8-15-2017 12:00 AM

Transcriptional Regulation of Cell-type Specific Expression in the Arabidopsis Root

Keegan M. Leckie, *The University of Western Ontario*

Supervisor: Dr. Ryan Austin, *The University of Western Ontario*
Joint Supervisor: Dr. David Smith, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Biology

Abstract

Characterizing transcription factor interactions with their corresponding binding sites is crucial for understanding how gene expression is regulated by DNA sequence. A more comprehensive understanding of this process could have benefits in synthetic promoter design and creation of genetically modified organisms. Herein, the promoters of genes exhibiting cell-type specific expression within a single layer of the Arabidopsis root are analyzed to identify *cis*-regulatory motifs implicated in cell-type specific expression. *De novo* motif prediction identifies multiple motif candidates overly represented in the promoter sequences of co-expressed genes specific for epidermal, cortex, and endodermal expression. Several endodermal specific putative motifs are further analyzed for positional biases and tested *in planta*. *A priori* mapping of known *cis*-regulatory motifs catalogued in publicly available databases is also performed. Results show that cell-types contain different statistically significant enrichment patterns of both predicted and known *cis*-regulatory motifs. These results will help future research in designing cell-type specific synthetic promoters.

Acknowlegements

Firstly, I would like to acknowledge my supervisor Dr. Ryan Austin for not only being an excellent mentor during my time in grad school, but for the previous two years as a intern and undergraduate. Over the last four and a half years he has helped refine my skills as a researcher and realize my passion for the sciences. I would like to thank Dr. Austin for his guidance and support the entire time. Without the opportunity given to me to become an intern in his lab, I would be on a very different path today. I could not imagine a better supervisor, or friend during this very exciting period of my life. I will miss our discussions.

Besides my supervisor, I would like to thank the professors on my thesis committee whose advice over the last two years has been invaluable. Thank you Dr. David Smith, my co-supervisor, and my advisors Dr. Kathleen Hill and Dr. Yuhai Cui. I would also like to thank Dr. Susanne Kohalmi and Dr. Krzysztof Szczyglowski for their help over the years. I appreciate all the advice and resources provided to me by them during graduate school.

I would also like to thank everyone apart of the Austin lab, both past and present. Thank you Fiona Bergin, Brianne Robinson, Tina Homayouni, Ellen Hamilton, and Shown Hoogstra. My time spent researching with you guys has been most enjoyable. I would also like to thank our lab technician Mana Croft. I could not have asked for a better friend in the lab.

Last, but far from least, I would like to thank my loving parents. If it wasn't for your unconditional support over the years, both financially and emotionally, I would not have been able to pursue the level of higher education I've enjoyed so much over the last seven years.

Keywords: Cell-type expression, *cis*-regulatory motifs, *Arabidopsis* root, gene regulation

Contents

List of Figures

List of Tables

List of Appendices

List of Abbreviations, Symbols, and Nomenclature

AGI	*Arabidopsis* Genome Initiative
bHLH	basic Helix-Loop-Helix
bZIP	basic Leucine Zipper
CBF	C-repeat Binding Factor
COR	Cold-regulated
CRE	*Cis*-Regulatory Element
CRM	*Cis*-Regulatory Module
DBD	DNA-Binding-Domain
EM	Expectation Maximization
EMC	Endodermal Minor Clade
ESM	Endodermal Specific Motif
FACS	Fluorescence-Activated Cell Sorting
FD	Functional Depth
GBA	Guilt by Association
GC	Guanine-Cytosine
GO	Gene Ontology
HAT	Histone Acetyltransferase
HDAC	Histone Deacetylase
HOPACH	Hierarchical Ordered Partitioning And Collapsing Hybrid
IC	Information Content
NGS	Next Generation Sequencing
PCC	Pearson Correlation Coefficient
PcG	Polycomb-Group Proteins
PIC	Transcription Preinitiation Complex
PRC	Polycomb Repressive Complex
PRE	Polycomb Response Elements
PSSM	Position Specific Scoring Matrix
PWM	Position Weight Matrix
ROC	Receiver Operating Characteristic
TF	Transcription Factor
TRE	Trithorax Response Elements
TrxG	Trithorax-Group Proteins
TSS	Transcriptional Start Site
UTR	Untranslated Region

Chapter 1

Introduction and pertinent scientific literature

Current understanding of how DNA sequences regulate gene transcription remains incomplete. Specific DNA sequence patterns found in proximity to coding sequence can control when a gene is expressed and in what tissue or cell-type it is expressed in. Identifying DNA sequence patterns that confine gene transcription within a single cell-type would be beneficial for targeting transgene expression within genetically engineered organisms. DNA sequence patterns that control cell-type specific expression within *Arabidopsis thaliana* have so far not been identified and is therefore the main focus of this research.

1.1 Transcription in *Arabidopsis*

The *Arabidopsis* genome contains exactly 33,602 genes encoded within 120 Mb of genomic DNA sequence[1] (Berardini et al., 2015). The expression of these genes is tightly regulated to maintain biological functions and development. Gene expression is regulated by proximal DNA sequences found upstream of a gene's coding sequence in regions known as gene promoters. Changes in DNA expression are induced by nuclear proteins called transcription factors (TF), which contain DNA binding domains to interact with gene promoters and activate or suppress transcription. The *Arabidopsis* genome encodes over 1,500 TFs involved in regulat-

[1]This gene number includes all known and predicted genes, including transposable elements, pseudogenes and non-protein coding RNA species.

ing its genome (Palaniswamy et al., 2006). Understanding how gene expression is regulated by TF-promoter interactions can help elucidate larger complex regulatory networks and their function. While advancements have been made in understanding gene transcription in *Arabidopsis*, a complete understanding of transcriptional control at the molecular level remains incomplete.

1.1.1 *Cis*-regulatory motifs in gene promoters

Cis-regulatory motifs, also known as *cis*-regulatory elements (CREs), are short, specific DNA sequences which act as the binding sites for TFs. Promoters act to regulate gene transcription via functional CREs within their primary sequence (Wellmer and Riechmann, 2005)(Figure 1.1). Thus, when and where a gene is transcribed depends on whether the appropriate TFs are present to occupy their corresponding binding sites. Protein-DNA interactions between gene promoters and TFs produce a favourable environment for RNA polymerase to initiate transcription. Gene expression is therefore dependent on the proper CREs being present in a gene's promoter, where different CRE combinations, also known as *cis*-regulatory modules (CRM), are able to produce different expression patterns. CREs are typically 8-16 bps in length in eukaryotes (Matys, 2006). They are also degenerate, meaning that motif sequences can vary to a certain degree while still remaining genetically active (D'haeseleer, 2006). This proves challenging for characterizing motifs as it is often difficult to determine whether similar sequences of known motifs are in fact degenerate versions of motifs instead of inactive sequence patterns. In *Arabidopsis*, many CREs have been discovered (Weirauch et al., 2014). Because of their sequence degeneracy, many CREs can often be bound by one or more TFs. These TFs however, are typically closely related and often found within the same family (Weirauch and Hughes, 2011; Weirauch et al., 2014).

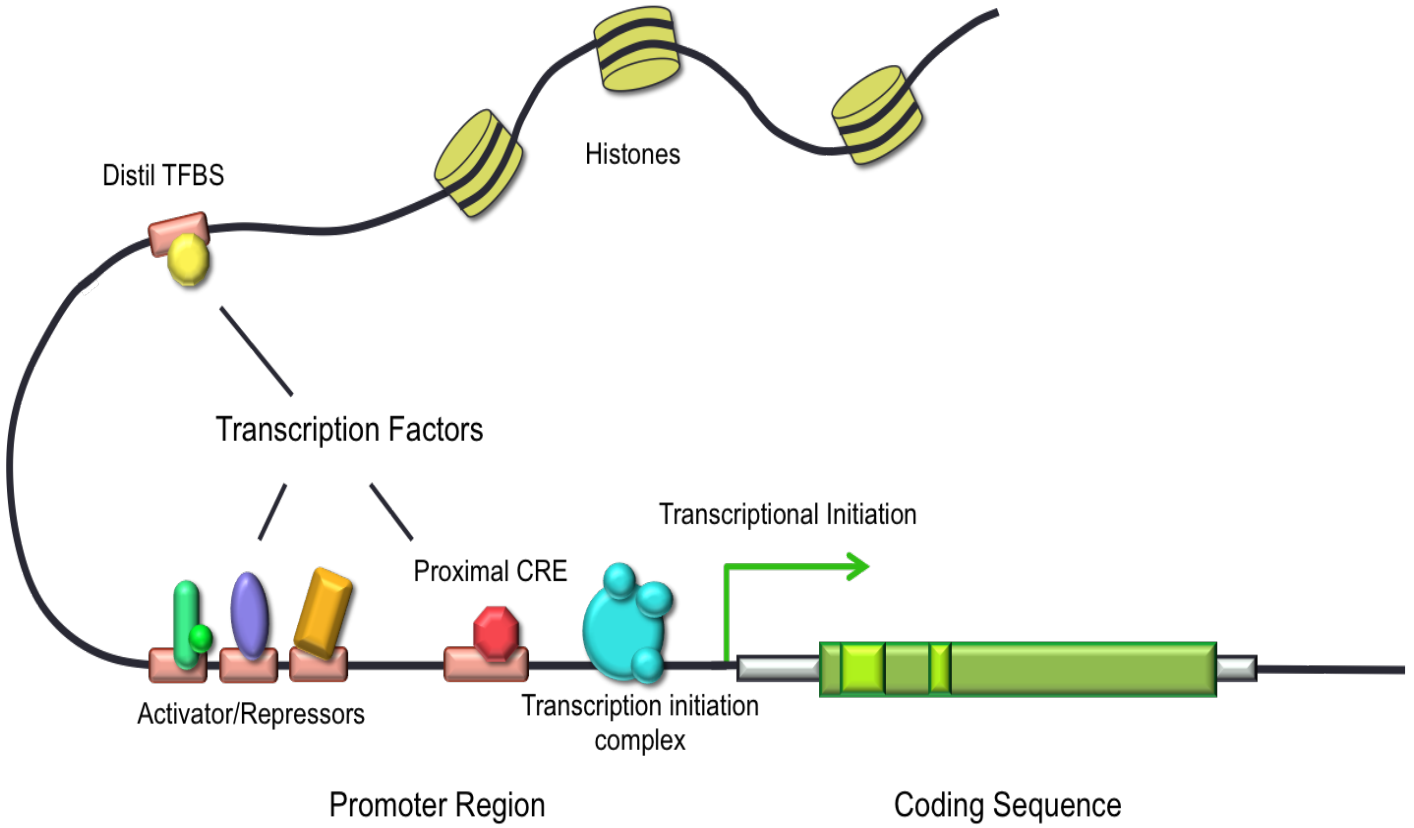


Figure 1.1: Gene transcriptional regulation by TF-DNA interactions. *Cis*-regulatory elements can exist in both proximal and distal regions upstream of the transcriptional start site (TSS). *Cis*-regulatory motifs can also produce different functional outcomes when bound by their corresponding TF. These include motifs that cause transcriptional repression and activation. RNA polymerase along with other proximal TFs form the transcriptional initiation complex (PIC), initiating transcription allowing RNA polymerase to start synthesizing RNA. Figure adapted from Wasserman and Sandelin (2004).

1.1.2 Transcription factor families

Plants regulate their genomes with a variety of TFs categorized into different families based on their DNA-binding-domains (DBDs) (Jin et al., 2014). There are 50 known TF families within *Arabidopsis* (Palaniswamy et al., 2006). These families account for 1,690 TFs encoded within the genome, amounting to approximately 6.1% of all protein coding genes (Palaniswamy et al., 2006; Berardini et al., 2015). Some of the largest families include the Myb/Sant, bHLH (basic helix-loop-helix), bZIP (basic leucine zipper), homeodomain, C_2H_2 zinc finger, MADS box, B3, Whirly, WRKY, SBP, Dof, AP2 and NAC families (Weirauch and Hughes, 2011). The last eight families, MADS box through NAC, are predominantly plant specific families, however, small numbers of MADS box members exist in nearly all eukaryotes, and other TF families have been found to share similarities between non-plant TF families (Weirauch and Hughes, 2011). The explanation of which has been hypothesized to be the result of horizontal gene transfer (Yamasaki et al., 2008) and ancient divergent evolution (Ülker and Somssich, 2004; Babu et al., 2006). The largest TF families are AP2 and NAC (Weirauch and Hughes, 2011). AP2 TFs are involved in disease resistance (Gutterson and Reuber, 2004) and abiotic stress (Dietz et al., 2010), most notably in cold and drought response (Sakuma et al., 2002; Shinozaki and Yamaguchi-Shinozaki, 2000; Liu et al., 1998; Stockinger et al., 1997). NAC TFs control a variety of plant processes, including shoot and root development (Takada et al., 2001; Xie et al., 2000; Aida et al., 1997). Gene regulation has classically been described as the action of TFs interacting with corresponding CREs within promoters. We know now that this model is an over simplification and that there are additional layers of information such as epigenetics that control gene expression.

1.1.3 Epigenetic factors

Epigenetic factors are heritable chemical modifications to DNA or histones that alter gene expression without changes to the genetic code (Goldberg et al., 2007). One DNA modification regulating gene transcription is the methylation of cytosine residues either directly adjacent to

a guanine (CG site) or in the proceeding sequence patterns, CHG and CHH, where H is either A, C or T (Meyer et al., 1994; Ingelbrecht et al., 1994; Gruenbaum et al., 1981). The inverse relationship between DNA methylation and transcription, where highly methylated genes are repressed from transcription, has long been known, indicating its involvement in gene regulation (Goll and Bestor, 2005). In mammals, CG rich sequences are known as CpG islands and are typically found in gene promoter sequences (Gardiner-Garden and Frommer, 1987). Methylation of CpG islands causes stable genetic silencing (Bird, 2002). While not as commonly associated with plants compared to mammals, CpG islands have been identified in *Arabidopsis*, with the majority of CpG island methylation occurring within promoters and coding sequences (Ashikawa, 2001). Interestingly, DNA methylation patterning in plants has been shown to differ between tissue types (Ashikawa, 2001). Genes expressed within one tissue can be found methylated and repressed in other tissue types. Similar forms of gene regulation by DNA methylation between cell-types has also been observed in humans (Bloushtain-Qimron et al., 2008) but is less studied in plants.

While DNA methylation is an epigenetic modification of the DNA molecule directly, other epigenetic modifications exist that chemically alter histones, the proteins that form nucleosomes and wraps DNA. Histone modifications represent a diverse range of different chemical markers on specific amino acid residues (typically lysine and arginine) within the four histone subunits (H2A, H2B, H3 and H4) (Pfluger and Wagner, 2007). Chemical modifications include ubiquitination, phosphorylation, acetylation, and methylation (Pfluger and Wagner, 2007). Histone ubiquitination influences gene expression activation through ring-type E3 ligases and deubiquitinases (Fleury et al., 2007; Liu et al., 2007; Pfluger and Wagner, 2007; Sridhar et al., 2007). Phosphorylation of both serine and threonine residues within histones by kinases and phosphatases has been shown to induce gene expression activation (Ashtiyani et al., 2011; Houben et al., 2007). Activation is also influenced by histone acetylation levels (Pfluger and Wagner, 2007). In *Arabidopsis*, histone acetylation is known to regulate flowering (Guyomarc'h et al., 2006; He et al., 2003), light response (Benhamed et al., 2006),

pathogen response (Zhou et al., 2005), root epidermal patterning (Xu et al., 2005) and elongation (Krichevsky et al., 2009). Gene activation *via* histone acetylation is a reversible process controlled by acting enzymes, acetyltransferases (HATs) and deacetylases (HDACs) (Chen and Tian, 2007). Methylation of histones is diverse in possibilities, with methylation occurring specific to histone subunit, amino acid, and methylation saturation (mono, di, and tri-methylation) (Pfluger and Wagner, 2007). Like histone acetylation, histone methylation is reversible, regulated by methyltransferases and demethylases (Liu et al., 2010). The diverse number of histone methylation modifications control both activation and repression of gene expression, some of which control gene expression through chromatin remodelling.

Before TFs can bind to promoter CREs and induce gene expression changes, regulatory regions of DNA must first be free of histones and accessible to TF binding. Promoter regions bound by histones are effectively silenced due to their inaccessibility to regulatory proteins. Gene regulation through chromatin remodelling is a dynamic process controlled on a cellular level. Chromatin accessibility has been found to play a vital role in development and cell identity within *Arabidopsis* (Aichinger et al., 2009). Chromatin remodeling complexes such as polycomb-group (PcG) proteins and trithorax-group (TrxG) proteins were first described in *Drosophila melanogaster* and are currently an active field of study within *Arabidopsis*. Together, PcGs and TrxGs work antagonistically through histone modifications controlling nucleosome eviction (Simon and Tamkun, 2002). In *Arabidopsis*, PcG proteins repress transcription though either H3K27 tri-methylation, as in the case of polycomb repressive complex 2 (PRC2) or monoubiquitinating histone H2A, as is for PRC1 (Pien and Grossniklaus, 2007). The recruitment of these chromatin remodeling complexes is mediated by specific CREs. These are polycomb response elements (PREs) for PcGs and trithorax response elements (TREs) for TrxGs. A few PREs have been discovered in *Arabidopsis* (Deng et al., 2013). However, knowledge about TREs in *Arabidopsis* remains limited. Due to their involvement in cell differentiation and identity (Bratzel et al., 2010; Aichinger et al., 2009; Schubert et al., 2005), PREs and possibly TREs likely play an important role in cell-type specific expression. Indeed, *Arabidopsis* mu-

tants lacking functional PRC2 have been reported to produce immortalized callus-like tissue of de-differentiated cell-types (Schubert et al., 2005).

1.2 *Arabidopsis* as a model for root cell-type specific expression

The *Arabidopsis* root offers an excellent model for studies focused on individual cell-types such as those looking at cell-identity (Dinneny et al., 2008; Birnbaum et al., 2003) and development (Aida et al., 1997; Benfey and Schiefelbein, 1994). The root anatomy consists of four main cell layers that run the length of the root: epidermis, cortex, endodermis, and stele, where vasculature tissue in the form of phloem and xylem are encased (Dolan et al., 1993) (Figure 1.2). With the exception of the stele, in *Arabidopsis*, cell layers are only a single cell thick, making identifying individual cell lines more tractable than other model plants (soybean or tobacco) (Dolan et al., 1993). Cell layers form a simple radial design with all cell files emerging from the root meristem (Taiz and Zeiger, 2010). The linear growth of roots, coupled with the meristem acting as the single source of cell division, means that the distance of a cell from the root meristem is directly related to the cell's age, despite how old the plant may be (Birnbaum et al., 2003; Taiz and Zeiger, 2010). This allows researchers to accurately study the development of cell-types across different growth stages. The root tip is divided into three distinct stages of development[2] (Figure 1.2). The first is the apical meristem, composed of the root stem cell niche and immediate surrounding cells up to the point where the root reaches its maximum radius (Birnbaum et al., 2003). Between this region and the zone of elongation lies the basal meristem (also known as the transition zone) composed of fully differentiated cell layers (Verbelen et al., 2014; Birnbaum et al., 2003). Finally, as cells begin to extend in length, the zone of elongation is reached (Verbelen et al., 2014; Dolan et al., 1993).

[2]Four developmental zones if you include the growth terminating zone above the zone of elongation (Verbelen et al., 2006). However, this stage was not defined in the Birnbaum et al. (2003) root cell-type microarray data used in this study.

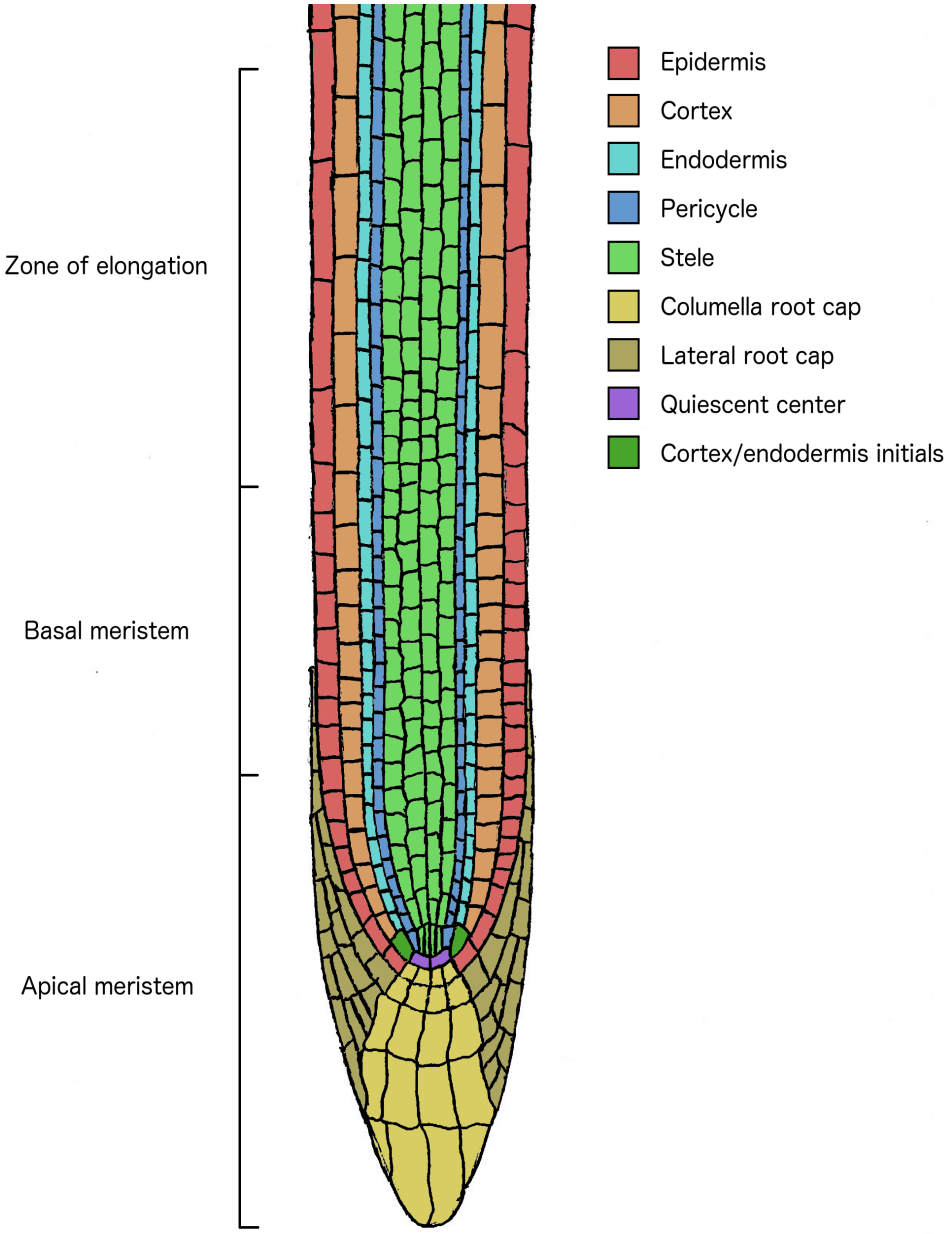


Figure 1.2: Anatomy of the *Arabidopsis* root with individual cell layers highlighted in colour. All cell lines emerge from the root meristem housing the quiescent cells and cortex/endodermal initials. Developmental stages defined by Birnbaum et al. (2003) are noted on the left.

Besides these unique features of the *Arabidopsis* root that make it an ideal model for studying individual cell-types, *Arabidopsis* is the most well-studied model organism in plant research. It has a relatively short life cycle, approximately 6 weeks from germination to senescence, small size for growth space, and produces an abundance of seed in a single generation. Moreover, the use of *Arabidopsis* as a model organism has produced a plethora of primary literature and genomic data on its biology, allowing for more accurate hypotheses and informed interpretations. The volume of research on *Arabidopsis* has resulted in numerous plant transformation methods that have been refined over the years (Bent, 2000; Zhang et al., 2006).

1.2.1 Genetic transformation of *Arabidopsis*

Continued research within *Arabidopsis* over the last two decades has developed several genetic transformation procedures with varying degrees of efficiency. The majority of these techniques utilize *Agrobacterium* as a means of gene delivery, however, other methods like particle bombardment exist (Bent, 2000; Seki et al., 1991). *Agrobacterium tumeficiens* is a gram-negative soil bacterium with the unique ability to copy and integrate a region of its genome into the genome of an infected host plant cell (Thomashow et al., 1980). In the wild, the transferred genetic material, referred to as Transfer DNA (T-DNA), contains a series of opine producing genes that are ultimately translated into a source of nutrition for the bacterium (Ellis et al., 1979). Scientists use *Agrobacterium* as a tool for plant genetic transformation by replacing the native virulence genes with transgene(s) and selectable marker genes, which allow clean integration of exogenous genetic material. The most facile *Agrobacterium* based method of transformation is the floral dip method (Zhang et al., 2006). Briefly, this method involves submerging *Arabidopsis* flowers into a culture of *Agrobacterium* harbouring a transgene vector. This then brings *Agrobacterium* in contact to immature oocytes within the *Arabidopsis* ovule where the oocytes are transfected with the *Agrobacterium's* T-DNA (Desfeux et al., 2000). T-

DNA is incorporated into the oocyte genome, producing fully transgenic seed upon maturity[3]. The floral dip's reliability allows for relatively quick generation of transgenic lines making it an ideal method for studies dealing with multiple transgenes. A transformation efficiency of up to 1% can be achieved depending on the *Agrobacterium* strain used (Zhang et al., 2006). Due to the short life cycle and high fecundity of *Arabidopsis*, multiple transgenic generations are often studied simultaneously. For the purpose of this study, transgenic lines are denoted with a "T", followed by the number of generations since transformation. The transgenic offspring of a wild type plant is referred to as T1 lines, with successive generation seeds denoted as T2, T3, *etc*.

1.2.2 Cell layers of the *Arabidopsis* root

Beginning from the outermost cell layer of the *Arabidopsis* root, the epidermis provides an nutrient absorbing tissue while simultaneously acting as a protective barrier from the outer environment (Esau, 1977). In mature epidermal cells, some cells develop protruding tubular extensions known as root hairs. Root hairs extend the absorbing surface of the root increasing water and nutrient uptake. As with all root cell layers, the epidermis differentiates from the root apical meristem, a collection of organized mitotically active cells (Taiz and Zeiger, 2010). The root meristem is composed of three cell layers. The outmost cell layer (L1), differentiates primarily through anticlinal divisions forming the epidermis. The remaining L2 and L3 meristem layers differentiated into the internal cell layers (Dolan et al., 1993).

Underneath the root epidermis lies the cortex cell layer. In *Arabidopsis*, the cortex layer is a single cell thick, but is thicker in many other plant species (Smith and De Smet, 2012). The cortex is easily identified as individual cells are larger than cells of other root cell layers. Their size is the result of large vacuoles found within cortical cells. Furthermore, plastids of cortical cells typically accumulate starch as a form of energy storage. Cortical cells develop

[3]The floral dip method has the benefit of producing fully transgenic plants, as opposed to chimeric plants, and are often heterozygous for the transgene.

with intercellular spaces between them, which assist with gas exchange and act as reservoirs of oxygen (Esau, 1977).

Below the cortex cell layer lies the endodermis, a unique cell layer which serves as a boundary between ground tissue (epidermis and cortex) and vascular tissue (stele, phloem, and xylem). Within *Arabidopsis*, and most plant species, the endodermis is a highly specialized cell layer typically a single cell in thickness. The endodermis regulates the movement of water, ions, and hormones between the ground tissue and vascular tissue (Esau, 1977). Its ability to control fluid movement is in part due to the water tight barrier formed by lignin polymer[4] deposits on the endodermal cell walls (Naseer et al., 2012). This barrier, called the Casparian strip, forms a band like region around the radial and transverse cell walls (Taiz and Zeiger, 2010). It forces fluids to pass through the selectively permeable membrane of the endodermal protoplast, instead of the apoplastic pathway which is composed of the inner space between the cell protoplast and cell wall. During drought stress, the endodermis is crucial in preserving water by preventing water and nutrients from diffusing out of the vascular tissue and into the soil (Taiz and Zeiger, 2010).

The next cell layer within the *Arabidopsis* root is the stele, also known as the vascular bundle as it comprises multiple cell-types involved in vertical fluid transport. The two main cell-types involved in fluid transport are the xylem and phloem. Xylem is the primary site of water and mineral transport, while phloem facilitates the transport of nutrients like carbohydrates from leaves to storage organs in the roots. The non-vascular cells within the stele are referred to as the pericycle which encompasses the vascular cell-types (Taiz and Zeiger, 2010). The pericycle is also the site of lateral root development, where a new meristem forms allowing secondary roots to grow (Péret et al., 2009).

All of these cell layers are found through the entire length of the root. At the root tip, the apical meristem is protected by the root cap, a cell layer consisting of living parenchyma cells that differentiate away from the root apical meristem and downward into the soil (Taiz

[4]Most sources will report that the Casparian strip is made of lignin polymer and suberin, however Naseer et al. (2012) showed that suberin production starts too late in *Arabidopsis* to be involved in Casparian strip formation.

and Zeiger, 2010). For ease of burrowing into soil, root cap cells are coated with a mucilage allowing the roots to slide through soil (Russell et al., 1977). Additionally, the root cap is constantly being replenished with new cells, allowing older cells on the outside of the root cap to shed off reducing friction between the root and soil. The root cap also contains specialized starch filled amyloplasts called statoliths (Esau, 1977). Statoliths respond to gravity allowing the plant to gain a sense of direction when extending into soil (Taiz and Zeiger, 2010). The root cap is divided into two sections, the lateral root cap, comprising the cell layer around the sides of the root cap, and the collumella, which is defined as the cells at the root tip (Birnbaum et al., 2003). The classification of root cell layers was largely pioneered with detailed microscopy work (Dolan et al., 1993). Current molecular and genetic techniques now allow scientists to further study cell-type differences in greater detail.

1.2.3 Cell-type analysis and isolation within *Arabidopsis*

Advancements in genome technologies are rapidly increasing our knowledge about gene regulation. Next generation sequencing (NGS) technologies allow for fast and reliable whole genome sequencing. Chromatin immunoprecipitation methods such as ChIP-chip and ChIP-seq allow researchers to study protein-DNA interactions and have even been used to identify TF binding at the cell-type level (Pique-Regi et al., 2011). RNA-seq and microarray technologies can quantify mRNA levels in real time providing transcriptomes of individual cell-types (Birnbaum et al., 2003; Islam et al., 2011; Jaitin et al., 2014). These methods are particularly valuable for studying gene regulatory networks. Analyzing cell-type transcriptomes allows for the identification of co-expressed genes under various environmental conditions. This in turn can be used to elucidate larger transcriptional networks, such as TF cascades of facultative genes responding to external stimuli, or constitutive genes that maintain cell identity and homeostasis (Rombauts, 2003). In plants, transcriptome analysis at a cell-type resolution remains limited due to difficulties isolating homogeneous cultures from tissues compared to their mammalian counterparts. However, fluorescence-activated cell sorting (FACS) (Bonner et al.,

1972) and other comparable techniques (Deal and Henikoff, 2010) offer a practical solution.

FACS utilizes cell-type specific promoters driving a fluorescent reporter gene to microfluidically sort individual cell-types. This has so far been applied to the *Arabidopsis* root, where protoplast cells of major root cell-types are isolated and analyzed by microarray to quantify cell-type transcriptomes (Birnbaum et al., 2003). An alternative method of cell-type isolation was developed by Deal and Henikoff Deal and Henikoff (2011). Briefly, this method used cell-type specific promoters to drive a fusion protein composed of a reporter gene, nuclear localization signal, and biotin ligase peptide. Biotinylation of the nuclear membrane bound fusion protein then allows cell-type specific nuclei to be isolated magnetically using streptavidin coded metallic beads which covalently bind to biotin (Deal and Henikoff, 2010). This technique was originally developed in *Arabidopsis* and provides an elegant way to isolate DNA of individual cell-types without the need for expensive cell sorting equipment (Deal and Henikoff, 2010).

Cell-type isolation studies are important as they allow researchers to study plant cells at a system's level. This in turn can be used to determine genetic differences between multiple cell-types. Furthermore, differentially expressed genes among multiple cell-types can identify regulatory networks and molecular processes that occur in a cell-type specific manner (Shen-Orr et al., 2010; Bryant et al., 1999). Cell-type specific genes and transcriptional cascades are of particular importance, as they represent specific molecular interactions that contribute to a cells unique identity and function. To date, little is known about how gene expression can be regulated to a single cell-type. As such, the primary objective of this study is to examine the promoter architecture of cell-type specific genes in order to identify CREs and CRMs responsible for cell-type specific expression.

1.2.4 Cell-type specific expression within *Arabidopsis*

Various *cis*-regulatory motifs have been characterized regulating genes involved in a variety of plant cellular functions including stress response (Yamaguchi-Shinozaki and Shinozaki, 2005),

development (Winter et al., 2011), and chromatin accessibility (Berger et al., 2011). Specific motifs identified as regulating the expression of genes in individual cell-types in plants remains limited but a tissue specific context has been observed, as in the case of the RY repeat necessary for seed expression (Inz and Wobus, 1992). In *C. elegans* and various human cell lines, several CREs have been identified directly responsible for gene expression within a single cell-type (Ernst et al., 2011; Wenick and Hobert, 2004). Several studies have reported cell-type specific responses to environmental stresses including salinity, drought, and osmotic shock in *Arabidopsis* (Dinneny et al., 2008, Kiegle et al., 2000). This implies that entire gene cascades can be activated in a cell-type specific context. Given our current understanding of gene control, one or more CREs could be responsible for regulating cell-type specific expression states. By considering multiple constitutively expressed cell-type specific genes, promoter sequences can be analyzed to identify shared regulatory elements possibly contributing to cell-type specific expression. This approach follows the guilt-by-association (GBA) heuristic whereby co-expressed genes are likely to be associated with common functional regulatory modules (Wolfe et al., 2005). This GBA heuristic is widely invoked in functional genomics and has been shown to accurately reflect functional gene cascades and expression networks (Harmer et al., 2000; Wolfe et al., 2005).

1.3 Microarray analysis

The use of microarray technology has allowed researchers to effectively quantify mRNA levels of thousands of genes simultaneously. Briefly, microarray chips are coated with specifically placed DNA oligonucleotides or probes. Probe sequences are designed to hybridize to different mRNA molecules expressed within the genome of the species under study. Messenger RNA samples are fluorescently labeled before hybridization with microarray probs. Relative abundance of mRNA species can be detected by analyzing the florescence intensity of probes bound to labeled mRNA (Hoheisel, 2006). A disadvantage of microarrays is that florescence

intensity produces background signals that can drown biologically real signals. A considerable amount of bioinformatic processing is required to interpret expression data results (Hoheisel, 2006). An additional draw back is that the physical size of microarray chips limits the number of unique sequence probes that chips can contain. This reduces the number of mRNA species that can be quantified at once. Large genomes may therefore only be partially covered by the microarray chip. The development of RNA-seq improves on these issues and has quickly become the standard for transcriptome analysis. However, the popular use of microarrays over the decade has generated a large volume of expression profile data available for researchers to draw on.

1.3.1 Hierarchal clustering and differential gene expression

Determining expression patterns from microarray data is a central process for identifying co-expressed genes and biologically meaningful patterns. A number of different methods have been developed to achieve this including K-means clustering (Tavazoie et al., 1999), partitioning around medoids (PAM) (Rousseeuw and Kaufman, 1990; Van der Laan et al., 2003), self-organizing maps (SOM) (Tamayo et al., 1999), clustering affinity search techniques (CAST) (Ben-Dor et al., 1999), and hierarchical clustering (Eisen et al., 1998). Hierarchical clustering is a popular method used to group genes with similar expression patterns by applying distance measures between gene expression profiles. A commonly used distance measure is Euclidean distance. An advantage of using Euclidean distance for gene expression profiles is that Euclidean distance will group genes by expression pattern and not absolute expression level. Two main methods of hierarchal clustering exist, divisive and agglomerative. For divisive clustering, all observations are grouped into a single cluster and are recursively split into the hierarchy. Agglomerative clustering performs the opposite, where all observations begin in separate clusters and are combined as one moves up the hierarchy. Alternative hierarchal methods exist, including algorithms that combine both divisive and agglomerative approaches such as the Hierarchical Ordered Partitioning And Collapsing Hybrid (HOPACH) algorithm (van der Laan

and Pollard, 2003).

1.3.2 Custom expression baits for cell-type gene targeting

A more direct method of identifying co-expressed genes is by designing an artificial expression profile reflecting a desired expression pattern (Austin et al., 2016). The expression profile of a gene can be represented as a vector over a set of conditions (examples being a point in a time series or tissue or cell-type). The Pearson Correlation Coefficient (PCC) can then be calculated between gene expression profiles and the artificial bait vector. Genes scoring high PCC values therefore have expression patterns similar to the bait. The primary advantage of this method is that genes with a specifically desired expression pattern can then be retrieved from an expression data collection by simply designing a bait profile mimicking the desired expression pattern. This technique is a key strategy used in this thesis to isolate cell-type specific expressing promoters from microarray data.

