, 2006). The expression of the *GmSBP* proteins 1 and 2 within soybean is similar to what was seen with *Le1* and *Le3* in this study.

Two lectins in *D. biflorus*, the seed lectin and DB58 (stem and leaf lectin), are an example of lectins that are differentially expressed similar to the soybean lectins *Le1* and *Le3*. The *D. biflorus* lectins have 94% sequence identity at the nucleotide level and 88% at the amino acid level, and the 5' and 3' regions also have over 90% nucleotide sequence identity with the exception if a 116bp fragment missing in the 5' region of DB58, the stem and leaf lectin. This fragment is thought to be responsible for the difference in expression between the seed and vegetative lectins (Harada *et al.*, 1990). The promoter

regions of *Le1* and *Le3* do not show as much similarity in their promoter sequences, but like the *Dolichos* lectins, there may be very few, small important regions in the soybean promoters directing specific expression.

The promoters of known vegetative proteins were selected based on similar protein expression profiles. The expression profile of the soybean vegetative lectin (SVL/LE3), is found in the vegetative soybean including the leaves, stems, root, petiole, and at relatively low levels in the cotyledon and seed pods, (Spilatro *et al.*, 1996). Soybean vegetative storage proteins VSPα and VSPβ are also found in vegetative tissues including stems, petioles, pods, roots, nodules and cotyledons after germination, but not in seeds (Wittenbach, 1983; Staswick, 1988). Like the soybean lectins *Le1* and *Le3*, the *D. biflorus* seed lectin and *DB58* are also found in the seed-specific and vegetative lectins, respectively (Harada *et al.*, 1990). Based on their shared expression profiles, many shared motifs between the promoters of the vegetative genes were expected, however, this was not seen. While all the vegetative proteins did contain motifs that were classified as vegetative, none were shared only in the vegetative proteins (i.e. the motif was also present in *Le1, Le2* or other seed promoters analyzed).

A slightly larger lectin family from black locust (*Robina pseudoacacia*) whose members have very similar predicted amino acid sequences, showed a range of expression profiles as well. *Rplec2* mRNA was detected mainly in the inner bark, while *Rplec5* was found in the inner bark, seeds and root, and *Rplec6* was not found anywhere, despite having CAAT and TATA boxes at the same general location in the 5' region as *Rplec2* and *Rplec5* (Yoshida and Tazaki, 1999). Like *Le1* and *Le3* in soybean, a strong

similarity between coding regions (96.1% between *Rplec2* and *Rplec5* amino acid sequences) did not correspond to similar expression profiles.

Genome duplications resulting in gene families allow for changes in the expression profile and function of genes, as long as at least one copy remains to serve the original function, as was seen with the PAL genes in parlsey (Logemann *et al.*, 1995). In soybean, several gene families contain genes that code for proteins whose amino acid sequences and promoter regions are highly conserved, but whose expression profiles vary. The Kunitz trypsin inhibitors, sucrose binding proteins, and *Robinia* lectin families all contain proteins with internally overlapping expression profiles to varying degrees.

The expression profiles of a given gene promoter can be difficult to predict, especially in transgenic plants. The complicated combinatorial nature of promoters may lead to position effects and interactions in the terminator (3' UTR) sequences, that alter the gene expression seen in the native state of the promoter.

In a study involving promoters from four endosperm-specific genes from maize, and using the nos terminator, 22% of first generation stably transformed maize plants showed improper expression, meaning transgenic lines expressed the GUS reporter gene in all tissues tested (flag leaf, developed anther and seed) (Russell and Fromm, 1997). Two of the promoters led to some lines with GUS activity in all the organs tested, and the same pattern of expression occurred in the next (T_2) generation as well, where roots, leaves, embryo, pollen and endosperm were tested (Russell and Fromm, 1997). Although these "all-expressing" lines were exceptional, they show that sometimes, seed-specific promoters can direct expression in vegetative and root tissue, as was seen in a few lines with the *Le1* promoter constructs. Previous studies used the *Le1* terminator (Okamuro *et*

al., 1986; Lindstrom *et al*., 1990; Cho *et al*., 1995; Philip *et al*., 2001) but since our objective was to compare the 5' regions of the three lectin genes, we used the conventional nopaline synthase terminator sequence (Tnos) for all three gene constructs. However, a first set of *Le1* constructs were made with the *Le1* terminator region. All of these lead to very clear seed-specific reporter gene expression (data not shown). The equivalent construct with the Tnos (*Le1* 5'::*gusA*::Tnos) in the majority of the plants lead to clear seed-specific gene expression, but surprisingly it sometimes led to strong GUS activity throughout the seedling. Since the only difference is the native *Le1* 3'UTR, we believe there may be regions, such as the RY-motif, in the *Le1* terminator region that act as silencers in vegetative tissues. Using the *Le1* terminator region with the *Le1* promoter may eliminate this effect seen using the Tnos terminator. Futher studies are planned to test these hypotheses.

The soybean *Le2* gene, though known from genome sequences, was long thought to be a pseudogene because *Le2* mRNA could not be detected, and because a frameshift mutation in the coding region would cause a premature stop in the gene product (Goldberg *et al*., 1983; Vodkin *et al*., 1983, Copley *et al*., unpublished data). However, our study shows that 1.3 kb of the *Le2* 5' region can drive gene expression and thus is a fully functional promoter. The activity of the *Le2* promoter region in transformed *Arabidopsis* does not correlate with *Le2* expression seen thus far in soybean, where nearly no expression was seen. Because of the premature termination codon, nonsense-mediated mRNA decay (NMD) may play a role in silencing *Le2*. NMD is mechanism by which mRNAs that have premature termination codons are degraded to protect the organism from improperly coded gene products (Hori and Watanabe, 2007). A recent study on

nonsense-mediated mRNA decay showed that mRNAs that have 3'UTRs that are over 200-300bp, and or mRNA termination codons more than 50bp upstream of that last exon-exon junction are targeted by the NMD system. The *Le2* sequence in soybean does not have any exon-exon junctions, however, the early termination codon is 384bp from the "proper" termination codon, which will lead to a 499bp long 3'UTR and consequently, it would be a good candidate for NMD. This would explain why the detection of *Le2* mRNA in soybean has been difficult to see under normal conditions, even though, as seen in this study, the promoter is functional. In our study, the nosT was used as terminator, and a properly coded gene was used (as a reporter gene). Another possibility is also RNA mediated gene silencing, die to possible target sites in the *Le2* terminus. The reporter gene expression in *Arabidopsis* due to the *Le2* promoter does not correspond to *Le2* mRNA detection results in soybean, however, using the *Le2* terminator, together with the *Le2* promoter in transgenic *Arabidopsis* may eliminate the expression seen with the Tnos.

Because the SVL (soybean vegetative lectin) protein is found in the vegetative tissues (Spilatro *et al.*, 1996), and the transcripts for *Le3* were found in EST data from vegetative tissues (Strömvik *et al.*, 2004), the same expression profile was expected with the GUS reporter gene driven by the *Le3* promoter. Gus activity using the *Le3* promoter, was found in vegetative tissues, as seen in Figure 4.9, 4.10 and 4.11, which confirmed earlier studies.

Promoter motif analysis

Promoters, like those for *Le1*, *Le2*, and *Le3*, contain many known transcription factor binding motifs, however, the presence of a certain motif does not guarantee that it

is affecting transcription, and presence or absence of its associated binding proteins (transcription factors) is just as critical to the gene regulation (Potenza *et al*., 2004). In addition to this, motifs work in a combinatorial fashion, which can make it difficult to determine which motif plays a more important role than others. To find motifs of interest, a group of genes expressed under the same conditions are searched for sequences that are common to all, or most of them, and then deletion analyses can be performed to test the functionality. Based on this, we have chosen to discuss a few motifs that are relevant to the known and anticipated gene expression, although there may be other interesting motifs present in the lectin promoter sequences. The *Le1, Le2* and *Le3* sequences were analysed against the PLACE database (Higo *et al.*, 1999) to locate known motifs within the promoter sequence, as can be seen in Table 4.1.

Because of economic reasons and because seed-specific promoters are a part of very stringent gene regulation, they have been well-studied in numerous plants and several motifs related to seed-specificity have been found in a wide range of plants.

The -295bp of the 5' upstream region of the β-phaseolin promoter from *Phaseolus vulgaris* contained enough c*is*-elements to confer a high level of seed-specific expression. In this region, 23 *cis*-elements were found to bind proteins during embryogenesis, indicating a complex system was being used to confer seed-specific activity (Li and Hall, 1999). Of these 23 motifs, ten were chosen in a later study for a more detailed functional analysis (Chandrasekharan *et al*., 2003). One of these was a G box (CACGTG), which is also present in the *Le1* promoter sequence, as seen in Figure 4.6 and Table 4.1. In the β-phaseolin gene, one of the G boxes was determined to be the major ABRE motif (ABA-responsive element), and was linked to an E box motif, also found in *Le1*. The E box

motif was determined to be a coupling element, which is a *cis*-element that is only active when combined with an ABRE (Busk and Pagès, 1998; Chandrasekharan *et al.*, 2003). Further experimental analysis, could determine which of the G box motifs found in the *Le1* or *Le2* promoters are functional. However, six E box motifs were found in the *Le1* sequence, and four in the *Le2*, suggesting they play a role in seed-specific expression of the lectins in soybean, as in the bean β-phaseolin. Within the promoter of the latter gene, the CCAAAT box was found to be very important to seed-specificity through site-directed mutagenesis (Chandrasekharan *et al.*, 2003), however it was missing from the *Le1* promoter sequence. This is an example of an important motif present in one promoter that may be absent in others despite their similar expression profiles. The β-phaseolin study showed that a number of elements were redundant and the interactions between *cis*-elemens was the most essential factor in determining seed-specific activity (Chandrasekharan *et al.*, 2003).

Several earlier studies have shown that RY motifs (also known as the legumin box, or *Sph* element), commonly found in seed-specific promoters of both monocots and dicots, are essential to seed-specific promoter activity (Chandrasekharan *et al.*, 2003). The deletion of an RY motif in the seed-specific *Vicia faba* legumin *LeB4* gene promoter caused seed-specific expression to be lost, and reporter gene expression was instead seen at low levels in the leaf. The conclusion was that the motif promoted high activity in the seed and repressed activity in leaves (Baumlein *et al.*, 1992). Later studies confirmed these results and suggested that the RY motifs increase the seed-specific expression, while repressing expression in leaves (Forster *et al.*, 1994; Fujiwara and Beachy, 1994). Of the four RY sequences present in the β-phaseolin promoter, the three most distal RY

boxes appear to repress expression in the radicle region of the embryo, however, when all four RY boxes were mutated, GUS activity was only 11.6% of that seen when all four were in their wild-type form (Chandrasekharan *et al.*, 2003). The soybean β-conglycinin α subunit is another seed-specific promoter containing several RY sequences (Yoshino *et al.*, 2001). The stepwise deletion of RY motifs in this promoter led to a decrease in the GUS activity seen in transgenic *Arabidopsis* seeds, demonstrating how these motifs increased transcriptional activation in seeds (Yoshino *et al.*, 2001). In the soybean *Le1* sequence, at least three RY boxes were present, whereas there were only one each in *Le2* and *Le3*.