1.4 Motif prediction

The advent of gene expression technologies combined with previously discussed analysis strategies has been used to identify networks of co-expressed genes. Gene co-expression analysis has been used to associate genes of unknown function to biological processes. The heuristic that genes with similar expression patterns should also share similar promoter architecture has been successfully used in identifying common CREs shared among co-expressed genes (Sharma et al., 2015; Vandepoele et al., 2009; Lenka et al., 2008; Harmer et al., 2000). Motif prediction therefore depends on pattern finding programs capable of identifying re-occurring patterns followed by a statistical scoring method to access significance of potential motifs. Five motif predicting programs specifically designed to identify statistically significant sequence patterns within the promoters of co-expressed genes are employed in this study. These programs are MEME (Bailey and Elkan, 1995), AlignAce (Hughes et al., 2000), Bioprospector (Liu et al.,

2001), Weeder (Pavesi et al., 2001), and Motif Sampler (Thijs et al., 2001). Motif prediction programs use different strategies to identify significant sequence patterns which are discussed below. The "wrapper" program *Cister* (Austin et al., 2016; Winter et al., 2011) can be used to manage the output of these five motif prediction programs into a common format for ease of downstream motif analysis.

1.4.1 Motif prediction through alignment based strategies

Motif prediction programs MEME (Bailey and Elkan, 1995), AlignAce (Hughes et al., 2000), Bioprospector (Liu et al., 2001), and Motif Sampler (Thijs et al., 2001) all use sequence alignment strategies to identify significant sequence patterns within promoters of co-expressed genes. However, the ability to identify CREs is complicated by the tendency for motif degeneracy within binding sites. Functional CREs may therefore be composed of many similar sequence patterns each with affinity to their CRE's corresponding TF. Because of this, CREs were traditionally summarized as "consensus" sequences, where the most frequent residue(s)[5] at each position is reported. This method however, is fundamentally flawed, as there is no way of identifying motifs in novel sequences, except for using the most common matching base pairs of the consensus sequence (Staden, 1984). Furthermore, consensus sequences fail to report the level of degeneracy at each residue position, as many motifs contain positions that will accept 3 or 4 base pairs at varying frequencies (Staden, 1984).

The use of matrices resolves the degeneracy issue in motif representation by using an A×L matrix, where L is the length of the motif and A the sequence alphabet size (4 for DNA), to represent the residue frequency at every position of a motif. Such matrices are referred to as positional specific scoring matrices (PSSM) (Stormo et al., 1982) and are produced by tallying the residue counts of multiple motif sequences. While more practical than consensus sequences, PSSMs do have limitations including not being able to record base-to-base dependencies. An-

[5]IUPAC symbols representing two or more base pairs are commonly used in motif consensus sequences. This method however, still falls short in representing residue frequencies.

other matrix often employed in motif analysis is the positional weight matrix (PWM). PSSM residue frequencies are converted to log-odd probabilities to assign weights reflecting the frequency biases observed in degenerate motifs. PWMs also take into account the GC content of the genome in which co-expressed gene promoters originate (Schneider et al., 1986; Hertz and Stormo, 1999). PWMs are used for a variety of applications including *de novo* motif prediction (Sinha, 2006) and scanning motif matches within sequence (Stormo, 2000). An example of a PSSM is provided in Figure 1.3 along with its conversion to a PWM.

The degeneracy of a motif can also be represented using an information content (IC) statistic (Stormo and Hartzell, 1989). Briefly, this statistic can be used to access the relative entropy within a matrix, also known as the Kullback-Leibler distance. In biological terms, the IC content of any residue along a motif can be regarded as the relative binding energy that the residue contributes to the overall motif (Stormo, 2000). Note, that IC is a site-wise calculation specific to each residue position in a motif. Therefore, to acquire the average IC for the whole motif, one must normalize the IC sum of each residue position by the total length of the motif. The resulting average can be used as an approximation of a motif's overall degeneracy.

Due to CRE degeneracy, log-likelihoods of PWMs are utilized to determine sequence motifs enriched within promoter sequences of co-expressed genes. This is achieved by maximizing the sum of the site-wise IC for putative PWMs, which in turn is used as a probability of motif expectancy (Hertz and Stormo, 1999; Stormo, 2000). Predicted motifs are therefore sequence patterns with the highest IC sum and the lowest probability of occurring by random chance (Stormo and Hartzell, 1989). Alignment based motif prediction programs use a variety of algorithms to maximize the IC sum within a putative motif PWM. MEME uses an expectation maximization (EM) approach described by Lawrence and Reilly (1990). This approach uses log-likelihood scores to determine an optimal start position to begin the alignment build between sequences. When a sequence match is found, the result is stored and the alignment process is reimplemented to find additional motifs (Bailey and Elkan, 1995). A draw back of EM alignment approaches is that they suffer from local maxima problems whereby premature

(a) Position Specific Scoring Matrix

	1	**2**	**3**	**4**	**5**	**6**	**7**
A	5	0	3	9	16	6	3
C	10	0	0	0	0	5	13
G	5	0	3	11	0	4	4
T	0	20	14	0	4	5	0

(b) Position Weight Matrix

	1	**2**	**3**	**4**	**5**	**6**	**7**
A	-0.36	-∞	-1.09	0.49	1.32	-0.09	-1.09
C	1.47	-∞	-∞	-∞	-∞	0.47	1.85
G	0.47	-∞	-0.26	1.61	-∞	0.15	0.15
T	-∞	1.64	1.13	-∞	-0.68	-0.36	-∞

(c) Sequence Logo

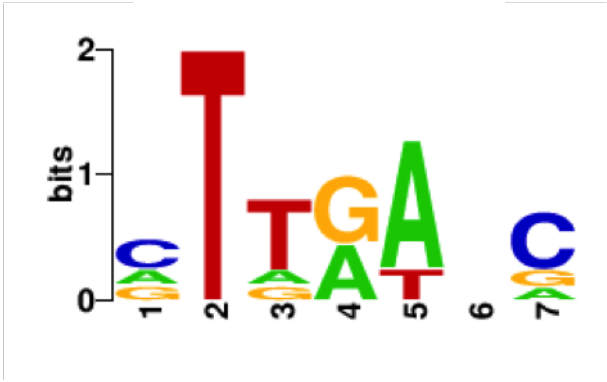


Figure 1.3: Example of a degenerate motif signal represented by PSSM, PWM, and sequence logo. **A**.) PSSM of example motif depicting the residue frequencies in each motif position. **B**.) Log-odds of PSSM residue frequencies in PWM format with an *Arabidopsis* GC content of 36%. **C**.) Sequence logo of example motif based on PWM. The height of each column indicates the site-wise IC. The proportion of the letter size in each column represents residue frequency.

IC maximums are fixated, ignoring other possibilities. This is the result of the EM algorithms selecting non-random start positions from which to build sequence alignments for testing. A solution to this problem is the use of Gibbs sampling as a stochastic implementation of EM (Lawrence et al., 1993). Here, multiple start positions are randomly chosen within sequence subsets to build alignments. The use of Gibbs sampling for motif prediction has the added benefit of being computationally faster than EM methods alone. AlignAce (Hughes et al., 2000), Bioprospector (Liu et al., 2001), and Motif Sampler (Thijs et al., 2001) all used Gibbs sampling as a method to maximize IC of putative PWM.

1.4.2 Motif prediction through enumerative based strategies

Enumerative based prediction strategies, such as the one employed by Weeder (Pavesi et al., 2001), function by searching for statistically over-represented motif sequences from a collection of permuted sequences. Because enumerative approaches consider all possible sequence combinations, best fit motifs are guaranteed to be found within a set of co-expressed promoter sequences. Unlike the previously discussed alignment based approaches, Weeder does not use PWMs to identify degenerate sequences, but rather uses suffix trees with a predetermined number of mismatches allowed during sequence alignment (Pavesi et al., 2001). The down fall of enumerative approaches however, is that the exhaustive number of iterative calculations needed to test all possible motif sequences of a predetermined length is computationally taxing. As such, the number of co-expressed promoter sequences one can start with is limited by computational resources. However, technological improvements in computer speed and performance are improving this limitation.

1.4.3 Motif statistical enrichment and mapping with *Cismer*

The identification of reoccurring sequence patterns in a given subset of promoters by prediction software produces a myriad of possible motifs. The large number of putative motif signals makes biologically validating prediction results as CREs near impossible. Moreover, the ob-

servation of a conserved pattern within a collection of sequences is not enough to definitively conclude biological function. As such, statistical tests have been designed to filter prediction results down to the most probable motifs (Bailey et al., 2010; Eden et al., 2007; Sinha and Tompa, 2000). Discriminative models have been used to assess motif significance (Redhead and Bailey, 2007; Grau et al., 2013) and are designed so that the probability of a motifs significance is conditional, usually based on factors regarding the nature of the motif. For example, the program *Cismer* (Austin et al., 2016; Winter et al., 2011) employs a commonly used strategy to determine a motif's significance by comparing its mean count distribution in a positive data set to a null data set for significant differences. The positive data set, or sometimes referred to as the foreground, contains a set of sequences that are believed to share a common motif. The null data set is a collection of randomly sampled sequences selected from the genomic background. Discriminative approaches utilize bootstrapping techniques to determine motif count distributions. In other words, motif counts within randomly selected promoter sequences are recorded and repeated for thousands of iterations. This process builds up a background distribution of motif counts which can be compared to the foreground mean counts. The main objective of significance testing is to isolate motifs whose enrichments within a set of sequences is far higher than what would be expected by random chance. In *Cismer*, statistical enrichment is determined with a Z-score statistic:

$$Z_{(x)} = \frac{Obs_{(x)} - Exp_{(x)}}{\delta_{(x)}} \tag{1.1}$$

Where $Obs_{(\mathrm{x})}$ are the mean motif counts observed in foreground sequences, $Exp_{(\mathrm{x})}$ are the mean motif counts observed in the genomic background, and δ_{x} the standard deviation of the background distribution.

Additional discriminative models have been designed to access motif significance that rely on different approaches including linear regressions (Pessiot et al., 2010), logistic regressions (Yao et al., 2014) and background distributions not generated from whole genome sequences (Patel and Stormo, 2014). These newer methods however, have been specifically designed for

dealing with ChIP-seq peaks which, unlike the foreground promoter clusters used in this study, are usually quite large.

Besides assessing statistical significance to putative motifs, *Cismer* is an effective tool for mapping putative motif PWMs to DNA sequence. Motif scanning is achieved through a scoring system whereby PWMs are aligned to subsequences to access their fit. The alignment score is determined as the sum of relevant log-odds values in a PWM, such that for subsequence s, the alignment score is calculated as:

$$Score_{(s)} = \sum_{l=1}^{L} \sum_{a=0}^{A} \omega_{l,a} \cdot S_{l,a} \tag{1.2}$$

Where $\omega_{\mathrm{l,a}} = 1$ if base a occurs at position l of the subsequence and 0 if it does not (Gribskov et al., 1987; Stormo and Fields, 1998). $S_{\mathrm{l,a}}$ is the log-odds probability of residue type a at position l of the aligned PWM. In other terms, each residue of the target sequence that the PWM is being aligned to must be a possible base pair option for that position within the PWM. For example, the PWM shown in Figure 1.3 would not match the sequence GTCGACG, because the third residue C, is not a possible option for the third position in the aligning PWM, even though all other residues in that sequence do fit.

PWMs of degenerate motif patterns can align to promoter sequences that are not functional CREs, producing false positive mappings. This is inherent to the scoring system used to map PWMs in genomic sequence (Equation 1.2). For highly degenerate motifs, where two or more base pairs are excepted at each motif position in varying frequencies, alignment matches may occur against sequences where most base pairs match low frequency residues in putative PWMs. The consequence of these alignments is that matched sequences poorly resemble the CRE's consensus sequence and are often not true *cis*-regulatory sites. These same challenges are faced when mapping PWMs of known CREs, where it is often difficult to determine if a matching sequence is a low affinity variant of a CRE or a similar non-active sequence. The functional depth (FD) statistic provides a means of setting thresholds to the level of degeneracy

tolerated when aligning PWMs to target subsequences[6] (Schones et al., 2007). In a biological context, the FD statistic is an empirical estimate of the TF binding affinity for CREs. Functional depth is defined as:

$$FD = \frac{Score_s - Score_{min}}{Score_{max} - Score_{min}} \tag{1.3}$$

Where $Score_{max}$ and $Score_{min}$ are the maximum and minimum potential alignment scores for a PWM, while $Score_s$ remains the alignment score derived between the PWM and the subsequence (Equation 1.2) (Schones et al., 2007).

1.4.4 Sequence logos

A more practical method of visualizing motif sequences is most often done with sequence logos (Figure 1.3c) (Schneider and Stephens, 1990). Here, the frequency of base pair residues at any specific site is represented by the proportional height of the base pair letter while the total height of the column indicates the site-wise IC. Sequence logos provide a visual representation of motif degeneracy as opposed to viewing numerical matrices.

1.4.5 *Cis*-regulatory element positional biases

The exact position of a CRE within a promoter can affect a gene's expression. For example, the distance between the GC-box (consensus GGGCGG) motif and the TATA box of the conserved E1B gene promoter of adenoviruses directly affects the expression levels of the E1B gene (Wu and Berk, 1988). Several other examples of CRE positional dependencies have since been observed (Senger et al., 2004; Spek et al., 1999; Sugiyama et al., 1998) including tissue specific promoters in *Phaseolus vulgaris* (common bean)(Grace et al., 2004). Positional biases of CREs have therefore been exploited for motif discovery (Berendzen et al., 2006; Vardhanabhuti et al., 2007). In humans, sequence motifs with positional biases have been observed in the promoters

[6]While a threshold scoring system based on PWM alignments was adopted by Staden (1984), our current definition of this statistic was refined by Schones and collogues (2007).

of co-expressed genes (Vardhanabhuti et al., 2007). Due to positional biases being the result of conserved evolution, which usually implies functionality (Thomas et al., 2003), identification of putative motif positional biases can be an indication of a functionally active CRE. However, positional disequilibriums of CREs are not necessarily required for gene regulation, as CREs may also function in a non-positional manner.

1.5 Advantages of decoding cell-type specific regulation in genetic engineering

Understanding the *cis*-regulatory mechanisms involved in cell-type specific expression has far reaching applications in biotechnology. For example, expression of transgenes within genetically modified organisms (GMOs) is commonly achieved using constitutive promoters that confer expression within the whole plant (Corrado and Karali, 2009). However, it would be practical to express transgenes within specific tissues and cell-types as to reduce any chances of unwanted molecular interactions. Utilizing cell-type specific promoters could be advantageous for economically important crops plagued by pests that feed on tissues other than the harvested fruit. For example, the Western Corn Rootworm (*Diabrotica virgifera*) is one of the most devastating rootworm species in North America. It has been estimated to be responsible for over 1 billion dollars of loss revenue each year in the United States (Mitchell et al., 2004). Moreover, the Western Corn Rootworm spread to Europe in the early 1990's where it continues to be a pest in southern and central Europe (Gray et al., 2009). Larvae of the Western Corn Rootworm feed on root hairs effecting overall nutrient and water uptake. Mature rootworms preferentially feed on corn silk and leaves over the kernel. As such, expressing endogenous proteins that confer resistance within the roots and leaves would prevent rootworm damage while leaving the edible corn cob free of transgenic material (with the exception of the transgene itself). This practice could be a more attractive GMO approach for European markets, where the use and consumption of GMOs is discouraged (Thayyil, 2012).

A complete and comprehensive knowledge about CREs, their DNA binding counterparts, and the expression states produced, will allow researchers to design effective synthetic promoters. Transgenes could be engineered to be active in one or more targeted cell-types. Alternatively, designing synthetic promoters that respond to time of day, or external stimuli like temperature could all be designed by understanding the *cis*-regulatory logic used by nature. In the future, economically important crops may have to be more extensively engineered than current GMOs. This will have to be accomplished to maintain high crop yields in environments rapidly changing by climate change. Crops of the future will have to tolerate more extreme temperatures, soil pollutants, high salinity, flooding, and drought. Furthermore, genetic engineering offers a more direct and faster method than selective breeding. While this study focuses only on the regulatory mechanisms involved in cell-type specific expression, it is hoped that the findings will contribute to our overall knowledge of gene regulation. Doing so could help future scientists design synthetic promoters capable of producing any desired expression pattern.

1.6 Research objective

This study looked to identify and characterize CREs involved in cell-type specific gene expression. Based on our current understanding of gene regulation, it is hypothesized that several CREs would be involved in restricting gene expression to a single cell-type. Moreover, because most motifs function as part of regulatory modules (CRM), it is expected that one or more specific motif combinations could direct cell-type expression. With the observed involvement of chromatin remodeling in cell differentiation, it is possible that PREs and TREs are enriched within cell-type specific promoters and contribute to their unique expression patterns. In this study, CREs are identified by analyzing the promoter sequences of cell-type specific genes expressed in the *Arabidopsis* root. Cell-type specific gene promoters are identified through bioinformatic analysis of root cell-type microarray data by Birnbaum et al. (2003). Motif

prediction software is used to identify over represented sequence patterns within groups of cell-type specific promoters. Putative motifs are then tested for their ability to control gene expression within transgenic *Arabidopsis* plants. In addition to identifying enrichment of putative motifs, enrichment of previously known CREs are used to determine their prominence in cell-type specific expression. The resulting findings and methods of this study have been used to postulate a possible strategy for designing synthetic promoters with specific expression targets (see Chapter 5).

Chapter 2

Materials and Methods

2.1 Microarray analysis

2.1.1 Preprocessing and hierarchal clustering

Publicly available microarray data published by Birnbaum et al. (2003) were downloaded from Science under the paper's supporting online material section (`http://science.sciencemag.org.proxy1.lib.uwo.ca/content/302/5652/1956/tab-figures-data`). Processing microarray data to remove genes irrelevant to this study was performed using a custom R script designed to implement the procedures described by Gentleman et al. (2006). This script reorders microarray data by (1) the sum of gene expression rates over all conditions and (2) the degree of expression change between conditions, removing genes falling within the lower quartile for both. Hierarchal clustering of the processed microarray data was performed in R using the HOPACH v2.28.0 package. Clustering with alternative methods was done using an agglomerative nesting approach with AGNES and a divisive approach with DIANA, both built in the Cluster v2.0.3 package in R.

2.1.2 Identification of root cell-type specific gene clusters

A custom R script was written to identify and rank cell-type specific genes found within co-regulated gene clusters. Gene expression profiles were tested for high PCC with an artificial bait vector designed to mimic perfect cell-type specific expression. Artificial expression baits consisted of vectors of equal length to gene expression profiles containing either 1's or 0's (Austin, 2016). A value of 1 indicated full expression and 0 for no expression. Genes with the highest PCC (r) were ranked as most cell-type specific. Vectors used as expression baits were "0,0,0,1,1,1,0,0,0,0,0,0,0,0,0" for endodermis, "0,0,0,0,0,0,1,1,1,0,0,0,0,0,0" for cortex, and "0,0,0,0,0,0,0,0,0,1,1,1,0,0,0" for epidermis across the 15 microarray data conditions.

2.2 Motif prediction

Motif prediction was performed using a bash script (*Cister*) (Austin et al., 2016) controlling the execution of five independent motif prediction programs run at various motif widths (5-9, 12 and 15 bp). The programs used along with the settings were: AlignAce 4.0 (Hughes et al., 2000): "-numcols 5-9,12,15 bp"; Bioprospector v5/14/01 (Liu et al., 2001): "-T 10 -w 5-9,12,15 bp"; MEME 3.5.4 (Bailey and Elkan, 1995): "-dna -mod anr -revcomp -nmotifs 10 -w 5-9,12,15 bp"; MotifSampler 3.2 (Thijset al., 2001): "-s 1 -n 3 -w 5-9,12,15 bp"; Weeder (Pavesi et al., 2001): "Medium/Extended scans". The commands used for this bash script, referred to as *Cister* (Austin et al., 2016, Winter et al., 2011), are "cister -x -p -f <cluster.fasta>", with the "cluster.fasta" file being the users lists of co-expressed promoter sequences in FASTA format. Testing for statistical enrichment was done using the *Cismer* program (Austin et al., 2016) with the following command "cismer -p <PSSM> -g <background.FASTA> -f <cluster.FASTA> -d 0.0" where "PSSM" is a list of motifs in PSSM format, "background.FASTA" is the *Arabidopsis* TAIR10 upstream 500 bp genome file in FASTA format (Berardini et al., 2015, `https://www.arabidopsis.org/download/index-auto.jsp?dir=`

%2Fdownload_files%2FSequences%2FTAIR10_blastsets%2Fupstream_sequences, file TAIR10_upstream_500_20101028) and "cluster.FASTA" the list of co-expressed promoter sequences of equal sequence length to the background file in FASTA format. Motifs with a Z-score < 3 were dropped from the study. Motifs were further filtered by removing highly degenerate motifs with the *Cistome* (Austin et al., 2016) command "cistome -f <cluster.FASTA> -m <file.PSSM> -w 6 -W 25 -Z 3 -l 10 -i 1.0 -S -F 5" where "cluster.fasta" is again a list of co-expressed promoter sequences and "file.PSSM" a list of PSSMs generated by *Cister* and *Cismer* programs. Distance matrices produced between motifs and the resulting dendrograms were generated with *Cistome* using the following command "cat <motif.PSSM> | cistome -N -R". Mapping motif PSSMs at various functional depth cutoffs was carried out with the following pipeline "cistome -f <cluster.fasta> -m <motif.PSSM> -Z 0.0 -l 0.0 -i 0.0 -p 0.0 -d x -F 5 | cismer -I -g <background.fasta> -f <cluster.fasta> -z 0.0" where x is the functional depth cut off ranging from 0 to 1 in 0.1 increments. The previous command was then repeated using the desired functional depth cutoff to refine motif PSSMs. Lastly, mapping of refined motifs was done with *Cismer*, using the command "cismer -p <motif.PSSM> -f <cluster.fasta> -d 0.0 -m" where "motif.PSSM" is a list of refined motifs and "cluster.fasta" file a collection of co-expressed upstream 1000 bp promoter sequences (Berardini et al., 2015, https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FSequences%2FTAIR10_blastsets%2Fupstream_sequences, file TAIR10_upstream_1000_20101104).

2.3 Recombinant DNA and molecular cloning

PCR primers were designed to amplify the upstream 1000 bp region plus the 5'UTR of endodermal specific promoters selected for genes expression assays (Table 2.1). Primers contain either *Apa*I or *Xma*I restriction endonucleases sites upstream of their hybridization sites. For promoter truncations, new forward primers were designed and used in conjunction with their

corresponding reverse primer (Table 2.2). Genomic DNA used for PCR was extracted from 150 mg of fresh *Arabidopsis* plant matter using the DNeasy Plant Mini kit (Qiagen). DNA concentration was then determined using a Qubit 2.0 fluorometer (Invitrogen) using high sensitivity buffers. PCR amplification products were digested with *Apa*I and *Xma*I (New England Biolabs) and purified using the Qiaquick PCR purification kit (Qiagen). Purified inserts were then ligated into a modified pCambia 2300 plasmid (referred to as pINTACT)[1], downstream of a *GFP* reporter gene containing a nuclear membrane localization signal (Figure 2.1). *E. coli* colonies passing 50 μg/ml kanamycin selection were grown to culture and stored as glycerol stocks for later use. Plasmid isolation of promoter constructs was purified using a QIAprep Spin Miniprep kit (Qiagen).

DNA fragments were amplified under the following PCR conditions: denaturation for 5 minutes at 98°C; 35 amplification cycles consisting of 30 s denaturation at 98°C, 30 s of primer annealing, and approximately 1 minute per 1 kb of extension at 72°C. Annealing temperatures of PCR products are listed in Tables 2.1 and 2.2.

2.4 Transgenic *Arabidopsis*

2.4.1 Plant growth conditions

Arabidopsis thaliana (Col-0) wild type (WT) seedlings were sterilized by chlorine gas exposure. Seeds were imbibed on agar plates containing ½ MS salts (Murashige and Skoog, 1962) for 10 minutes and put to 4°C for 3 days to vernalize. After which seeds were germinated under 24 hour light at 24°C. One week after germination seedlings were either used for genomic DNA isolation or transplanted to soil to provide adult plants for transformations. Plantlets for transformation purposes were transplanted into soil supplemented with 20-20-20 fertilizer (Plant Products Co. Ltd). Plants were grown in growth chambers (24°C) with a 20 hour photoperiod.

[1] pINTACT is a construct in the Austin lab used to isolate cell-type specific nuclei using the cell sorting method described by Deal and Henikoff (2010).

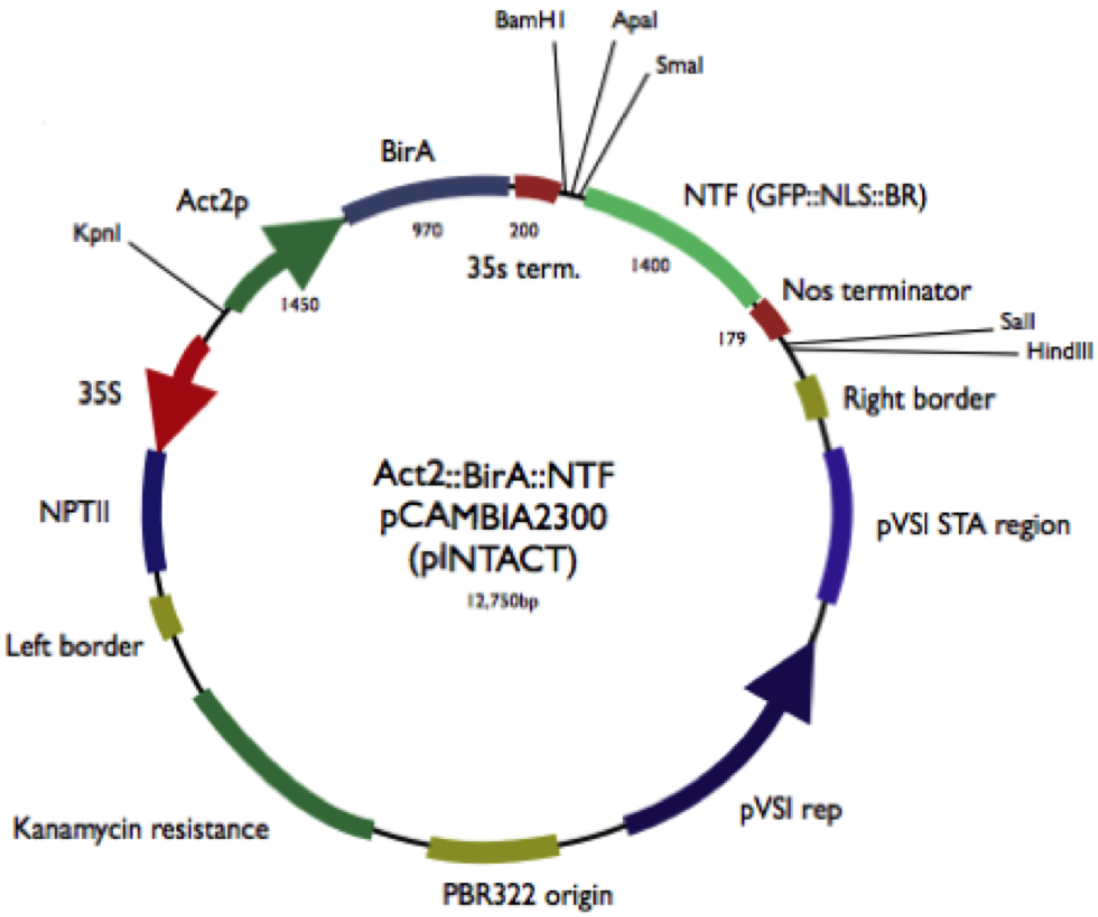


Figure 2.1: Diagram depicting pINTACT plasmid used for both gene expression assays testing cell-type specific motifs and for isolating cell-type specific nuclei. Designed from the pCambia2300 vector backbone. Sequences between left and right border repeats are transferred and integrated into the *Arabidopsis* genome. NTF region translates into a fusion protein consisting of a GFP reporter, nuclei localization sequence, and biotin ligase recognition peptide. *Sma*I restriction cut site is recognized and cut by *Xma*I. BirA encodes a *Escherichia coli* biotin ligase and NPT2 encodes for plant kanamycin resistance for selection of transgenic plants.

Transgenic seed, T1 or T2, was germinated and grown under the same conditions as wild type, with the exception that ½ MS agar plates were supplemented with 50 μg/ml kanamycin as a selective agent for plants harbouring transgene constructs.

2.4.2 Plant transformations

Plant transformations were performed *via* the flora-dip method described by Zhang et al. (2006). Transgene constructs were transformed into *Agrobacterium tumefaciens* (GV3101) (Van Larebeke et al., 1974) using the freeze-thaw method (Holsters et al., 1978) and plated on LB plates supplemented with kanamycin (50 μg/ml) for selection. Single colonies were picked from plates and used to inoculate LB (lysogeny broth) media (Bertani, 1951). Cultures were grown at 28°C to an O.D. of 1.2-1.5. Dipped plants were kept on their side for three days for recovery, stored at 24°C under a 20 hour photoperiod. After recovery, plants were placed upright, sprayed with water, and grown under the same temperature and photoperiod until senescence.

2.5 Nuclei isolation and chromatin accessibility profiling

Arabidopsis lines were transformed with the pINTACT plasmid. Lines contained either a endodermal specific promoter ligated into pINTACT for endodermal cell layer nuclei isolation or and epidermal specific promoter for epidermal cell layer nuclei isolation. Nuclei isolation for epidermal and endodermal cell layers were conducted based off the protocol by Deal and Henikoff (2011). Chromatin from isolated nuclei were digested with *DNase*I to remove regions of accessible chromatin. After digestion, DNA was then sequenced with Illumina NGS on a Mi-Seq bench top sequencer. Measuring of accessible regions of chromatin was performed using a custom designed program written by Shawn Hoogstra, master's candidate within the lab of Dr. Ryan Austin of Agriculture and Agri-Food Canada, adjunct professor of the University of Western Ontario. For more information on chromatin digestion and analysis on *DNase*I hypersensitivity sites, refer to the upcoming thesis by Shawn Hoogstra (2017)(to be published

by the University of Western Ontario).

2.6 Microscopy

All images were taken using a Nikon florescent microscope, model Eclipse Ni - U. For stained images, roots were emerged in 1X propidium iodide for 5 minutes and thoroughly washed with water before being mounted to slides.

Table 2.1: Forward (For) and reverse (Rev) primers used for cloning endodermal specific promoters.

Gene name	Primer tag	A. T.	Sequence
AT1G33055	Endo-1 (For)	62°C	GCG CGC GGG CCC GAA TTT TGA TGT CGT TTA CAT TCT
AT1G33055	Endo-1 (Rev)		GCG CGC CCC GGG TTT TTA GTT TCT TTA GCT TCA ATG GAA TCT TC
AT3G09390	Endo-2 (For)	62°C	GCG CGC GGG CCC AAT CTT TTG AAG AGA CTC TTA CTA AAG TAA CTT
AT3G09390	Endo-2 (Rev)		GCG CGC CCC GGG TTT TCT CGA GAA AAT TCA AAT TGA
AT3G21720	Endo-3 (For)	62°C	GCG CGC GGG CCC TTG TTT ACA CTA ATC AAA AAT AAT TCC CAT TGT
AT3G21720	Endo-3 (Rev)		GCG CGC CCC GGG GGC TTT AAC TTT TAT AAA TTG GAA ATG
AT5G09570	Endo-4 (For)	62°C	GAG CAC GGG CCC CAA CCT GAA CTA GAA CCC AAT TGG TTT
AT5G09570	Endo-4 (Rev)		GCG CGC CCC GGG TTT GAA TTT CAG ATG TTG AAG TGT
AT1G13440	Endo-5 (For)	62°C	GCG CGC GGG CCC CTA TTT ATT GGA TGT ATA AAG ATC CAT
AT1G13440	Endo-5 (Rev)		GAG CAC CCC GGG TTT GCG AAA TTG AGA TCG AGA GAG ATT
AT2G36460	Endo-6 (For)	62°C	GCG CAC GGG CCC AAC ATA ACT CTT CGT TGT CTG AAA
AT2G36460	Endo-6 (Rev)		GAG CAC CCC GGG GGT TGA AGA AGA ATC GAT TTG GTG AAG AAA
AT4G09150	Endo-7 (For)	62°C	GCG CAC GGG CCC GAT GTT GTT CGT ATT GTT GGG TTT CTC
AT4G09150	Endo-7 (Rev)		GCG CAC CCC GGG TGA TTT CAC GAC CAC ACA AAT GTA GAT
AT2G47180	Endo-8 (For)	63°C	GCG CGC GGG CCC TTA AAA AGG ACT TGT GGA AAA TGT GAC
AT2G47180	Endo-8 (Rev)		GAG CAC CCC GGG GTG ATT AGC ACG TGA TCT GCT GTG
AT4G39900	Endo-9 (For)	62°C	GCG CGC GGG CCC TGT TGT TGA ATT AGT TGT TTA GTG
AT4G39900	Endo-9 (Rev)		GAA CAC CCC GGG TGC AGA GAG AAT CGT GAA CCA A
AT5G10040	Endo-10 (For)	66°C	GCG CGC GGG CCC GCC AAG TTG ATT AAT TAA TCG AAG
AT5G10040	Endo-10 (Rev)		GCG CAC CCC GGG TGT TGG ATC TTC CTA TAT CAG TTG TCT
AT2G06430	Endo-11 (For)	66°C	GCG CAC GGG CCC GCA ATT AAG GGA CTG TAA CCG AAA
AT2G06430	Endo-11 (Rev)		GCG CAC CCC GGG CTG CAA AAC ATG TTA TGA AAC AGC GTC
AT2G15890	Endo-12 (For)	62°C	GCG CGC GGG CCC TAC GTA AAT GTC CTT TGT AAC ATG
AT2G15890	Endo-12 (Rev)		GCG CGC CCC GGG CTC TAG TGA AGA TAT TTT TCA AAA CTC

A. T. stands for annealing temperature for PCR conditions.

Table 2.2: Forward primers designed for endodermal-specific promoter truncations.

Gene name	Primer tag	A. T.	Sequence
AT1G33055	Endo-1: Truncation-1	62°C	GCG CGC GGG CCC AAT TGT CTC GTG ATT TTC ACT AGA
AT3G09390	Endo-2: Truncation-1	62°C	GCG CGC GGG CCC TTT TCC TGC TAA TTT TAT CTA GTG
AT3G09390	Endo-2: Truncation-2	62°C	GCG CGC GGG CCC CGA TTA TTA TAA AAA CAC GAA ATA CCA
AT3G21720	Endo-3: Truncation-1	62°C	GAG CAC GGG CCC GAA ACC GAC AGC AAC AAA ATG
AT3G21720	Endo-3: Truncation-2	62°C	GAG CAC GGG CCC CAA ACG TTT GAG ATT GGT AGG ATG
AT5G09570	Endo-4: Truncation-1	62°C	GAG CAC GGG CCC CAG TGA ACC ACC ACT AGT AAT CTA
AT5G09570	Endo-4: Truncation-2	62°C	GCG CGC GGG CCC CAA AAG TTT AGA TTC GCT CTT TAC
AT1G13440	Endo-5: Truncation-1	62°C	GAT CAA GGG CCC CTT TTG GCA AAA GCC AAG AGT CAT
AT4G13440	Endo-5: Truncation-2	62°C	GCG CGC GGG CCC CAA CAA TAA TAC AAA ATC TTA TGA
AT2G36460	Endo-6: Truncation-1	63°C	GCG CAC GGG CCC CTC AAA GTC TTC ACT TCT CAT TTT
AT4G09150	Endo-7: Truncation-1	62°C	GAG CAC GGG CCC ACT CAT CGA ACT AAA GAG AGG TCA
AT4G09150	Endo-7: Truncation-2	62°C	GCG CGC GGG CCC ATA TTT TGG GTA TTT TTG GTT ACC
AT2G47180	Endo-8: Truncation-1	62°C	GCG CGC GGG CCC CAT TTC TAC CGA CAT CTG AGA AGA
AT4G39900	Endo-9: Truncation-1	62°C	GAT TAT GGG CCC TAC AGA CAG AGG CCC GTG AGC TCT GTA TGT
AT4G39900	Endo-9: Truncation-2	62°C	GCG CGC GGG CCC GTT TTG ATT TTG TGA TTT TTA CAG
AT5G10040	Endo-10: Truncation-1	63°C	GAG CAC GGG CCC TTA AAA TCG TCG TGG AAC GAG ACC
AT5G10040	Endo-10: Truncation-2	61°C	GCG CGC GGG CCC TGA GAA ATA ATA ATT TAT GGC CCA
AT5G10040	Endo-10: Truncation-3	61°C	GCG CGC GGG CCC CGA TAA TGA GGA ATA AAA TGA TTT CAC
AT2G06430	Endo-11: Truncation-1	63°C	GCG CGC GGG CCC AAT GAC TGT CTT ACC CTC TGA ATG
AT2G06430	Endo-11: Truncation-2	63°C	GCA CAC GGG CCC TCG CCG ACG AAT TTT CTC TCT GAG
AT2G15890	Endo-12: Truncation-1	61°C	GCG CGC GGG CCC GGT TAT TTA CTA ATA TTA CCC
AT2G15890	Endo-12: Truncation-2	61°C	GAG CAC GGG CCC GGC CCA AAG TGT ACC AAT CTA AGA TAT

Complementary reverse primers used were the same as reverse primers used in cloning the non-truncated control promoters. A. T. stands for annealing temperature for PCR conditions.