The *Le1* promoter contained a variety of motifs, however, it contained more seed-related motifs than the other two promoters, after filtering out motifs that were common to all three promoters. The next largest group was the vegetative-tissue related motifs. Motifs for vegetative expression that were found in the *Le1* sequence were also found in the other seed-specific promoters, including the core motif from a potato MYB homologue gene (*MybSt1*) (Baranowskij *et al.*, 1994), and the pyrimidine box found in the barley *EPB-1* promoter. The latter being involved with expression in germinating seeds (Cercós *et al.*, 1999). This re-emphasizes the point that promoters that drive tissue-specific expression, certainly do contain motifs that are related to gene expression elsewhere in the plant.

The soybean vegetative lectin (SVL/LE3), soybean vegetative storage proteins VSPα and VSPβ and the *D. biflorus* DB58 lectin are all found in vegetative tissues (Wittenbach, 1983; Staswick, 1988; Spilatro and Anderson, 1989; Harada *et al.*, 1990). Because of this, they were expected to contain motifs that were unique to them,

conferring vegetative-specific expression, however, none were seen that were not found in at least one of the non-vegetative promoters examined as well. This would suggest that vegetative expression is the "default" expression of proteins in plants, requiring no specific promoters, and that instead, motifs, such as the RY-motif might be required to silence, rather than induce, expression in vegetative tissue.

In comparison to the *Le1* and *Le2* promoter regions, the *Le3* promoter was expected to contain more vegetative motifs based on its vegetative-specific profile. However, all three promoters had similar numbers of motifs and number of motifs occurrences in the vegetative motif category, despite having different levels of expression in the vegetative tissue. Since vegetative-related motifs were also commonly found in the other seed-specific promoters, it suggests that vegetative-motifs are overall commonly found in promoters. This supports the idea that vegetative expression is much more general/generic than other expression profiles and non-vegetative expression is the result of silencing motifs in vegetative sequences. (i.e. not a vegetative-motif so much as a not-in-vegetative-motif). Our analysis of the lectin promoters of in *D. biflorus* supports this idea.

As discussed earlier, *Le2* was thought to be a pseudogene (Goldberg *et al.*, 1983; Vodkin *et al.*, 1983). Like *Le2*, the coding region of a tomato class I basic chitinase gene contains a frameshift mutation in the open reading frame causing it to code for a truncated protein (Goldberg *et al.*, 1983; Baykal *et al.*, 2006). Both are thought to be pseudogenes, which are considered defective copies of functional genes, *Le1* and *LECHI9*, respectively (Baykal *et al.*, 2006). In an analysis of 303 149 publicly available EST sequences from soybean, only one *Le2* mRNA was been found, while 111 ESTs representing *Le3* were

found (Strömvik *et al.*, 2004). By comparison, spliced ψ*BCH* transcripts in tomato were present at approximately half the level of its homologue, *LECHI9* (Baykal *et al.*, 2006). Transgenic tobacco plants containing the GUS gene driven by the ψ*BCH* promoter showed GUS expression in all wounded tissues tested (leaves, petioles, stems, roots), independently of developmental regulation (Baykal *et al.*, 2006).

Promoter analysis using the PLACE database revealed several stress and defense-related motifs that were in *Le2*, but not *Le1* or *Le3*, including the MYB consensus I binding site, core DRE motif site, the box-L-like sequence and GCC-box core which were also all in the ψ*BCH* promoter sequence.

Despite the varied vegetative GUS activity seen in the experimental results, all three promoters had similar amounts of vegetative motifs, proportional to the total motifs in their respective sequences. The protein sequence for *Le2* is more similar to *Le1* than to *Le3*, but its promoter activity profile seems to fall in between the two. Interestingly, *Le2* has more motifs related to defense than have either *Le1* or *Le3*. *Le2* had over three times as many stress motifs as *Le1*, and nearly twice as many stress motifs as *Le3*. The *Le2* 5' region contained fewer seed-specific motifs than *Le1* or *Le3*. Based on these results, it would be interesting to test *Le2* for stress and defense related inducibility as several lectins in other plants have been shown to be toxic to insects, such as the mannose-binding lectin *GNA* (*Galanthus nivulis* agglutinin) from *Galanthus nivulis* (snowdrop) (Wool *et al.*, 1992; Pusztai *et al.*, 1993; Down *et al.*, 1996; Nagadhara *et al.*, 2004).

Section 6:
Conclusions and suggestions for future studies

The versatility and popularity of soybean as an agricultural crop highlights the importance of conducting research related to its genome and genome functions. To date, very few soybean promoters have been well-characterized and further research into non-constitutive promoters is crucial to increase basic knowledge of plant gene regulation and discovery of very specific promoters for fine-tuned biotechnology applications. Perhaps the most well-known soybean promoter is that of the soybean lectin, *Le1*, which is specific to the seed. In this study, we have isolated the gene and promoter regions for two *Le1* gene homologues, *Le2* and *Le3*, *in silico* characterized the content of regulatory motifs of their promoters, as well as functionally characterized the reporter gene expression profiles resulting from the promoters in transgenic *Arabidopsis thaliana.* Our study confirms previous *in silico* analysis, which predicted that the *Le2* and *Le3* promoters drive gene expression in a specific developmental and tissue specific manner, different from the *Le1* promoter. We show that the *Le2* promoter is a weak driver of GUS reporter gene expression in *Arabidopsis* rosette and cauline leaves, flowers, siliques and seeds but not in roots, whereas the *Le3* promoter is a strong driver of gene expression in all tissues, including roots, but excluding mature seeds. A bioinformatics analysis of the three soybean lectin promoters using the online PLACE database reveals many motifs consistent with the expression profiles. However, because many "contradictory" motifs were also found in the promoter sequences (e.g. seed motifs in the *Le3* promoter), it is clear that a bioinformatics promoter motif analysis alone at this point in time is not enough to predict the expression profile of a promoter sequence.

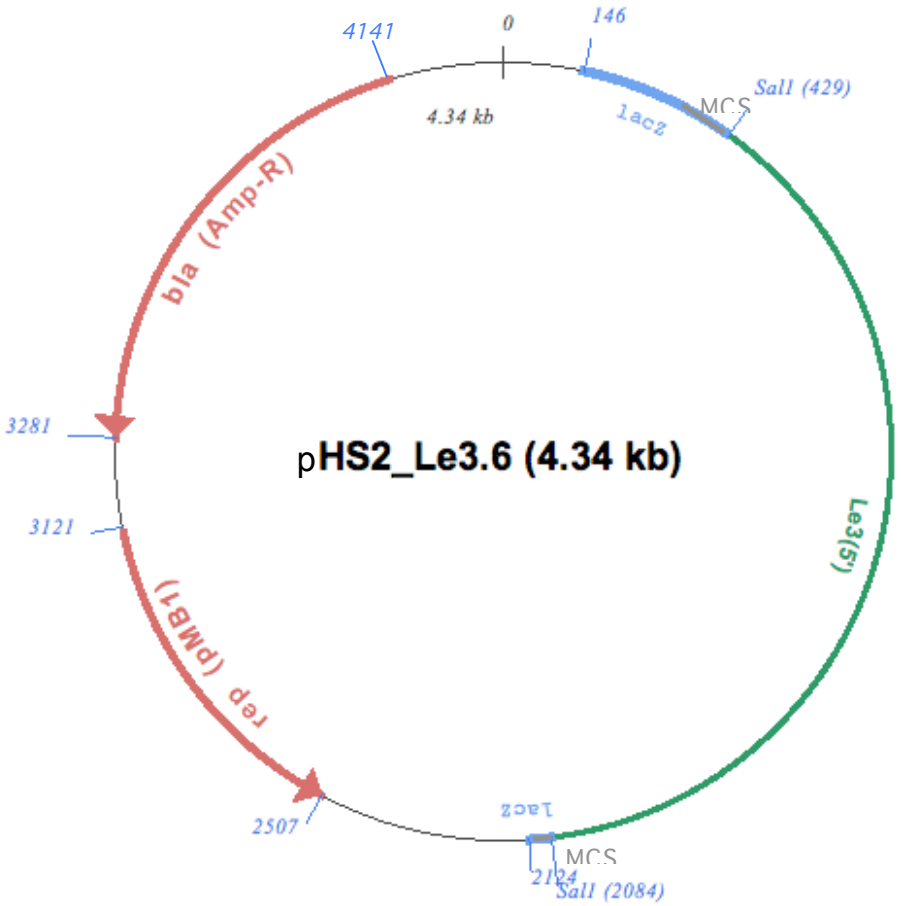
To functionally pinpoint the most essential regions of the promoters and to correlate these with the motif data, future studies should include deletion series of the *Le2* and *Le3* promoters. In addition, future studies should further examine the role of the 3' regions of the genes. With an exploding amount of publications on RNA interference (RNAi), it is becoming increasingly clear that the 3'UTR, or terminator sequence, can be of high importance for tissue specific gene regulation. So far, virtually no *Le2* expression has been seen in soybean, which is contrary to our evidence that the *Le2* promoter if fully functional, albeit in a heterologous plant. Using the *Le2* terminator in lieu of the nopaline synthase terminator with the *Le2* promoter in transgenic *Arabidopsis* may thus eliminate the effect we see and in effect silence the reporter gene expression there. Constructs and *Arabidopsis* transformations testing the effects of the *Le1*, *Le2* and *Le3* 3'UTR terminator sequences would provide new information into the control of gene expression extending beyond the 5' region of the genes.

Whereas *Le3* is likely the gene for the soybean vegetative lectin (SVL), previously described only as a protein, the possible function of *Le2*, or alternatively its pseudogene status, is yet to be investigated in soybean. If *Le2* is expressed in soybean under some conditions not yet tested (stress or defense), despite yielding a truncated protein, it could potentially retain some lectin function and be an energy-saving response to the stress condition. Plans are being made to express *Le2*, driven by a stronger promoter, in soybean, in order to detect whether a protein will be produced. Tests for nonsense mediated decay (NMD) activity in soybean could also be carried out, and the *Le2* terminator sequence for microRNA binding sites.