Chapter 3

Results

A collection of cell-type specific microarray data (Birnbaum et al., 2003) for the *Arabidopsis* root cell layers was analyzed to identify genes with cell-type specific expression profiles. Cell-type specific gene clusters were identified for endodermis, cortex, epidermis, stele, and lateral root cap cell layers. Large co-expressed gene clusters were also identified across 3 developmental stages of the *Arabidopsis* root. Promoter sequences of endodermal, cortex, and epidermal specific genes were analyzed with motif prediction software and statistical testing to identify putative motifs with potential for driving cell-type specific expression. Motif analysis focused primarily on the endodermal prediction results. Endodermal specific promoters were found to be enriched with 6 motif patterns, 4 novel and 2 previously described. Promoter enrichment of known TF binding sites were also assessed in endodermal, cortex, and epidermal cell specific gene promoters. Endodermal specific promoters are dominantly enriched with the binding sites associated with the AP2 TF family binding domain. Epidermal specific promoters are enriched with bZIP sites, specifically G-boxes, and cortex specific promoters were dominantly enriched with Myb/SANT binding sites. To assess biological activity of predicted motifs, truncations of endodermal specific promoters, with different putative motifs removed, were used in GFP expression assays in transgenic *Arabidopsis*. The promoter truncations of *ICL* (*ISOCITRATE LYASE*) produced cell-type specific ectopic expression in epidermis and stele cell-layers when

removing putative motif patterns. Epigenetic profiles of endodermal specific promoters were also examined within both the endodermis and epidermis cell layers. Endodermal specific promoters show a greater degree of chromatin accessibility within the endodermis compared to the epidermis, while CpG methylation patterns shown no observable difference between cell layers.

3.1 Co-expressed gene clusters in five root cell-layers

To identify CREs that confer cell-type specific expression within the *Arabidopsis* root, a collection of cell-type specific genes was generated using publicly available microarray data by Birnbaum et al. (2003). Briefly, Birnbaum et al. (2003) used promoters of well-documented cell-type specific genes to drive GFP expression suitable for FACS. This allowed for the accurate separation of protoplast root cells into their respective cell types: epidermis, cortex, endodermis, stele, and lateral root cap. In addition, cells were also separated into three stages of development: apical meristem, basal meristem, and zone of elongation. All together, root protoplast cells were separated into fifteen cell-type/developmental stage conditions. RNA extraction for microarray analysis was then performed on each protoplast pool. The resulting study provided the unique transcriptomes of the five main root cell-types covering expression profiles of 22,748 genes of the *Arabidopsis* genome. These data was mined to obtain a list of target genes with cell-type specific expression.

Before cell-type specific genes could be isolated, preprocessing of the raw microarray data was performed in order to remove genes without significant expression changes, making further data mining easier. This procedure is carried out in two steps and is based on similar methodologies described by Gentleman et al. (2006). First, genes with expression levels too low to be confidently discerned from microarray background noise were removed. This was accomplished by ordering genes by the sum of their expression value over all conditions and excluding the lower quartile, or 25th percentile from this study. Secondly, genes were ordered

based on their degree of expression change by subtracting the lowest expression rate from the highest across conditions. Again, genes falling within the lower quartile of this order were excluded. The purpose of this filter is to remove potential housekeeping genes that aren't differentially expressed between cell-types. In total, 10,045 genes were removed from the data set.

The remaining 12,703 genes were hierarchically clustered to identify cell-type specific co-expressed genes. Hierarchical clustering involves grouping genes together with similar expression patterns. Both divisive and agglomerative clustering algorithms were tested, including DIANA (DIvisive ANAlysis Clustering), AGNES (AGglomerative NESting), and HOPACH (Hierarchical Ordered Partitioning And Collapsing Hybrid). HOPACH, a hybrid of divisive and agglomerative clustering methods proved to be the most effective in identifying co-expressed gene clusters (Figure 3.1), while both DIANA and AGNES approaches produced undesirable results (data not shown)[1]. Results from HOPACH clustering indicate that the majority of genes show developmental stage specificity, with the apical meristem and the zone of elongation comprising the major developmental stages (Figure 3.1). To better identify cell-type specific gene clusters, these developmentally stage dependent genes were removed from the data. Figure 3.2 shows the gene expression heatmap after the 7,245 stage dependent genes were removed. Cell-type specific gene clusters are highlighted for stele, endodermis, cortex, epidermis, and lateral root cap (Figure 3.2a-e). Within cell-type specific clusters, Pearson correlation was calculated between gene expression profiles and an artificial expression profile, or bait gene designed to mimic perfect cell-type specific expression. Genes with a correlation coefficient of $r > 0.75$ to the respective bait were considered cell-type specific and used for motif prediction and mapping. Based on these criteria, the *Arabidopsis* root was found to have 250 stele, 255 endodermal, 76 cortex, 175 epidermal, and 466 lateral root cap specific genes. Lists of cell-type specific genes for endodermal, epidermal, and cortex cell layers are provided in Appendix

[1]Unlike DIANA and AGNES, the combination of both agglomerative and divisive clustering implemented by HOPACH, where clusters are split into two or more sub-clusters with the two closest sub-clusters collapsed and merged, is far superior in identifying expression patterns of large data sets such as the one used in this thesis.

A.

A gene ontology (GO) (Ashburner et al., 2000) analysis of cell-type specific gene sets indicated significant over representation of various biological processes within cell-types compared to the whole *Arabidopsis* genome (Figure 3.3)[2]. Stele specific genes were found to be significantly overrepresented in DNA/RNA metabolism, energy pathways, and with genes involved in unknown biological processes. Endodermal specific genes were enriched in cell organization and biogenesis, transportation, and known and unknown biological processes. Cortex specific genes were not overrepresented in any category compared to the whole genome, with the exception of a significant enrichment in genes involved in unknown biological processes. Epidermal specific genes were found to only be significantly enriched in genes involved in DNA/RNA metabolism. Finally, lateral root cap specific genes were observed to be significantly overrepresented in developmental processes, DNA/RNA metabolism, energy pathways, protein metabolism, along with other biological processes, known and unknown.

3.2 Promoter analysis reveals enrichment of putative motifs

Motif prediction was performed against endodermal, epidermal, and cortex cell-type specific gene promoter regions (Figures 3.4-3.6). Within each cell-type specific co-expression cluster, as determined by hierarchal clustering, gene expression patterns were correlated against an artificial cell-type specific bait vector. Forty genes with the highest correlation coefficient, and therefore most cell-type specific, were selected for motif predictions on their promoters sequences (Appendix B). Forty gene promoters were chosen to provide a large enough subset of genes to accurately reflect the cell-type specific cluster, while small enough not to exceed computational limits. Subsets of cell-type specific promoters were examined before motif prediction to identify and remove gene duplicates using information from TAIR and promoter alignments. Gene duplicates cause unwanted biases in motif predictions due to their shared

[2]Thanks to Shawn Hoogstra for help generating the bar chart seen in Figure 3.3

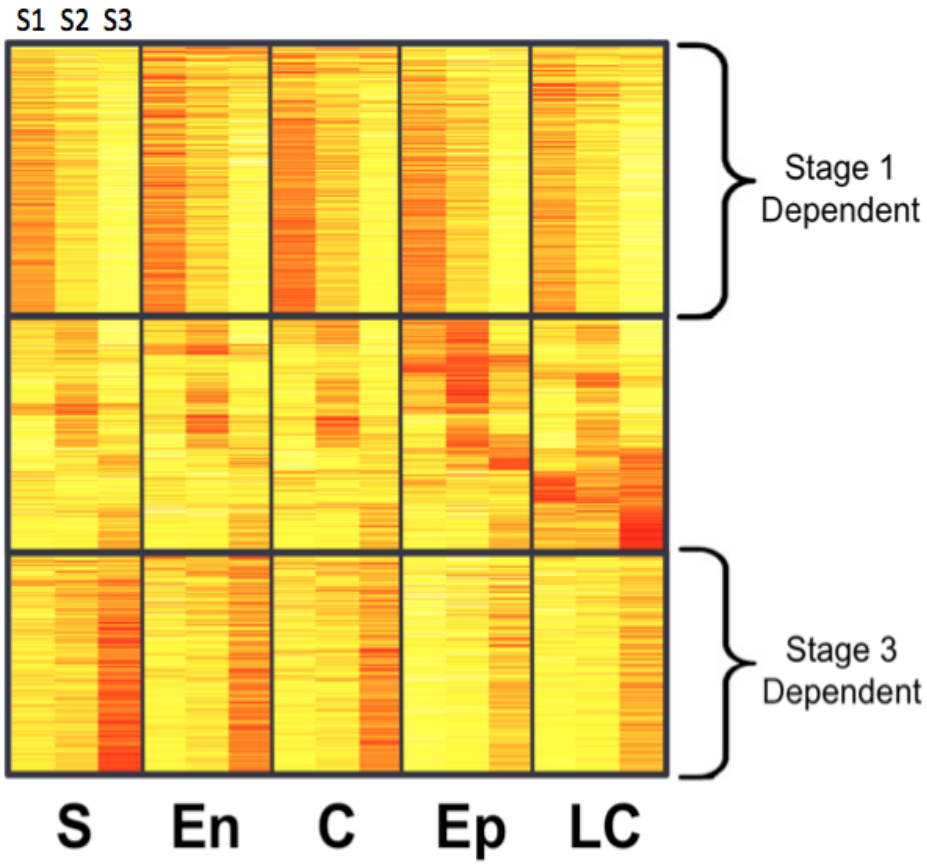


Figure 3.1: Heatmap depicting expression profiles of 12 703 genes from the *Arabidopsis* root. Root cell-type specific microarray data was (Birnbaum et al., 2003) hierarchically clustered using the HOPACH algorithm. Cell layers are denoted as S (stele), En (endodermis), C (cortex), Ep (epidermis) and LC (lateral root cap). Cell layer expression profiles are further subdivided into apical meristem (stage 1, S1), basal meristem (stage 2, S2), and elongation zone (stage 3, S3) developmental stages. Approximately 1 third of genes show stage 1 specificity across all cell-types with another third stage 3 specificity.

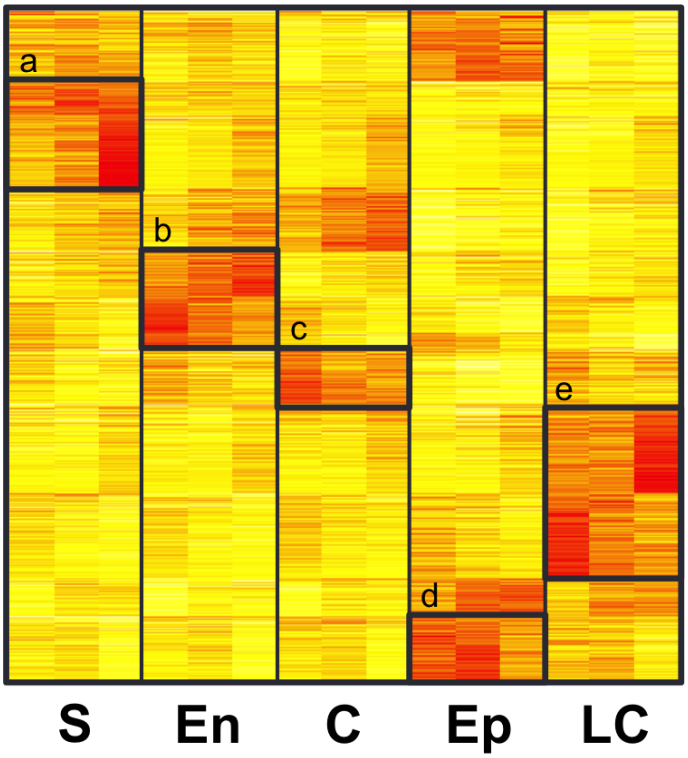


Figure 3.2: Heatmap depicting expression profiles of 5 458 genes after removal of developmental stage specific genes. Cell layers are denoted as S (stele), En (endodermis), C (cortex), Ep (epidermis) and LC (lateral root cap). Cell-type specific gene clusters are shown for (a) stele, (b) endodermis, (c) cortex, (d) epidermis, and (e) lateral root cap.

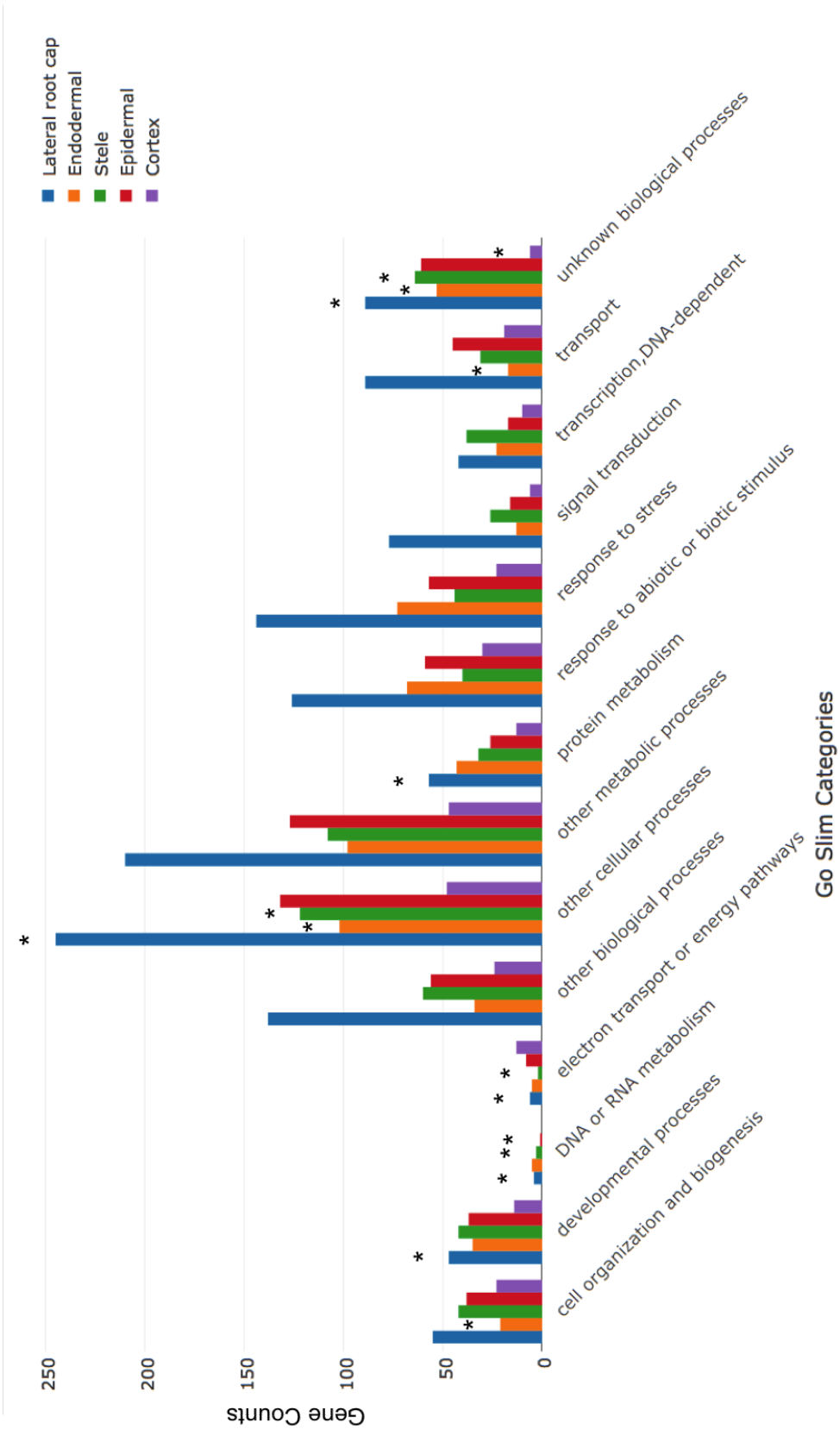


Figure 3.3: Counts of cell-type specific genes classified by the Gene Ontology Consortium's GO slim categories. Cell-types included are stele, endodermal, epidermal, cortex and lateral root cap. Asterisks (*) indicated significant over enrichment of gene counts in one category compared to the whole genome as indicated by a hypergeometric test (alpha = 0.05).

sequences.

Five separate motif finding algorithms were used and include MEME (Bailey and Elkan, 1995), AlignAce (Hughes et al., 2000), Bioprospector (Liu et al., 2001), Weeder (Pavesi et al., 2001), and Motif Sampler (Thijs et al., 2001). These programs all use slightly different approaches for motif finding and offer a level of redundancy in identifying putative motifs. Prediction was performed on the upstream 500 bp promoter region of gene candidates. As motif prediction software usually produces an abundance of putative motifs, to which the majority are false positives, statistical analysis is required to reduce results to a workable number of probable CREs. A non-parametric discriminative algorithm was used to determine if motifs were statistically enriched within cell-type specific promoters (Austin et al., 2016; Winter et al., 2011). Altogether, motif prediction algorithms produced 256 PSSMs for endodermal-specific, 270 for epidermal-specific, and 176 for cortex-specific genes. After testing for statistical enrichment ($Z \geq 3$), PSSM counts were reduced to 131, 105, and 42 PSSMs, respectively. Further filtering of PSSMs was performed to removed highly degenerate motifs that otherwise do not make biological sense (see Methods). Final putative PSSM counts were 88 for endodermal specific genes, 70 for epidermal, and 31 for cortex (Appendix C).

Motif analysis and biological validation was focused mainly on identifying endodermal specific motifs. Distance matrices between PSSMs were generated by *Cistome* (Austin et al., 2016) and processed in R to group highly similar motifs into discernible clades. Endodermal specific motifs were found to group into 3 major clades, along with a variety of minor ones (Figure 3.7). The first major clade produced a GAAGA signal and contains 9 PSSMs (Figure 3.7a). Due to the similarity these motifs share with the well studied GAGA motif (Deng et al., 2013; Horard et al., 2000), motifs falling into this clade are referred to as GAGA-like motifs. The second major clade contains 8 motifs and has a GATC sequence at their core (Figure 3.7b). The third major clade contains 8 motifs (Figure 3.7c) that resemble the TBF1 (Telobox Factor 1) binding site, a conserved sequence found repeated in telomeric regions in yeast, plants, and humans (Bilaud et al., 1996). Three minor clades were also investigated and are annotated

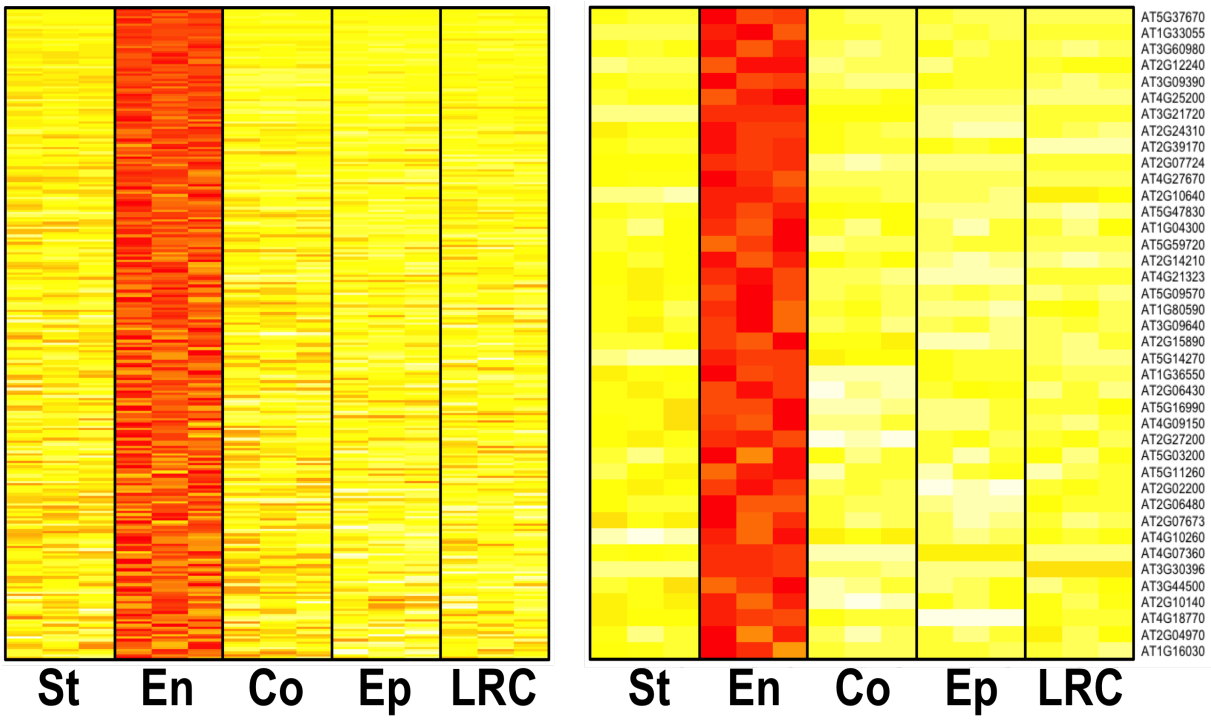


Figure 3.4: Expression profiles of endodermal specific genes ordered by Pearson correlation against artificial expression bait vector. **A**.) Expression profiles of endodermal specific genes with a correlation coefficient $r > 0.75$ to the endodermal specific bait (255 genes in total). **B**.) Expression profiles of 40 endodermal specific genes ($r > 0.89$) used in motif prediction.

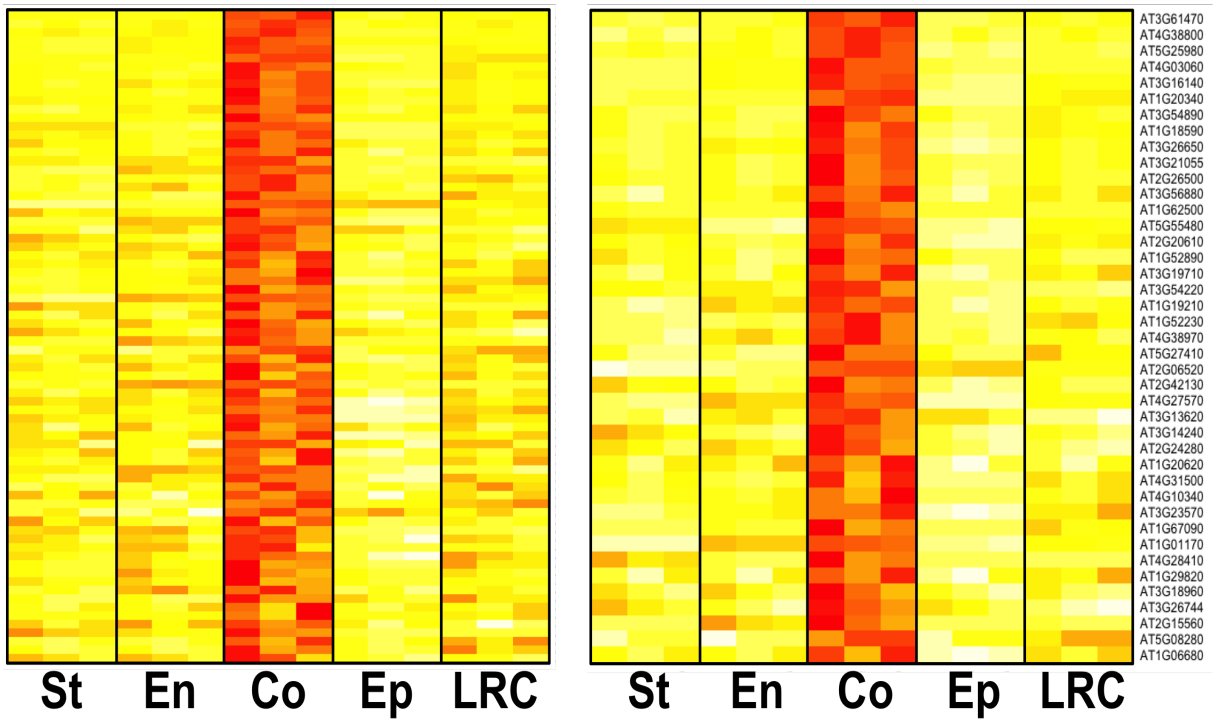


Figure 3.5: Expression profiles of cortex specific genes ordered by Pearson correlation against artificial expression bait vector. A.) Expression profiles of cortex specific genes with a correlation coefficient $r > 0.75$ to the cortex specific bait. 76 genes in total. B.) Expression profiles of 40 cortex specific genes ($r > 0.87$) used in motif prediction.

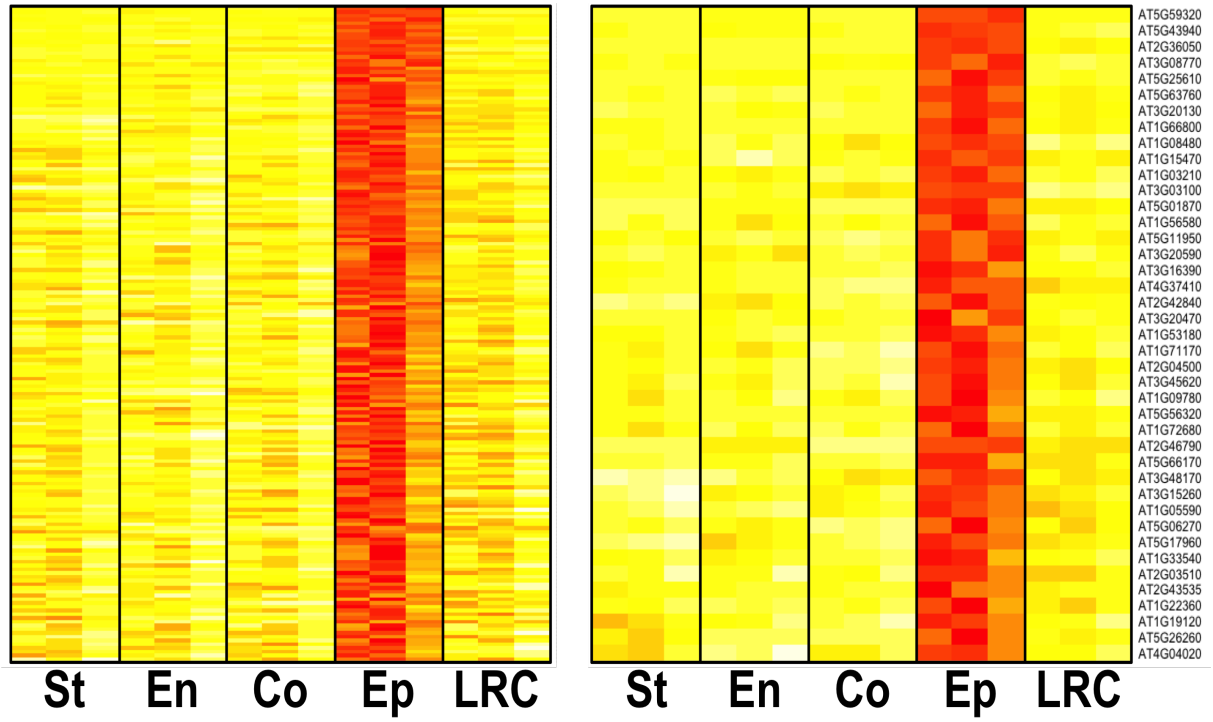


Figure 3.6: Expression profiles of epidermal specific genes ordered by Pearson correlation against artificial expression bait vector. A.) Expression profiles of epidermal specific genes with a correlation coefficient $r > 0.75$ to the epidermal specific bate. 175 genes in total. **B**.) Expression profiles of 40 epidermal specific genes ($r > 0.90$) used in motif prediction.

as Endodermal Minor Clades (EMC) 1-3. The first (Figure 3.7d), contains 3 members which share a thymine core flanked by guanine bases (EMC1). The second (Figure 3.7e), contains just two members with AC rich sequences (EMC2). The last minor clade (Figure 3.7f) contains GT rich motifs with 2 members (EMC3). Additional minor clades investigated were either too degenerate or shared poor sequence similarity among clade members and were deemed unlikely candidates for being biologically functional CREs.

Due to the degeneracy of some motifs, mapping counts to endodermal specific promoters can vary greatly depending on the functional depth (FD) cutoff used. It is therefore imperative to determine an optimal FD to map motifs with, in order to reduce overall false positive rates. This was achieved by comparing the relationship between motif significance and cluster enrichment proportion by mapping motifs at multiple FD cutoffs. This was done for all motifs within clades and helped achieve a baseline FD cutoff for mapping motif occurrences. Figure 3.8 shows that for motifs within clades, increased FD cutoffs leads to a decrease in total enriched promoters and an increase in significance. Note that the scales used for promoter enrichment proportion and Z-score are not proportional. Functional depth cutoffs for each motif were selected to maximize the proportion of enriched endodermal specific promoters while maintaining a high degree of motif significance (Z-score > 4). Functional depth cut offs selected for each motif, along with enrichment significance are presented in Table 3.1 for all 6 motif clades examined.

Re-mapping motifs at their optimal FD depth cutoffs reduces the total number of mapping sites. This reduced set of motif mapping positions can then be used to adjust a motif's PSSM in motif refinement. Motif refinement was found to remove degeneracy and simplify subsequent mapping. Note that non-degenerate motifs remain the same as their mapping positions are fixed and do not vary in sequence.

Since motifs found in the same clade contain a high degree of sequence similarity, and therefore a high degree of overlap in their mapping positions to endodermal specific promoters, a representative motif from each refined clade was selected to represent the overall sequence

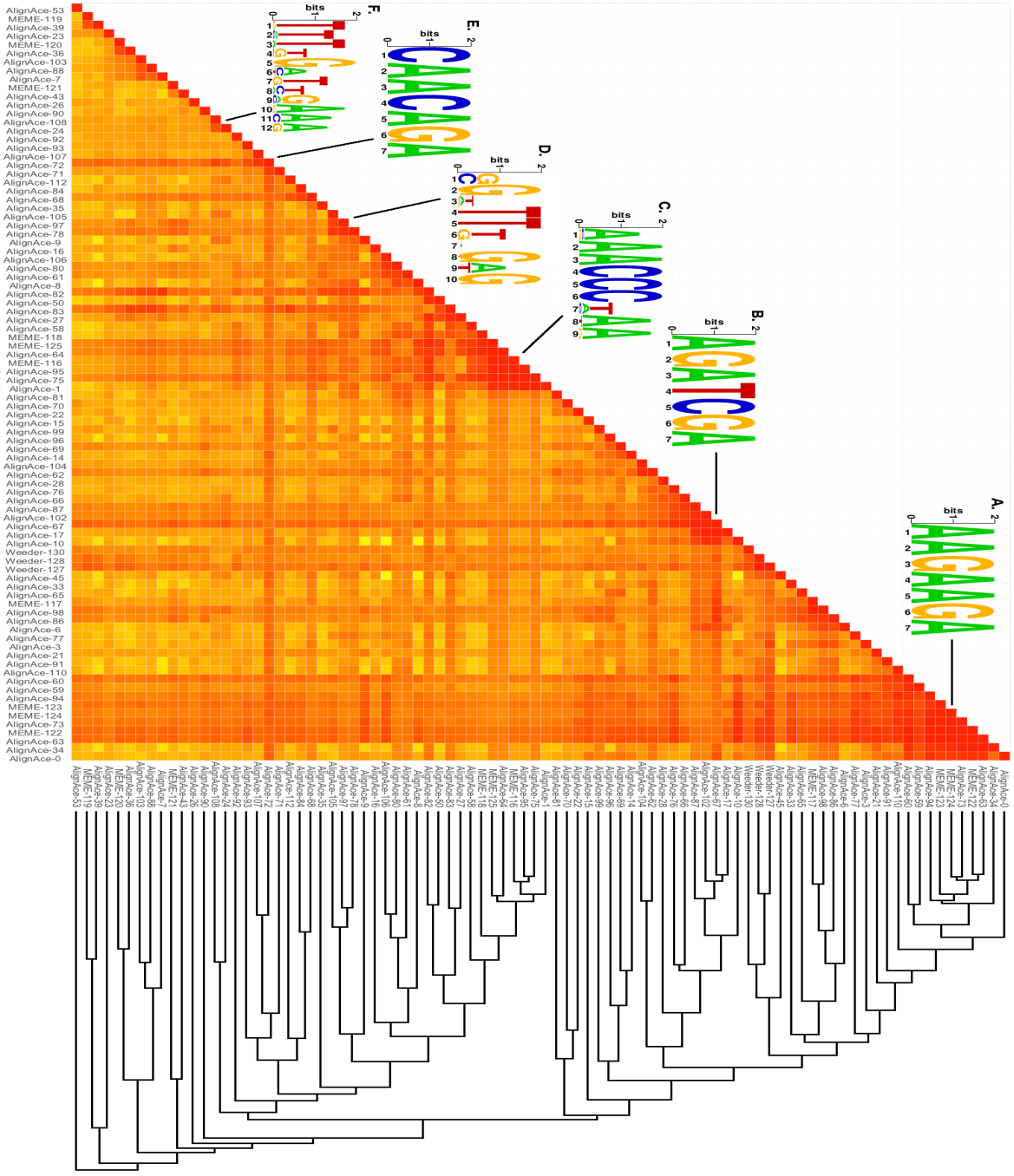


Figure 3.7: Heatmap and dendrogram depicting distance matrixes between 88 endodermal specific motifs. Clades are shown with sequence logos representing the dominating sequence signal within each clade. Dark shades of red indicate strong sequence similarity between motifs, while yellow indicates a weak relationship. **A**.) Clade containing motifs with strong GAGA-like signal. **B**.) Clade containing core GATC sequences. **C**.) Telobox clade. **D-F**.) Minor clades with 2 to 3 members. **D**.) Endodermal minor clade 1 (EMC1). **E**.) Endodermal minor clade 2 (EMC2). **F**.) Endodermal minor clade 3 (EMC3).

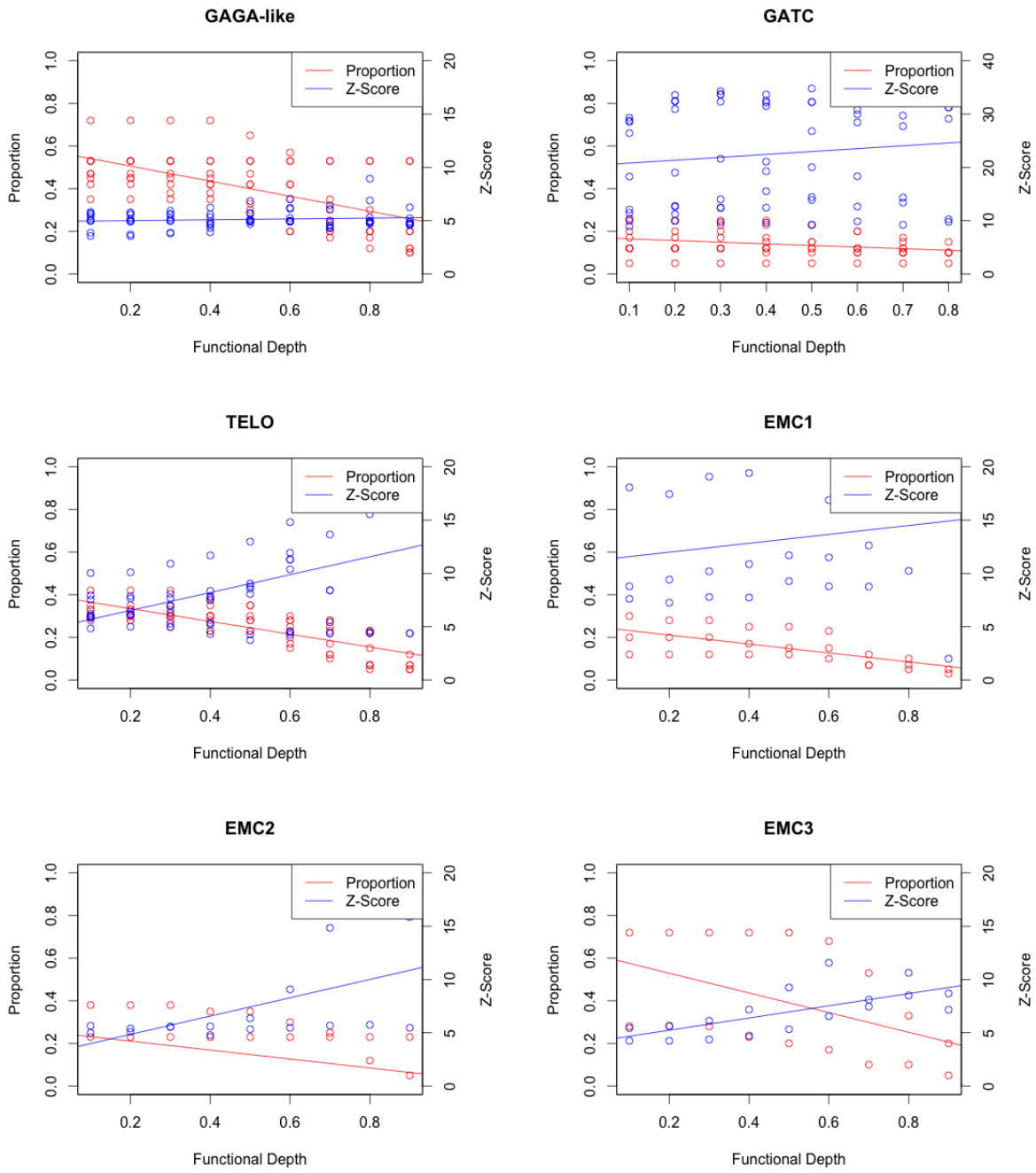


Figure 3.8: Changes in motif significance and counts over degrees of functional depth cutoffs. Scatter plots depicting the relationship between the functional depth (FD) cutoffs motifs are mapped at and their resulting enrichment significance (right y axis, red) and cluster proportion (left axis, blue). Proportion is calculated as the fraction of endodermal specific promoters possessing at least one instance of a motif. As FD cutoffs increase, motifs map to less promoters but tend to become more significant compared to background genome enrichment.