A forth lectin gene homologue, *Le4* has been identified in soybean EST data. It is predicted *in silico* that this is a close homologue of *Le3* and would possibly have a very similar expression profile and promoter. Future studies should confirm this experimentally in soybean.

This study of the promoters from the small gene family of the soybean legume lectins has provided important information regarding differential gene regulation in homologous genes. It is clear that much is still to be learned from promoter motif analysis and that experimentation coupled with *in silico* predictions will yield the best knowledge of gene regulation. As the soybean genome becomes available within the next few years, studies like this will be important references for gene and promoter annotation.

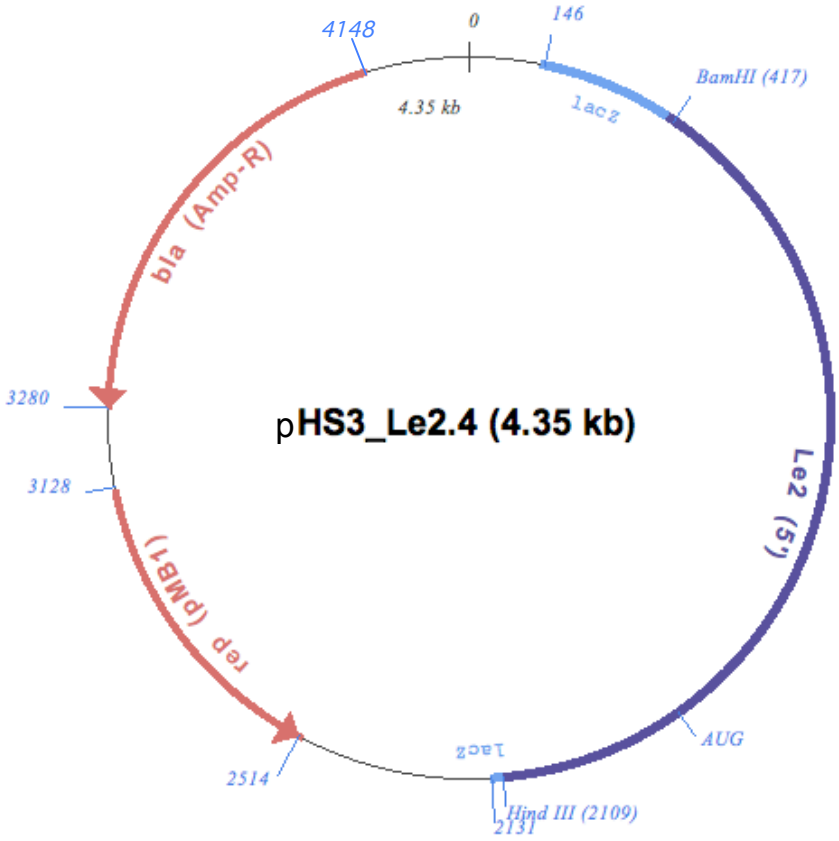
Appendix I



(http://www.changbioscience.com/res/resmap.html)

Plasmid Name:	HS2_Le3.6
Plasmid Size:	4341 bp
Lab:	Dr. M. V. Stromvik
Constructed by:	Hanaa Saeed
Construction date:	January 14, 2005
Comments/ References:	• The 1655 bp soybean Le3 5'-upstream region was isolated from the GenomeWalker DL2 library and blunt-end cloned into a pUC19 vector cut with SalI. • This vector was used as the template for the sequencing of the soybean Le3 5'-upstream region • Total vector length is 4341bp • Le3(5') insert length is 1655bp

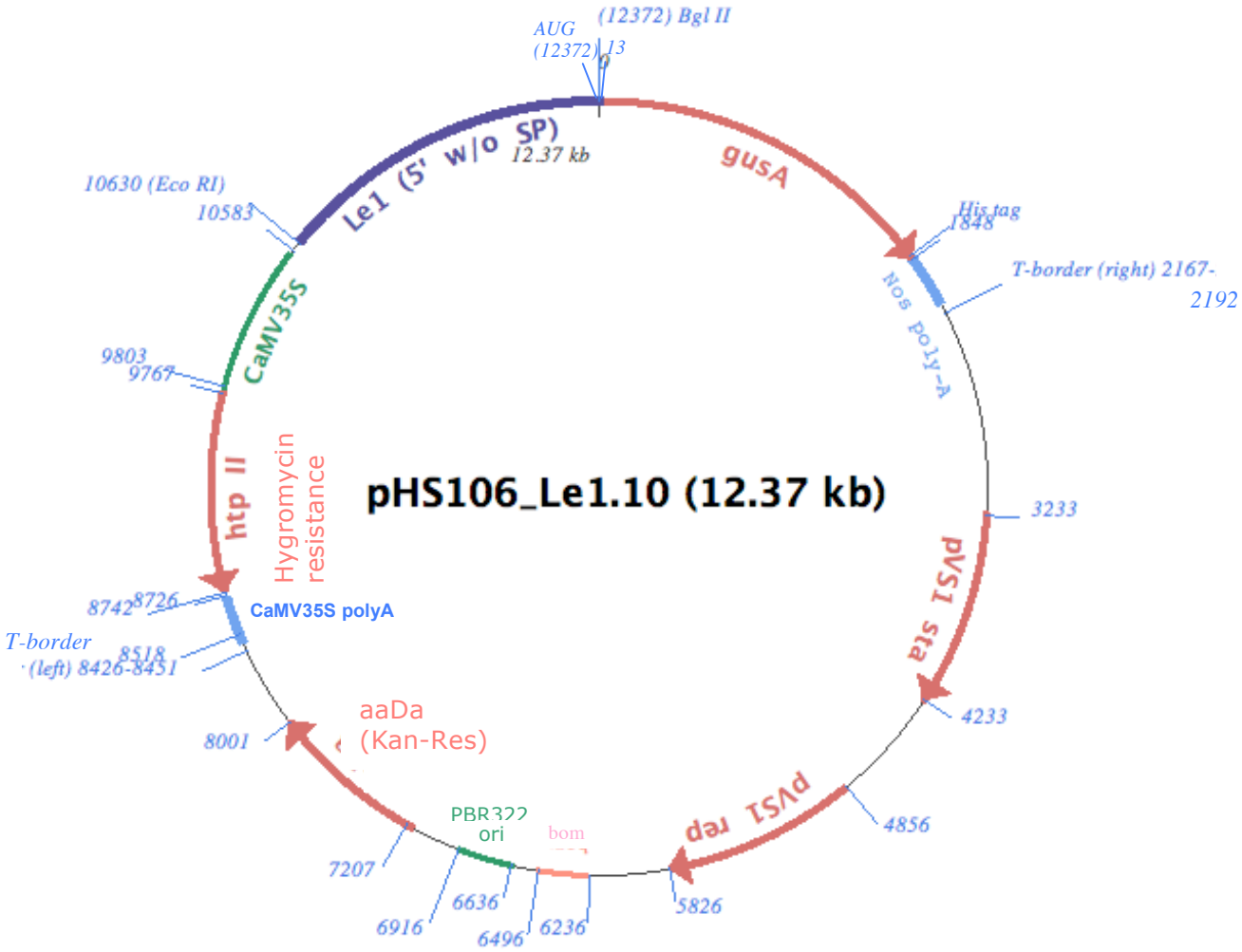
Appendix II



(http://www.changbioscience.com/res/resmap.html)

Plasmid Name:	HS3_Le2.4
Plasmid Size:	4348 bp
Lab:	Dr. M. V. Stromvik
Constructed by:	Hanaa Saeed
Construction date:	January 24, 2005
Comments/ References:	• The 1662 bp soybean Le2 5'-upstream region was isolated from the GenomeWalker DL3 library and blunt-end cloned into a pUC19 vector cut with BamHI and HindIII. • This vector was used as the template for the sequencing of the soybean Le2 5'-upstream region, and for cloning the 5' region, with and without the signal peptide in later experiments • Vectors HS3_Le2.8 and HS3_Le2.10 are identical to this one, but have ot been sequenced • Total vector length is 4348bp • Le3(5') insert length is 1662bp

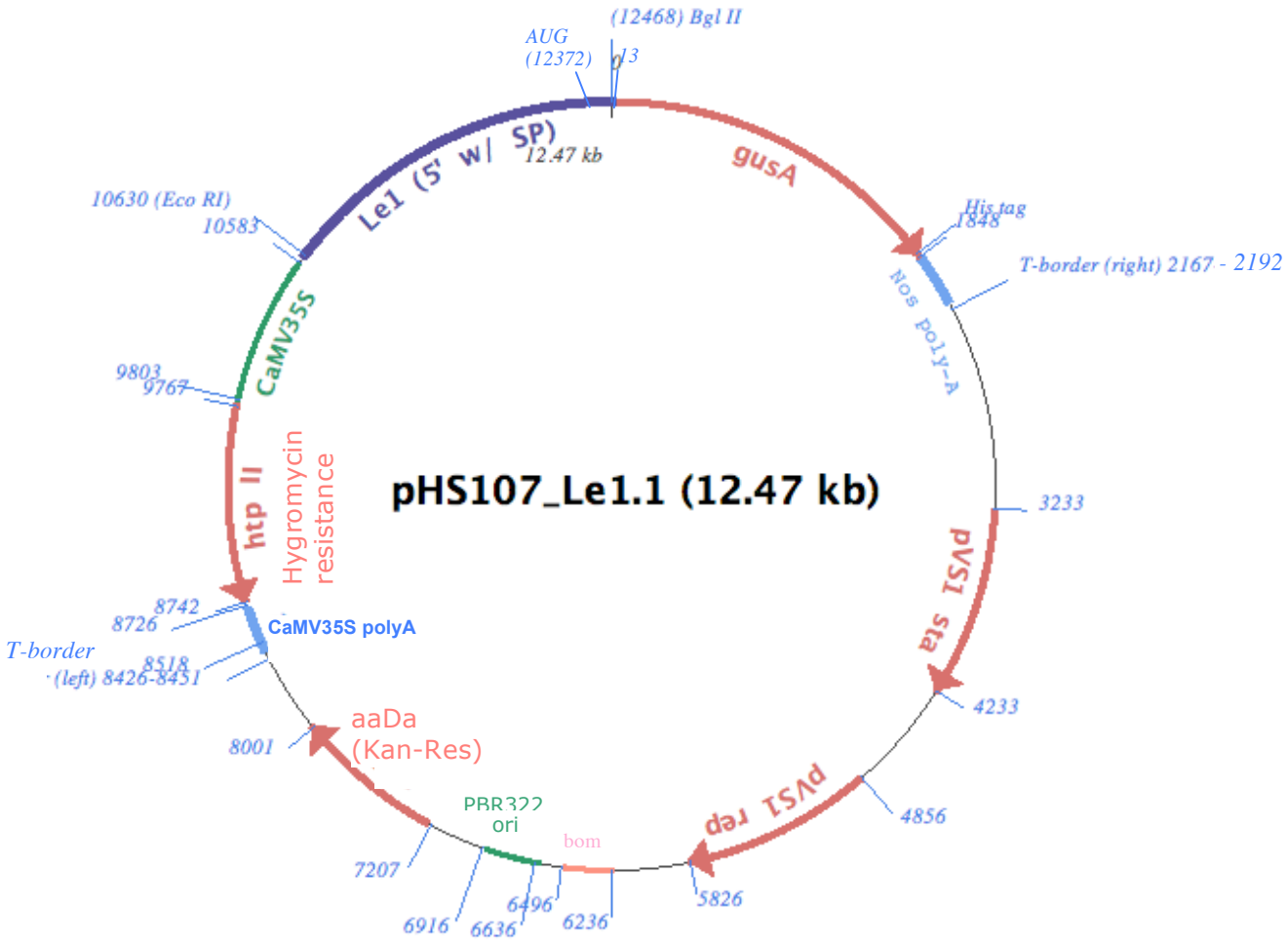
Appendix III



(http://www.changbioscience.com/res/resmap.html)

Plasmid Name:	HS106_Le1.10
Plasmid Size:	12372 bp
Lab:	Dr. M. V. Stromvik
Constructed by:	Hanaa Saeed
Construction date:	October 31, 2006
Comments/ References:	• The 1771 bp soybean *Le1* 5'-upstream region, not including the signal peptide was isolated from the pGLGUS-21 vector (Vodkin lab, contains signal peptide) and *EcoRI/BglII* cloned into a pCAMBIA 1391Xa vector cut with the same enzymes. • This vector was used to transform Agrobacterium to make the HS131_Le1.1 stock used for plant transformations

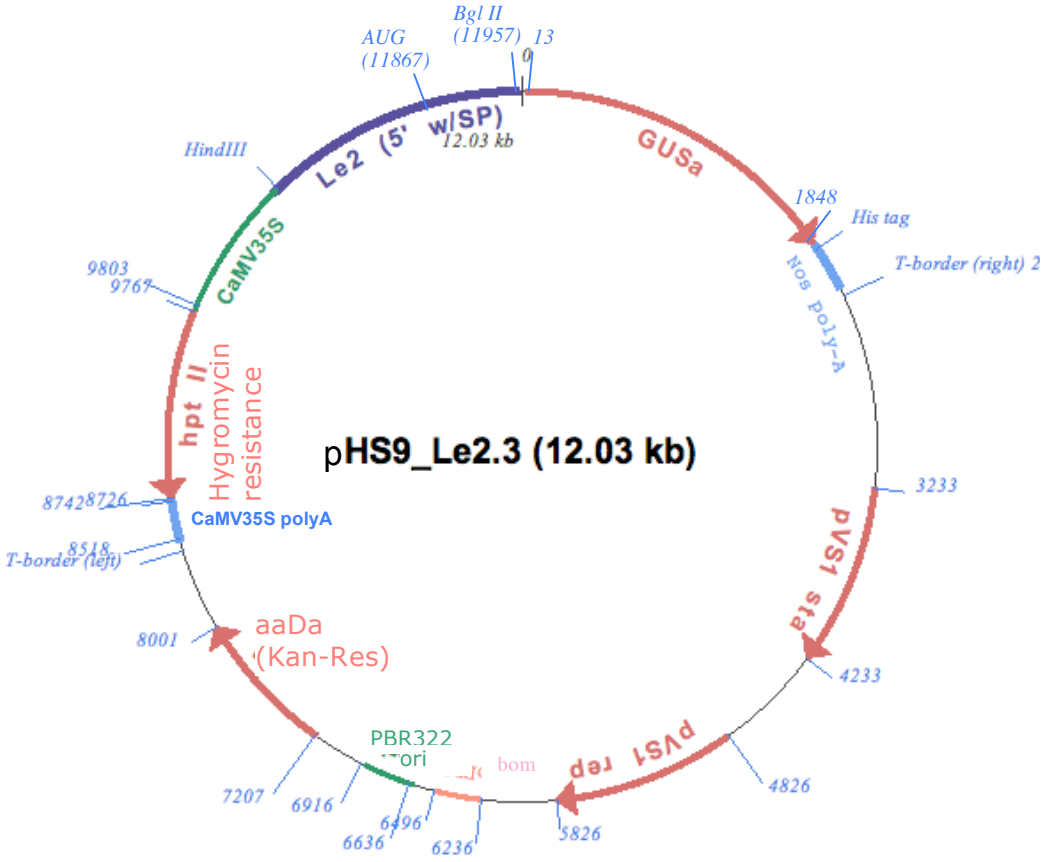
Appendix IV



http://www.changbioscience.com/res/resmap.html)

Plasmid Name:	HS107_Le1.1
Plasmid Size:	12468 bp
Lab:	Dr. M. V. Stromvik
Constructed by:	Hanaa Saeed
Construction date:	November 2, 2006
Comments/ References:	• The 1771 bp soybean *Le1* 5'-upstream region, including the signal peptide (96 bp) was isolated from the pGLGUS-21 vector (Vodkin lab, contains signal peptide) and *EcoRI/BglII* cloned into a pCAMBIA 1391Xa vector cut with the same enzymes. • This vector was used to transform Agrobacterium to make the HS130_Le1.1 stock used for plant transformations

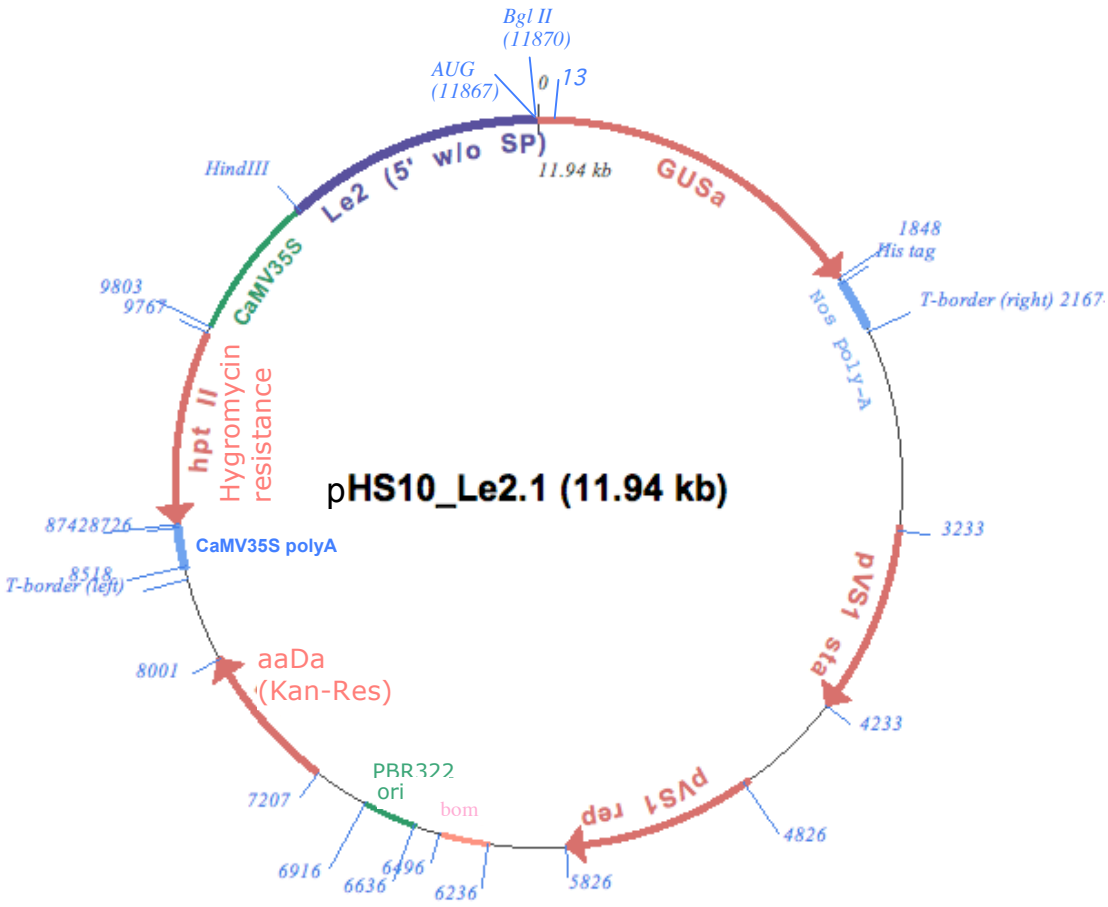
Appendix V



(http://www.changbioscience.com/res/resmap.html)

Plasmid Name:	HS9_Le2.3
Plasmid Size:	12028 bp
Lab:	Dr. M. V. Stromvik
Constructed by:	Hanaa Saeed
Construction date:	May 20, 2005
Comments/ References:	• The 1383 bp soybean Le2 5'-upstream region, including the signal peptide was isolated from the HS3_Le2.4 vector and *HindIII/BglII* cloned into a pCAMBIA 1391Xa vector cut with the same enzymes. • This vector was used to transform Agrobacterium to make the HS17_Le2.1 stock used for plant transformations