Table 3.1: Optimal functional depth (FD) cutoffs for endodermal specific motifs.

Motif	Optimal FD cut off	% Cluster enriched	Ave. Z-score
AlignAce-0	0.8	30	8.81
AlignAce-59	0.6	35	5.27
AlignAce-60	0.5	33	5.92
AlignAce-63	0.3	53	5.22
AlignAce-73	0.3	45	5.89
AlignAce-94	0.6	42	4.88
MEME-122	0.1	53	4.87
MEME-123	0.3	53	5.28
MEME-124	0.2	42	5.69
AlignAce-10	0.3	5	12.34
AlignAce-17	0.1	5	19.22
AlignAce-28	0.5	20	11.81
AlignAce-66	0.4	12	15.29
AlignAce-67	0.1	7	8.54
AlignAce-76	0.5	15	9.78
AlignAce-87	0.3	10	13.10
AlignAce-102	0.3	10	15.11
AlignAce-1	0.5	35	12.98
AlignAce-58	0.6	30	10.38
AlignAce-64	03	25	6.27
AlignAce-75	0.2	28	6.11
AlignAce-95	0.5	30	8.06
MEME-116	0.4	30	7.84
MEME-118	0.6	28	11.31
MEME-125	0.7	28	5.42
AlignAce-78	0.5	25	9.27
AlignAce-97	0.3	20	10.20
AlignAce-105	0.5	12	22.76
AlignAce-72	0.1	23	5.65
AlignAce-107	0.7	25	14.85
AlignAce-24	0.8	33	10.63
AlignAce-108	0.3	28	6.12

Endodermal specific motifs grouped by clade with the FD cutoff used to optimize enrichment significance and cluster percentage. Cluster percentage is defined as the proportion of endodermal specific promoters significantly enriched by a motif. Z-score is defined as the average Z-score of 3 independent enrichment significance tests of motifs mapped at a given FD cutoff. Motif mapping for significance testing was on the upstream 500 bp promoter regions of genes.

signal seen in a clade. Selection of a representative motif from each clade was accomplished through the consideration of a variety of factors: including the significance score of each motif (Z-score), its degeneracy measured by information content (IC), the number of mapping sites found within endodermal specific promoters, and the proportion of endodermal specific promoters significantly enriched for each motif. Note, that while previous mappings of motifs were performed against the upstream 500 bp promoter region of endodermal specific genes, mapping of refined motifs was done against the upstream 1000 bp promoter region. This was done to include any distal motif sites. Table 3.2 shows these results for all motifs found within selected clades. For the GAGA-like clade, motif MEME-122 was selected primarily because it is non-degenerate and has a high degree of enrichment within endodermal specific promoters. Within the TELO clade, MEME-125 was selected for the same reasons. AlignAce-67 was selected in the GATC clade for its non-degeneracy and because it consists only of the AGATCGA sequence seen in the core of other GATC clade motifs. AlignAce-78, AlignAce-72 and AlignAce-24 were selected as representative motifs for minor clades 1-3. For simplicity, these motifs will be referred to as Endodermal Specific Motifs (ESM) 1 through 3, respectively. Sequence logos of selected refined motifs are shown in Figure 3.9.

As a secondary measure to assure an appropriate FD cutoff has been selected, receiver operating characteristic (ROC) curves were generated from counts of unrefined representative motifs mapped at increasing FD cutoffs (Figure 3.10). The results from these curves were able to validate the choice of functional depth cutoffs used to refine degenerate motifs within clades. The functional depth cutoffs suggested by the ROC results agreed with what was used for motif refinement; with the exception of the Telobox motif (MEME-125). The ROC curve for MEME-125 indicated that a FD cutoff of 0.7 used to refine the Telobox motif was too stringent (Figure 3.10a). However, comparing Telobox sites mapped at a functional depth of 0.7 and a less stringent cut off of 0.3, revealed no difference in Telobox sites in promoters used for biological validation of putative motifs.

Table 3.2: Selection of representative motifs within clades.

Motif	Z-score	I.C.	40 Gene set		255 Gene set	
			Sites	% Cluster enriched	Sites	% Cluster enriched
AlignAce-0	3.73	1.24	195	95	1127	96.4
AlignAce-59	3.69	1.40	108	62.5	604	69
AlignAce-60	5.81	1.73	60	42.5	309	47
AlignAce-63	5.13	1.87	70	65	338	60.4
AlignAce-73	5.72	1.78	67	50	297	50.2
AlignAce-94	5.83	1.60	89	67.5	419	64
MEME-122*	5.13	2.00	70	65	338	60.4
MEME-123	5.09	1.71	69	57.5	338	57.3
MEME-124	5.79	1.68	59	50	236	44
AlignAce-10	15.36	1.99	5	5	5	1
AlignAce-17	14.83					
AlignAce-28	7.79	1.29	18	20	50	15
AlignAce-66	12.85	1.32	18	15	26	5.5
AlignAce-67*	7.75	2.00	21	12.5	49	11.4
AlignAce-76	8.07	1.3	20	22.5	73	19.6
AlignAce-87	11.23					
AlignAce-102	12.79	1.96	15	10	25	5
AlignAce-1	10.30	1.19	36	35	79	19.2
AlignAce-58	7.72	1.10	35	50	125	33
AlignAce-64	5.98	1.70	30	27.5	85	21.2
AlignAce-75	4.92	1.68	31	27.5	93	23.2
AlignAce-95	6.02	1.41	28	30	81	22.4
MEME-116	5.84	1.41	30	32.5	87	24.3
MEME-118	7.45	1.20	33	35	93	27
MEME-125*	4.91	1.70	38	37.5	140	35
AlignAce-78*	7.21	1.49	26	35	104	26.6
AlignAce-97	9.15	1.48	19	22.5	76	15.9
AlignAce-105	18.66	1.40	16	12.5	44	6.3
AlignAce-72*	5.53	2.00	24	32.5	47	14
AlignAce-107	4.82					
AlignAce-24*	4.16	1.10	71	80	421	80
AlignAce-108	6.07	1.26	21	32.5	102	22

Z-scores generated for enrichment of refined motifs against the upstream 500bp promoter regions of endodermal specific genes. Degeneracy of motifs is measured as information content (IC). Number of mapped motif sites and proportion of enriched endodermal specific promoters are indicated for the 40 promoter set used for prediction and the larger set of endodermal specific promoters. Asterisks (*) denote motifs selected to represent the overall motif signal seen in each clade. Missing values are for motifs which were not tested due to their high degeneracy.

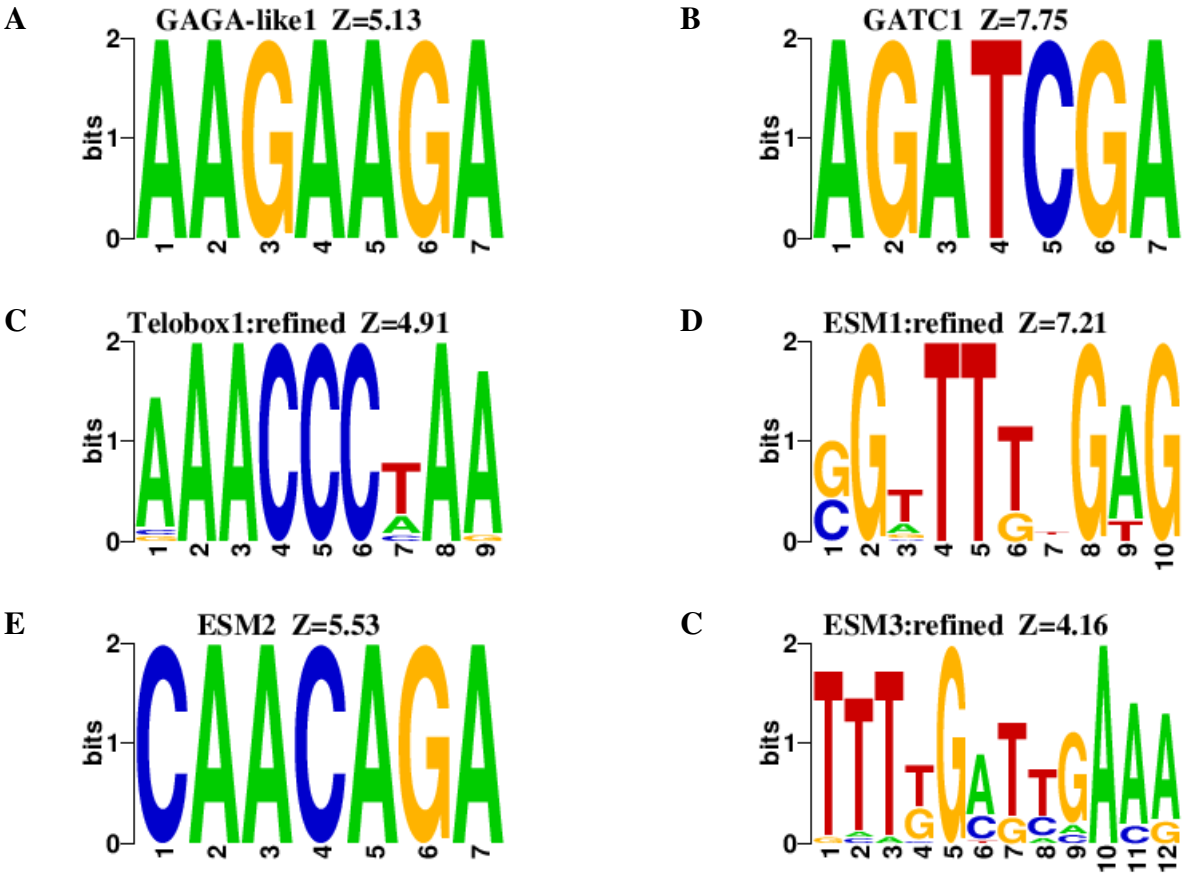
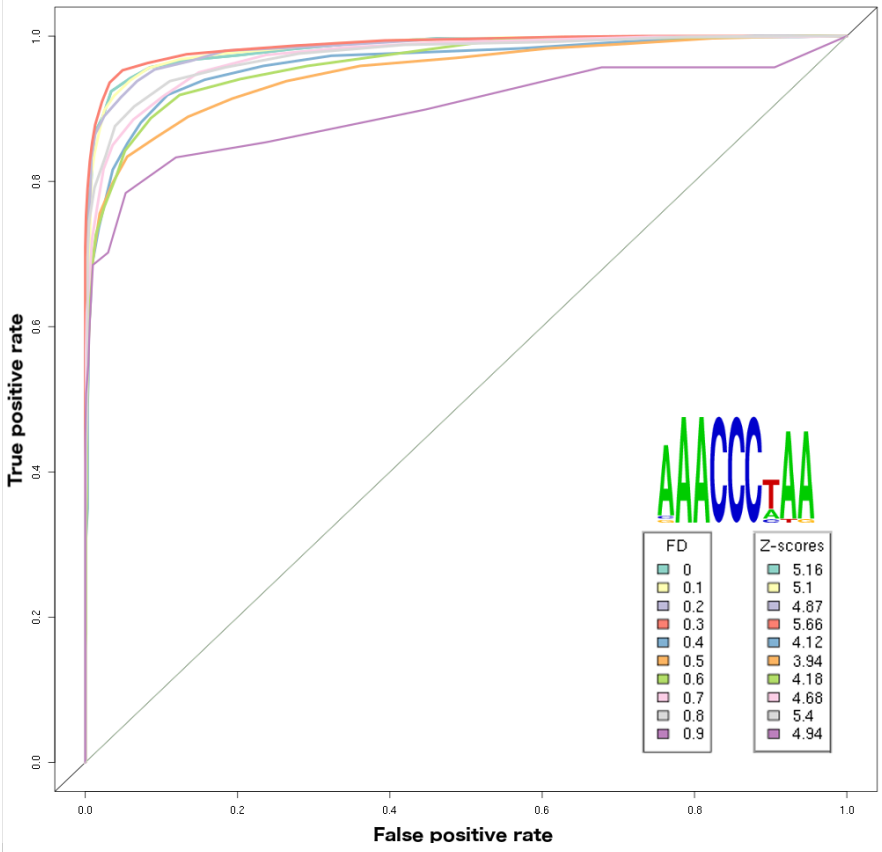
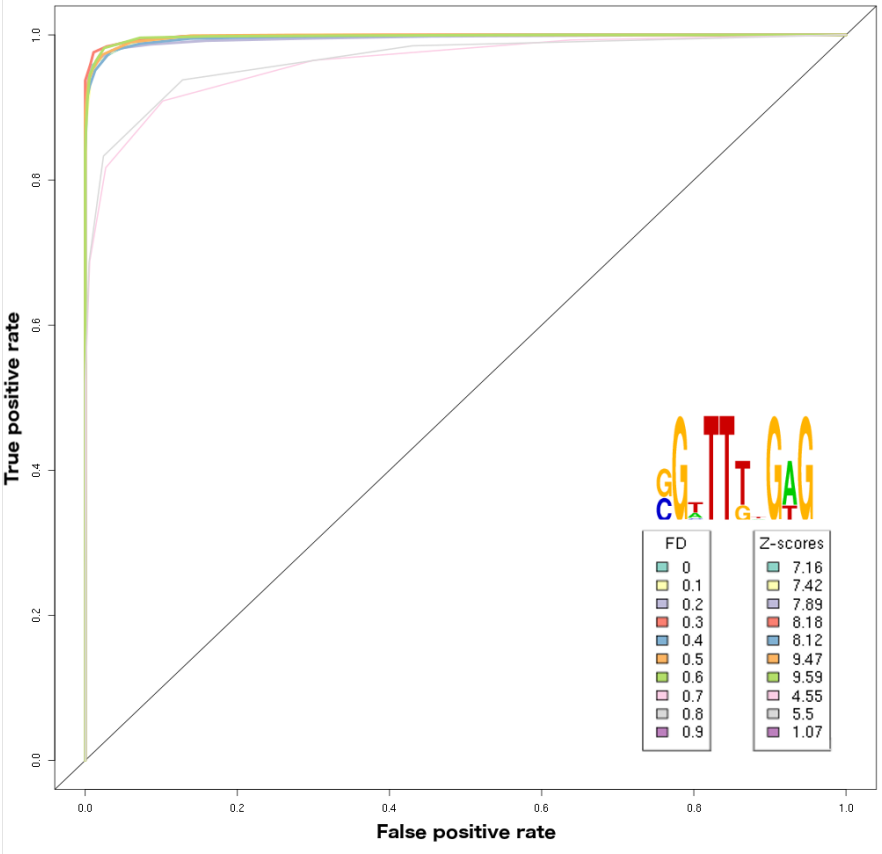


Figure 3.9: Refined sequence logos for motifs significantly enriched within endodermal cell-type specific promoters. **A**.) GAGA-like motif, non-degenerate and unrefined. **B**.) GATC motif, non-degenerate and unrefined. **C**.) Telobox motif refined at a FD cutoff of 0.7. **D**.) Endodermal specific motif 1 (ESM1) refined at a FD cutoff of 0.5. **E**.) Endodermal specific motif 2 (ESM2), non-degenerate with no refinement. **F**.) Endodermal specific motif 3 (ESM3) refined at a FD cutoff of 0.8. Significant scores are calculated based off motif enrichment to the upstream 500 bp promoter region of endodermal specific genes.

A.


True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
FD
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Z-scores
5.16
5.1
4.87
5.66
4.12
3.94
4.18
4.68
5.4
4.94

B.


True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
FD
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Z-scores
7.16
7.42
7.89
8.18
8.12
9.47
9.59
4.55
5.5
1.07

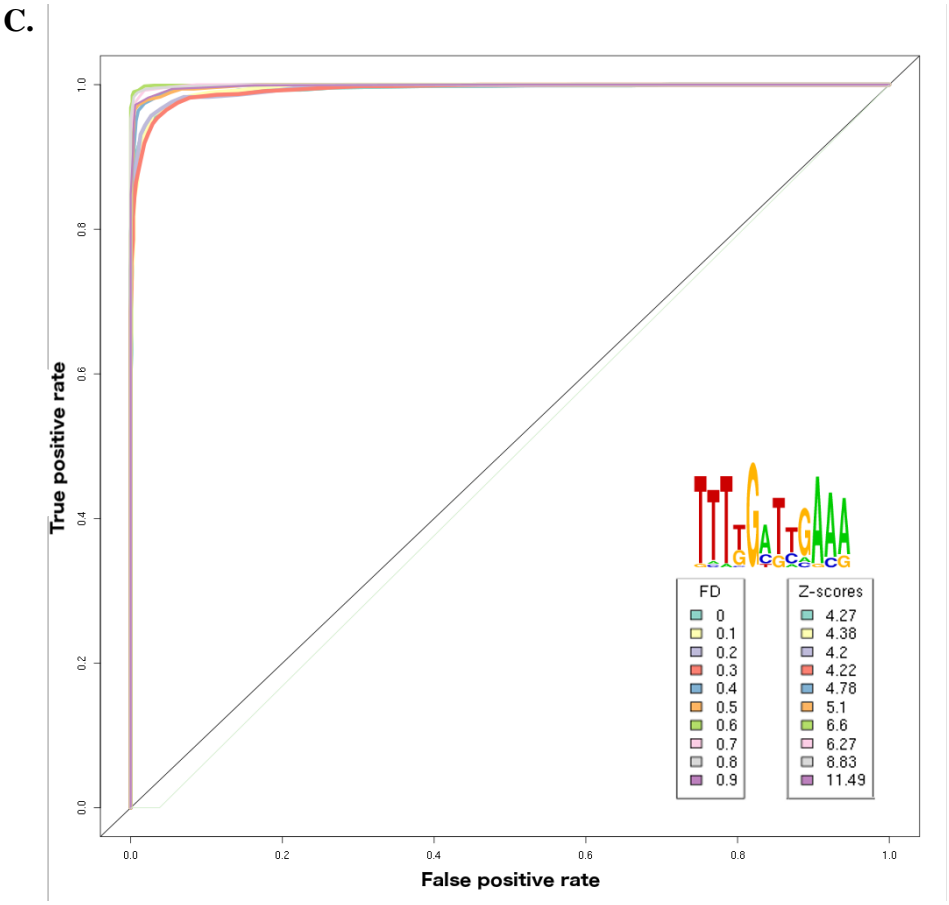


Figure 3.10: Receiver operator characteristic (ROC) curves for endodermal enriched motifs. ROC curves generated from mapping counts across various function depth (FD) cutoffs for the three degenerate motifs enriched within endodermal specific promoters: Telobox, ESM1, and ESM3. **A**), ROC curve for Telobox motif. FD of 0.3 produced the best true positive to false positive ratio, however a more stringent cutoff of 0.7 was used. **B**), ROC curve for ESM1. Function depth cutoffs of 0.3-0.6 produce the best ratio curves with 0.5 being selected. **C**), ROC curve for ESM3. Functional depth cutoffs of 0.6-0.8 produce the best ratio curves with 0.8 being selected.

3.3 Positional disequilibriums in motif occurrences

Motifs were mapped to the upstream 1000 bp promoter region of all 255 endodermal specific genes. Of this set, 154 promoters (60%) were significantly enriched with the GAGA-like motif. The GAGA-like motif was also found to contain a positional disequilibrium when mapped to both promoter and upstream sequences flanking the TTS's of endodermal specific genes (Figure 3.11a). GAGA-like enrichment increased toward the TSS from both sides with a reduction in frequency in sequence directly adjacent to the TSS. Due to the sequence similarity observed in the GAGA-like motif compared to the GAGA motif, positional disequilibriums of the GAGA motif were additionally analyzed (Figure 3.11b). Indeed, GAGA enrichment bias was seen to emulate that of the GAGA-like motif with positional biases occurring directly up and downstream of a given gene's TSS. This trend however was not exclusive to endodermal specific promoters, as the GAGA-like motif was found to contain this enrichment pattern just as frequently in non-endodermal specific genes. Furthermore, positional frequencies of the GAGA motif are on average 3 fold higher than the GAGA-like motif. The Telobox motif was found to be significantly enriched within 89 endodermal specific promoters (35%). No readily discernible positional bias was seen in Telobox motif positions. However, a sharp spike in frequency at approximately -400 bp is seen (Figure 3.11c). Telobox positional frequencies remained constant around the TSS in endodermal specific genes and had a slight increase in frequency within the genome as a whole. The GATC motif was found to only be enriched within 29 of 255 (11%) endodermal specific promoters with no observable positional disequilibriums within upstream promoter sequences (Figure 3.11d). There was however, a slight increase in GATC motif positions downstream of the TSS observed in both endodermal specific promoters and within the background genome. ESM1 (AlignAce-78) is significantly enriched within 68 (27%) promoters. A positional disequilibrium was seen with increased enrichment around -250 bp from the TSS while remaining invariable in other regions included the TSS flanking sequences (Figure 3.11e). ESM2 (AlignAce-72) is significantly enriched within 36 promoters and contains no positional biases (Figure 3.11f). Lastly, ESM3 was found to be highly enriched

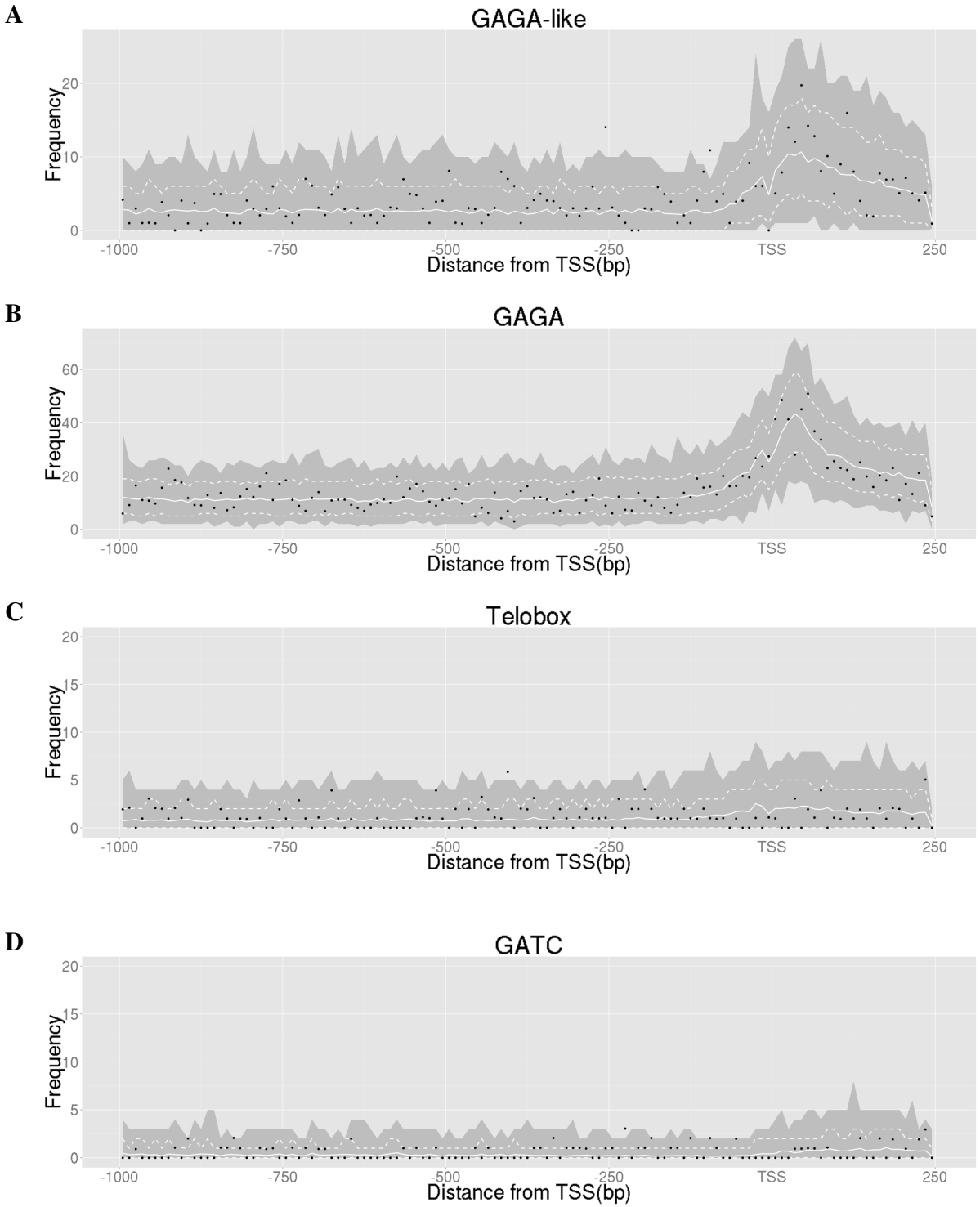
within endodermal specific promoters. Mapping results show that 204 of the 255 endodermal specific promoters contain significant enrichment. In addition, a noticeable positional disequilibrium was detected with increased enrichment occurring -400 bp downstream of the TSS (Figure 3.11g). Frequency positions flanking TSS sequence were uniform for ESM3.

3.4 ESM1/ESM3 motifs are necessary for endodermal expression

Gene expression assays were conducted within transgenic *Arabidopsis* to assess whether putative motifs are functional CREs. Identified endodermal specific promoters were cloned in front of a GFP reporter gene fused to a nuclear membrane localization signal.[3] Constructs were then transformed into *Arabidopsis* to confirm endodermal specific expression of promoters. All transgenic lines were identified by kanamycin resistance selection. Twelve endodermal specific gene promoters were chosen for cloning. These were selected based on their motif placement and high expression in cell-type microarray data (Birnbaum et al., 2003) (Table 3.3). Due to the low number of motif counts, ESM2 was not investigated further as the few endodermal specific promoters that did contain ESM2 enrichment weren't suitable candidates for gene expression assays. Twenty-three additional GFP constructs were produced. These contain truncated versions of the 12 endodermal specific promoters. Each truncation was designed to systematically remove motifs to test their involvement in promoter activity (Figure 3.12). In total, 35 promoter constructs were designed, cloned, and transformed into *Arabidopsis* to assess CRM involvement in cell-type specific expression.

Among the 12 endodermal specific constructs, GFP expression was confirmed in two lines, Endo-1 (AT1G33055), an unknown protein coding gene, and Endo-3, which encodes a glyoxylate cycle enzyme *ISOCITRATE LYASE* (*ICL*). Endo-1 GFP expression was confirmed in over

[3]The GFP reporter gene originates from Deal and Henikoff (2010). The nuclear localization signal allows GFP to be localized on the nuclear membrane making identification of cell-layers easier and reduces GFP bleaching.


A
GAGA-like
Frequency
0
10
20
-1000
-750
-500
-250
TSS
250
Distance from TSS(bp)
B
GAGA
Frequency
0
20
40
60
-1000
-750
-500
-250
TSS
250
Distance from TSS(bp)
C
Telobox
Frequency
0
5
10
15
20
-1000
-750
-500
-250
TSS
250
Distance from TSS(bp)
D
GATC
Frequency
0
5
10
15
20
-1000
-750
-500
-250
TSS
250
Distance from TSS(bp)

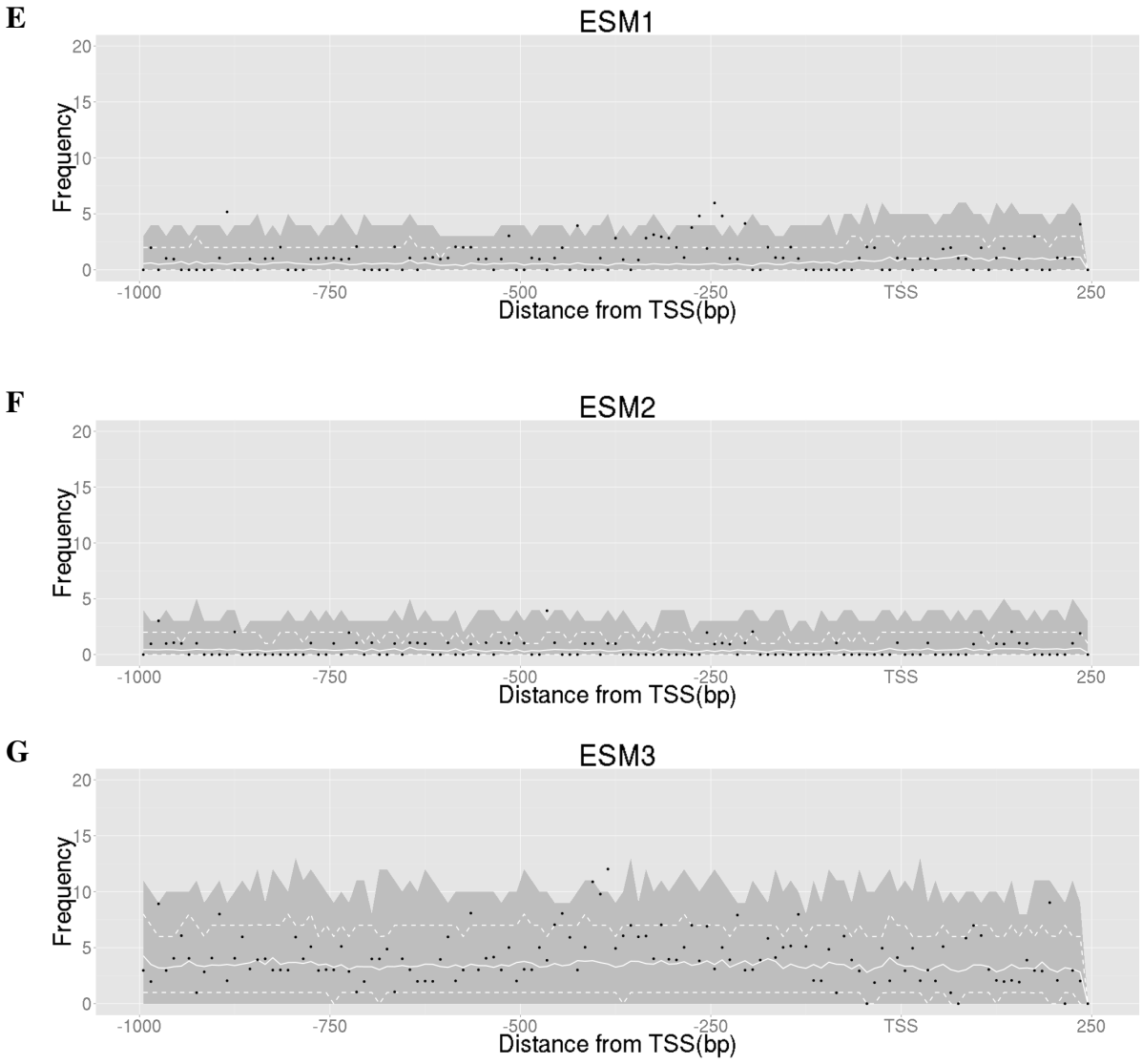


Figure 3.11: Positional frequencies of motif enrichment within endodermal specific promoters compared to the background genome. Positional frequencies of motifs enriched within endodermal specific promoters compared to the *Arabidopsis* background genome for both the upstream 1000 bp and downstream 250 bp regions from transcriptional starts sites (TSS). Black dots indicate positional frequencies for motifs found within endodermal specific promoters. Solid white lines indicate mean positional frequencies of motifs within the *Arabidopsis* genome with dashed lines indicating 5th and 95th percentile. Grey edges indicate maximum and minimum frequency counts seen in in the background genome. Positional frequencies presented for GAGA-like and GAGA motifs **(A-B)**, Telobox **(C)**, GATC **(D)**, ESM1 **(E)**, ESM2 **(F)**, and ESM3 **(G)** motifs.

Table 3.3: Gene promoters selected for gene expression assays.

Promoter ID	AGI †	Gene/class ⋄	Fluorescence ‡
Endo-1	AT1G33055	Unkonwn	448.93
Endo-2	AT3G09390	METALLOTHIONEIN 2A	492.35
Endo-3	AT3G21720	ISOCITRATE LYASE	861.38
Endo-4	AT5G09570	Cox19-like CHCH family protein	385.69
Endo-5	AT1G13440	GLYCERALDEHYDE-3-PHOSPHATE	5842.81
Endo-6	AT2G36460	FRUCTOSE-BISPHOSPHATE ALDOLASE 6	1151.7
Endo-7	AT4G09150	T-complex protein 11	294.35
Endo-8	AT2G47180	GALACTINOL SYNTHASE 1	243.1
Endo-9	AT4G39900	adenine deaminase	211.34
Endo-10	AT5G10040	transmembrane protein	205.46
Endo-11	AT2G06430	Ulp1 protease family	129.56
Endo-12	AT2G15890	MATERNAL EFFECT EMBRYO ARREST 14	173.18

‡ Fluorescence intensity values taken from Birnbaum et al. (2003) microarray data. ⋄ Gene names and description are taken from the *Arabidopsis* information resource (TAIR). † AGI stands for Arabidopsis Genome Initiative and represent unique identity tags for each gene within the *Arabidopsis* genome.

6 independent transgenic lines with expression concentrated in both the endodermis and cortex (Figure 3.13). The truncated promoter for Endo-1 was designed to remove 3 placements of ESM3 starting approximately 370 bp downstream of the TSS (Figure 3.12). No GFP expression was detected within transgenic plants possesing this construct, indicating that the removal of this region interrupted promoter activity (Figure 3.14). Transformation efficiency was low for Endo-3 constructs but two independent transformant lines were isolated and confirmed for GFP expression. GFP expression within Endo-3 constructs was less intense than that of Endo-1, and was also confined to both the endodermal and cortex cell layers (Figure 3.15). Two promoter truncations were designed for Endo-3. The first removed a distal GAGA-like and ESM3 motif found in close proximity to each other approximately 830 bp downstream of the TSS (Figure 3.12). This resulted in GFP expression observed in the root tip epidermis, with no expression in the endodermis or cortex (Figure 3.16b-d). GFP tagged nuclei were also detected in small patches in mature regions of roots (Figure 3.16a) across multiple cell-types. The second promoter truncation of Endo-3 removed a further 428 bp region containing one placement of ESM1 (Figure 3.12). GFP expression for this construct was shifted to the stele cell-layer of roots (Figure 3.17). Unlike the previous expression patterns of Endo-3, stele expression in this construct was observed throughout the entire root persisting towards the hypocotyl. The Endo-3 promoter additionally contains two placements of ESM3, however these were too close to the TSS to design promoter truncations. GFP expression for Endo-3 truncations 1 and 2 were confirmed in 3 and 4 independent lines, respectively.

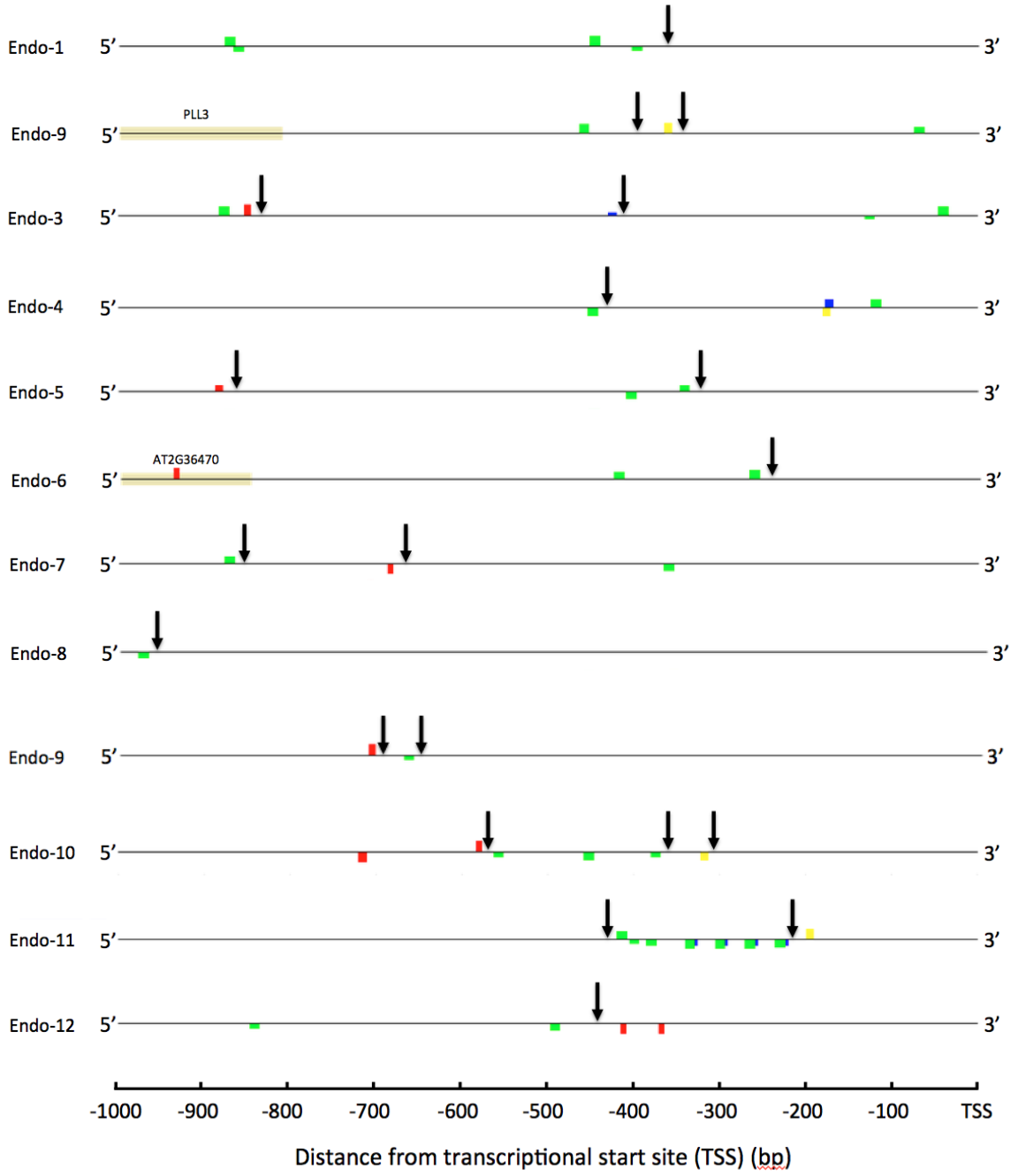


Figure 3.12: Positional mappings of motifs within endodermal specific promoters used for motif biological validation. Figure depicts motif positions found 1000 bp downstream of transcriptional starts (TSSs). Yellow markers indicated Telobox motif positions, red markers for GAGA-like motifs, blue for ESM1, and green for ESM3 motifs. Yellow blocks denote neighbouring genes found within 1000 bp from the corresponding gene's TSS. Arrows mark positions of the 5' end of truncated versions of promoters used to assess motif involvement in endodermal specific expression.