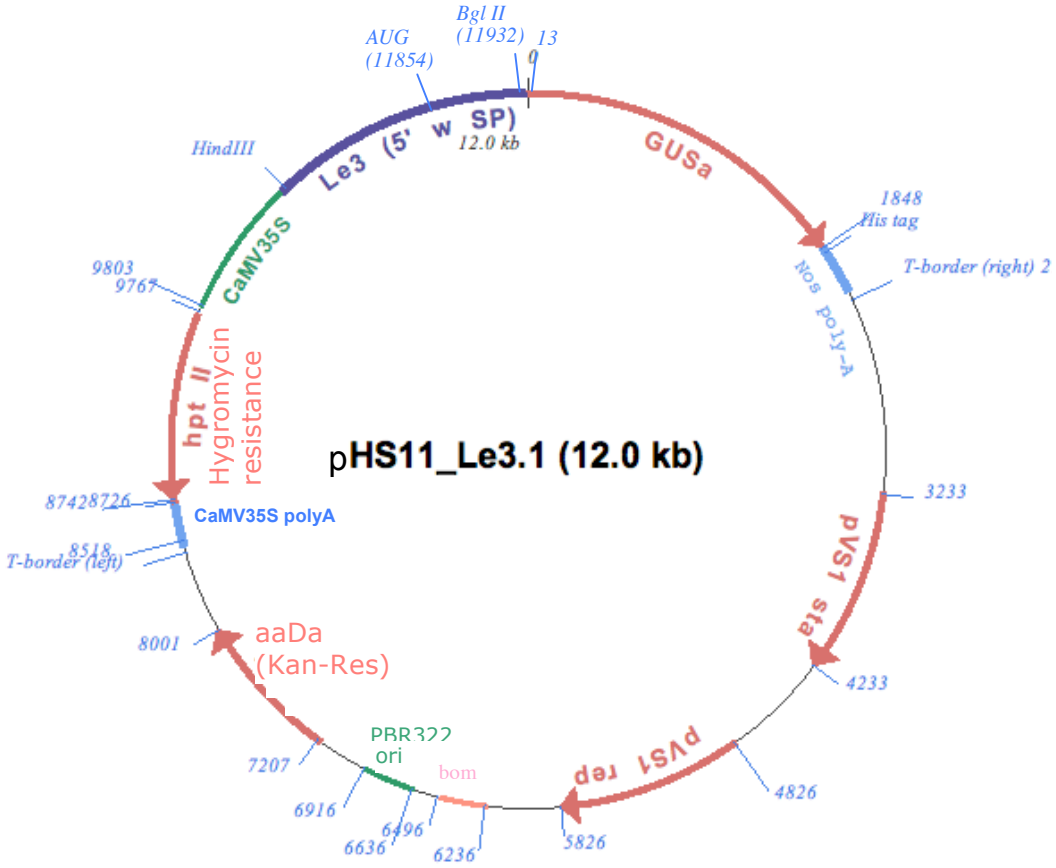
Appendix VI



(http://www.changbioscience.com/res/resmap.html)

Plasmid Name:	HS10_Le2.1
Plasmid Size:	11941 bp
Lab:	Dr. M. V. Stromvik
Constructed by:	Hanaa Saeed
Construction date:	May 20, 2005
Comments/ References:	• The 1383 bp soybean Le2 5'-upstream region, including the start codon was isolated from the HS3_Le2.4 vector and *HindIII/BglII* cloned into a pCAMBIA 1391Xa vector cut with the same enzymes. • This vector was used to transform Agrobacterium to make the HS18_Le2.1 stock used for plant transformations

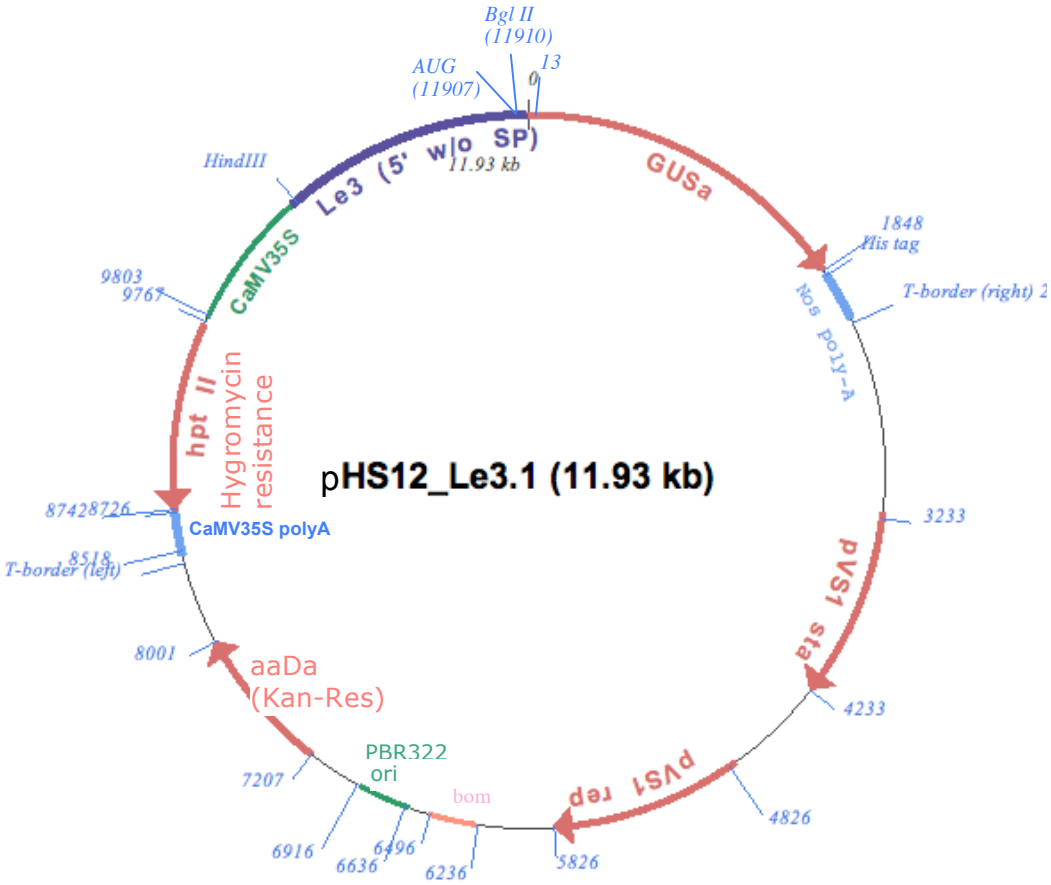
Appendix VII



(http://www.changbioscience.com/res/resmap.html)

Plasmid Name:	HS11_Le3.1
Plasmid Size:	12003 bp
Lab:	Dr. M. V. Stromvik
Constructed by:	Hanaa Saeed
Construction date:	May 20, 2005
Comments/ References:	• The 1280 bp soybean Le3 5'-upstream region, including the signal peptide was isolated from the HS2_Le3.6 vector and *HindIII/BglII* cloned into a pCAMBIA 1391Xa vector cut with the same enzymes. • This vector was used to transform Agrobacterium to make the HS19_Le3.1 stock used for plant transformations

Appendix VIII



(http://www.changbioscience.com/res/resmap.html)

Plasmid Name:	HS12_Le3.1
Plasmid Size:	11925 bp
Lab:	Dr. M. V. Stromvik
Constructed by:	Hanaa Saeed
Construction date:	May 20, 2005
Comments/ References:	• The 1280 bp soybean Le3 5'-upstream region, including the start codon was isolated from the HS2_Le3.6 vector and *HindIII/BglII* cloned into a pCAMBIA 1391Xa vector cut with the same enzymes. • This vector was used to transform Agrobacterium to make the HS20_Le3.1 stock used for plant transformations

References

Abe, H., Yamaguchi-Shinozaki, K., Urao, T., Iwasaki, T., Hosokawa, D. and Shinozaki, K. (1997) Role of *Arabidopsis* MYC and MYB homologs in drought-and abscisic acid-regulated gene expression. *Plant Cell*, **9**, 1859-1868.

Akella, V. and Lurquin, P.F. (1993) Expression in cowpea seedlings of chimeric transgenes after electroporation into seed-derived embryos. *Plant Cell Reports*, **12**, 110-117.

Alvarado, M.C., Zsigmond, L.M., Kovacs, I., Cseplo, A., Koncz, C. and Szabados, L.M. (2004) Gene trapping with firefly luciferase in *Arabidopsis*. Tagging of stress-responsive genes. *Plant Physiology*, **134**, 18-27.

Arumuganathan, K. and Earle, E. (1991) Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter*, **9**, 208-219.

Atchison, M.L. (1988) Enhancers: Mechanisms of action and cell specificity. *Annual Review of Cell Biology*, **4**, 127-153.

Baranowskij, N., Frohberg, C., Prat, S. and Willmitzer, L. (1994) A novel DNA binding protein with homology to Myb oncoproteins containing only one repeat can function as a transcriptional activator. *EMBO Journal*, **13**, 5383-5392.

Bauchrowitz, M.A., Barker, D.G. and Truchet, G. (1996) Lectin genes are expressed throughout root nodule development and during nitrogen-fixation in the *Rhizobium-Medicago* symbiosis. *The Plant Journal*, **9**, 31-43.

Baumlein, H., Nagy, I., Villarroel, R., Inze, D. and Wobus, U. (1992) *cis*-analysis of a seed protein gene promoter: the conservative RY repeat CATGCATG within the legumin box is essential for tissue-specific expression of a legumin gene. *The Plant Journal*, **2**, 233-239.

Baykal, U., Moyne, A.-L. and Tuzun, S. (2006) A frameshift in the coding region of a novel tomato class I basic chitinase gene makes it a pseudogene with a functional wound-responsive promoter. *Gene*, **376**, 37-46.

Bechtold, N., Ellis, J. and Pelletier, G. (1993) *In planta Agrobacterium* mediated gene transfer by infiltration of adult *Arabidopsis thaliana* plants. *Comptes rendus de l'Académie des sciences. Série III, Sciences de la vie*, **316**, 1194-1199.

Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology*, **340**, 783-795.

Benfey, P.N. and Chua, N.-H. (1990) The cauliflower mosaic virus 35S promoter: Combinatorial regulation of transcription in plants. *Science*, **250**, 959-966.

Berrocal-Lobo, M., Molina, A. and Solano, R. (2002) Constitutive expression of *ETHYLENE-RESPONSE-FACTOR1* in *Arabidopsis* confers resistance to several necrotrophic fungi. *The Plant Journal*, **29**, 23-32.

Britten, R.J. and Davidson, E.H. (1969) Gene regulation for higher cells: A theory. *Science*, **165**, 349-357.

Busk, P.K. and Pagès, M. (1998) Regulation of abscisic acid-induced transcription. *Plant Molecular Biology*, **37**, 425-435.

Castillo-Davis, C.I., Hartl, D.L. and Achaz, G. (2004) *cis*-regulatory and protein evolution in orthologous and duplicate genes. *Genome Research*, **14**, 1530-1536.

Cercós, M., Gomez-Cadenas, A. and Ho, T.-H.D. (1999) Hormonal regulation of a cysteine proteinase gene, *EPB-1*, in barley aleurone layers: *cis*- and *trans*-acting elements involved in the co-ordinated gene expression regulated by gibberellins and abscisic acid. *The Plant Journal*, **19**, 107-118.

Chandrasekharan, M.B., Bishop, K.J. and Hall, T.C. (2003) Module-specific regulation of the β-phaseolin promoter during embryogenesis. *The Plant Journal*, **33**, 853-866.

Cho, M.-J., Widholm, J. and Vodkin, L.O. (1995) Cassettes for seed-specific expression tested in transformed embryogenic cultures of soybean. *Plant Molecular Biology Reporter*, **13**, 255-269.

Chong, J., Baltz, R., Schmitt, C., Beffa, R., Fritig, B. and Saindrenan, P. (2002) Downregulation of a pathogen-responsive tobacco UDP-Glc:phenylpropanoid glucosyltransferase reduces scopoletin glucoside accumulation, enhances oxidative stress, and weakens virus resistance. *Plant Cell*, **14**, 1093-1107.

Chrispeels, M.J. and Raikhel, N.V. (1991) Lectins, lectin genes, and their role in plant defense. *Plant Cell*, **3**, 1-9.

Christou, P., McCabe, D.E. and Swain, W.F. (1988) Stable transformation of soybean callus by DNA-coated gold particles. *Plant Cell Physiology*, **87**, 671-674.

Christou, P., McCabe, D.E., Martinell, B.J. and Swain, W.F. (1990) Soybean genetic engineering - commercial production of transgenic plants. *Trends in Biotechnology*, **8**, 145-151.

Clough, S.J. and Bent, A.F. (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *The Plant Journal*, **16**, 735-743.

Contim, L.A.S., Waclawovsky, A.J., Delu-Filho, N., Pirovani, C.P., Clarindo, W.R., Loureiro, M.E., Carvalho, C.R. and Fontes, E.P.B. (2003) The soybean sucrose binding protein gene family: genomic organization, gene copy number and tissue-specific expression of the *SBP2* promoter. *Journal of Experimental Botany*, **54**, 2643-2653.

Cornish-Bowden, A. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research*, **13**, 3021-3030.

Darnowski, D. and Vodkin, L.O. (2002) A soybean lectin-GFP fusion labels the vacuoles in developing *Arabidopsis thaliana* embryos. *Plant Cell Reports*, **20**, 1033-1038.

de Castro, E., Sigrist, C.J.A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A. and Hulo, N. (2006) ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research*, **34**, W362-365.

Diaz, C.L., Melchers, L.S., Hooykaas, P.J.J., Lugtenberg, B.J.J. and Kijne, J.W. (1989) Root lectin as a determinant of host-plant specificity in the *Rhizobium*-legume symbiosis. *Nature*, **338**, 579-581.

DiRita, V.J. and Gelvin, S.B. (1987) Deletion analysis of the mannopine synthase gene promoter in sunflower crown gall tumors and *Agrobacterium tumefaciens*. *Molecular and General Genetics MGG*, **207**, 233-241.

Doebley, J. and Lukens, L. (1998) Transcriptional regulators and the evolution of plant form. *Plant Cell*, **10**, 1075-1082.

Down, R.E., Gatehouse, A.M.R., Hamilton, W.D.O. and Gatehouse, J.A. (1996) Snowdrop lectin inhibits development and decreases fecundity of the glasshouse potato aphid (*Aulacorthum solani*) when administered in vitro and via transgenic plants both in laboratory and glasshouse trials. *Journal of Insect Physiology*, **42**, 1035-1045.

Elmayan, T. and Vaucheret, H. (1996) Expression of single copies of a strongly expressed 35S transgene can be silenced post-transcriptionally. *The Plant Journal*, **9**, 787-797.

Elmer, A., Chao, W. and Grimes, H. (2003) Protein sorting and expression of a unique soybean cotyledon protein, GmSBP, destined for the protein storage vacuole. *Plant Molecular Biology*, **52**, 1089-1106.

Etzler, M.E. (1985) Plant lectins: Molecular and biological aspects. *Annual Review of Plant Physiology*, **36**, 209-234.

Etzler, M.E. (1998) From Structure to activity: New insights into the functions of legume lectins. *Trends in Glycoscience and Glycotechnology*, **10**, 247-255.

Finnegan, E.J., Genger, R.K., Peacock, W.J. and Dennis, E.S. (1998) DNA methylation in plants. *Annual Review of Plant Physiology and Plant Molecular Biology*, **49**, 223-247.

Fischer, R.L. and Goldberg, R.B. (1982) Structure and flanking regions of soybean seed protein genes. *Cell*, **29**, 651-660.

Forster, C., Arthur, E., Crespi, S., Hobbs, S.L.A., Mullineaux, P. and Casey, R. (1994) Isolation of a pea (*Pisum sativum*) seed lipoxygenase promoter by inverse polymerase chain reaction and characterization of its expression in transgenic tobacco. *Plant Molecular Biology*, **26**, 235-248.

Fujiwara, T. and Beachy, R.N. (1994) Tissue-specific and temporal regulation of a β-conglycinin gene: Roles of the RY repeat and other *cis*-acting elements. *Plant Molecular Biology*, **24**, 261-272.

Gachon, C., Baltz, R. and Saindrenan, P. (2004) Over-expression of a scopoletin glucosyltransferase in *Nicotiana tabacum* leads to precocious lesion formation during the hypersensitive response to tobacco mosaic virus but does not affect virus resistance. *Plant Molecular Biology*, **54**, 137-146.

Gallie, D.R. and Walbot, V. (1992) Identification of the motifs within the tobacco mosaic virus 5'-leader responsible for enhancing translation. *Nucleic Acids Research*, **20**, 4631-4638.

Geisler, M., Kleczkowski, L.A. and Karpinski, S. (2006) A universal algorithm for genome-wide in silicio identification of biologically significant gene promoter putative *cis*-regulatory-elements; identification of new elements for reactive oxygen species and sucrose signaling in *Arabidopsis*. *The Plant Journal*, **45**, 384-398.

Goldberg, R.B., Hoschek, G., Tam, S.H., Ditta, G.S. and Breidenbach, R.W. (1981) Abundance, diversity, and regulation of mRNA sequence sets in soybean embryogenesis. *Developmental Biology*, **83**, 201-217.

Goldberg, R.B., Hoschek, G. and Vodkin, L.O. (1983) An insertion sequence blocks the expression of a soybean lectin gene. *Cell*, **33**, 465-475.

Goldberg, R.B., Barker, S.J. and Perez-Grau, L. (1989) Regulation of gene expression during plant embryogenesis. *Cell*, **56**, 149-160.

Goldstein, I.J., Hughes, R.C., Monsigny, M., Osawa, T. and Sharon, N. (1980) What should be called a lectin? *Nature*, **285**, 66.

Gonzales, M.D., Archuleta, E., Farmer, A., Gajendran, K., Grant, D., Shoemaker, R., Beavis, W.D. and Waugh, M.E. (2005) The legume information system (LIS): An

integrated information resource for comparative legume biology. *Nucleic Acids Research*, **33**, D660-665.

Goy, P.A., Signer, H., Reist, R., Aichholz, R., Blum, W., Schmidt, E. and Kessmann, H. (1993) Accumulation of scopoletin is associated with the high disease resistance of the hybrid *Nicotiana glutinosa* x *Nicotiana debneyi*. *Planta*, **191**, 200-206.

Grace, M.L., Chandrasekharan, M.B., Hall, T.C. and Crowe, A.J. (2004) Sequence and spacing of TATA box elements are critical for accurate initiation from the β-phaseolin promoter. *Journal of Biological Chemistry*, **279**, 8102-8110.

Hadley, H.H. and Hymowitz, T. (1973) Speciation and cytogenetics. In B.E. Calldwell, Editor. *Agronomy Monographs*, Vol. 16, ASA, CSSA, SSSA, Madison, Wisconsin: pp. 97-116.

Hajdukiewicz, P., Svab, Z. and Maliga, P. (1994) The small, versatilep PZP family of *Agrobacterium* binary vectors for plant transformation. *Plant Molecular Biology*, **25**, 989-994.

Harada, J.J., Spadoro-Tank, J., Maxwell, J.C., Schnell, D.J. and Etzler, M.E. (1990) Two lectin genes differentially expressed in *Dolichos biflorus* differ primarily by a 116-base pair sequence in their 5' flanking regions. *Journal of Biological Chemistry*, **265**, 4997-5001.

Haseloff, J., Siemering, K.R., Prasher, D.C. and Hodge, S. (1997) Removal of a cryptic intron and subcellular localization of green fluorescent protein are required to mark transgenic *Arabidopsis* plants brightly. *Proceedings of the National Academy of Sciences*, **94**, 2122-2127.

Hightower, R.C. and Meagher, R.B. (1985) Divergence and differential expression of soybean actin genes. *EMBO Journal*, **4**, 1-8.

Higo, K., Ugawa, Y., Iwamoto, M. and Korenaga, T. (1999) Plant *cis*-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Research*, **27**, 297-300.

Hofer, J. and Ellis, N. (2002) Conservation and diversification of gene function in plant development. *Current Opinion in Plant Biology*, **5**, 56-61.

Holsters, M., Waele, D., Depicker, A., Messens, E., Montagu, M. and Schell, J. (1978) Transfection and transformation of *Agrobacterium tumefaciens*. *Molecular and General Genetics MGG*, **163**, 181-187.

Hong, J.K., Lee, S.C. and Hwang, B.K. (2005) Activation of pepper basic PR-1 gene promoter during defense signaling to pathogen, abiotic and environmental stresses. *Gene*, **356**, 169-180.

Hori, K. and Watanabe, Y. (2007) Context analysis of termination codons in mRNA that are recognized by plant NMD. *Plant Cell Physiology*, pcm075.

Huang, X. (1994) On global sequence alignment. *Computer Applications in the Biosciences*, **10**, 227-235.

Jefferson, R.A., Kavanagh, T.A. and Bevan, M.V. (1987) GUS fusions: β-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO Journal*, **6**, 3901-3907.

Jofuku, K.D. and Goldberg, R.B. (1989) Kunitz trypsin inhibitor genes are differentially expressed during the soybean life cycle and in transformed tobacco plants. *Plant Cell*, **1**, 1079-1093.

Joshi, C.P. (1987) An inspection of the domain between putative TATA box and translation start site in 79 plant genes. *Nucleic Acids Research*, **15**, 6643-6653.

Kariola, T., Brader, G., Helenius, E., Li, J., Heino, P. and Palva, E.T. (2006) *EARLY RESPONSIVE TO DEHYDRATION 15*, a negative regulator of abscisic acid responses in *Arabidopsis*. *Plant Physiology*, **142**, 1559-1573.

Kay, R., Chan, A., Daly, M. and McPherson, J. (1987) Duplication of CaMV 35S promoter sequences creates a strong enhancer for plant genes. *Science*, **236**, 1299-1302.

Kazan, K. (2003) Alternative splicing and proteome diversity in plants: The tip of the iceberg has just emerged. *Trends in Plant Science*, **8**, 468-471.