Figure 3.13: One-week-old transgenic *Arabidopsis* roots expressing GFP under Endo-1 (AT1G33055) promoter. **(A-B)**, GFP expression within cortex (i.) and endodermis (ii.) cell layers of two independent transgenic lines. **(C)**, light GFP expression seen in cortex. Roots stained with propidium iodide and imaged at 200X magnification.

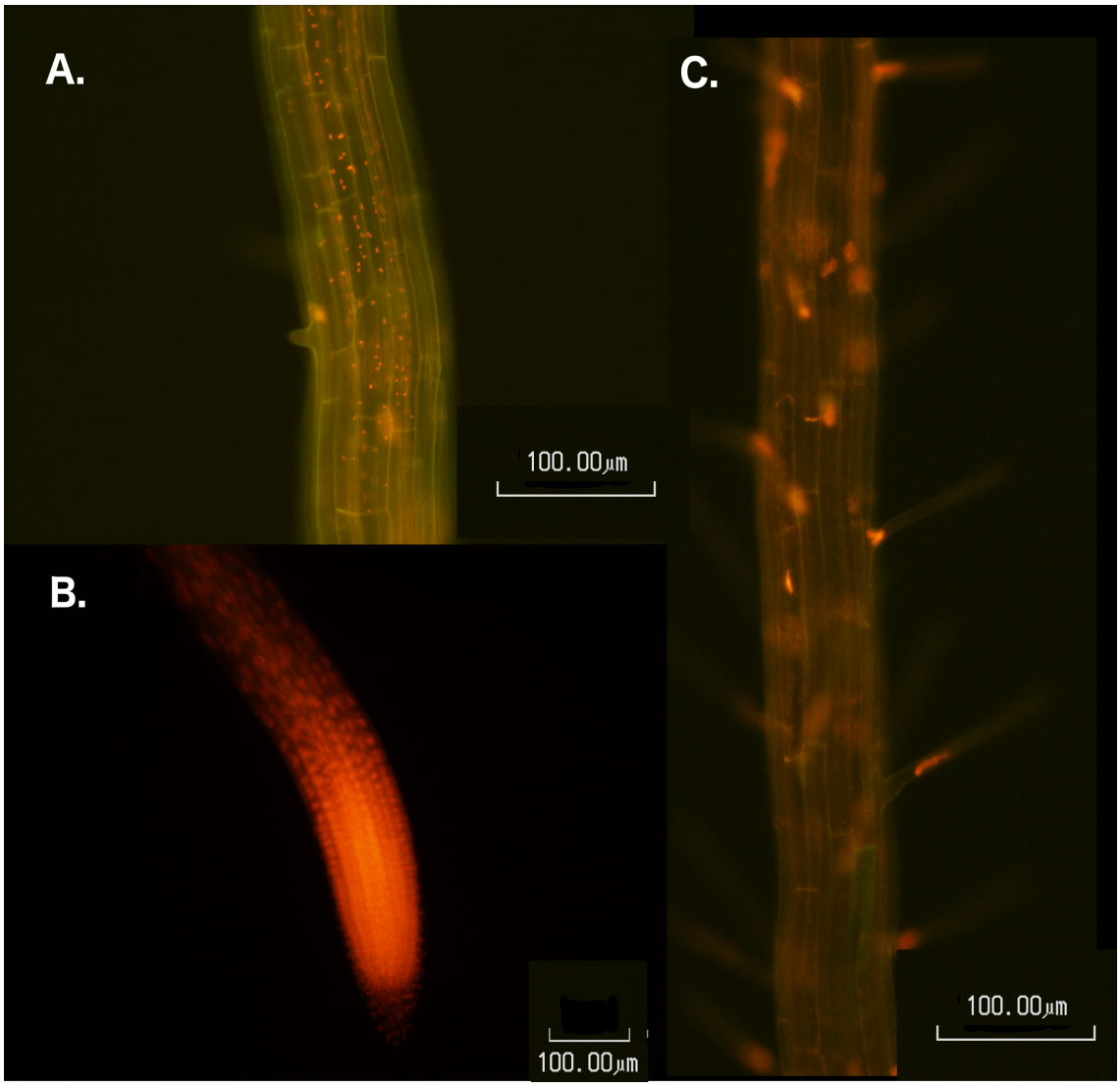


Figure 3.14: Transgenic root expression of truncated Endo-1 promoter. No visible GFP expression was detected within roots. Images of regions **(A)** above zone of elongation, **(B)** root cap and meristem, and **(C)** mature root. One week old roots stained with propidium iodide and imaged at 100X **(B)** and 200X **(A and C)**.

Figure 3.15: One-week-old transgenic *Arabidopsis* roots expression GFP under Endo-3 (*ISOCITRATE LYASE, ICL*) promoter. GFP expression seen in cortex and endodermis (i). Images taken at 200X magnification with with **(A)** propidium iodide stain and **(B)** unstained root.

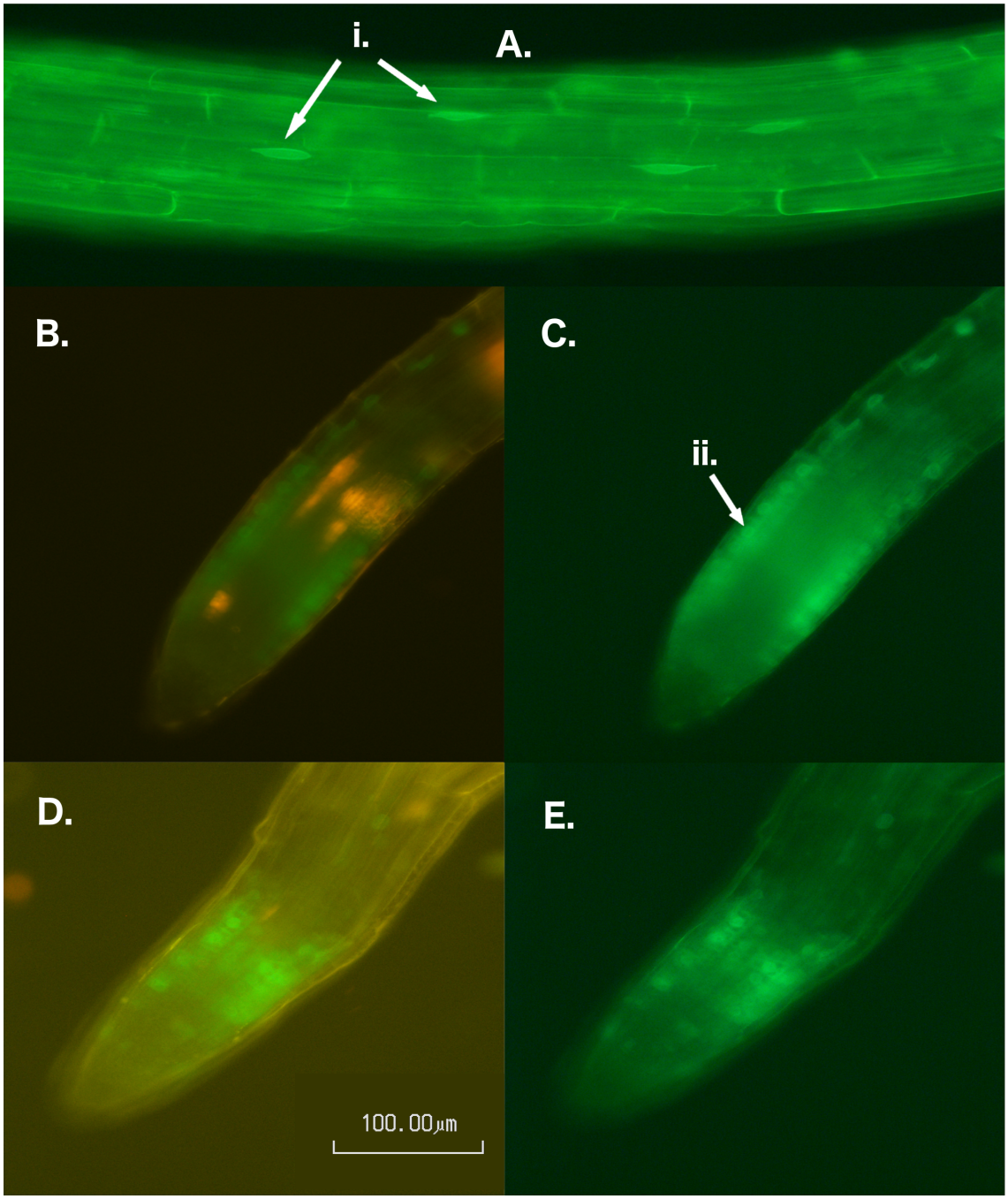


Figure 3.16: Transgenic root expression of truncated Endo-3 (*ICL*) promoter. Promoter designed to remove distil ESM3 and GAGA-like motifs. Slight GFP expression was detected in random regions of the mature root with GFP bound nuclei (i). GFP expression seen in epidermal layers (ii) of root tips **(B/C and D/E)**. Propidium iodine staining for panels **(B)** and **(D)**. All images taken at 200X magnification. Roots approximately 1 week old.

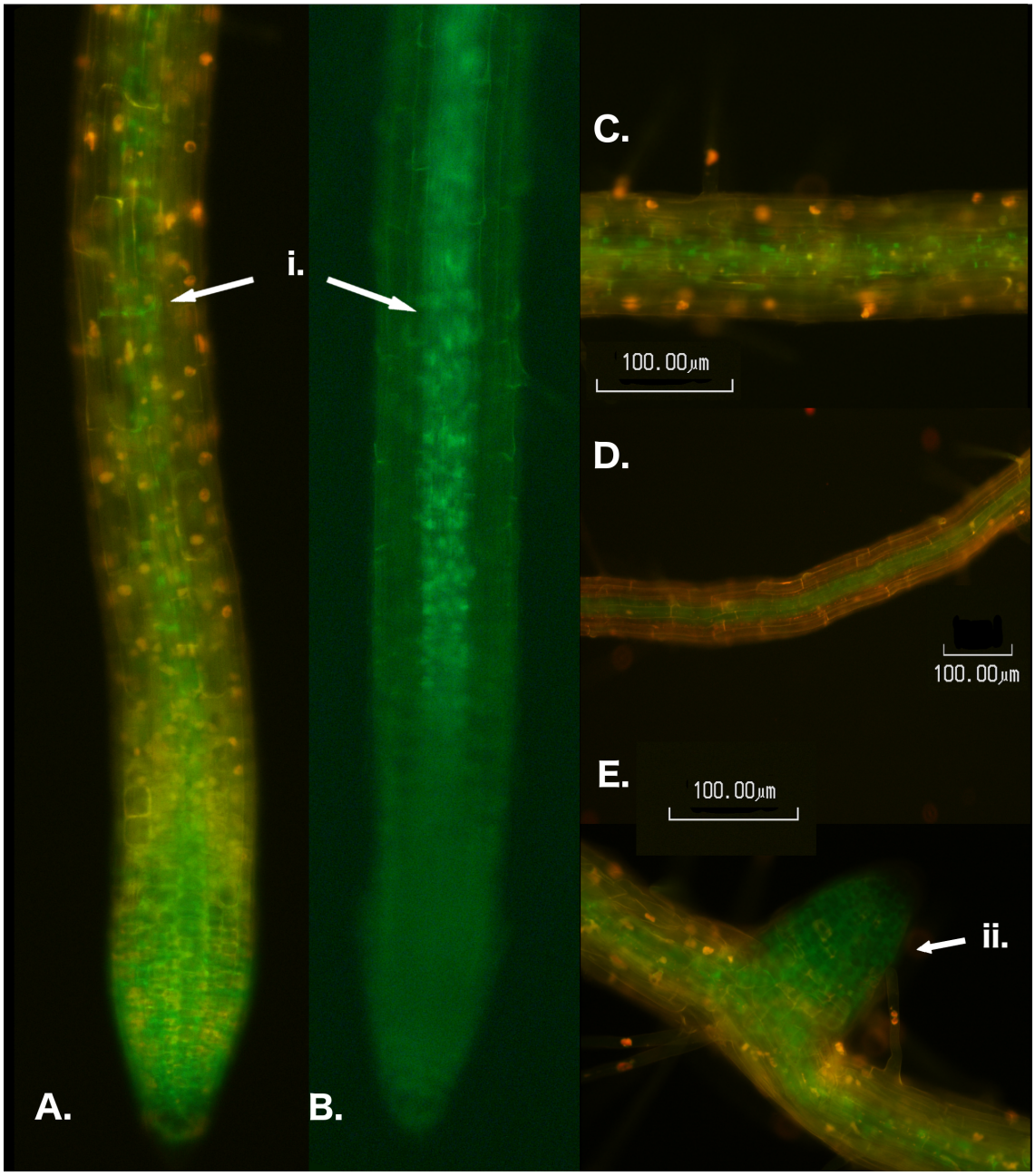


Figure 3.17: Transgenic root expression of second truncated Endo-3 (*ICL*) promoter. Truncation removes downstream ESM3, GAGA-like and Telobox motifs. GFP expression detected within the stele cell layer (i) and meristem (ii) of roots. **(A)**, propidium iodine stained root with GFP expression in both stele and root meristem. **(B)**, unstained root with stele GFP expression only. **(C-D)**, zone of elongation with stele expression. **(E)**, lateral root emerging with meristem GFP expression. Images taken at 100X **(D)** and 200X **(A-C, E)** magnification. Approximately 1 week old seedlings.

3.5 *A priori* mapping identifies enrichment of DNA binding domains in three root cell layers

Protein binding microarray (PBM) studies aided by current databases[4] of known transcriptional binding sites were used to determine significant enrichment of motifs not predicted by pattern finding programs. Binding domains of numerous *Arabidopsis* TF families had been recently characterized by Weirauch et al. (2014). This was a large study where the binding preferences of over 1000 transcription factors encompassing 54 DBD classes were determined for 131 eukaryotic organisms. Mapping PWMs generated from TF PBM binding peaks by Weirauch et al. (2014) revealed significant enrichment of three DNA binding motifs within endodermal specific promoters (Figure 3.18). These include motifs for the AP2 ($Z > 3.41$), Myb-SANT ($Z = 3.02$), and B3 ($Z = 3.32$) TF family (Table 3.4) binding domains[5]. AP2 binding sequences were found to be the most abundant motifs being enriched within all 255 endodermal specific genes. The AP2 family of TFs binds to two known recognition sequences, CCGAC and CAACA. Of the 3000 plus AP2 sites found in all 255 endodermal specific promoters, 2551 of them were AP2 sites for CCGAC. The second most abundant motif enriched within endodermal specific promoters was the B3 DNA-binding-domain (DBD) motif which has a canonical consensus sequence of GCATGCA. However, the B3 motif sequences found enriched within endodermal promoters represent a non-canonical variant with the consensus sequence NCCGACANN, which closely resembles the CCGAC AP2 variant. The non-canonical B3 motif was observed 279 times within 164 of the 255 endodermal specific promoters. Lastly, binding sites for Myb/SANT domains were observed 44 times in 37 endodermal promoters. Myb/SANT TFs bind to the consensus sequence TTATC.

Interestingly, many of the *a priori* mapping motif sites were found in close proximity to

[4]While the JASPAR (Sandelin, 2004) database contains a large collection of eukaryotic DNA binding motifs, it remains rather limited in plants. The PLACE (Higo et al., 1999) database was also found to be outdated and uninformative.

[5]Note, that there are multiple PSSMs representing the same consensus sequences within the Weirauch et al. (2014) data set. Therefore, the Z-score of the least significant PSSM is reported. Since these PSSMs are representing the same consensus sequence, their significance scores vary marginally.

predicted motif patterns. For endodermal specific promoters Endo-1 and Endo-3, several DNA binding domain motifs, including AP2 and B3, were observed in close proximity with predicted motifs (mainly ESM3) in distal regions far from the TSS, possibly acting as distal control modules (Figure 3.18c). Truncated promoters designed to elucidate the involvement of putative motifs in endodermal expression had also removed groups of *a priori* mapped motifs found in conjunction with putative motifs, possibly contributing to the change in expression states. Both JASPAR (Sandelin, 2004) and PLACE (Higo et al., 1999) databases did not contain any known motifs found to be enriched in endodermal specific genes. However, with the JASPAR data base, a GATA-type zinc finger motif did contain a consensus sequence of AGATCT, very similar to the consensus sequence of AGATCGA seen in the putative GATC motifs.

A priori mapping was also performed on epidermal and cortex cell-type specific promoters. Results indicated that cell-type motif enrichment is uniquely different between promoters of endodermal, cortex, and endodermal specific promoters (Table 3.4). While endodermal promoters were dominantly enriched with AP2 sites, epidermal specific promoters were significantly enriched ($Z > 4$) with motifs associated with basic leucine zipper (bZIP) binding which recognizes a conserved ACGT sequence. Of the 175 epidermal specific gene promoters, 120 were significantly enriched with bZIP binding motifs. Promoters contained on average four bZIP motifs each, totalling 510 sites. As bZIP domains are categorized into several classes based on variations in their binding sequence, bZIP motif matches were further analyzed to assess their exact motif sequence and bZIP class. It was found that the bZIP binding sites enriched within epidermal specific promoters belonged to the G-box class, consisting of a CACGTG binding sequence. Epidermal promoters were also enriched with motif sites associated with basic helix-loop-helix binding (bHLH) ($Z = 3.56$). Seventy-three bHLH were identified in 54 epidermal specific promoters. Basic helix-loop-helix binding typically recognizes a consensus sequence of CANNTG, which has a close similarity to the CACGTG G-Box.

Cortex specific promoters were found to be predominantly enriched with binding sites for Myb/SANT domains ($Z > 3.01$) which recognize a TTATC consensus sequence. A total of 60

Myb/SANT motif sites were observed in 33 of the 74 cortex specific promoters. G-box enrichment was also observed in cortex promoters ($Z > 3.29$), although at a much lower frequency. Twenty G-box sites were observed in 14 cortex promoters. Interestingly, G-box enrichment was concentrated to the 3' end of promoters near the TSS.

3.6 Chromatin accessibility is involved in maintaining endodermal specific expression

To determine whether epigenetic modifications are involved in maintaining cell-type specific expression, endodermal specific promoter sequences were measured for levels of chromatin accessibility and CpG methylation. Nuclei from the *Arabidopsis* endodermal and epidermal cell layers were isolated *via* the INTACT (isolation of nuclei tagged in specific cell types) method described by Deal and Henikoff (2011). Cell layer nuclei isolates were digested with *DNase*I to removed regions of accessible chromatin. Extracted DNA was then sequence to determine regions of open chromatin. Chromatin accessibility data for both endodermal and epidermal cell-layers were mapped to a 2 kb sequence region flanking all ESM3 sites, covering 80% of endodermal specific promoters (Figure 3.19a). Endodermal specific promoters show an increase in accessibility within the endodermis as opposed to the epidermis, indicating that chromatin remodelling is involved in maintaining a state of TF accessibility for endodermal promoters while remaining less accessible in other cell layers. Cell-type specific CpC methylation data from Kawakatsu et al. (2016) was also used to determine differences in methylation patterns for endodermal specific promoters in the endodermis and epidermis cell layers (Figure 3.19b). No noticeable differences of promoter methylation between endodermal and epidermal cell-types were observed. These result show that, for at least the epidermal and endodermal cell layers, CpG methylation is not involved in cell-type specific expression.

Priori Mapping2.png

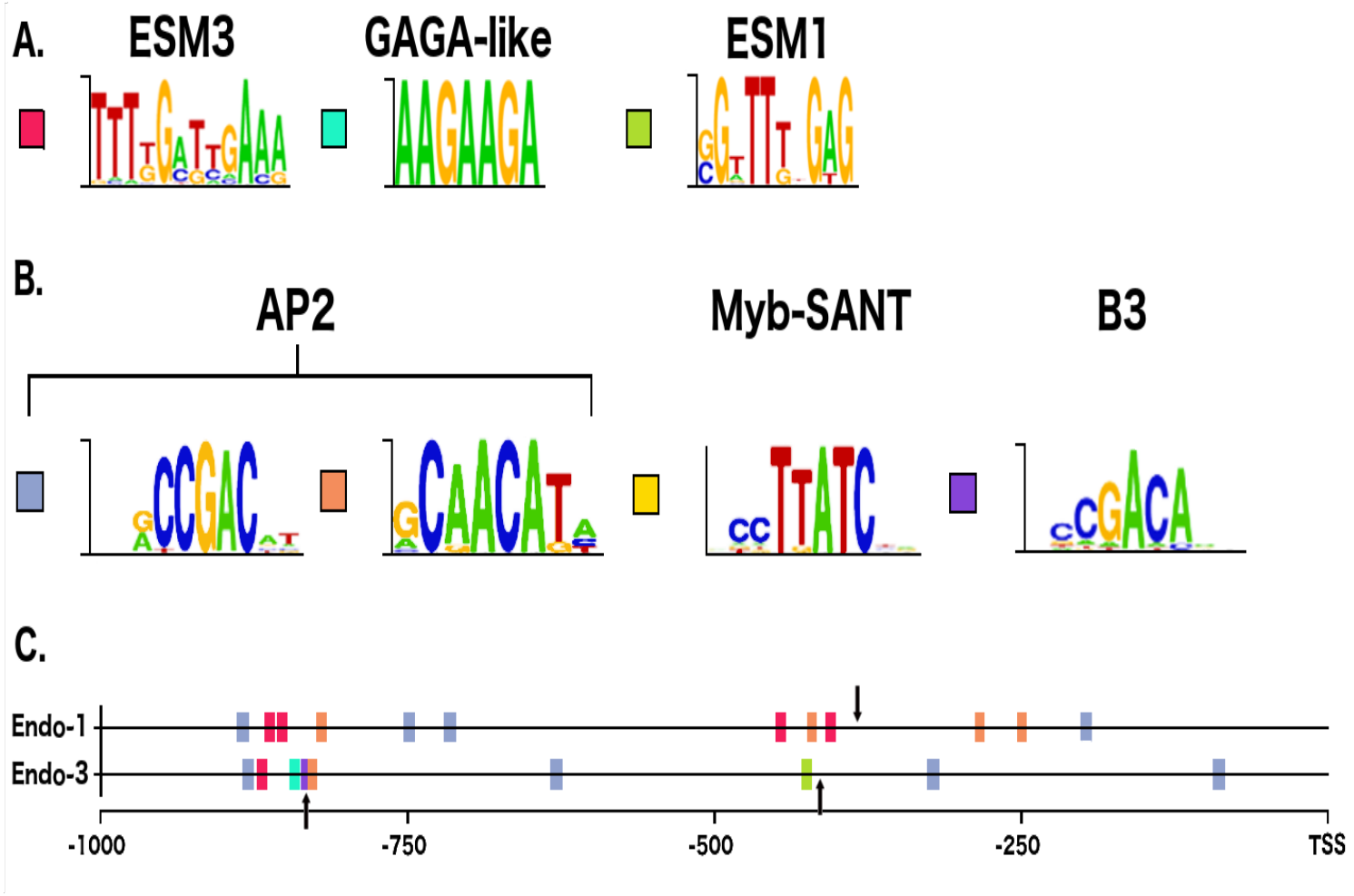


Figure 3.18: Enrichment of *Arabidopsis* DNA binding domain (DBD) motif sites in endodermal specific promoters. **(A)** Three putative motifs found in Endo-1 and Endo-3 promoters. **(B)** DBD motifs significantly enriched within endodermal specific promoters. **(C)** Motif mapping positions within the downstream 1000 bp promoter regions of Endo-1 and Endo-3. Arrows indicate positions of 5' end of truncated version of Endo-1 and Endo-3 promoters.

Table 3.4: *A priori* results of ChIP-seq determined CREs mapped to promoters of cell-type specific gene clusters for endodermis, epidermis, and cortex cell layers.

Cell layer	*Cis* regulatory element ‡	Number of motif sites in gene clusters ⋄	Gene cluster enrichment significance (Z) †
Endodermis	AP2	2551	3.41
	Myb/SANT	279	3.02
	B3	44	3.32
Epidermis	bZIP (G-box)	510	4.00
	bHLH	73	3.56
Cortex	Myb/SANT	60	3.01
	bZIP (G-box)	20	3.29

‡ PWMs of CREs are provided by Weirauch et al. (2014). ⋄ Number of motif sites mapped for AP2 and B3 CREs are the non-canonical variants with consensus sequences CCGAC and NCCGACANN respectively. † Z-scores indicate the statistical enrichment of motifs present in promoters of cell-type specific gene clusters.

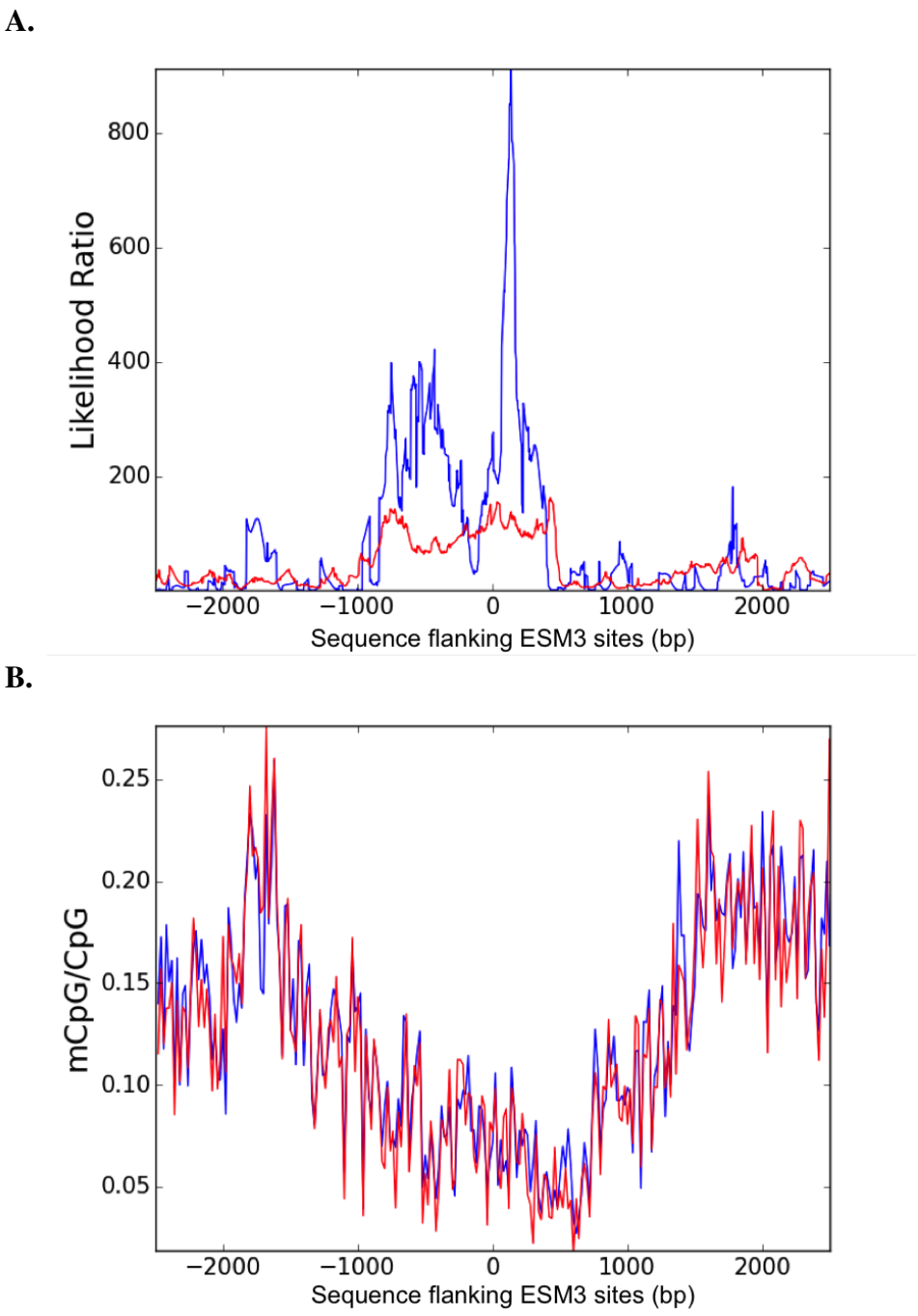


Figure 3.19: Epigenetic profiles around ESM3 motif sites within endodermal specific promoters. A. The likelihood ratio for open chromatin around ESM3 sites within endodermal specific promoters for the endodermal cell layer (blue), compared to the epidermal cell layer (red). B. Site-wise methylation frequency flanking ESM3 motifs present in endodermal specific promoters for endodermal (blue) and epidermal (red) cell layers. Methylation is measured as the fraction of methylated sequence reads to total reads mapped.

Chapter 4

Discussion

A central question in biology is understanding how gene expression can be tightly regulated on a spatiotemporal level. A prime example of this is in the development of the root, in which progenitor cells found in proximity to meristemic stem cells begin to differentiate and form the many cell-types comprising the root. Differences between cell-types can be characterized by how their genomes are regulated, with individual cell-types maintaining unique transcriptomes (Schrader, 2004; Birnbaum et al., 2003). This study identified large sets of genes whose gene expression is limited to a single cell-type. How these gene sets, which can be several hundred genes in some cell-types, maintain high expression states in one root cell-type while only being marginally expressed in nearby cell-types is unknown. This study attempts to elucidate this problem by examining the motif composition of promoter sequences for cell-type specific expressed genes in the *Arabidopsis* root, with particular emphasis on the endodermal cell layer.

Motif prediction identified six putative *de novo* motifs significantly enriched within endodermal promoters. Putative motifs were examined for their abundance within endodermal specific gene promoters and their positional patterning within promoter sequences. Twelve endodermal specific genes whose promoters were enriched with putative *de novo* motifs were tested *in planta* using a GFP reporter protein. Two promoters were capable of driving endodermal specific GFP expression, and their resulting truncations provided insight into endo-

dermal specific expression involving putative motifs. Truncations of the *ICL* (*ISOCYTRATE LYASE*, Endo-3) promoter produced ectopic GFP expression that remained localized to single cell-types. A potential explanation as to why ectopic expression remains cell-type specific is discussed, possibly through the actions of antagonistic TFs competing for the same binding sites (Sparks et al., 2017).

4.1 Cell-type specific expression is likely complex and multi-faceted

While a single obvious transcriptional mechanism governing endodermal specific expression was not apparent, two putative motifs, GAGA-like and ESM3, were found to be present in over half of all endodermal specific promoters. ESM3 was found to be present within 80% of endodermal specific promoters with significant enrichment compared to its distribution within randomly sampled promoter sets. Further, it showed pronounced positional disequilibrium in mapping positions within endodermal promoters (Figure 3.11). ESM3 positional biases were found around the -400 bp region of endodermal specific promoters. *Cis*-regulatory element positional biases have been previously observed in most eukaryotic organisms studied (Zou et al., 2011; Smith et al., 2007; Berendzen et al., 2006). Bellora et al. (2007) identified several positional biases in tissue specific CREs found in mice, most notably for liver and testis specific motifs. Taken together, these observations give strong support for ESM3 being a biologically active CRE. Furthermore, as the gene set used to predict ESM3 contained endodermal specific gene promoters, this suggests that ESM3 could be involved in regulating endodermal specific gene expression.

ESM3 may be promoting endodermal specific expression through TF-DNA interactions. However, ESM3 could also act with other CREs to form working modules (CRMs). Because of the different enrichment combinations observed with ESM3 (Figure 3.18), there could be a variety of CRMs involving EMS3 that may produce endodermal specific expression. Research

into identifying CREs involved in *Arabidopsis* stress response have found that there is no master combinatorial rule for specific stress responses. Instead, multiple possible CRMs governing small subsets of stress responsive genes have been found (Zou et al., 2011). It is possible that endodermal specific expression may function in a similar manner, relying on many different motif combinations, possibly involving ESM3. One possible motif combination was between the putative ESM3 and AP2 motifs. While biological validation of putative motifs was hindered due to GFP expression being confirmed in only two endodermal lines, motif placement within Endo-1 and Endo-3 suggests a possible CRM between ESM3 and AP2. Within both Endo-1 and Endo-3 promoters, EMS3 placements are found in close proximity to AP2 motif sites (Figure 3.18). Indeed, endodermal specific expression was disrupted when these module sites were removed during promoter truncations (Figure 3.14, 3.16).

Another interesting result was the strong stele expression observed in the 584 bp truncation of Endo-3 (Figure 3.17). The remaining 556 bp segment, which is composed of the 136 bp 5'UTR of *ICL* plus an additional 418 bps of downstream promoter sequence, contains two AP2 motifs and two EMS3 motifs. The arrangement of these motifs was more evenly spaced throughout this region, unlike the tight modular sites previously described. It is conclude from the truncations of the Endo-3 promoter, that the first 418 bp downstream of the TSS plus 5'UTR contains the necessary architecture for stele expression. It would be interesting to see if similar motif composition and arrangement is observed in stele specific promoters.

The two promoter truncations for Endo-3 each resulted in expression changes limited to a single cell-type (Figure 3.16, 3.17). This could suggest a possible *cis*-regulatory mechanism that maintains expression to individual cell-types, as opposed to ectopic expression in two or more cell-types. One possible explanation for this was described by Sparks et al. (2017) who looked at the establishment of cell-type transcriptional cascades within the *Arabidopsis* root. Their research focused on two TFs, *SHORTROOT* (*SHR*) and *SCARECROW* (*SCR*), which are required for determining endodermis and cortex cell fates. *SHR* functions at the top of this cascade and is expressed only within the stele layer of the root. They show that *SHR's*

stele specific expression is maintained by opposing transcriptional activators and repressors competing for the same binding sites. Within the stele, *SHR* activators outcompete repressors allowing expression. Within other cell-types, repressors are more abundant and function to silence *SHR* expression. Synthetic promoters designed with repressor and activator motifs from the *SHR* promoter are successfully able to mimic stele expression as well as alter expression to other individual cell-layers like the epidermis (Sparks et al., 2017). The changes observed in Endo-3 expression could be explained by a similar mechanism whereby various TFs compete for promoter binding to determine cell layer expression. Promoter truncations may have removed motif binding sites and altered the number of activators or repressors contributing towards Endo-3 cell-type expression.

An additional finding of Sparks et al. (2017) was that no single mechanism was responsible for maintaining *SHR* expression. Instead, multiple enhancer and repressor motifs contributed to confining *SHR* expression to the stele cell layer. Similar findings have been reported whereby gene regulation is determined by multiple CREs in various combinations (Zou et al., 2011). The findings in this study suggest a similar mechanism, where cell-type specific expression appears to involve multiple CREs in varying functional combinations.

4.2 Developmental stage specific genes display expression patterns reminiscent of gradient hormonal signaling

Applying hierarchal clustering to cell-type specific microarray data was able to positively identify groups of cell-type specific co-expressed genes. A finding of this analysis was that gene expression in the root is largely controlled in a manner specific to the level of development. A majority (57%) of genes included in the hierarchal clustering analysis were found to be expressed specifically in the apical meristem or zone of elongation. This pattern closely reflects what is seen in hormone controlled signaling, where gene expression cascades are controlled by hormone gradients along the root axis (Petersson et al., 2009; Sabatini et al., 1999). One

of the most well-documented examples of gradient acting phytohormones is auxin, which has been shown to be crucial for proper root development (Friml et al., 2002; Xie et al., 2000; Tian and Reed, 1999). In turn, auxin has been shown to also regulate expression of transcription factors (Li et al., 2016), some of which indirectly control large transcriptional cascades necessary for root development (Galinha et al., 2007). Similar mechanisms, involving auxin or other phytohoromes, could account for the observed expression states. Many genes with high expression levels in the apical root meristem may be driven by cascades activated through hormone signaling. In plants, stem cell niches are known for releasing signals to regulate cell division and differentiation (Van den Berg et al., 1997; Galinha et al., 2007). As cell division proceeds, newly formed daughter cells are pushed away from the meristem and exposed to lower gradients of stem cell niche derived signals (Galinha et al., 2007). In this study, the low number of stage specific genes observed in the basal meristem could reflect the transcriptional changes caused by this action. Also, the large number of stage specific genes in the zone of elongation could be explained by stem cell niche signaling being at too low of a dosage to effect cells at that distance. Because phytohormones like auxins play a crucial role in root development and induction signaling (Overvoorde et al., 2010), the large number of stage specific genes observed in this study may reflect the gradient dependent manner in which many phytohoromes act on transcriptional regulation. This was also suggested by Birnbaum et al. (2003), who took a different strategy in identifying dominant expression patterns within the root.