Klein, T.M., Wolf, E.D., Wu, R. and Sanford, J.C. (1987) High-velocity microprojectiles for delivering nucleic acids into living cells. *Nature*, **327**, 70-73.

Kluth, A., Sprunck, S., Becker, D., Lörz, H. and Lütticke, S. (2002) 5′ deletion of a *gbss1* promoter region from wheat leads to changes in tissue and developmental specificities. *Plant Molecular Biology*, **49**, 665-678.

Koncz, C. and Schell, J. (1986) The promoter of TL-DNA gene 5 controls the tissue-specific expression of chimaeric genes carried by a novel type of *Agrobacterium* binary vector. *Molecular and General Genetics MGG*, **204**, 383-396.

Lam, E. and Chua, N.H. (1989) ASF-2: A factor that binds to the cauliflower mosaic virus 35S promoter and a conserved GATA motif in Cab promoters. *Plant Cell*, **1**, 1147-1156.

Lescot, M., Dehais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., Rouze, P. and Rombauts, S. (2002) PlantCARE, a database of plant *cis*-acting regulatory

elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Research*, **30**, 325-327.

Li, G. and Hall, T.C. (1999) Footprinting in vivo reveals changing profiles of multiple factor interactions with the β-phaseolin promoter during embryogenesis. *The Plant Journal*, **18**, 633-641.

Li, Z., Jayasankar, S. and Gray, D.J. (2004) Bi-directional duplex promoters with duplicated enhancers significantly increase transgene expression in grape and tobacco. *Transgenic Research*, **13**, 143-154.

Lindstrom, J.T., Vodkin, L.O., Harding, R.W. and Goeken, R.M. (1990) Expression of soybean lectin gene deletions in tobacco. *Developmental Genetics*, **11**, 160-167.

Logemann, E., Parniske, M. and Hahlbrock, K. (1995) Modes of expression and common structural features of the complete phenylalanine ammonia-lyase gene family in parsley. *Proceedings of the National Academy of Sciences*, **92**, 5905-5909.

Maeda, K., Kimura, S., Demura, T., Takeda, J. and Ozeki, Y. (2005) DcMYB1 acts as a transcriptional activator of the carrot phenylalanine ammonia-lyase gene (*DcPAL1*) in response to elicitor treatment, UV-B irradiation and the dilution effect. *Plant Molecular Biology*, **59**, 739-752.

McElroy, D., Zhang, W., Cao, J. and Wu, R. (1990) Isolation of an efficient actin promoter for use in rice transformation. *Plant Cell*, **2**, 163-171.

Morikami, A., Matsunaga, R., Tanaka, Y., Suzuki, S., Mano, S. and Nakamura, K. (2005) Two *cis*-acting regulatory elements are involved in the sucrose-inducible expression of the sporamin gene promoter from sweet potato in transgenic tobacco. *Molecular Genetics and Genomics*, **272**, 690-699.

Nagadhara, D., Ramesh, S., Pasalu, I.C., Rao, Y.K., Sarma, N.P., Reddy, V.D. and Rao, K.V. (2004) Transgenic rice plants expressing the snowdrop lectin gene (*gna*) exhibit high-level resistance to the whitebacked planthopper (*Sogatella furcifera*). *TAG Theoretical and Applied Genetics*, **109**, 1399-1405.

Nagaya, Nagaya, S., Yoshida, Yoshida, K., Kato, Kato, K., Akasaka, Akasaka, K., Shinmyo and Shinmyo, A. (2001) An insulator element from the sea urchin *Hemicentrotus pulcherrimus* suppresses variation in transgene expression in cultured tobacco cells. *Molecular Genetics and Genomics*, **265**, 405-413.

Nelson, R.T. and Shoemaker, R. (2006) Identification and analysis of gene families from the duplicated genome of soybean using EST sequences. *BMC Genomics*, **7**, 204.

Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering Design and Selection*, **10**, 1-6.

Nielsen, N.C., Dickinson, C.D., Cho, T.J., Thanh, V.H., Scallon, B.J., Fischer, R.L., Sims, T.L., Drews, G.N. and Goldberg, R.B. (1989) Characterization of the glycinin gene family in soybean. *Plant Cell*, **1**, 313-328.

Odell, J.T., Nagy, F. and Chua, N.-H. (1985) Identification of DNA sequences required for activity of the cauliflower mosaic virus 35S promoter. *Nature*, **313**, 810-812.

Ohler, U. and Niemann, H. (2001) Identification and analysis of eukaryotic promoters: Recent computational approaches. *Trends in Genetics*, **17**, 56-60.

Okamuro, J.K., Jofuku, K.D. and Goldberg, R.B. (1986) Soybean seed lectin gene and flanking nonseed protein genes are developmentally regulated in transformed tobacco plants. *Proceedings of the National Academy of Sciences*, **83**, 8240-8244.

Olhoft, Olhoft, P., Somers and Somers, D. (2001) *L*-cysteine increases *Agrobacterium*-mediated T-DNA delivery into soybean cotyledonary-node cells. *Plant Cell Reports*, **20**, 706-711.

Ouyang, S. and Buell, C.R. (2004) The TIGR plant repeat databases: A collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Research*, **32**, D360-363.

Pauli, S., Rothnie, H.M., Chen, G., He, X. and Hohn, T. (2004) The cauliflower mosaic virus 35S promoter extends into the transcribed region. *The Journal of Virology*, **78**, 12120-12128.

Philip, R., Darnowski, D.W., Sundararaman, V., Cho, M.-J. and Vodkin, L.O. (1998) Localization of β-glucuronidase in protein bodies of transgenic tobacco seed by fusion to an amino terminal sequence of the soybean lectin gene. *Plant Science*, **137**, 191-204.

Philip, R., Darnowski, D.W., Maughan, P.J. and Vodkin, L.O. (2001) Processing and localization of bovine β-casein expressed in transgenic soybean seeds under control of a soybean lectin expression cassette. *Plant Science*, **161**, 323-335.

Potenza, C., Aleman, L. and Sengupta-Gopalan, C. (2004) Targeting transgene expression in research, agricultural, and environmental applications: Promoters used in plant transformation. *In Vitro Cellular & Developmental Biology - Plant*, **40**, 1-22.

Pusztai, A., Grant, G., Spencer, R.J., Duguid, T.J., Brown, D.S., Ewen, S.W., Peumans, W.J., Van Damme, E.J.M. and Bardocz, S. (1993) Kidney bean lectin-

induced *Escherichia coli* overgrowth in the small intestine is blocked by GNA, a mannose-specific lectin. *Journal of Applied Bacteriology*, **75**, 360-368.

Reich, T.J., Iyer, V.N. and Miki, B.L. (1986) Efficient transformation of alfalfa protoplasts by the intranuclear microinjection of Ti plasmids. *Nature Biotechnology*, **4**, 1001-1004.

Rizzi, C., Galeoto, L., Zoccatelli, G., Vincenzi, S., Chignola, R. and Peruffo, A.D.B. (2003) Active soybean lectin in foods: Quantitative determination by ELISA using immobilised asialofetuin. *Food Research International*, **36**, 815-821.

Rombauts, S., Dehais, P., Van Montagu, M. and Rouze, P. (1999) PlantCARE, a plant *cis*-acting regulatory element database. *Nucleic Acids Research*, **27**, 295-296.

Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouze, P. and Van de Peer, Y. (2003) Computational approaches to identify promoters and *cis*-regulatory elements in plant genomes. *Plant Physiology*, **132**, 1162-1176.

Rüdiger, H. (1984) On the physiological role of plant lectins. *BioScience*, **34**, 95-99.

Rüdiger, H. and Gabius, H.-J. (2001) Plant lectins: Occurrence, biochemistry, functions and applications. *Glycoconjugate Journal*, **18**, 589-613.

Russell, D.A. and Fromm, M.E. (1997) Tissue-specific expression in transgenic maize of four endosperm promoters from maize and rice. *Transgenic Research*, **6**, 157-168.

Schäffner, A.R. and Sheen, J. (1991) Maize *rbcS* promoter activity depends on sequence elements not found in dicot *rbcS* promoters. *Plant Cell*, **3**, 997-1012.

Scharf, S.J., Horn, G.T. and Erlich, H.A. (1986) Direct cloning and sequence analysis of enzymatically amplified genomic sequences. *Science*, **233**, 1076-1078.

Schlueter, J.A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J.J. and Shoemaker, R.C. (2004) Mining EST databases to resolve evolutionary events in major crop species. *Genome*, **47**, 868-876.

Shahmuradov, I.A., Gammerman, A.J., Hancock, J.M., Bramley, P.M. and Solovyev, V.V. (2003) PlantProm: A database of plant promoter sequences. *Nucleic Acids Research*, **31**, 114-117.

Shahmuradov, I.A., Solovyev, V.V. and Gammerman, A.J. (2005) Plant promoter prediction with confidence estimation. *Nucleic Acids Research*, **33**, 1069-1076.

Shoemaker, R., Keim, P., Vodkin, L.O., Retzel, E., Clifton, S.W., Waterston, R., Smoller, D., Coryell, V., Khanna, A., Erpelding, J., Gai, X., Brendel, V., Raph-Schmidt, C., Shoop, E.G., Vielweber, C.J., Schmatz, M., Pape, D., Bowers, Y.,

Theising, B., Martin, J., Dante, M., Wylie, T. and Granger, C. (2002) A compilation of soybean ESTs: Generation and analysis *Genome*, **45**, 329-338.

Shoemaker, R.C., Polzin, K., Labate, J., Specht, J., Brummer, E.C., Olson, T., Young, N., Concibido, V., Wilcox, J., Tamulonis, J.P., Kochert, G. and Boerma, H.R. (1996) Genome duplication in soybean (*Glycine* subgenus *soja*). *Genetics*, **144**, 329-338.

Shultz, J.L., Kurunam, D., Shopinski, K., Iqbal, M.J., Kazi, S., Zobrist, K., Bashir, R., Yaegashi, S., Lavu, N., Afzal, A.J., Yesudas, C.R., Kassem, M.A., Wu, C., Zhang, H.B., Town, C.D., Meksem, K. and Lightfoot, D.A. (2006) The soybean genome database (SoyGD): A browser for display of duplicated, polyploid, regions and sequence tagged sites on the integrated physical and genetic maps of *Glycine max*. *Nucleic Acids Research*, **34**, D758-765.

Singh, K.B. (1998) Transcriptional regulation in plants: The importance of combinatorial control. *Plant Physiology*, **118**, 1111-1120.

Slightom, J.L., Sun, S.M. and Hall, T.C. (1983) Complete nucleotide sequence of a french bean storage protein gene: Phaseolin. *Proceedings of the National Academy of Sciences*, **80**, 1897-1901.