4.3 Chromatin remodelling may control cell-type specificity

Analysis of chromatin accessibility within endodermal specific promoters show that promoters are more open and accessible to TF binding within the endodermal cell layer as opposed to the epidermis. The differences between these two cell layers indicated that chromatin remodelling may be involved in maintaining cell-type specific expression by closing off promoters to TFs in specific cell-types. Of the 6 predicted motif signals enriched within endodermal specific pro-

moters, two of them, GAGA-like and Telobox motifs, have been implicated in recruiting chromatin remodelling proteins (Deng et al., 2013). The Telobox motif was originally identified as tandem repeats found in telomere regions of chromosomes (Richards and Ausubel, 1988). They have been classically studied for their role in protecting chromosome integrity during replication (O'Sullivan and Karlseder, 2010) and their own unique method of repeat extension involving various enzymes such as telomerase (Autexier and Lue, 2006). While comprising the main sequence repeat in telomeres, the Telobox motif is overly represented within interstitial regions of the genome (Regad et al., 1994; Stoll et al., 1993). Within the *Arabidopsis* genome, Teloboxes are observed in tandem repeats of 1-3 units in both transcribed and untranscribed sequence regions (Regad et al., 1994). The Telobox motif has been shown to form modules with other CREs regulating shoot branching (Tatematsu, 2005) and gene expression in root meristems (Manevski et al., 2000; Tremousaygue et al., 1999). These results suggest that the interstitial Telobox may act as a general regulator element involved in a variety of biological processes including cell-type identity.

More recently, the Telobox motif has been linked to chromatin remodeling (Wang et al., 2016; Zhou et al., 2015). Telobox motifs are enriched within ChIP-seq peaks analyzing DNA binding sites of *FIE* (*FERTLIZATION INDEPENDENT ENDOSPERM*), a protein component of Polycomb Repressive Complex 2 (PRC2) (Deng et al., 2013). PRC2 is essential for gene regulation by maintaining gene repression through trimethylation of histone H3 lysine 27 (H3K27me3). Intriguingly, this same study found enrichment of the GAGA motif with similar distribution patterns to the Telobox, suggesting a synergetic module incorporating the Telobox motif. GAGA-like motif enrichment was also observed, however it wasn't statistically significant among FIE binding sites, but may play a larger role with other chromatin remodelers (Deng et al., 2013). Recently, the connection between chromatin remodeling and the Telobox and GAGA motifs was confirmed when a ChIP-seq based study reported that the Telobox motif, in conjunction with the GAGA motif, was sufficient in recruiting PRC2, although residual activity in H3K27me3 suggest involvement of additional motifs (Xiao et al., unpublished).

Telobox and GAGA-like motifs are not unique in regulating endodermal specific genes, given that they have been observed regulating a variety of biological processes. However, their role in chromatin remodeling could be an important part of maintaining cell-type specific expression, where promoters must first be opened in order to interact with TFs. Chromatin remodeling has been shown to be vital for maintaining cell identity during development in *Arabidopsis* (Bratzel et al., 2010). Indeed, double mutants for PRC2 proteins *clf* and *swn* have even produced immortalized callus-like tissue lacking proper cell differentiation (Schubert et al., 2005). As such, it's possible that many endodermal specific genes, especially those involved in development, are regulated in this manner.

Besides the obvious similarity in sequence pattern, this study identified parallels between the GAGA motif and the putative GAGA-like motif enriched within endodermal specific promoters. For one, both contain similar positional patterning within gene sequences with increased enrichment flanking the TSS. This enrichment pattern has been described by Yamamoto et al. (2007) as the Y-patch, characterized as GA rich sequences flanking the TSS and is found in 21.6% of all genetic promoters in the *Arabidopsis* genome (Yamamoto et al., 2009). One hypothesis, is that the GAGA and GAGA-like motifs may be two motif variants acting as functional components of the Y-patch. Since the GAGA motif is found in a higher frequency then the GAGA-like, the GAGA-like motif could be the lesser of the two Y-patch variants. Or more simply, the consequence of the GA rich nature of the Y-patch.

4.4 Unique motif enrichment between cortex, epidermal, and endodermal specific promoters

Research on CREs has produced a collection of functionally known CRE sequences that have been catalogued in various scientific databases (Higo et al., 1999; Sandelin, 2004). With recent advancements in genomic tools such as ChIP-seq and protein binding microarrays (Stormo and Zhao, 2010), our understanding of CREs and their function are rapidly increasing. Plant

genomes are predominantly populated by TFs belonging to the Myb/SANT, B3, AP2, NAC, MADS box and WRKY families, classified by their DNA binding domains (Weirauch and Hughes, 2011). The binding sites for these TF families and others were recently determined experimentally in *Arabidopsis* (Weirauch et al., 2014) using ChIP-seq to isolate TF bound DNA. TF binding sites determined by Weirauch et al. (2014) were used in this thesis to perform an *a priori* mapping of motifs within cell-type specific promoters and identify trends in motif occurrences between cell-types.

4.4.1 TF binding motif enrichment in endodermal specific promoters

Endodermal specific promoters were found to contain significant enrichment of the AP2, B3, and Myb/SANT TF family binding sites, with AP2 accounting for the largest share of motif sites. This is partially owing to the short length of the motif and overall abundance genome wide. One of the two known binding sequences for the AP2 TF family match the CAACA consensus sequence seen in putative ESM2. The AP2 DNA binding domain consists of a basic helix-loop-helix and was first described in the *APETALA*2 gene, a TF involved in flowering morphology in *Arabidopsis* (Jofuku et al., 1994). Since then, AP2 binding domains have been observed in a number of transcription factors, many of them involved in ABA independent stress responses (Sakuma et al., 2002) and disease resistant pathways (Gutterson and Reuber, 2004). The AP2 TF family also contains a subfamily called ERFs (ethylene response factors), as the AP2 binding domain is found conserved in EREBPs (ethylene-responsive element binding proteins)(Dietz et al., 2010)[1]. AP2 enrichment could be reflected by the large number of endodermal stress-responsive genes observed, however, this category wasn't statistically overrepresented among endodermal specific promoters, so likely isn't the only reason for AP2 enrichment (Figure 3.3). Cell-type specific stress responses have been observed, with endodermal cells responding to osmotic stress more vigorously than other cell layers (Dinneny et al.,

[1]Other ethylene response elements are known, including the GCC-box (consensus AGCCGCC) which is bound by ethylene response factors involved in pathogen attack response; see Deikman (1997) and Ohme-Takagi and Shinshi (1995).

2008). Furthermore, AP2 containing TFs in maize similar to *APETALA*2 have been shown to maintain leaf epidermal cell identity (Moose and Sisco, 1996). As AP2 domains are conserved in a variety of TFs governing a wide range of biological processes, it's possible that the AP2 TF family may also be involved in maintaining endodermal specific expression and cell identity within the root.

Of the two known AP2 binding motif sequences (CAACA and CCGAC), (Weirauch et al., 2014), the vast majority of AP2 sites enriched within endodermal specific promoters were CCGAC sequences. The CCGAC sequence is a known drought responsive element (DRE), first identified in the promoters of cold response genes (COR) (Sinha et al., 2015; Yamaguchi-Shinozaki and Shinozaki, 1994; Baker et al., 1994). Indeed, of the 17 TFs identified binding to the CCGAC motif, 16 of them belonged to the dehydration response element-binding protein (DREB) family (Weirauch et al., 2014). These included 4 C-repeat binding factor (CBF) TFs of which 3 (*CBF*1, *CBF*2, and *CBF*3) are known to regulate COR genes (Medina et al., 1999; Jaglo-Ottosen, 1998; Liu et al., 1998; Stockinger et al., 1997). Both drought and cold stress responses are regulated by DREs. Plant injury from freezing has largely been revealed as consequences of freeze-induced dehydration (Steponkus et al., 1998). Enrichment of DRE motifs within endodermal specific promoters could be explained by the endodermis' role in controlling water and nutrient uptake. The endodermis, in conjunction with the Casparian strip — a lignin polymer/suberin lamellae — form a barrier preventing water flow and free diffusion of solutes taken up from the soil. In this manner, movement of water and nutrients is actively regulated by the plant. More importantly, in times of drought, the barriers formed by the endodermis prevent water diffusing from the stele to the outer root. As an overrepresentation of endodermal specific genes were found to be involved in transportation of water and nutrients (see Results, Figure 3.3), the high enrichment of DREs in endodermal specific promoters could be explained by the unique role the endodermis plays in cold/drought stress and selective transport of water and nutrients.

4.4.2 TF binding motif enrichment in epidermal specific promoters

A priori mapping of known TF binding motifs in epidermal specific promoters revealed a unique enrichment pattern of motifs different to endodermal promoters. Epidermal specific promoters are heavily enriched with binding sites for basic leucine zipper (bZIP) domains, with less enrichment for basic helix-loop-helix domain binding sites. bZIP domains bind to a core ACGT sequence (Izawa et al., 1993). Extensive expansion of the bZIP TF family in plants has resulted in preferential binding for related bZIP containing transcription factors (Corrêa et al., 2008; Izawa et al., 1993) to ACGT variants: G-box, CACGTG; C-box, GACGTC; and A-box, TACGTA. Of these three variants, epidermal specific motifs were exclusively enriched with G-boxes. Similar G-box enrichment was observed in epidermal specific genes unregulated in response to salinity stress (Dinneny et al., 2008). Basic leucine zipper binding has also been shown to regulate a wide range of plant biological functions such as cell differentiation (Silveira et al., 2007; Abe et al., 2005; Chuang et al., 1999), pathogen defense (Kaminaka et al., 2006; Pontier et al., 2001) light response (Stracke et al., 2010), osmotic control (Xu et al., 2013; Weltmeier et al., 2006), hormone and sugar signaling (Matiolli et al., 2011; Nieva et al., 2005), and protein denaturation response (Iwata and Koizumi, 2005). G-box sites could act as a common element among motif modules governing various epidermal specific functions, much like AP2 enrichment in endodermal specific promoters. Indeed, G-box sites have been shown to function in distinct modules with other CREs (Yamaguchi-Shinozaki and Shinozaki, 2005; Menkens et al., 1995), most notably in ABA responsive signaling (Shen and Ho, 1995; Shen et al., 1996).

A final observation in epidermal promoter enrichment is the lack of Myb/SANT domain binding sites. Myb/SANT containing TFs are well documented in their roles pertaining to epidermal molecular functions (Du et al., 2009). For example, *Arabidopsis* Myb/SANT containing TFs *WEREWOLF* (*WER*) (Lee and Schiefelbein, 1999) and *AtMYB23* (Matsui, 2005) both control epidermal cell differentiation. Indeed, many epidermal specific genes did contain Myb/SANT sites, however the presence of the motif wasn't statistically overrepresented among

genes with epidermal specific promoters. A possible explanation for the lack of Myb/SANT enrichment could be that Myb/SANT motifs regulate a small select group of high level TF like *WER*, and are not necessarily directly involved in maintaining cell-type specific expression.

4.4.3 TF binding motif enrichment in cortex specific promoters

Within some cortex specific promoters, G-box enrichment was observed. However, Myb/SANT motifs sites are more frequently enriched within cortex specific promoters. As one of the largest TF families in plants, Myb-SANT containing TFs, much like G-box containing TFs, are involved in a wide range of biological functions. Some of these include cell morphology (Higginson et al., 2003), meristem formation (Schmitz et al., 2002), cell cycle (Araki et al., 2004), and others (Du et al., 2009). Reasons for strong enrichment of Myb/SANT binding sites among cortex specific promoters is uncertain. As most Myb related studies focusing on root expression and development are centered around the epidermis (Kurata, 2005; Lee and Schiefelbein, 1999; Wada et al., 1997), no research currently exists connecting a possible function for Myb/SANT motif enrichment in cortex specific expression.

In conclusion, the *a priori* mapping of known CREs revealed distinct patterns of enrichment for different TF family binding sites. Endodermal specific gene promoters tend to be enriched with CREs of AP2, B3, and Myb/SANT motifs with the majority of gene promoters containing an abundance of AP2 sites. For epidermal specific gene expression, G-box motifs, and to a lesser extent bHLH motifs, are significantly enriched within promoters. Finally, for cortex specific expressing genes, few promoters were also found to be enriched with G-boxes. A greater number of cortex specific gene promoters however, contained enrichment of Myb/SANT binding sites.

Chapter 5

Conclusions and future perspectives

5.1 Cell-type *cis*-regulation in the *Arabidopsis* root

This thesis examined *Arabidopsis* cell-type specific microarray data (Birnbaum et al., 2003) to identify promoters of cell-type specific genes for epidermis, cortex, endodermis, stele, and lateral root cap cell-layers. These five main cell-layers each contain between 76 and 466 co-expressed genes each. The analysis isolating these gene clusters revealed that the majority of genes (approximately 7,245 genes analyzed by hierarchal clustering), showed developmental stage specificity. Intriguingly, the expression patterns observed in stage-specific genes is reminiscent of hormone gradient signalling. For the epidermis, cortex and endodermis cell layers, cell-type specific gene promoters were analyzed for potential CREs responsible for driving cell-type specific expression. Motif prediction and statistical significance testing was performed on the promoters of 40 cell-type specific genes from the above three cell layers. Prediction results found numerous putative motifs of varying sequence patterns enriched in cell-type specific promoters. Putative motifs enriched in endodermal promoters were extensively examined to identify CREs possibly regulating endodermal expression. Six different motif patterns were significantly enriched ($Z > 3.0$) within endodermal promoters. Two of these motif patterns, ESM3 and GAGA-like motifs, were present in over half of all endodermal specific gene pro-

moters (n=255). Both these motifs contained an interesting positional disequilibrium in motif occurrence. ESM3 motifs cluster around -400 bp downstream of the TSS, and GAGA-like enrichment gradually increases towards the TSS from both flanks. Telobox motifs were also found to be enriched within many endodermal promoters, and given their involvement in chromatin remodeling with the GAGA motif, Teloboxes could help to regulate cell-type specific expression through chromatin dynamics.

In addition to *de novo* motif prediction, the *a priori* scanning of known DNA-binding-domain (DBD) sites was applied to cell-type specific promoters of root cell layers. Some of the TF binding sites identified were also observed in *de novo* motif prediction, however most were not. Unique motif enrichments were observed for all three cell layers, with each layer being dominantly enriched with one type of known DBD site. Epidermal promoters are predominantly enriched with basic leucine zipper binding motifs (bZIP), specifically G-boxes. Cortex promoters are enriched with Myb/SANT binding sites. Finally, along with strong enrichment of predicted motifs, endodermal specific promoters typically contain multiple AP2 motif sites, often in close proximity to putative EMS3 motifs.

The final phase of this study sought to biologically validate putative motif involvement in endodermal specific expression. Two selected promoters exhibited GFP expression specific for endodermal/cortex cell layers. Promoter truncations removing putative motifs enriched in *ICL* (*ISOCITRATE LYASE*, Endo-3) resulted in cell-type specific ectopic expression in the epidermis and stele. Epidermal expression was achieved by removing distal motif sites of ESM3, GAGA-like, B3 and AP2, while stele expression was achieved by removing another upstream AP2 site and a ESM1 site. The fact that ectopic expression was confined to single cell-types indicates a possible underlying mechanism controlling gene expression to single cell layers, possibly through antagonistic TFs as observed by Sparks et al. (2016). In conclusion, cell-type specific expression within the *Arabidopsis* root is a complex process that likely involves both *cis*-regulatory motifs and other epigenetic factors to confine transcription to a single cell layer.

5.2 Study limitations

While the current study was successfully able to identify unique patterns of motif enrichment within *Arabidopsis* cell-type specific promoters, there are limitations to the study. Many of these limitations are inherent to the challenges faced in making sense of highly complex systems. While statistical enrichment and positional disequilibriums are strong indications of biological function, until their characterization *in planta*, motifs are highly putative. The mere presence of a known TF binding motif within a promoter is not enough to infer its involvement in gene regulation. Additional layers of regulatory information can determine a motif's context, like whether or not a motif is found in an accessible region of the genome. In this study, only the primary DNA sequence was used in identifying possible CREs. Cell-type specific chromatin accessibility and methylation data could be used to further identify cell-type specific regulatory mechanisms. There is also the challenge of interpreting mathematically derived results to their biological importance. For example, significance testing provides an excellent means of determining motif over-representation, but must be interpreted in respect to the biology. This study encountered several motif signals with very high significant scores. While this at first would indicate positive results, such motifs were found to be enriched in only two or three cell-type specific promoters. The rarity of their sequence patterns elsewhere in the genome however, is responsible for their high significance and does not necessarily imply functionality. Further work will be needed to improve upon existing methods of motif significance testing for more accurate motif prediction.

The use of microarray data to identify cell-type specific co-expressed gene clusters comes with its own limitations. Besides the bioinformatic challenges faced with discerning microarray background noise from biological signals, microarray chips are limited by the number of hybridization probes they can contain. This reduces the total number of genes that can be measured for expression. The root cell-type specific microarray data used in this thesis contains expression profiles for 22, 744 genes, covering only 68% of the *Arabidopsis* genome. As a consequence, many cell-type specific genes were likely not included in this study. *De novo* motif

prediction was also performed on the promoter sequences of the 40 highest correlated genes to cell-type specific baits (see Methods). Cell-type specific gene clusters for the endodermis, epidermis, and cortex however contained well over 40 cell-type specific genes. Motif prediction on promoter clusters composed of different cell-type specific promoter sequences could have generated different prediction results. Computational complexity limited the consideration of all cell-type specific promoters for motif prediction. For *a priori* motif mapping, public data bases of catalogued CREs are incomplete, meaning unknown CREs could have been missed in promoter analyses. Furthermore, mapping of either *a priori* or *de novo* predicted motifs can produce false positives or miss true positives if an appropriate functional depth (FD) is not selected. The choice of which can be problematic.

Another limitation of this study was in biologically validating putative motifs through gene expression assays in transgenic *Arabidopsis*. Out of the 12 endodermal specific genes selected to be studied, only 2 promoters were successfully confirmed for endodermal specific expression of GFP. It is possible that for many genes, regulatory motifs necessary for endodermal specific expression lie more distal from the TSS and were not included in the cloned promoter region. Perhaps longer promoter segments downstream of the TSS should have been used for cloning despite that most endodermal specific promoters used had downstream neighbouring genes within 1000 bp from their TSS. Further, running motif predictions on longer promoter sequences requires greater computational power. Another major limitation for biologically testing putative motifs was inherent in the use of promoter truncations.

Truncations had to be designed in a way where instances of one motif type/pattern could be removed by a single truncation without removing additional motif types. This way changes in gene expression could be accounted by a single CRE instead of multiple. The exception is that multiple motifs in close proximity could be removed together to test possible CRMs. Many highly expressed endodermal specific genes were not selected for biological testing as their motif enrichment was not suitable for truncation experiments. The large number of promoter truncations that had to be cloned, transformed, and confirmed for GFP expression was also very

laborious and time intensive. Recent advancements in DNA editing technologies could provide an alternative approach. CRISPR/Cas9 systems (Cong et al., 2013; Mali et al., 2013) could be used to remove or interfere with CREs without having to remove sequence downstream of the motif site. This way, only the motif in question is altered allowing researchers to better draw conclusions about a motifs effect on gene expression. Additionally, recombinant DNA methods such as golden gate cloning (Engler and Marillonnet, 2014) and multi-guide RNA containing CRISPR constructs could drastically reduce time spent functionally characterizing putative CREs.

5.3 Future directions

The approaches used to analyze the promoter structure of cell-type specific genes suggests a possible approach for designing cell-type specific synthetic promoters. The basis of this method would be to replicate the average overall CRE positional placement observed in the promoters of co-expressed genes clusters. As an example, endodermal specific promoters were found to be significantly enriched with 3 motifs that were present in over 65% of promoters. These motifs were the binding sites of the AP2 DBD and putative motifs GAGA-like and EMS3. By designing a synthetic promoter sequence enriched with these three motifs, it may be possible to emulate endodermal specific expression. Furthermore, positional biases seen in many enriched motifs could be reflected in the their choice of placement within a synthetic promoter sequence. For instance, ESM3 motifs would be placed approximately -400 bp from the TSS, as is observed in its positional disequilibrium in native endodermal specific promoters (Figure 3.11). Similarly, GAGA-like motifs could be placed in greater numbers flanking the TSS (Figure 3.11). Indeed, CREs have been observed in 5'UTRs and therefore should also be included in the design of synthetic promoters. An overall average number of motifs found among endodermal promoters would also be reflected in the design where around 8 to 10 AP2 sites would be evenly dispersed throughout the upstream 1000 bp promoter region. Synthetic

promoters would therefore reflect a similar amount of AP2 motif instances as observed in endodermal promoters.

Current technological standards for artificial DNA synthesis is capable of complete reconstructions of viral and bacterial genomes (Kosuri and Church, 2014). As such, the ability to generate a 1-3 kb size strand of DNA is very possible. The greatest hindrance to designing synthetic promoters in the above proposed way is our incomplete collection of known regulatory motifs and their functional combinations. The design of synthetic promoters could unknowingly incorporate functional motifs that cause undesired expression. As a consequence, DNA sequence between functional motifs should be designed to be as inert as possible. One potential solution would be to randomly generate sequences with a GC content matching what is observed in promoters of the biological system being transformed into. An *a priori* scanning of known CREs could then detect unwanted motifs for removal. This method still requires a comprehensive catalogue of TF binding sites further highlighting the importance of regulatory studies identifying *cis*-regulatory sites and their characterized promoter functions.

Another aspect to consider when designing synthetic promoters, is the placement of motifs regulating chromatin remodeling. To ensure that a promoter remains accessible to transcriptional machinery, motifs involved in nucleosome eviction like TREs could be used (Li et al., 2016). Unfortunately, unlike PREs, our knowledge of TrxG proteins in plants remains rudimentary, with no TREs identified to date (Pien and Grossniklaus, 2007) and only a couple of known TrxG homologues identified (Alvarez-Venegas and Avramova, 2001, Alvarez-Venegas et al. 2003).

Continued research is required in all areas of gene regulation so that economically important crops may be better engineered to serve human needs. This is more important now than ever, with challenges like climate change presenting an imposing threat to the next generation. The design of synthetic promoters that can precisely regulate specific target genes in a flexible and reliable way will likely be a significant breakthrough in achieving better genetically engineered economically important crops.

Bibliography

Abe, M., Kobayashi, Y., Yamamoto, S., Daimon, Y., Yamaguchi, A., Ikeda, Y., Ichinoki, H., Notaguchi, M., Goto, K., and Araki, T. (2005). Fd, a bzip protein mediating signals from the floral pathway integrator ft at the shoot apex. *Science*, 309(5737):1052–1056.

Aichinger, E., Villar, C. B. R., Farrona, S., Reyes, J. C., Hennig, L., and Köhler, C. (2009). CHD3 proteins and polycomb group proteins antagonistically determine cell identity in *Arabidopsis*. *PLoS Genetics*, 5(8):e1000605–12.

Aida, M., Ishida, T., Fukaki, H., Fujisawa, H., and Tasaka, M. (1997). Genes involved in organ separation in *Arabidopsis*: an analysis of the cup-shaped cotyledon mutant. *The Plant Cell*, 9(6):841–857.

Araki, S., Ito, M., Soyano, T., Nishihama, R., and Machida, Y. (2004). Mitotic cyclins stimulate the activity of c-Myb-like factors for transactivation of G2/M phase-specific genes in tobacco. *Journal of Biological Chemistry*, 279(31):32979–32988.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Ashikawa, I. (2001). Gene-associated CpG islands in plants as revealed by analyses of genomic sequences. *The Plant Journal*, 26(6):617–625.

Ashtiyani, R. K., Moghaddam, A. M. B., Schubert, V., Rutten, T., Fuchs, J., Demidov, D.,

Blattner, F. R., and Houben, A. (2011). AtHaspin phosphorylates histone H3 at threonine 3 during mitosis and contributes to embryonic patterning in *Arabidopsis*. *The Plant Journal*, 68(3):443–454.

Austin, R. S., Hiu, S., Waese, J., Ierullo, M., Pasha, A., Wang, T. T., Fan, J., Foong, C., Breit, R., Desveaux, D., Moses, A., and Provart, N. J. (2016). New BAR tools for mining expression data and exploring *cis*-elements in *Arabidopsis thaliana*. *The Plant Journal*, 88(3):490–504.

Autexier, C. and Lue, N. F. (2006). The structure and function of telomerase reverse transcriptase. *Annual Review of Biochemistry*, 75:493–517.

Babu, M. M., Iyer, L. M., Balaji, S., and Aravind, L. (2006). The natural history of the WRKY–GCM1 zinc fingers and the relationship between transcription factors and transposons. *Nucleic Acids Research*, 34(22):6505–6520.

Bailey, T. L., Bodén, M., Whitington, T., and Machanick, P. (2010). The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics*, 11(1):179.

Bailey, T. L. and Elkan, C. (1995). The value of prior knowledge in discovering motifs with meme. In *Ismb*, volume 3, pages 21–29.

Baker, S. S., Wilhelm, K. S., and Thomashow, M. F. (1994). The 5'-region of *Arabidopsis thaliana* COR15a has cis-acting elements that confer cold-, drought-and ABA-regulated gene expression. *Plant Molecular Biology*, 24(5):701–713.

Bellora, N., Farré, D., and Albà, M. M. (2007). Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters. *BMC Genomics*, 8(1):459–13.

Ben-Dor, A., Shamir, R., and Yakhini, Z. (1999). Clustering gene expression patterns. *Journal of Computational Biology*, 6(3-4):281–297.

Benfey, P. N. and Schiefelbein, J. W. (1994). Getting to the root of plant development: the genetics of *Arabidopsis* root formation. *Trends in Genetics*, 10(3):84–88.

Benhamed, M., Bertrand, C., Servet, C., and Zhou, D. X. (2006). *Arabidopsis* GCN5, HD1, and TAF1/HAF2 interact to regulate histone acetylation required for light-responsive gene expression. *The Plant Cell*, 18(11):2893–2903.

Bent, A. F. (2000). *Arabidopsis in planta* transformation. Uses, mechanisms, and prospects for transformation of other species. *Plant Physiology*, 124(4):1540–1547.

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., and Huala, E. (2015). The *Arabidopsis* information resource: making and mining the "gold standard" annotated reference plant genome. *Genesis*, 53(8):474–485.

Berendzen, K. W., Stüber, K., Harter, K., and Wanke, D. (2006). *Cis*-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves. *BMC Bioinformatics*, 7(1):522–19.

Bertani, G. (1951). Studues on lysogenesis i: The mode of phage liberation by lysogenic *Escherichia coli*1. *Journal of Bacteriology*, 62(3):293.

Bilaud, T., Koering, C. E., Binet-Brasselet, E., Ancelin, K., Pollice, A., Gasser, S. M., and Gilson, E. (1996). The telobox, a Myb-related telomeric DNA binding motif found in proteins from yeast, plants and human. *Nucleic Acids Research*, 24(7):1294–1303.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes and Development*, 16(1):6–21.

Birnbaum, K., Shasha, D. E., Wang, J. Y., Jung, J. W., Lambert, G. M., Galbraith, D. W., and Benfey, P. N. (2003). A gene expression map of the *Arabidopsis* root. *Science*, 302(5652):1956–1960.

Bloushtain-Qimron, N., Yao, J., Snyder, E. L., Shipitsin, M., Campbell, L. L., Mani, S. A., Hu, M., Chen, H., Ustyansky, V., Antosiewicz, J. E., et al. (2008). Cell type-specific DNA methylation patterns in the human breast. *Proceedings of the National Academy of Sciences*, 105(37):14076–14081.

Bonner, W., Hulett, H., Sweet, R., and Herzenberg, L. (1972). Fluorescence activated cell sorting. *Review of Scientific Instruments*, 43(3):404–409.

Bratzel, F., López-Torrejón, G., Koch, M., Del Pozo, J. C., and Calonje, M. (2010). Keeping cell identity in *Arabidopsis* requires PRC1 RING-finger homologs that catalyze H2A monoubiquitination. *Current Biology*, 20(20):1853–1859.

Bryant, Z., Subrahmanyan, L., Tworoger, M., LaTray, L., Liu, C.-R., Li, M.-J., Van Den Engh, G., and Ruohola-Baker, H. (1999). Characterization of differentially expressed genes in purified *Drosophila* follicle cells: toward a general strategy for cell type-specific developmental analysis. *Proceedings of the National Academy of Sciences*, 96(10):5559–5564.

Chen, Z. J. and Tian, L. (2007). Roles of dynamic and reversible histone acetylation in plant development and polyploidy. *Biochimica et Biophysica Acta - Gene Structure and Expression*, 1769(5-6):295–307.

Chuang, C.-F., Running, M. P., Williams, R. W., and Meyerowitz, E. M. (1999). The PERIANTHIA gene encodes a bZIP protein involved in the determination of floral organ number in *Arabidopsis thaliana*. *Genes and Development*, 13(3):334–344.

Corrado, G. and Karali, M. (2009). Inducible gene expression systems and plant biotechnology. *Biotechnology Advances*, 27(6):733–743.

Corrêa, L. G. G., Riaño-Pachón, D. M., Schrago, C. G., Vicentini dos Santos, R., Mueller-Roeber, B., and Vincentz, M. (2008). The role of bZIP transcription factors in green plant evolution: adaptive features emerging from four founder genes. *PLoS ONE*, 3(8):e2944–16.

Deal, R. B. and Henikoff, S. (2010). A simple method for gene expression and chromatin profiling of individual cell types within a tissue. *Developmental Cell*, 18(6):1030–1040.

Deal, R. B. and Henikoff, S. (2011). The INTACT method for cell type-specific gene expression and chromatin profiling in *Arabidopsis thaliana*. *Nature protocols*, 6(1):56–68.

Deikman, J. (1997). Molecular mechanisms of ethylene regulation of gene transcription. *Physiologia Plantarum*, 100(3):561–566.

Deng, W., Buzas, D. M., Ying, H., Robertson, M., Taylor, J., Peacock, W. J., Dennis, E. S., and Helliwell, C. (2013). *Arabidopsis* Polycomb Repressive Complex 2 binding sites contain putative GAGA factor binding motifs within coding regions of genes. *BMC Genomics*, 14(1):593.

Desfeux, C., Clough, S. J., and Bent, A. F. (2000). Female reproductive tissues are the primary target of *Agrobacterium*-mediated transformation by the *Arabidopsis* floral-dip method. *Plant Physiology*, 123(3):895–904.

D'haeseleer, P. (2006). What are DNA sequence motifs? *Nature biotechnology*, 24(4):423–425.

Dietz, K.-J., Vogel, M. O., and Viehhauser, A. (2010). AP2/EREBP transcription factors are part of gene regulatory networks and integrate metabolic, hormonal and environmental signals in stress acclimation and retrograde signalling. *Protoplasma*, 245(1-4):3–14.

Dinneny, J. R., Long, T. A., Wang, J. Y., Jung, J. W., Mace, D., Pointer, S., Barron, C., Brady, S. M., Schiefelbein, J., and Benfey, P. N. (2008). Cell identity mediates the response of *Arabidopsis* roots to abiotic stress. *Science*, 320(5878):942–945.

Dolan, L., Janmaat, K., Willemsen, V., Linstead, P., Poethig, S., Roberts, K., and Scheres, B. (1993). Cellular organisation of the *Arabidopsis thaliana* root. *Development*, 119(1):71–84.

Du, H., Zhang, L., Liu, L., Tang, X.-F., Yang, W.-J., Wu, Y.-M., Huang, Y.-B., and Tang, Y.-X. (2009). Biochemical and molecular characterization of plant MYB transcription factor family. *Biochemistry*, 74(1):1–11.

Eden, E., Lipson, D., Yogev, S., and Yakhini, Z. (2007). Discovering motifs in ranked lists of DNA sequences. *PLoS Computational Biology*, 3(3):e39–15.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868.

Ellis, J. G., Kerr, A., Tempé, J., and Petit, A. (1979). Arginine catabolism: a new function of both octopine and nopaline ti-plasmids of *Agrobacterium*. *Molecular and General Genetics*, 173(3):263–269.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49.

Esau, K. (1977). *Anatomy of Seed Plants. 2'ed*. New York.

Fleury, D., Himanen, K., Cnops, G., Nelissen, H., Boccardi, T. M., Maere, S., Beemster, G. T. S., Neyt, P., Anami, S., Robles, P., Micol, J. L., Inze, D., and Van Lijsebettens, M. (2007). The *Arabidopsis thaliana* homolog of yeast BRE1 has a function in cell cycle regulation during early leaf and root growth. *The Plant Cell*, 19(2):417–432.

Friml, J., Benková, E., Blilou, I., Wisniewska, J., and Hamann, T. (2002). AtPIN4 mediates sink-driven auxin gradients and root patterning in *Arabidopsis*. *Cell Reports*, 108(5):661–673.

Galinha, C., Hofhuis, H., Luijten, M., Willemsen, V., Blilou, I., Heidstra, R., and Scheres, B. (2007). PLETHORA proteins as dose-dependent master regulators of *Arabidopsis* root development. *Nature*, 449(7165):1053–1057.

Gardiner-Garden, M. and Frommer, M. (1987). CpG islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2):261–282.

Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. (2006). *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer Science & Business Media.

Goldberg, A. D., Allis, C. D., and Bernstein, E. (2007). Epigenetics: a landscape takes shape. *Cell*, 128(4):635–638.

Goll, M. G. and Bestor, T. H. (2005). Eukaryotic cytosine methyltransferases. *Annual Review of Biochemistry*, 74:481–514.

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 *Phaseolin* promoter. *Journal of Biological Chemistry*, 279(9):8102–8110.

Grau, J., Posch, S., Grosse, I., and Keilwagen, J. (2013). A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Research*, 41(21):e197–e197.

Gray, M. E., Sappington, T. W., Miller, N. J., Moeser, J., and Bohn, M. O. (2009). Adaptation and invasiveness of Western Corn Rootworm: Intensifying research on a worsening pest. *Annual Review of Entomology*, 54(1):303–321.

Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences*, 84(13):4355–4358.

Gruenbaum, Y., Naveh-Many, T., Cedar, H., and Razin, A. (1981). Sequence specificity of methylation in higher plant DNA. *Nature*, 292(5826):860–862.

Gutterson, N. and Reuber, T. L. (2004). Regulation of disease resistance pathways by AP2/ERF transcription factors. *Current Opinion in Plant Biology*, 7(4):465–471.

Guyomarc'h, S., Benhamed, M., Lemonnier, G., Renou, J.-P., Zhou, D.-X., and Delarue, M. (2006). MGOUN3: evidence for chromatin-mediated regulation of FLC expression. *Journal of Experimental Botany*, 57(9):2111–2119.

Harmer, S. L., Hogenesch, J. B., Straume, M., Chang, H.-S., Han, B., Zhu, T., Wang, X., Kreps, J. A., and Kay, S. A. (2000). Orchestrated transcription of key pathways in *Arabidopsis* by the circadian clock. *Science*, 290(5499):2110–2113.

He, Y., Michaels, S. D., and Amasino, R. M. (2003). Regulation of flowering time by histone acetylation in *Arabidopsis*. *Science*, 302(5651):1751–1754.

Hertz, G. Z. and Stormo, G. D. (1999). Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7):563–577.

Higginson, T., Li, S. F., and Parish, R. W. (2003). AtMYB103 regulates tapetum and trichome development in *Arabidopsis thaliana*. *The Plant Journal*, 35(2):177–192.

Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. (1999). Plant *cis*-acting regulatory dna elements (place) database: 1999. *Nucleic Acids Research*, 27(1):297–300.

Hoheisel, J. D. (2006). Microarray technology: beyond transcript profiling and genotype analysis. *Nature Reviews Microbiology*, 7(3):200–210.

Holsters, M., Waele, D. D., Depicker, A., Messens, E., Montagu, M. V., and Schell, J. (1978). Transfection and transformation of *Agrobacterium tumefaciens*. *Molecular and General Genetics*, 163(2):181–187.

Horard, B., Tatout, C., Poux, S., and Pirrotta, V. (2000). Structure of a polycomb response element and in vitro binding of polycomb group complexes containing gaga factor. *Molecular and Cellular Biology*, 20(9):3187–3197.

Houben, A., Demidov, D., Caperta, A. D., Karimi, R., Agueci, F., and Vlasenko, L. (2007). Phosphorylation of histone H3 in plants—A dynamic affair. *Biochimica et Biophysica Acta - Gene Structure and Expression*, 1769(5-6):308–315.

Hughes, J. D., Estep, P. W., Tavazoie, S., and Church, G. M. (2000). Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology*, 296(5):1205–1214.

Ingelbrecht, I., Van Houdt, H., Van Montagu, M., and Depicker, A. (1994). Posttranscriptional silencing of reporter transgenes in tobacco correlates with DNA methylation. *Proceedings of the National Academy of Sciences*, 91(22):10502–10506.

Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J. B., Lonnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, 21(7):1160–1167.

Iwata, Y. and Koizumi, N. (2005). An *Arabidopsis* transcription factor, AtbZIP60, regulates the endoplasmic reticulum stress response in a manner unique to plants. *Proceedings of the National Academy of Sciences*, 102(14):5280–5285.

Izawa, T., Foster, R., and Chua, N. H. (1993). Plant bZIP protein DNA binding specificity. *Journal of Molecular Biology*, 230(4):1131–1144.

Jaglo-Ottosen, K. R. (1998). *Arabidopsis* CBF1 overexpression induces COR genes and enhances freezing tolerance. *Science*, 280(5360):104–106.

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A.,

Cohen, N., Jung, S., Tanay, A., and Amit, I. (2014). Massively parallel single-cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779.

Jin, J., Zhang, H., Kong, L., Gao, G., and Luo, J. (2014). Planttfdb 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Research*, 42(D1):D1182–D1187.

Jofuku, K. D., Den Boer, B. G., Van Montagu, M., and Okamuro, J. K. (1994). Control of *Arabidopsis* flower and seed development by the homeotic gene APETALA2. *The Plant Cell*, 6(9):1211–1225.

Kaminaka, H., Näke, C., Epple, P., Dittgen, J., Schütze, K., Chaban, C., Holt, B. F., Merkle, T., Schäfer, E., Harter, K., et al. (2006). bZIP10-LSD1 antagonism modulates basal defense and cell death in *Arabidopsis* following infection. *The EMBO Journal*, 25(18):4400–4411.

Kawakatsu, T., Stuart, T., Valdes, M., Breakfield, N., Schmitz, R. J., Nery, J. R., Urich, M. A., Han, X., Lister, R., Benfey, P. N., et al. (2016). Unique cell-type-specific patterns of dna methylation in the root meristem. *Nature Plants*, 2:16058.

Krichevsky, A., Zaltsman, A., Kozlovsky, S. V., Tian, G.-W., and Citovsky, V. (2009). Regulation of root elongation by histone acetylation in *Arabidopsis*. *Journal of Molecular Biology*, 385(1):45–50.

Kurata, T. (2005). Cell-to-cell movement of the CAPRICE protein in *Arabidopsis* root epidermal cell differentiation. *Development*, 132(24):5387–5398.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., Wootton, J. C., et al. (1993). Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science-New York then Washington*, 262:208–208.

Lawrence, C. E. and Reilly, A. A. (1990). An expectation maximization (EM) algorithm for

the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Structure, Function, and Bioinformatics*, 7(1):41–51.

Lee, M. M. and Schiefelbein, J. (1999). WEREWOLF, a MYB-related protein in *Arabidopsis*, is a position-dependent regulator of epidermal cell patterning. *Cell Reports*, 99(5):473–483.

Lenka, S. K., Lohia, B., Kumar, A., Chinnusamy, V., and Bansal, K. C. (2008). Genome-wide targeted prediction of ABA responsive genes in rice based on over-represented cis-motif in co-expressed genes. *Plant molecular biology*, 69(3):261–271.

Li, S.-B., Xie, Z.-Z., Hu, C.-G., and Zhang, J.-Z. (2016). A review of auxin response factors (ARFs) in plants. *Frontiers in Plant Science*, 7(742):137–7.

Liu, C., Lu, F., Cui, X., and Cao, X. (2010). Histone methylation in higher plants. *Annual Review of Plant Biology*, 61(1):395–420.

Liu, Q., Kasuga, M., Sakuma, Y., Abe, H., Miura, S., Yamaguchi-Shinozaki, K., and Shinozaki, K. (1998). Two transcription factors, DREB1 and DREB2, with an EREBP/AP2 DNA binding domain separate two cellular signal transduction pathways in drought- and low-temperature-responsive gene expression, respectively, in *Arabidopsis*. *The Plant Cell*, 10(8):1391–1406.

Liu, X., Brutlag, D. L., Liu, J. S., et al. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In *Pacific symposium on biocomputing*, volume 6, pages 127–138.

Liu, Y., Koornneef, M., and Soppe, W. J. J. (2007). The absence of histone H2B monoubiquitination in the *Arabidopsis* hub1 (rdo4) mutant reveals a role for chromatin remodeling in seed dormancy. *The Plant Cell*, 19(2):433–444.

Manevski, A., Bertoni, G., Bardet, C., Tremousaygue, D., and Lescure, B. (2000). In synergy

with various cis-acting elements, plant insterstitial telomere motifs regulate gene expression in *Arabidopsis* root meristems. *FEBS Letters*, 483(1):43–46.

Matiolli, C. C., Tomaz, J. P., Duarte, G. T., Prado, F. M., Del Bem, L. E. V., Silveira, A. B., Gauer, L., Correa, L. G. G., Drumond, R. D., Viana, A. J. C., Di Mascio, P., Meyer, C., and Vincentz, M. (2011). The *Arabidopsis* bZIP gene AtbZIP63 is a sensitive integrator of transient abscisic acid and glucose signals. *Plant Physiology*, 157(2):692–705.

Matsui, K. (2005). A chimeric AtMYB23 repressor induces hairy roots, elongation of leaves and stems, and inhibition of the deposition of mucilage on seed coats in *Arabidopsis*. *Plant and Cell Physiology*, 46(1):147–155.

Matys, V. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34(90001):D108–D110.

Medina, J., Bargues, M., Terol, J., Pérez-Alonso, M., and Salinas, J. (1999). The *Arabidopsis CBF* gene family is composed of three genes encoding AP2 domain-containing proteins whose expression is regulated by low temperature but not by abscisic acid or dehydration. *Plant Physiology*, 119(2):463–470.

Menkens, A. E., Schindler, U., and Cashmore, A. R. (1995). The G-box: a ubiquitous regulatory DNA element in plants bound by the GBF family of bZIP proteins. *Trends in Biochemical Sciences*, 20(12):506–510.

Meyer, P., Niedenhof, I., and ten Lohuis, M. (1994). Evidence for cytosine methylation of non-symmetrical sequences in transgenic *Petunia hybrida*. *The EMBO Journal*.

Mitchell, P. D., Gray, M. E., and Steffey, K. L. (2004). A composed-error model for estimating pest-damage functions and the impact of the western corn rootworm soybean variant in illinois. *American Journal of Agricultural Economics*, 86(2):332–344.

Moose, S. P. and Sisco, P. H. (1996). Glossy15, an APETALA2-like gene from maize that regulates leaf epidermal cell identity. *Genes and Development*, 10(23):3018–3027.

Murashige, T. and Skoog, F. (1962). A revised medium for rapid growth and bio assays with tobacco tissue cultures. *Physiologia Plantarum*, 15(3):473–497.

Naseer, S., Lee, Y., Lapierre, C., Franke, R., Nawrath, C., and Geldner, N. (2012). Casparian strip diffusion barrier in *Arabidopsis* is made of a lignin polymer without suberin. *Proceedings of the National Academy of Sciences*, 109(25):10101–10106.

Nieva, C., Busk, P. K., Domínguez-Puigjaner, E., Lumbreras, V., Testillano, P. S., Risueño, M.-C., and Pagès, M. (2005). Isolation and functional characterisation of two new bZIP maize regulators of the ABA responsive gene rab28. *Plant Molecular Biology*, 58(6):899–914.

Ohme-Takagi, M. and Shinshi, H. (1995). Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element. *The Plant Cell*, 7(2):173–182.

O'Sullivan, R. J. and Karlseder, J. (2010). Telomeres: protecting chromosomes against genome instability. *Nature Reviews Molecular Cell Biology*, 11(3):171–181.

Overvoorde, P., Fukaki, H., and Beeckman, T. (2010). Auxin control of root development. *Cold Spring Harbor Perspectives in Biology*, 2(6):a001537–a001537.

Palaniswamy, S. K., James, S., Sun, H., Lamb, R. S., Davuluri, R. V., and Grotewold, E. (2006). Agris and atregnet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiology*, 140(3):818–829.

Patel, R. Y. and Stormo, G. D. (2014). Discriminative motif optimization based on perceptron training. *Bioinformatics*, 30(7):941–948.

Pavesi, G., Mauri, G., and Pesole, G. (2001). An algorithm for finding signals of unknown length in dna sequences. *Bioinformatics*, 17(1):S207–S214.

Péret, B., De Rybel, B., Casimiro, I., Benková, E., Swarup, R., Laplaze, L., Beeckman, T., and Bennett, M. J. (2009). *Arabidopsis* lateral root development: an emerging story. *Trends in plant science*, 14(7):399–408.

Pessiot, J.-F., Kim, Y.-M., Amini, M. R., and Gallinari, P. (2010). Improving document clustering in a learned concept space. *Information Processing and Management*, 46(2):180–192.

Petersson, S. V., Johansson, A. I., Kowalczyk, M., Makoveychuk, A., Wang, J. Y., Moritz, T., Grebe, M., Benfey, P. N., Sandberg, G., and Ljung, K. (2009). An auxin gradient and maximum in the *Arabidopsis* root apex shown by high-resolution cell-specific analysis of IAA distribution and synthesis. *The Plant Cell*, 21(6):1659–1668.

Pfluger, J. and Wagner, D. (2007). Histone modifications and dynamic regulation of genome accessibility in plants. *Current Opinion in Plant Biology*, 10(6):645–652.

Pien, S. and Grossniklaus, U. (2007). Polycomb group and trithorax group proteins in *Arabidopsis*. *Biochimica et Biophysica Acta - Gene Structure and Expression*, 1769(5-6):375–382.

Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 21(3):447–455.

Pontier, D., Miao, Z.-H., and Lam, E. (2001). Trans-dominant suppression of plant TGA factors reveals their negative and positive roles in plant defense responses. *The Plant Journal*, 27(6):529–538.

Redhead, E. and Bailey, T. L. (2007). Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC Bioinformatics*, 8(1):385–19.

Regad, F., Lebas, M., and Lescure, B. (1994). Interstitial telomeric repeats within the *Arabidopsis thaliana* genome. *Journal of molecular biology*, 239(2):163–169.

Richards, E. J. and Ausubel, F. M. (1988). Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*. *Cell Reports*, 53(1):127–136.

Rombauts, S. (2003). Computational approaches to Identify promoters and cis-regulatory elements in plant genomes. *Plant Physiology*, 132(3):1162–1176.

Rousseeuw, P. J. and Kaufman, L. (1990). *Finding Groups in Data*. Wiley Online Library.

Russell, R. S. et al. (1977). *Plant root systems: their function and interaction with the soil.* McGraw-Hill Book Company (UK) Limited.

Sabatini, S., Beis, D., Wolkenfelt, H., Murfett, J., and Guilfoyle, T. (1999). An auxin-dependent distal organizer of pattern and polarity in the *Arabidopsis* root. *Cell Reports*, 99(5):463–472.

Sakuma, Y., Liu, Q., Dubouzet, J. G., Abe, H., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2002). DNA-binding specificity of the ERF/AP2 domain of *Arabidopsis* DREBs, transcription factors involved in dehydration- and cold-inducible gene expression. *Biochemical and Biophysical Research Communications*, 290(3):998–1009.

Sandelin, A. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(90001):91–94.

Schmitz, G., Tillmann, E., Carriero, F., Fiore, C., Cellini, F., and Theres, K. (2002). The tomato Blind gene encodes a MYB transcription factor that controls the formation of lateral meristems. *Proceedings of the National Academy of Sciences*, 99(2):1064–1069.

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100.

Schneider, T. D., Stormo, G. D., and Gold, L. (1986). Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology*, 188(3):415–431.

Schones, D. E., Smith, A. D., and Zhang, M. Q. (2007). Statistical significance of *cis*-regulatory modules. *BMC Bioinformatics*, 8(1):19–11.

Schrader, J. (2004). A high-resolution transcript profile across the wood-forming meristem of poplar identifies potential regulators of cambial stem cell identity. *The Plant Cell*, 16(9):2278–2292.

Schubert, D., Clarenz, O., and Goodrich, J. (2005). Epigenetic control of plant development by polycomb-group proteins. *Current Opinion in Plant Biology*, 8(5):553–561.

Seki, M., Shigemoto, N., Komeda, Y., Imamura, J., and Yamada, Y. (1991). Transgenic *Arabidopsis thaliana* plants obtained by particle-bombardment-mediated transformation. *Applied Microbiology and Biotechnology*, 36(2):228–230.

Senger, K., Armstrong, G. W., Rowell, W. J., and Kwan, J. M. (2004). Immunity regulatory DNAs share common organizational features in *Drosophila*. *Molecular cell*, 13(1):19–32.

Sharma, A. et al. (2015). *In silico* identification of regulatory motifs in co-expressed genes under osmotic stress representing their co-regulation. *Plant Gene*, 1:29–34.

Shen, Q. and Ho, T. H. (1995). Functional dissection of an abscisic acid (ABA)-inducible gene reveals two independent ABA-responsive complexes each containing a G-box and a novel *cis*-acting element. *The Plant Cell*, 7(3):295–307.

Shen, Q., Zhang, P., and Ho, T. H. (1996). Modular nature of abscisic acid (ABA) response complexes: composite promoter units that are necessary and sufficient for ABA induction of gene expression in barley. *The Plant Cell*, 8(7):1107–1119.

Shen-Orr, S. S., Tibshirani, R., Khatri, P., Bodian, D. L., Staedtler, F., Perry, N. M., Hastie, T., Sarwal, M. M., Davis, M. M., and Butte, A. J. (2010). Cell type–specific gene expression differences in complex tissues. *Nature Publishing Group*, 7(4):287–289.

Shinozaki, K. and Yamaguchi-Shinozaki, K. (2000). Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways. *Current Opinion in Plant Biology*, 3(3):217–223.

Silveira, A. B., Gauer, L., Tomaz, J. P., Cardoso, P. R., Carmello-Guerreiro, S., and Vincentz, M. (2007). The *Arabidopsis* AtbZIP9 protein fused to the VP16 transcriptional activation domain alters leaf and vascular development. *Plant Science*, 172(6):1148–1156.

Sinha, S. (2006). On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics*, 22(14):e454–e463.

Sinha, S., Kukreja, B., Arora, P., Sharma, M., Pandey, G. K., Agarwal, M., and Chinnusamy, V. (2015). The omics of cold stress responses in plants. In *Elucidation of Abiotic Stress Signaling in Plants*, pages 143–194. Springer.

Sinha, S. and Tompa, M. (2000). A statistical method for finding transcription factor binding sites. In *ISMB*, volume 8, pages 344–354.

Smith, B., Fang, H., Pan, Y., Walker, P. R., Famili, A. F., and Sikorska, M. (2007). Evolution of motif variants and positional bias of the cyclic-AMP response element. *BMC Evolutionary Biology*, 7(Suppl 1):S15–9.

Smith, S. and De Smet, I. (2012). Root system architecture: insights from *Arabidopsis* and cereal crops. *Philosophical Transactions of the Royal Society*, 367:1441–1452.

Sparks, E. E., Drapek, C., Gaudinier, A., Li, S., Ansariola, M., Shen, N., Hennacy, J. H., Zhang, J., Turco, G., Petricka, J. J., Foret, J., Hartemink, A. J., Gordân, R., Megraw, M., Brady, S. M., and Benfey, P. N. (2017). Establishment of expression in the SHORTROOT-SCARECROW transcriptional cascade through opposing activities of both activators and repressors. *Developmental Cell*, pages 1–13.

Spek, C. A., Betina, R. M., and Reitsma, P. H. (1999). Unique distance-and DNA-turn-dependent interactions in the human protein C gene promoter confer submaximal transcriptional activity. *Biochemical Journal*, 340(2):513–518.

Sridhar, V. V., Kapoor, A., Zhang, K., Zhu, J., Zhou, T., Hasegawa, P. M., Bressan, R. A., and Zhu, J.-K. (2007). Control of DNA methylation and heterochromatic silencing by histone H2B deubiquitination. *Nature*, 447(7145):735–738.

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research*, pages 1–15.

Steponkus, P. L., Uemura, M., Joseph, R. A., Gilmour, S. J., and Thomashow, M. F. (1998). Mode of action of the COR15a gene on the freezing tolerance of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*, 95(24):14570–14575.

Stockinger, E. J., Gilmour, S. J., and Thomashow, M. F. (1997). *Arabidopsis thaliana* CBF1 encodes an AP2 domain-containing transcriptional activator that binds to the C-repeat/DRE, a cis-acting DNA regulatory element that stimulates transcription in response to low temperature andwaterdeficit. *Proceedings of the National Academy of Sciences*, 94(3):1035–1040.

Stoll, S., Zirlik, T., Maercker, C., and Lipps, H. J. (1993). The organization of internal telomeric repeats in the polytene chromosomes of the hypotrichous ciliate *Stylonychia lemnae*. *Nucleic Acids Research*, 21(8):1783–1788.

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23.

Stormo, G. D. and Fields, D. S. (1998). Specificity, free energy and information content in protein–DNA interactions. *Trends in Biochemical Sciences*, 23(3):109–113.

Stormo, G. D. and Hartzell, G. W. (1989). Identifying protein-binding sites from unaligned dna fragments. *Proceedings of the National Academy of Sciences*, 86(4):1183–1187.

Stormo, G. D., Schneider, T. D., and Gold, L. M. (1982). Characterization of translational initiation sites in *E. coli*. *Nucleic Acids Research*, 10(9):2971–2996.

Stormo, G. D. and Zhao, Y. (2010). Determining the specificity of protein–DNA interactions. *Nature Reviews Genetics*, 11(11):751–760.

Stracke, R., Favory, J.-J., Gruber, H., Bartelniewoehner, L., Bartels, S., Binkert, M., Funk, M., Weisshaar, B., and Ulm, R. (2010). The *Arabidopsis* bzip transcription factor hy5 regulates expression of the pfg1/myb12 gene in response to light and ultraviolet-b radiation. *Plant, Cell & Environment*, 33(1):88–103.

Sugiyama, T., Scott, D. K., Wang, J.-C., and Granner, D. K. (1998). Structural requirements of the glucocorticoid and retinoic acid response units in the phosphoenolpyruvate carboxykinase gene promoter. *Molecular Endocrinology*, 12(10):1487–1498.

Taiz, L. and Zeiger, E. (2010). *Plant physiology 5th Ed*. Sunderland, MA: Sinauer Associates.

Takada, S., Hibara, K.-i., Ishida, T., and Tasaka, M. (2001). The cup-shaped cotyledon1 gene of *Arabidopsis* regulates shoot apical meristem formation. *Development*, 128(7):1127–1135.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, 96(6):2907–2912.

Tatematsu, K. (2005). Identification of *cis*-elements that regulate gene expression during initiation of axillary bud outgrowth in *Arabidopsis*. *Plant Physiology*, 138(2):757–766.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature genetics*, 22(3):281–285.

Thayyil, K. N. (2012). *GMO's in Europe: Law, technology and public contestations*. PhD thesis, Tilburg University.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y.

(2001). A higher-order background model improves the detection of promoter regulatory elements by gibbs sampling. *Bioinformatics*, 17(12):1113–1122.

Thomas, J., Touchman, J., Blakesley, R., Bouffard, G., Beckstrom-Sternberg, S., Margulies, E., Blanchette, M., Siepel, A., Thomas, P., McDowell, J., et al. (2003). Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424(6950):788–793.

Thomashow, M. F., Nutter, R., Montoya, A. L., and Gordon, M. P. (1980). Integration and organization of Ti plasmid sequences in crown gall tumors. *Cell Reports*, 19(3):729–739.

Tian, Q. and Reed, J. W. (1999). Control of auxin-regulated root development by the *Arabidopsis thaliana* SHY2/IAA3 gene. *Development*, 126(4):711–721.

Tremousaygue, D., Manevski, A., Bardet, C., Lescure, N., and Lescure, B. (1999). Plant interstitial telomere motifs participate in the control of gene expression in root meristems. *The Plant Journal*, 20(5):553–561.

Ülker, B. and Somssich, I. E. (2004). WRKY transcription factors: from DNA binding towards biological function. *Current Opinion in Plant Biology*, 7(5):491–498.

Van den Berg, C., Willemsen, V., Hendriks, G., and Weisbeek, P. (1997). Short-range control of cell differentiation in the *Arabidopsis* root meristem. *Nature*, 390(6657):287–289.

Van der Laan, M., Pollard, K., and Bryan, J. (2003). A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73(8):575–584.

van der Laan, M. and Pollard, K. S. (2003). A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *Journal of Statistical Planning and Inference*, 117(2):275–303.

Van Larebeke, N., Engler, G., Holsters, M., Van den Elsacker, S., Zaenen, I., Schilperoort, R., and Schell, J. (1974). Large plasmid in *Agrobacterium tumefaciens* essential for crown gall-inducing ability. *Nature*, 252(5479):169–170.

Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L., and Van de Peer, Y. (2009). Unraveling transcriptional control in *Arabidopsis* using *cis*-regulatory elements and coexpression networks. *Plant Physiology*, 150(2):535–546.

Vardhanabhuti, S., Wang, J., and Hannenhalli, S. (2007). Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Research*, 35(10):3203–3213.

Verbelen, J.-P., Cnodder, T. D., Le, J., Vissenberg, K., and Baluška, F. (2014). The root apex of *Arabidopsis thaliana* consists of four distinct zones of growth activities. *Plant Signaling & Behavior*, 1(6):296–304.

Wada, T., Tachibana, T., Shimura, Y., and Okada, K. (1997). Epidermal cell differentiation in *Arabidopsis* determined by a Myb homolog, CPC. *Science*, 277(5329):1113–1116.

Wang, H., Liu, C., Cheng, J., Liu, J., Zhang, L., He, C., Shen, W.-H., Jin, H., Xu, L., and Zhang, Y. (2016). *Arabidopsis* flower and embryo developmental genes are repressed in seedlings by different combinations of polycomb group proteins in association with distinct sets of *cis*-regulatory elements. *PLoS Genetics*, 12(1):e1005771.

Wasserman, W. W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287.

Weirauch, M. T. and Hughes, T. R. (2011). A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. In *Introduction to "a handbook of transcription factors"*, pages 25–73. Springer Netherlands, Dordrecht.

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J. M., Bouget, F.-Y., Ratsch, G., Larrondo,

L. F., Ecker, J. R., and Hughes, T. R. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell Reports*, 158(6):1431–1443.

Wellmer, F. and Riechmann, J. L. (2005). Gene network analysis in plant development by genomic technologies. *The International Journal of Developmental Biology*, 49(5-6):745–759.

Weltmeier, F., Ehlert, A., Mayer, C. S., Dietrich, K., Wang, X., Schütze, K., Alonso, R., Harter, K., Vicente-Carbajosa, J., and Dröge-Laser, W. (2006). Combinatorial control of *Arabidopsis* proline dehydrogenase transcription by specific heterodimerisation of bzip transcription factors. *The EMBO Journal*, 25(13):3133–3143.

Wenick, A. S. and Hobert, O. (2004). Genomic *cis*-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in *C. elegans*. *Developmental Cell*, 6(6):757–770.

Winter, C. M., Austin, R. S., Blanvillain-Baufumé, S., Reback, M. A., Monniaux, M., Wu, M.-F., Sang, Y., Yamaguchi, A., Yamaguchi, N., Parker, J. E., Parcy, F., Jensen, S. T., Li, H., and Wagner, D. (2011). LEAFY target genes reveal floral regulatory logic, *cis* motifs, and a link to biotic stimulus response. *Developmental Cell*, 20(4):430–443.

Wolfe, C. J., Kohane, I. S., and Butte, A. J. (2005). Systematic survey reveals general applicability of guilt-by-association within gene coexpression networks. *BMC Bioinformatics*, 6(1):227–10.

Wu, L. and Berk, A. (1988). Constraints on spacing between transcription factor binding sites in a simple adenovirus promoter. *Genes and Development*, 2(4):403–411.

Xie, Q., Frugis, G., Colgan, D., and Chua, N.-H. (2000). *Arabidopsis* NAC1 transduces auxin signal downstream of TIR1 to promote lateral root development. *Genes & Development*, 14(23):3024–3036.

Xu, C. R., Liu, C., Wang, Y. L., Li, L. C., Chen, W. Q., Xu, Z. H., and Bai, S. N. (2005). Histone acetylation affects expression of cellular patterning genes in the *Arabidopsis* root epidermis. *Proceedings of the National Academy of Sciences*, 102(40):14469–14474.

Xu, Z. Y., Kim, S. Y., Hyeon, D. Y., Kim, D. H., Dong, T., Park, Y., Jin, J. B., Joo, S. H., Kim, S. K., Hong, J. C., Hwang, D., and Hwang, I. (2013). The *Arabidopsis* NAC transcription factor ANAC096 cooperates with bZIP-type transcription factors in dehydration and osmotic stress responses. *The Plant Cell*, 25(11):4708–4724.

Yamaguchi-Shinozaki, K. and Shinozaki, K. (1994). A novel *cis*-acting element in an *Arabidopsis* gene is involved in responsiveness to drought, low-temperature, or high-salt stress. *The Plant Cell*, 6(2):251–264.

Yamaguchi-Shinozaki, K. and Shinozaki, K. (2005). Organization of *cis*-acting regulatory elements in osmotic- and cold-stress-responsive promoters. *Trends in Plant Science*, 10(2):88–94.

Yamamoto, Y. Y., Ichida, H., Matsui, M., Obokata, J., Sakurai, T., Satou, M., Seki, M., Shinozaki, K., and Abe, T. (2007). Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics*, 8(1):67–23.

Yamamoto, Y. Y., Yoshitsugu, T., Sakurai, T., Seki, M., Shinozaki, K., and Obokata, J. (2009). Heterogeneity of *Arabidopsis* core promoters revealed by high-density TSS analysis. *The Plant Journal*, 60(2):350–362.

Yamasaki, K., Kigawa, T., Inoue, M., Watanabe, S., Tateno, M., Seki, M., Shinozaki, K., and Yokoyama, S. (2008). Structures and evolutionary origins of plant-specific transcription factor DNA-binding domains. *Plant Physiology and Biochemistry*, 46(3):394–401.

Yao, Z., MacQuarrie, K. L., Fong, A. P., and Tapscott, S. J. (2014). Discriminative motif analysis of high-throughput dataset. *Bioinformatics*, 30(6):775–783.

Zhang, X., Henriques, R., Lin, S.-S., Niu, Q.-W., and Chua, N.-H. (2006). *Agrobacterium*-mediated transformation of *Arabidopsis thaliana* using the floral dip method. *Nature protocols*, 1(2):641–646.

Zhou, C., Zhang, L., Duan, J., Miki, B., and Wu, K. (2005). HISTONE DEACETYLASE19 Is Involved in jasmonic acid and ethylene signaling of pathogen response in *Arabidopsis*. *The Plant Cell*, 17(4):1196–1204.

Zhou, Y., Hartwig, B., Velikkakam James, G., Schneeberger, K., and Turck, F. (2015). Complementary activities of TELOMERE REPEAT BINDING proteins and Polycomb Group complexes in transcriptional regulation of target genes. *The Plant Cell*, pages TPC2015–00787–RA–50.

Zou, C., Sun, K., Mackaluso, J. D., Seddon, A. E., Jin, R., Thomashow, M. F., and Shiu, S.-H. (2011). *Cis*-regulatory code of stress-responsive transcription in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*, 108(36):14992–14997.

Appendix A

Cell-type specific genes

The provided appendix contains lists of cell-type specific genes determined by Pearson correlation (r) against an artificial bait vector. All genes listed have $r \geq 0.75$. Annotation information is taking from the genomic sequence file TAIR10_cdna_20101214_updated provided by TAIR. Annotation information is a concise summary from the original TAIR file.

A.1 Epidermis

Table A.1: List of 175 epidermal specific genes

AT2G28390.1	SAND family protein
AT1G26110.1	decapping 5
AT2G46410.1	Homeodomain-like superfamily protein
AT3G01280.1	voltage dependent anion channel 1
AT1G75420.1	UDP-Glycosyltransferase superfamily protein
AT5G66800.1	unknown protein; BEST Arabidopsis protein match is AT3G50640.1
AT2G35010.1	thioredoxin O1
AT3G10630.1	UDP-Glycosyltransferase superfamily protein
AT5G63380.1	AMP-dependent synthetase and ligase family protein
AT5G63700.1	zinc ion binding;DNA binding
AT1G66260.1	RNA-binding (RRM/RBD/RNP motifs) family protein
AT5G52830.1	WRKY DNA-binding protein 27
AT1G66620.1	Protein with RING/U-box and TRAF-like domains
AT5G66460.1	Glycosyl hydrolase superfamily protein
AT1G03220.1	Eukaryotic aspartyl protease family protein
AT5G12950.1	Putative glycosyl hydrolase of unknown function (DUF1680)
AT3G04480.1	endoribonucleases
AT4G34160.1	CYCLIN D3

AT3G13450.1	Transketolase family protein
AT1G13060.1	20S proteasome beta subunit E1
AT3G53650.1	Histone superfamily protein
AT4G01410.1	Late embryogenesis abundant (LEA) hydroxyproline-rich glycoprotein family
AT4G00770.1	unknown protein
AT3G50520.1	Phosphoglycerate mutase family protein
AT2G40765.1	unknown protein
AT1G68490.1	unknown protein; BEST Arabidopsis protein match is: AT1G13390.2
AT3G23300.1	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
AT3G12230.1	serine carboxypeptidase-like 14
AT4G16710.1	glycosyltransferase family protein 28
AT4G33780.1	BEST Arabidopsis match is: short hypocotyl in white light AT1G69935.1
AT1G04360.1	RING/U-box superfamily protein
AT5G41000.1	YELLOW STRIPE like 4
AT2G37260.1	WRKY family transcription factor family protein
AT4G36360.1	beta-galactosidase 3
AT2G15490.1	UDP-glycosyltransferase 73B4
AT1G28490.1	syntaxin of plants 61
AT1G27950.1	glycosylphosphatidylinositol-anchored lipid protein transfer 1
AT1G06270.1	Pentatricopeptide repeat (PPR) superfamily protein
AT1G72970.1	Glucose-methanol-choline (GMC) oxidoreductase family protein
AT2G40316.1	CONTAINS InterPro DOMAIN/s: Autophagy-related 27
AT5G25040.1	Major facilitator superfamily protein
AT2G25980.1	Mannose-binding lectin superfamily protein
AT1G79360.1	organic cation/carnitine transporter 2
AT2G07050.1	cycloartenol synthase 1
AT1G47260.1	gamma carbonic anhydrase 2

AT2G19460.1	Protein of unknown function (DUF3511)
AT3G45430.1	Concanavalin A-like lectin protein kinase family protein
AT1G68560.1	alpha-xylosidase 1
AT3G11050.1	ferritin 2
AT1G56020.1	unknown protein; BEST Arabidopsis protein match is: TAIR:AT3G12970.1
AT4G11780.1	unknown protein; BEST Arabidopsis protein match is: TAIR:AT4G23020.2
AT4G01660.1	ABC transporter 1
AT3G27340.1	Gamma-butyrobetaine dioxygenase/Trimethyllysine dioxygenase
AT1G27530.1	InterPro DOMAIN/s: Ubiquitin-conjugating enzyme/RWD-like
AT2G25220.1	Protein kinase superfamily protein
AT2G17630.1	Pyridoxal phosphate (PLP)-dependent transferases superfamily protein
AT5G22570.1	WRKY DNA-binding protein 38
AT2G23670.1	homolog of Synechocystis YCF37
AT4G33090.1	aminopeptidase M1
AT1G14020.1	O-fucosyltransferase family protein
AT1G48900.1	Signal recognition particle, SRP54 subunit protein
AT3G56710.1	sigma factor binding protein 1
AT1G66680.1	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
AT3G61880.1	cytochrome p450 78a9
AT4G16670.1	Protein of unknown function (DUF828) with pleckstrin homology-like region
AT4G24890.1	purple acid phosphatase 24
AT4G22930.1	pyrimidin 4
AT3G55310.1	NAD(P)-binding Rossmann-fold superfamily protein
AT4G11410.1	NAD(P)-binding Rossmann-fold superfamily protein
AT4G29690.1	Alkaline-phosphatase-like family protein
AT2G05840.1	20S proteasome subunit PAA2
AT3G12320.1	unknown protein; BEST Arabidopsis protein match is: TAIR:AT5G06980.4

AT3G20940.1	cytochrome P450, family 705, subfamily A, polypeptide 30
AT3G09940.1	monodehydroascorbate reductase
AT5G66530.1	Galactose mutarotase-like superfamily protein
AT5G40330.1	myb domain protein 23
AT4G15370.1	baruol synthase 1
AT5G62340.1	Plant invertase/pectin methylesterase inhibitor superfamily protein
AT5G39220.1	alpha/beta-Hydrolases superfamily protein
AT4G16240.1	unknown protein
AT1G31950.1	Terpenoid cyclases/Protein prenyltransferases superfamily protein
AT5G18920.1	Cox19-like CHCH family protein
AT5G62810.1	peroxin 14
AT3G46720.1	UDP-Glycosyltransferase superfamily protein
AT2G34470.1	urease accessory protein G
AT1G74030.1	enolase 1
AT1G79840.1	HD-ZIP IV family of homeobox-leucine zipper with lipid-binding
AT5G47520.1	RAB GTPase homolog A5A
AT2G17370.1	3-hydroxy-3-methylglutaryl-CoA reductase 2
AT3G15760.1	BEST Arabidopsis thaliana protein match is: TAIR:AT1G52565.1
AT1G72470.1	exocyst subunit exo70 family protein D1
AT1G55260.1	Bifunctional inhibitor/lipid-transfer protein
AT1G08280.1	Glycosyltransferase family 29 (sialyltransferase) family protein
AT1G54870.1	NAD(P)-binding Rossmann-fold superfamily protein
AT3G02480.1	Late embryogenesis abundant protein (LEA) family protein
AT5G59250.1	Major facilitator superfamily protein
AT5G58710.1	rotamase CYP 7
AT1G15330.1	Cystathionine beta-synthase (CBS) protein
AT4G04470.1	Peroxisomal membrane 22 kDa (Mpv17/PMP22) family protein

AT4G21865.1	unknown protein
AT1G33490.1	BEST Arabidopsis thaliana protein match is: TAIR:AT4G10140.1
AT3G27570.1	Sucrase/ferredoxin-like family protein
AT3G15820.1	phosphatidic acid phosphatase-related / PAP2-related
AT3G45300.1	isovaleryl-CoA-dehydrogenase
AT5G57920.1	early nodulin-like protein 10
AT3G10920.1	manganese superoxide dismutase 1
AT4G33220.1	pectin methylesterase 44
AT5G43030.1	Cysteine/Histidine-rich C1 domain family protein
AT3G12290.1	Amino acid dehydrogenase family protein
AT4G11010.1	nucleoside diphosphate kinase 3
AT2G20420.1	ATP citrate lyase (ACL) family protein
AT4G35200.1	Arabidopsis protein of unknown function (DUF241)
AT4G32870.1	Polyketide cyclase/dehydrase and lipid transport superfamily protein
AT3G21160.1	alpha-mannosidase 2
AT1G44830.1	Integrase-type DNA-binding superfamily protein
AT1G64900.1	cytochrome P450, family 89, subfamily A, polypeptide 2
AT5G20550.1	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein
AT5G60840.1	unknown protein
AT4G22130.1	STRUBBELIG-receptor family 8
AT1G52260.1	PDI-like 1-5
AT4G21860.1	methionine sulfoxide reductase B 2
AT5G55610.1	unknown protein
AT5G42980.1	thioredoxin 3
AT1G48030.1	mitochondrial lipoamide dehydrogenase 1
AT5G63810.1	beta-galactosidase 10
AT1G01490.1	Heavy metal transport/detoxification superfamily protein

AT3G59760.1	O-acetylserine (thiol) lyase isoform C
AT1G27190.1	Leucine-rich repeat protein kinase family protein
AT5G20070.1	nudix hydrolase homolog 19
AT1G76620.1	Protein of unknown function, DUF547
AT4G29020.1	glycine-rich protein
AT1G26550.1	FKBP-like peptidyl-prolyl cis-trans isomerase family protein
AT3G53400.1	Arabidopsis match is: conserved peptide upstream ORF 47 AT5G03190.1
AT4G04020.1	fibrillin
AT5G26260.1	TRAF-like family protein
AT1G19120.1	Small nuclear ribonucleoprotein family protein
AT1G22360.1	UDP-glucosyl transferase 85A2
AT2G43535.1	Scorpion toxin-like knottin superfamily protein
AT2G03510.1	SPFH/Band 7/PHB domain-containing membrane-associated protein
AT1G33540.1	serine carboxypeptidase-like 18
AT5G17960.1	Cysteine/Histidine-rich C1 domain family protein
AT5G06270.1	BEST Arabidopsis thaliana protein match is: TAIR:AT3G11600.1
AT1G05590.1	beta-hexosaminidase 2
AT3G15260.1	Protein phosphatase 2C family protein
AT3G48170.1	aldehyde dehydrogenase 10A9
AT5G66170.1	sulfurtransferase 18
AT2G46790.1	pseudo-response regulator 9
AT1G72680.1	cinnamyl-alcohol dehydrogenase
AT5G56320.1	expansin A14
AT1G09780.1	Phosphoglycerate mutase, 2,3-bisphosphoglycerate-independent
AT3G45620.1	Transducin/WD40 repeat-like superfamily protein
AT2G04500.1	Cysteine/Histidine-rich C1 domain family protein
AT1G71170.1	6-phosphogluconate dehydrogenase family protein

AT1G53180.1	BEST Arabidopsis protein match is: TAIR:AT3G15115.1
AT3G20470.1	glycine-rich protein 5
AT2G42840.1	protodermal factor 1
AT4G37410.1	cytochrome P450, family 81, subfamily F, polypeptide 4
AT3G16390.1	nitrile specifier protein 3
AT3G20590.1	Late embryogenesis abundant (LEA) hydroxyproline-rich glycoprotein family
AT5G11950.1	Putative lysine decarboxylase family protein
AT1G56580.1	Protein of unknown function, DUF538
AT5G01870.1	Bifunctional inhibitor/lipid-transfer protein
AT3G03100.1	NADH:ubiquinone oxidoreductase, 17.2kDa subunit
AT1G03210.1	Phenazine biosynthesis PhzC/PhzF protein
AT1G15470.1	Transducin/WD40 repeat-like superfamily protein
AT1G08480.1	unknown protein
AT1G66800.1	NAD(P)-binding Rossmann-fold superfamily protein
AT3G20130.1	cytochrome P450, family 705, subfamily A, polypeptide 22
AT5G63760.1	RING/U-box superfamily protein
AT5G25610.1	BURP domain-containing protein
AT3G08770.1	lipid transfer protein 6
AT2G36050.1	ovate family protein 15
AT5G43940.1	GroES-like zinc-binding dehydrogenase family protein
AT5G59320.1	lipid transfer protein 3