Sojikul, P., Buehner, N. and Mason, H.S. (2003) A plant signal peptide-hepatitis B surface antigen fusion protein with enhanced stability and immunogenicity expressed in plant cells. *Proceedings of the National Academy of Sciences*, **100**, 2209-2214.

Solano, R., Nieto, C., Avila, J., Cañas, L., Diaz, I. and Paz-Ares, J. (1995) Dual DNA binding specificity of a petal epidermis-specific MYB transcription factor (MYB.Ph3) from *Petunia hybrida*. *EMBO Journal*, **14**, 1773-1784.

Somers, D.A., Samac, D.A. and Olhoft, P.M. (2003) Recent advances in legume transformation. *Plant Physiology*, **131**, 892-899.

Song, Q., Marek, L., Shoemaker, R., Lark, K., Concibido, V., Delannay, X., Specht, J. and Cregan, P. (2004) A new integrated genetic linkage map of the soybean. *TAG Theoretical and Applied Genetics*, **109**, 122-128.

Sparvoli, F., Lanave, C., Santucci, A., Bollini, R. and Lioi, L. (2001) Lectin and lectin-related proteins in lima bean (*Phaseolus lunatus* L.) seeds: Biochemical and evolutionary studies. *Plant Molecular Biology*, **45**, 587-597.

Spilatro, S.R. and Anderson, J.M. (1989) Characterization of a soybean leaf protein that Is related to the seed lectin and is increased with pod removal. *Plant Cell Physiology*, **90**, 1387-1393.

Spilatro, S.R., Cochran, G.R., Walker, R.E., Cablish, K.L. and Bittner, C.C. (1996) Characterization of a new lectin of soybean vegetative tissues. *Plant Physiology*, **110**, 825-834.

Springer, P.S. (2000) Gene traps: Tools for plant development and genomics. *Plant Cell*, **12**, 1007-1020.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G.R., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D. and Birney, E. (2002) The bioperl toolkit: Perl modules for the life sciences. *Genome Research*, **12**, 1611-1618.

Staswick, P.E. (1988) Soybean vegetative storage protein structure and gene expression. *Plant Physiology*, **87**, 250-254.

Stomp, A.M. (1992) Histochemical localization of β-glucuronidase. In: S.R. Gallagher, Editor, *GUS Protocols: Using the GUS gene as a reporter of gene expression*. Academic Press Inc., San Diego: pp. 103-113.

Strittmatter, G. and Chua, N.-H. (1987) Artificial combination of two *cis*-regulatory elements generates a unique pattern of expression in transgenic plants. *Proceedings of the National Academy of Sciences*, **84**, 8986-8990.

Strömvik, M.V., Sundararaman, V.P. and Vodkin, L.O. (1999) A novel promoter from soybean that is active in a complex developmental pattern with and without its proximal 650 base pairs. *Plant Molecular Biology*, **41**, 217-231.

Strömvik, M.V., Thibaud-Nissen, F. and Vodkin, L.O. (2004) Mining soybean expressed sequence tags and microarray data. In: J.T. Romero, Editor, *Recent Advances in Phytochemistry*, Vol. Secondary Metabolism in Model Systems. Elsevier, Pergamon, Oxford: pp. 177-195.

Sunilkumar, G., Connell, J.P., Smith, C.W., Reddy, A.S. and Rathore, K.S. (2002) Cotton α-globulin promoter: Isolation and functional characterization in transgenic cotton, *Arabidopsis*, and tobacco. *Transgenic Research*, **11**, 347-359.

Talbot, C.F. and Etzler, M.E. (1978) Isolation and characterization of a protein from leaves and stems of *Dolichos biflorus* that cross reacts with antibodies to the seed lectin. *Biochemistry*, **17**, 1474-1479.

Toki, S., Takamatsu, S., Nojiri, C., Ooba, S., Anzai, H., Iwata, M., Christensen, A.H., Quail, P.H. and Uchimiya, H. (1992) Expression of a maize ubiquitin gene promoter-bar chimeric gene in transgenic rice plants. *Plant Physiology*, **100**, 1503-1507.

Urao, T., Yamaguchi-Shinozaki, K., Urao, S. and Shinozaki, K. (1993) An *Arabidopsis myb* homolog is induced by dehydration stress and its gene product binds to the conserved MYB recognition sequence. *Plant Cell*, **5**, 1529-1539.

Vaghchhipawala, Z.E., Schlueter, J.A., Shoemaker, R.C. and Mackenzie, S.A. (2004) Soybean FGAM synthase promoters direct ectopic nematode feeding site activity. *Genome*, **47**, 404-413.

Van Damme, E.J.M., Barre, A., Smeets, K., Torrekens, S., Van Leuven, F., Rouge, P. and Peumans, W.J. (1995) The bark of *Robinia pseudoacacia* contains a complex mixture of lectins (characterization of the proteins and the cDNA clones). *Plant Physiology*, **107**, 833-843.

Van Damme, E.J.M., Peumans, W.J., Barre, A., Roug, eacute and Pierre (1998) Plant lectins: A composite of several distinct families of structurally and evolutionary related proteins with diverse biological roles. *Critical Reviews in Plant Sciences*, **17**, 575 - 692.

Vaucheret, H., Beclin, C., Elmayan, T., Feuerbach, F., Godon, C., Morel, J.-B., Mourrain, P., Palauqui, J.-C. and Vernhettes, S. (1998) Transgene-induced gene silencing in plants. *The Plant Journal*, **16**, 651-659.

Venter, M. (2007) Synthetic promoters: genetic control through *cis* engineering. *Trends in Plant Science*, **12**, 118-124.

Vickers, C., Xue, G. and Gresshoff, P. (2006) A novel *cis*-acting element, ESP, contributes to high-level endosperm-specific expression in an oat globulin promoter. *Plant Molecular Biology*, **62**, 195-214.

Vodkin, L.O., Rhodes, P.R. and Goldberg, R.B. (1983) A lectin gene insertion has the structural features of a transposable element. *Cell*, **34**, 1023-1031.

Vodkin, L.O. and Raikhel, N.V. (1986) Soybean lectin and related proteins in seeds and roots of Le+ and Le- soybean varieties. *Plant Physiology*, **81**, 558-565.

Vodkin, L.O., Khanna, A., Shealy, R., Clough, S., Gonzalez, D., Philip, R., Zabala, G., Thibaud-Nissen, F., Sidarous, M., Stromvik, M., Shoop, E., Schmidt, C., Retzel, E., Erpelding, J., Shoemaker, R., Rodriguez-Huete, A., Polacco, J., Coryell, V., Keim, P., Gong, G., Liu, L., Pardinas, J. and Schweitzer, P. (2004) Microarrays for global expression constructed with a low redundancy set of 27,500 sequenced cDNAs representing an array of developmental stages and physiological conditions of the soybean plant. *BMC Genomics*, **5**, 73.

Waclawovsky, A.J., Freitas, R.L., Rocha, C.S., Contim, L.A.S. and Fontes, E.P.B. (2006) Combinatorial regulation modules on *GmSBP2* promoter: A distal cis-regulatory domain confines the *SBP2* promoter activity to the vascular tissue in vegetative organs. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, **1759**, 89-98.

Wandelt, C.I., Khan, M.R., Craig, S., Schroeder, H.E., Spencer, D. and Higgins, T.J. (1992) Vicilin with carboxy-terminal KDEL is retained in the endoplasmic reticulum and accumulates to high levels in the leaves of transgenic plants. *The Plant Journal*, **2**, 181-192.

Wendel, J.F. (2000) Genome evolution in polyploids. *Plant Molecular Biology*, **42**, 225-249.

Wittenbach, V.A. (1983) Purification and characterization of a soybean leaf storage glycoprotein. *Plant Physiology*, **73**, 125-129.

Wool, I.G., Gluck, A. and Endo, Y. (1992) Ribotoxin recognition of ribosomal RNA and a proposal for the mechanism of translocation. *Trends in Biochemical Sciences*, **17**, 266-269.

Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V. and Romano, L.A. (2003) The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, **20**, 1377-1419.

Wu, C., Sun, S., Nimmakayala, P., Santos, F.A., Meksem, K., Springman, R., Ding, K., Lightfoot, D.A. and Zhang, H.-B. (2004) A BAC- and BIBAC-based physical map of the soybean genome. *Genome Research*, **14**, 319-326.

Xiao, K., Liu, J., Dewbre, G., Harrison, M. and Wang, Z.Y. (2006) Isolation and characterization of root-specific phosphate transporter promoters from *Medicago truncatula*. *Plant Biology*, **8**, 439-449.

Xu, B. and Timko, M. (2004) Methyl jasmonate induced expression of the tobacco putrescine N-methyltransferase genes requires both G-box and GCC-motif elements. *Plant Molecular Biology*, **55**, 743-761.

Yabor, L., Arzola, M., Aragón, C., Hernández, M., Arencibia, A. and Lorenzo, J. (2006) Biochemical side effects of genetic transformation of pineapple. *Plant Cell, Tissue and Organ Culture*, **86**, 63-67.

Yevtushenko, D.P., Romero, R., Forward, B.S., Hancock, R.E., Kay, W.W. and Misra, S. (2005) Pathogen-induced expression of a cecropin A-melittin antimicrobial peptide gene confers antifungal resistance in transgenic tobacco. *Journal of Experimental Botany*, **56**, 1685-1695.

Yoshida, K. and Tazaki, K. (1999) Expression patterns of the genes that encode lectin or lectin-related polypeptides in *Robinia pseudoacacia*. *Functional Plant Biology*, **26**, 495-502.

Yoshino, M., Kanazawa, A., Tsutsumi, K.-i., Nakamura, I. and Shimamoto, Y. (2001) Structure and characterization of the gene encoding α subunit of soybean β-conglycinin. *Genes & Genetic Systems*, **76**, 99-105.

Zhang, W., Peumans, W.J., Barre, A., Houles Astoul, C., Rovira, P., Rougé, P., Proost, P., Truffa-Bachi, P., Jalali, A.A.H. and Van Damme, E.J.M. (2000) Isolation and characterization of a jacalin-related mannose-binding lectin from salt-stressed rice (*Oryza sativa*) plants. *Planta*, **210**, 970-978.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W.L., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E. and Ecker, J.R. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell*, **126**, 1189-1201.

Zhu, K., Huesing, J.E., Shade, R.E., Bressan, R.A., Hasegawa, P.M. and Murdock, L.L. (1996) An insecticidal *N*-acetylglucosamine-specific lectin gene from *Griffonia simplicifolia* (Leguminosae). *Plant Physiology*, **110**, 195-202.