A.2 Cortex

Table A.2: List of 76 cortex specific genes

AT5G07200.1	gibberellin 20-oxidase 3
AT5G55120.1	galactose-1-phosphate guanylyltransferase (GDP)s
AT3G61190.1	BON association protein 1
AT1G03840.1	C2H2 and C2HC zinc fingers superfamily protein
AT2G16950.1	transportin 1
AT1G29910.1	chlorophyll A/B binding protein 3
AT4G03280.1	photosynthetic electron transfer C
AT5G51110.1	Transcriptional coactivator/pterin dehydratase
AT1G12100.1	Bifunctional inhibitor/lipid-transfer protein storage 2S albumin superfamily
AT1G67830.1	alpha-fucosidase 1
AT2G47450.1	chloroplast signal recognition particle component (CAO)
AT3G60920.1	CONTAINS InterPro DOMAIN/s: Beige/BEACH (InterPro:IPR000409)
AT1G73620.1	Pathogenesis-related thaumatin superfamily protein
AT1G76050.1	Pseudouridine synthase family protein
AT3G06450.1	HCO3- transporter family
AT2G46310.1	cytokinin response factor 5
AT2G24200.1	Cytosol aminopeptidase family protein
AT5G17880.1	disease resistance protein (TIR-NBS-LRR class)
AT4G27640.1	ARM repeat superfamily protein
AT4G15160.1	Bifunctional inhibitor/lipid-transfer protein storage 2S albumin superfamily
AT5G10270.1	cyclin-dependent kinase C;1
AT1G49480.1	related to vernalization1 1
AT4G39940.1	APS-kinase 2

AT3G59380.1	farnesyltransferase A
AT2G23700.1	Protein of unknown function, DUF547
AT2G30860.1	glutathione S-transferase PHI 9
AT5G64940.1	ABC2 homolog 13
AT2G25690.1	Protein of unknown function (DUF581)
AT2G10940.1	Bifunctional inhibitor/lipid-transfer protein storage 2S albumin superfamily
AT4G26480.1	RNA-binding KH domain-containing protein
AT2G22330.1	cytochrome P450, family 79, subfamily B, polypeptide 3
AT4G04180.1	P-loop containing nucleoside triphosphate hydrolases superfamily protein
AT1G06680.1	photosystem II subunit P-1
AT5G08280.1	hydroxymethylbilane synthase
AT2G15560.1	Putative endonuclease or glycosyl hydrolase
AT3G26744.1	basic helix-loop-helix (bHLH) DNA-binding superfamily protein
AT3G18960.1	AP2/B3-like transcriptional factor family protein
AT1G29820.1	Magnesium transporter CorA-like family protein
AT4G28410.1	Tyrosine transaminase family protein
AT1G01170.1	Protein of unknown function (DUF1138)
AT1G67090.1	ribulose bisphosphate carboxylase small chain 1A
AT3G23570.1	alpha/beta-Hydrolases superfamily protein
AT4G10340.1	light harvesting complex of photosystem II 5
AT4G31500.1	cytochrome P450, family 83, subfamily B, polypeptide 1
AT1G20620.1	catalase 3
AT2G24280.1	alpha/beta-Hydrolases superfamily protein
AT3G14240.1	Subtilase family protein
AT3G13620.1	Amino acid permease family protein
AT4G27570.1	UDP-Glycosyltransferase superfamily protein
AT2G42130.1	Plastid-lipid associated protein PAP / fibrillin family protein

AT2G06520.1	photosystem II subunit X
AT5G27410.1	D-aminoacid aminotransferase-like PLP-dependent enzymes superfamily
AT4G38970.1	fructose-bisphosphate aldolase 2
AT1G52230.1	photosystem I subunit H2
AT1G19210.1	Integrase-type DNA-binding superfamily protein
AT3G54220.1	GRAS family transcription factor
AT3G19710.1	branched-chain aminotransferase4
AT1G52890.1	NAC domain containing protein 19
AT2G20610.1	Tyrosine transaminase family protein
AT5G55480.1	SHV3-like 1
AT1G62500.1	Bifunctional inhibitor/lipid-transfer protein storage 2S albumin superfamily
AT3G56880.1	VQ motif-containing protein
AT2G26500.1	cytochrome b6f complex subunit (petM), putative
AT3G21055.1	photosystem II subunit T
AT3G26650.1	glyceraldehyde 3-phosphate dehydrogenase A subunit
AT1G18590.1	sulfotransferase 17
AT3G54890.1	photosystem I light harvesting complex gene 1
AT1G20340.1	Cupredoxin superfamily protein
AT3G16140.1	photosystem I subunit H-1
AT4G03060.1	AOP2 (ALKENYL HYDROXALKYL PRODUCING 2)
AT5G25980.1	glucoside glucohydrolase 2
AT4G38800.1	methylthioadenosine nucleosidase 1
AT3G61470.1	photosystem I light harvesting complex gene 2

A.3 Endodermis

Table A.3: List of 255 endodermal specific genes

AT5G14850.1	Alg9-like mannosyltransferase family
AT2G22490.1	Cyclin D2
AT2G30600.1	BTB/POZ domain-containing protein
AT1G21460.1	Nodulin MtN3 family protein
AT2G40350.1	Integrase-type DNA-binding superfamily protein
AT5G52760.1	Copper transport protein family
AT5G15420.1	unknown protein
AT1G19600.1	pfkB-like carbohydrate kinase family protein
AT5G13890.1	Family of unknown function (DUF716)
AT5G63200.1	tetratricopeptide repeat (TPR)-containing protein
AT3G13200.1	Cwf15 / Cwc15 cell cycle control family protein
AT5G01160.1	RING/U-box superfamily protein
AT4G23880.1	unknown protein
AT2G07713.1	unknown protein
AT3G45770.1	Polyketide synthase, enoylreductase family protein
AT2G23110.1	Late embryogenesis abundant protein, group 6
AT1G03200.1	unknown protein
AT1G03240.1	unknown protein
AT1G32830.1	transposable element gene
AT2G33740.1	Nitrogen regulatory PII-like, alpha/beta
AT5G52980.1	CONTAINS InterPro DOMAIN/s: ATPase
AT5G42850.1	Thioredoxin superfamily protein
AT3G29130.1	CONTAINS InterPro DOMAIN/s: Domain of unknown function KxDL

AT5G42290.1	transcription activator-related
AT2G42210.1	Mitochondrial import inner membrane translocase subunit
AT1G64250.1	transposable element gene
AT3G10040.1	sequence-specific DNA binding transcription factors
AT5G12110.1	Glutathione S-transferase, C-terminal-like
AT1G62960.1	ACC synthase 10
AT1G22770.1	gigantea protein (GI)
AT4G22740.1	glycine-rich protein
AT1G21610.1	wound-responsive family protein
AT5G43460.1	HR-like lesion-inducing protein-related
AT4G08780.1	Peroxidase superfamily protein
AT1G02850.1	beta glucosidase 11
AT4G39235.1	unknown protein; BEST Arabidopsis protein match is AT3G05570.1
AT5G52400.1	cytochrome P450, family 715, subfamily A, polypeptide 1
AT1G72360.1	Integrase-type DNA-binding superfamily protein
AT5G15450.1	casein lytic proteinase B3
AT1G09280.1	CONTAINS InterPro DOMAIN/s: Rhodanese-like, Serine hydrolase
AT2G43790.1	MAP kinase 6
AT1G50640.1	ethylene responsive element binding factor 3
AT3G22840.1	Chlorophyll A-B binding family protein
AT5G53850.1	haloacid dehalogenase-like hydrolase family protein
AT1G33130.1	transposable element gene
AT4G20310.1	Peptidase M50 family protein
AT4G09830.1	Uncharacterised conserved protein UCP009193
AT5G18110.1	novel cap-binding protein
AT1G53540.1	HSP20-like chaperones superfamily protein
AT1G23180.1	ARM repeat superfamily protein

AT4G08890.1	transposable element gene
AT2G43420.1	3-beta hydroxysteroid dehydrogenase/isomerase family protein
AT5G05100.1	Single-stranded nucleic acid binding R3H protein
AT3G60300.1	RWD domain-containing protein
AT3G50190.1	Plant protein of unknown function (DUF247)
AT2G27380.1	extensin proline-rich 1
AT5G63260.1	Zinc finger C-x8-C-x5-C-x3-H type family protein
AT3G51240.1	flavanone 3-hydroxylase
AT3G09850.1	D111/G-patch domain-containing protein
AT3G56290.1	unknown protein
AT1G32370.1	tobamovirus multiplication 2B
AT3G02550.1	LOB domain-containing protein 41
AT1G32840.1	transposable element gene
AT3G18980.1	EIN2 targeting protein1
AT4G31420.1	Zinc finger protein 622
AT4G24500.1	hydroxyproline-rich glycoprotein family protein
AT2G01960.1	tetraspanin14
AT1G65920.1	Regulator of chromosome condensation (RCC1) family
AT4G08770.1	Peroxidase superfamily protein
AT2G07672.1	BEST Arabidopsis thaliana protein match is: ATMG01050.1
AT1G61670.1	Lung seven transmembrane receptor family protein
AT1G33110.1	MATE efflux family protein
AT3G28310.1	Protein of unknown function (DUF677)
AT1G71690.1	Protein of unknown function (DUF579)
AT4G18170.1	WRKY DNA-binding protein 28
AT2G06390.1	transposable element gene
AT2G10880.1	transposable element gene

AT2G22080.1	unknown protein; Has 96314 Blast hits to 34847 proteins in 1702 species
AT2G07722.1	unknown protein; BEST Arabidopsis protein match is ATMG00620.1
AT2G07180.1	Protein kinase superfamily protein
AT4G15780.1	vesicle-associated membrane protein 724
AT3G27150.1	Galactose oxidase/kelch repeat superfamily protein
AT3G63460.1	transducin family protein / WD-40 repeat family protein
AT1G22940.1	thiamin biosynthesis protein, putative
AT3G28320.1	Protein of unknown function (DUF677)
AT1G24340.1	FAD/NAD(P)-binding oxidoreductase family protein
AT3G10670.1	non-intrinsic ABC protein 7
AT5G38820.1	Transmembrane amino acid transporter family protein
AT4G10270.1	Wound-responsive family protein
AT1G22490.1	basic helix-loop-helix (bHLH) DNA-binding superfamily protein
AT1G55980.1	FAD/NAD(P)-binding oxidoreductase family protein
AT2G07675.1	Ribosomal protein S12/S23 family protein
AT1G63060.1	unknown protein
AT5G07330.1	unknown protein; BEST Arabidopsis protein match is AT1G63060.1
AT1G18330.1	Homeodomain-like superfamily protein
AT2G29500.1	HSP20-like chaperones superfamily protein
AT5G24470.1	pseudo-response regulator 5
AT2G17850.1	Rhodanese/Cell cycle control phosphatase superfamily protein
AT1G73980.1	Phosphoribulokinase / Uridine kinase family
AT1G57550.1	Low temperature and salt responsive protein family
AT2G14140.1	transposable element gene
AT5G17060.1	ADP-ribosylation factor B1B
AT5G55060.1	unknown protein; BEST Arabidopsis protein match is AT5G58510.1
AT3G32000.1	transposable element gene

AT2G46900.1	CONTAINS InterPro DOMAIN/s: Basic helix-loop-helix, Nulp1-type
AT2G14650.1	transposable element gene
AT2G07706.1	unknown protein; BEST Arabidopsis protein match is ATMG00470.1
AT5G56290.1	peroxin 5
AT4G02560.1	Homeodomain-like superfamily protein
AT2G31830.1	endonuclease/exonuclease/phosphatase family protein
AT5G13110.1	glucose-6-phosphate dehydrogenase 2
AT1G78180.1	Mitochondrial substrate carrier family protein
AT3G50880.1	DNA glycosylase superfamily protein
AT2G21640.1	Encodes a protein of unknown function that is a marker for oxidative stress
AT5G51440.1	HSP20-like chaperones superfamily protein
AT4G27960.1	ubiquitin conjugating enzyme 9
AT2G18440.1	GUT15 (GENE WITH UNSTABLE TRANSCRIPT 15); other RNA
AT5G03690.1	Aldolase superfamily protein
AT2G37585.1	Core-2/I-branching beta-1,6-N-acetylglucosaminyltransferase family
AT1G27340.1	Galactose oxidase/kelch repeat superfamily protein
AT3G53540.1	unknown protein
AT5G49630.1	amino acid permease 6
AT4G03900.1	transposable element gene
AT5G12030.1	heat shock protein 17.6A
AT4G13730.1	Ypt/Rab-GAP domain of gyp1p superfamily protein
AT3G54660.1	glutathione reductase
AT5G40100.1	Disease resistance protein (TIR-NBS-LRR class) family
AT1G74310.1	heat shock protein 101
AT5G18040.1	unknown protein; BEST Arabidopsis protein match is AT4G29760.1
AT3G48070.1	RING/U-box superfamily protein

AT2G26210.1	Ankyrin repeat family protein
AT1G50290.1	unknown protein; Has 2 Blast hits to 2 proteins in 1 species
AT5G51020.1	crumpled leaf
AT2G40950.1	Basic-leucine zipper (bZIP) transcription factor family protein
AT1G28320.1	protease-related
AT3G16640.1	translationally controlled tumor protein
AT3G52300.1	ATP synthase D chain, mitochondrial
AT3G53340.1	nuclear factor Y, subunit B10
AT5G22600.1	FBD / Leucine Rich Repeat domains containing protein
AT2G34390.1	NOD26-like intrinsic protein 2;1
AT3G27310.1	plant UBX domain-containing protein 1
AT5G67380.1	casein kinase alpha 1
AT5G23380.1	Protein of unknown function (DUF789)
AT1G54050.1	HSP20-like chaperones superfamily protein
AT5G48570.1	FKBP-type peptidyl-prolyl cis-trans isomerase family protein
AT2G18550.1	homeobox protein 21
AT3G47610.1	transcription regulators;zinc ion binding
AT1G28330.1	dormancy-associated protein-like 1
AT2G47720.1	FUNCTIONS IN: molecular function unknown
AT2G26870.1	non-specific phospholipase C2
AT2G43970.1	RNA-binding protein
AT5G53190.1	Nodulin MtN3 family protein
AT2G07687.1	Cytochrome c oxidase, subunit III
AT5G12120.1	Ubiquitin-associated/translation elongation factor EF1B protein
AT2G15140.1	transposable element gene
AT2G10740.1	transposable element gene
AT1G59860.1	HSP20-like chaperones superfamily protein

AT5G38140.1	nuclear factor Y, subunit C12
AT2G36460.1	Aldolase superfamily protein
AT2G18670.1	RING/U-box superfamily protein
AT1G05840.1	Eukaryotic aspartyl protease family protein
AT3G47260.1	transposable element gene
AT5G10040.1	unknown protein; BEST Arabidopsis protein match is AT5G65207.1
AT1G55510.1	branched-chain alpha-keto acid decarboxylase E1 beta subunit
AT1G27370.1	squamosa promoter binding protein-like 10
AT5G13010.1	RNA helicase family protein
AT5G49580.1	Chaperone DnaJ-domain superfamily protein
AT1G13440.1	glyceraldehyde-3-phosphate dehydrogenase C2
AT2G32120.1	heat-shock protein 70T-2
AT4G12400.1	stress-inducible protein, putative
AT4G26270.1	phosphofructokinase 3
AT3G44470.1	transposable element gene
AT5G28590.1	DNA-binding family protein
AT3G62190.1	Chaperone DnaJ-domain superfamily protein
AT4G39900.1	unknown protein
AT3G23170.1	unknown protein; BEST Arabidopsis protein match is AT4G14450.1
AT5G58575.1	CONTAINS InterPro DOMAIN/s: Sgf11, transcriptional regulation
AT5G48250.1	B-box type zinc finger protein with CCT domain
AT3G31970.1	transposable element gene
AT2G31340.1	embryo defective 1381
AT4G02550.1	unknown protein; BEST Arabidopsis protein match is AT4G02210.2
AT5G44000.1	Glutathione S-transferase family protein
AT1G70480.1	Domain of unknown function (DUF220)
AT1G71000.1	Chaperone DnaJ-domain superfamily protein

AT2G04980.1	transposable element gene
AT1G76080.1	chloroplastic drought-induced stress protein of 32 kD
AT1G79790.1	Haloacid dehalogenase-like hydrolase (HAD) superfamily protein
AT2G03080.1	transposable element gene
AT2G17570.1	Undecaprenyl pyrophosphate synthetase family protein
AT5G35320.1	unknown protein; Has 1807 Blast hits to 1807 proteins in 277 species
AT3G46230.1	heat shock protein 17.4
AT2G38780.1	unknown protein
AT5G54350.1	BEST Arabidopsis thaliana protein match is: C2H2-like zinc finger
AT4G21320.1	Aldolase-type TIM barrel family protein
AT4G10250.1	HSP20-like chaperones superfamily protein
AT1G64105.1	NAC domain containing protein 27
AT3G13800.1	Metallo-hydrolase/oxidoreductase superfamily protein
AT2G47180.1	galactinol synthase 1
AT1G17300.1	unknown protein; BEST Arabidopsis protein match is AT1G17285.1
AT3G01560.1	Protein of unknown function (DUF1421)
AT3G29210.1	transposable element gene
AT4G19240.1	unknown protein; BEST Arabidopsis protein match is AT3G43280.1
AT2G16700.1	actin depolymerizing factor 5
AT2G22240.1	myo-inositol-1-phosphate synthase 2
AT2G07718.1	Cytochrome b/b6 protein
AT5G52640.1	heat shock protein 90.1
AT1G62770.1	Plant invertase/pectin methylesterase inhibitor superfamily protein
AT2G25140.1	casein lytic proteinase B4
AT2G46240.1	BCL-2-associated athanogene 6
AT1G37160.1	transposable element gene
AT1G16030.1	heat shock protein 70B

AT2G04970.1	transposable element gene
AT4G18770.1	myb domain protein 98
AT2G10140.1	transposable element gene
AT3G44500.1	transposable element gene
AT3G30396.1	transposable element gene
AT2G14130.1	transposable element gene
AT4G07360.1	transposable element gene
AT4G10260.1	pfkB-like carbohydrate kinase family protein
AT2G07673.1	unknown protein
AT2G06480.1	transposable element gene
AT2G02200.1	transposable element gene
AT5G11260.1	Basic-leucine zipper (bZIP) transcription factor family protein
AT5G03200.1	RING/U-box superfamily protein
AT2G27200.1	P-loop containing nucleoside triphosphate hydrolases superfamily
AT4G09150.1	T-complex protein 11
AT4G04010.1	transposable element gene
AT5G16990.1	Zinc-binding dehydrogenase family protein
AT3G32900.1	transposable element gene
AT2G06430.1	transposable element gene
AT1G36550.1	transposable element gene
AT5G14270.1	bromodomain and extraterminal domain protein 9
AT2G15890.1	maternal effect embryo arrest 14
AT3G09640.1	ascorbate peroxidase 2
AT1G80590.1	WRKY DNA-binding protein 66
AT3G30440.1	transposable element gene
AT5G09570.1	Cox19-like CHCH family protein
AT5G64400.1	CONTAINS InterPro DOMAIN/s: CHCH (InterPro:IPR010625)

AT4G21323.1	Subtilase family protein
AT2G14210.1	AGAMOUS-like 44
AT5G59720.1	heat shock protein 18.2
AT1G04300.1	TRAF-like superfamily protein
AT5G47830.1	unknown protein
AT2G10640.1	transposable element gene
AT4G27670.1	heat shock protein 21
AT2G07724.1	unknown protien
AT2G39170.1	unknown protein; CONTAINS InterPro DOMAIN/s: NEP
AT2G24310.1	unknown protein
AT3G21720.1	isocitrate lyase
AT4G25200.1	mitochondrion-localized small heat shock protein 23.6
AT3G09390.1	metallothionein 2A
AT1G58025.1	DNA-binding bromodomain-containing protein
AT2G12240.1	transposable element gene
AT3G60980.1	Tetratricopeptide repeat (TPR)-like superfamily protein
AT1G33055.1	unknown protein; FUNCTIONS IN: molecular function unknown
AT5G37670.1	HSP20-like chaperones superfamily protein

Appendix B

Cell-type specific genes used in motif prediction

The following appendix contains a list of cell-type specific genes for epidermal, cortex, and endodermal cell layers. The upstream 500bp promoter region of these listed genes was used for motif prediction in order to identify over represented motif patterns as possible CREs. Forty gene promoters with the highest Pearson correlation coefficient (PCC) against a cell-type specific bait were used for all predictions. Annotation information is taking from the genomic sequence file TAIR10_cdna_20101214_updated provided by TAIR. Annotation information is a concise summary from the original TAIR file.

B.1 Epidermis

Table B.1: Forty epidermal-specific genes used for motif prediction.

Gene AGI	PCC	Gene Name	Gene description
AT5G59320	0.99		lipid transfer protein 3
AT5G43940	0.99	ATGSNOR1	GroES-like zinc-binding dehydrogenase family protein
AT2G36050	0.98	ATOFP15	ovate family protein 15
AT3G08770	0.97	LTP6	lipid transfer protein 6
AT5G25610	0.97	RD22	BURP domain-containing protein
AT5G63760	0.97	ARI15	RING/U-box superfamily protein
AT3G20130	0.97	CYP705A22	cytochrome P450, family 705, subfamily A
AT1G66800	0.97		NAD(P)-binding Rossmann-fold superfamily protein
AT1G08480	0.96		unknown protein
AT1G15470	0.96		Transducin/WD40 repeat-like superfamily protein
AT1G03210	0.96		Phenazine biosynthesis PhzC/PhzF protein
AT3G03100	0.95		NADH:ubiquinone oxidoreductase, 17.2kDa subunit
AT5G01870	0.95		Bifunctional inhibitor/lipid-transfer protein storage 2S
AT1G56580	0.95	SVB	Protein of unknown function, DUF538
AT5G11950	0.95		Putative lysine decarboxylase family protein
AT3G20590	0.94		Late embryogenesis abundant (LEA) glycoprotein
AT3G16390	0.94	NSP3	nitrile specifier protein 3
AT4G37410	0.94	CYP81F4	cytochrome P450, family 81, subfamily F, polypeptide 4
AT2G42840	0.94	PDF1	protodermal factor 1
AT3G20470	0.93	GRP-5	glycine-rich protein 5
AT1G53180	0.93		unknown protein
AT1G71170	0.93		6-phosphogluconate dehydrogenase family protein
AT2G04500	0.93		Cysteine/Histidine-rich C1 domain family protein

AT3G45620	0.92		Transducin/WD40 repeat-like superfamily protein
AT1G09780	0.92		Phosphoglycerate mutase, 2,3-bisphosphoglycerate
AT5G56320	0.91	ATEXPA14	expansin A14
AT1G72680	0.91	ATCAD1	cinnamyl-alcohol dehydrogenase
AT2G46790	0.91	APRR9	pseudo-response regulator 9
AT5G66170	0.90	STR18	sulfurtransferase 18
AT3G48170	0.90	ALDH10A9	aldehyde dehydrogenase 10A9
AT3G15260	0.90		Protein phosphatase 2C family protein
AT1G05590	0.89	HEXO2	beta-hexosaminidase 2
AT5G06270	0.89		unknown protein
AT5G17960	0.89		Cysteine/Histidine-rich C1 domain family protein
AT1G33540	0.89	scpl18	serine carboxypeptidase-like 18
AT2G03510	0.89		SPFH/B and 7/PHB domain-containing protein
AT2G43535	0.89		Scorpion toxin-like knottin superfamily protein
AT1G22360	0.88	AtUGT85A2	UDP-glucosyl transferase 85A2
AT1G19120	0.88		Small nuclear ribonucleoprotein family protein
AT5G26260	0.88		TRAF-like family protein

B.2 Cortex

Table B.2: Forty cortex-specific genes used for motif prediction.

AT3G61470	0.99	LHCA2	photosystem I light harvesting complex gene 2
AT4G38800	0.98	ATMTN1, ATMTAN1	methylthioadenosine nucleosidase 1
AT4G03060	0.97	APO2	Alkenyl Hydroxalkyl Producing 2
AT5G25980	0.97	TGG2, BGLU37	glucoside glucohydrolase 2
AT3G16140	0.97	PSAH-1	photosystem I subunit H-1
AT1G20340	0.96	DRT112, PETE2	Cupredoxin superfamily protein
AT3G54890	0.95	LHCA1	photosystem I light harvesting complex gene 1
AT1G18590	0.94	SOT17, ATSOT17	sulfotransferase 17
AT3G26650	0.94	GAPA, GAPA-1	glyceraldehyde 3-phosphate dehydrogenase
AT3G21055	0.94	PSBTN	photosystem II subunit T
AT2G26500	0.94		cytochrome b6f complex subunit (petM), putative
AT3G56880	0.93		VQ motif-containing protein
AT1G62500	0.93		Bifunctional inhibitor/lipid-transfer protein
AT5G55480	0.92	SVL1	SHV3-like 1
AT2G20610	0.92	SUR1, HLS3, RTY	Tyrosine transaminase family protein
AT1G52890	0.92	ANAC019, NAC019	NAC domain containing protein 19
AT3G19710	0.91	BCAT4	branched-chain aminotransferase4
AT3G54220	0.91	SCR, SGR1	GRAS family transcription factor
AT1G19210	0.91		Integrase-type DNA-binding superfamily protein
AT1G52230	0.90	PSAH2, PSAH-2	photosystem I subunit H2
AT4G38970	0.90	FBA2	fructose-bisphosphate aldolase 2
AT5G27410	0.88		D-aminoacid aminotransferase-like
AT2G06520	0.87	PSBX	photosystem II subunit X
AT2G42130	0.87		Plastid-lipid associated protein PAP

AT4G27570	0.87		UDP-Glycosyltransferase superfamily protein
AT3G13620	0.86		Amino acid permease family protein
AT3G14240	0.86		Subtilase family protein
AT2G24280	0.86		alpha/beta-Hydrolases superfamily protein
AT1G20620	0.86	CAT3, SEN2, ATCAT3	catalase 3
AT4G31500	0.85	CYP83B1, SUR2	cytochrome P450, family 83
AT4G10340	0.85	LHCB5	light harvesting complex of photosystem II 5
AT3G23570	0.84		alpha/beta-Hydrolases superfamily protein
AT1G67090	0.84	RBCS1A	ribulose bisphosphate carboxylase small chain
AT1G01170	0.84		Protein of unknown function (DUF1138)
AT4G28410	0.84		Tyrosine transaminase family protein
AT1G29820	0.83		Magnesium transporter CorA-like family protein
AT3G18960	0.83		AP2/B3-like transcriptional factor family protein
AT3G26744	0.83	ICE1, ATICE1	basic helix-loop-helix (bHLH)
AT2G15560	0.82		Putative endonuclease or glycosyl hydrolase
AT5G08280	0.82	HEMC	hydroxymethylbilane synthase

B.3 Endodermis

Table B.3: Forty endodermal-specific genes used for motif prediction.

Gene AGI	PCC	Gene Name	Gene description
AT5G37670	0.99		HSP20-like chaperones superfamily protein
AT3G60980	0.98		Tetratricopeptide repeat (TPR)-like superfamily
AT3G09390	0.98	MT2A, ATMT-K	metallothionein 2A
AT2G12240	0.98		CACTA-like transposase family
AT1G58025	0.98		DNA-binding bromodomain-containing protein
AT1G33055	0.98		unknown protein
AT5G47830	0.97		unknown protein
AT4G27670	0.97	HSP21	heat shock protein 21
AT4G25200	0.97	ATHSP23.6-MITO	mitochondrion-localized small HSP
AT3G21720	0.97	ICL	isocitrate lyase
AT2G39170	0.97		unknown protein
AT2G24310	0.97		unknown protein
AT2G10640	0.97		CACTA-like transposase family
AT2G07724	0.97		unknown protein
AT5G59720	0.96	HSP18.2	heat shock protein 18.2
AT5G09570	0.96		Cox19-like CHCH family protein
AT4G21323	0.96		Subtilase family protein
AT3G09640	0.96	APX2, APX1B	ascorbate peroxidase 2
AT2G14210	0.96	ANR1, AGL44	AGAMOUS-like 44
AT1G80590	0.96	WRKY66	WRKY DNA-binding protein 66
AT1G04300	0.96		TRAF-like superfamily protein
AT5G14270	0.95	ATBET9, BET9	extraterminal domain protein 9
AT2G15890	0.95	MEE14	maternal effect embryo arrest 14

AT2G06430	0.95		Ulp1 protease family
AT1G36550	0.95		Transposable element gene
AT5G16990	0.94		Zinc-binding dehydrogenase family protein
AT5G11260	0.94	HY5, TED 5	Basic-leucine zipper transcription factor family
AT5G03200	0.94		RING/U-box superfamily protein
AT4G10260	0.94		pfkB-like carbohydrate kinase family protein
AT4G09150	0.94		T-complex protein 11
AT4G07360	0.94		Gypsy-like retrotransposon family
AT3G30396	0.94		CACTA-like transposase family
AT2G27200	0.94		P-loop nucleoside triphosphate hydrolases
AT2G07673	0.94		unknown protein
AT2G06480	0.94		Transposable element gene
AT2G02200	0.94		Transposable element gene
AT4G18770	0.93	MYB98, AtMYB98	myb domain protein 98
AT2G10140	0.93		CACTA-like transposase family
AT2G04970	0.93		Similar to heat shock protein binding
AT1G16030	0.93	Hsp70b	heat shock protein 70B

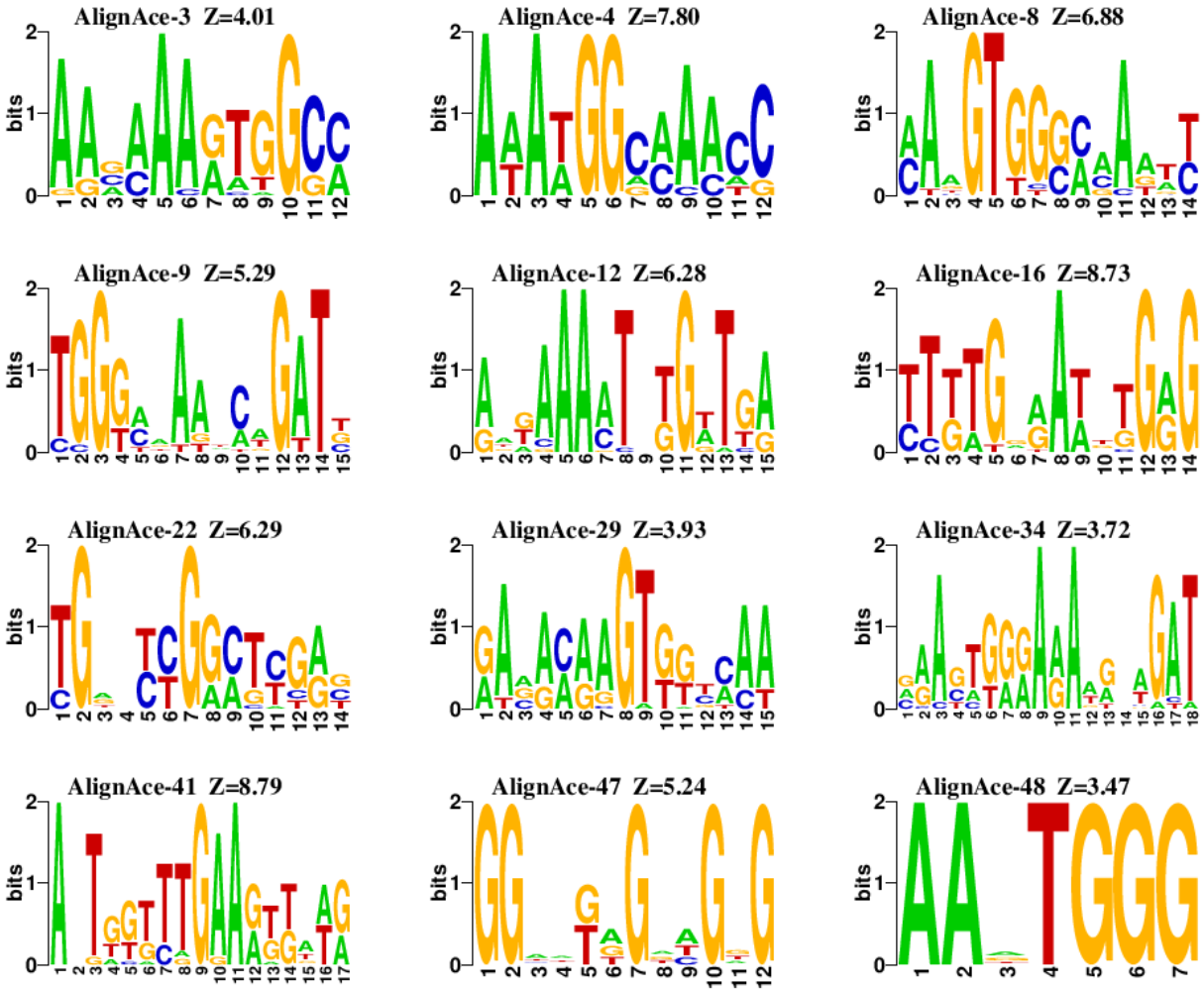
Appendix C

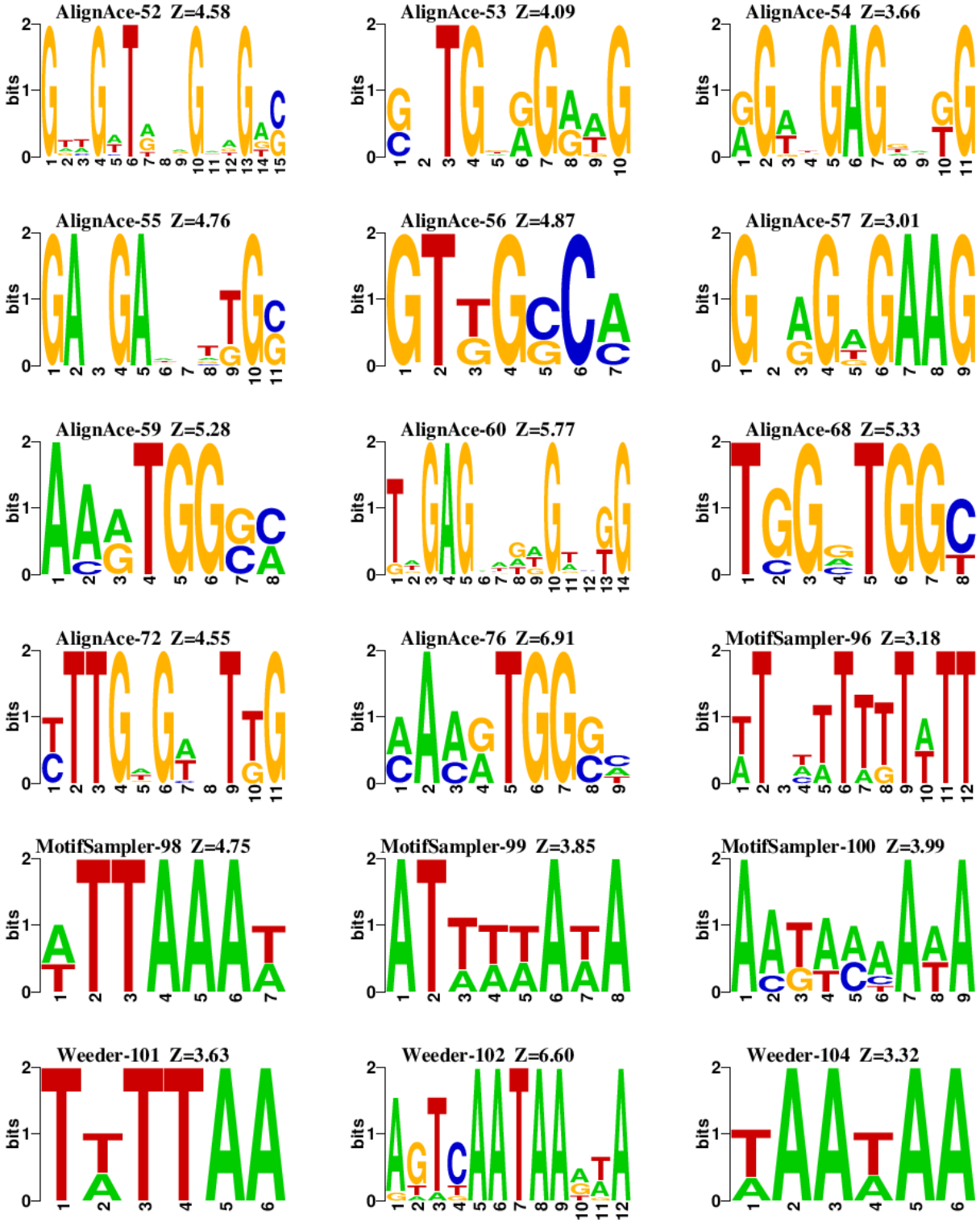
Cell-type specific putative motifs

The following appendix contains a list of significant ($Z \geq 3$) putative motifs predicted by the *Cister* associated programs (see Methods) against cell-type promoter sequences. Motifs are represented as sequence logos with their corresponding significance scores and prediction program used provided at the top of each logo. For large prediction sets, the first 30-33 motifs are provided.

C.1 Epidermis

Table C.1: Epidermal-specific motif sequence logos.

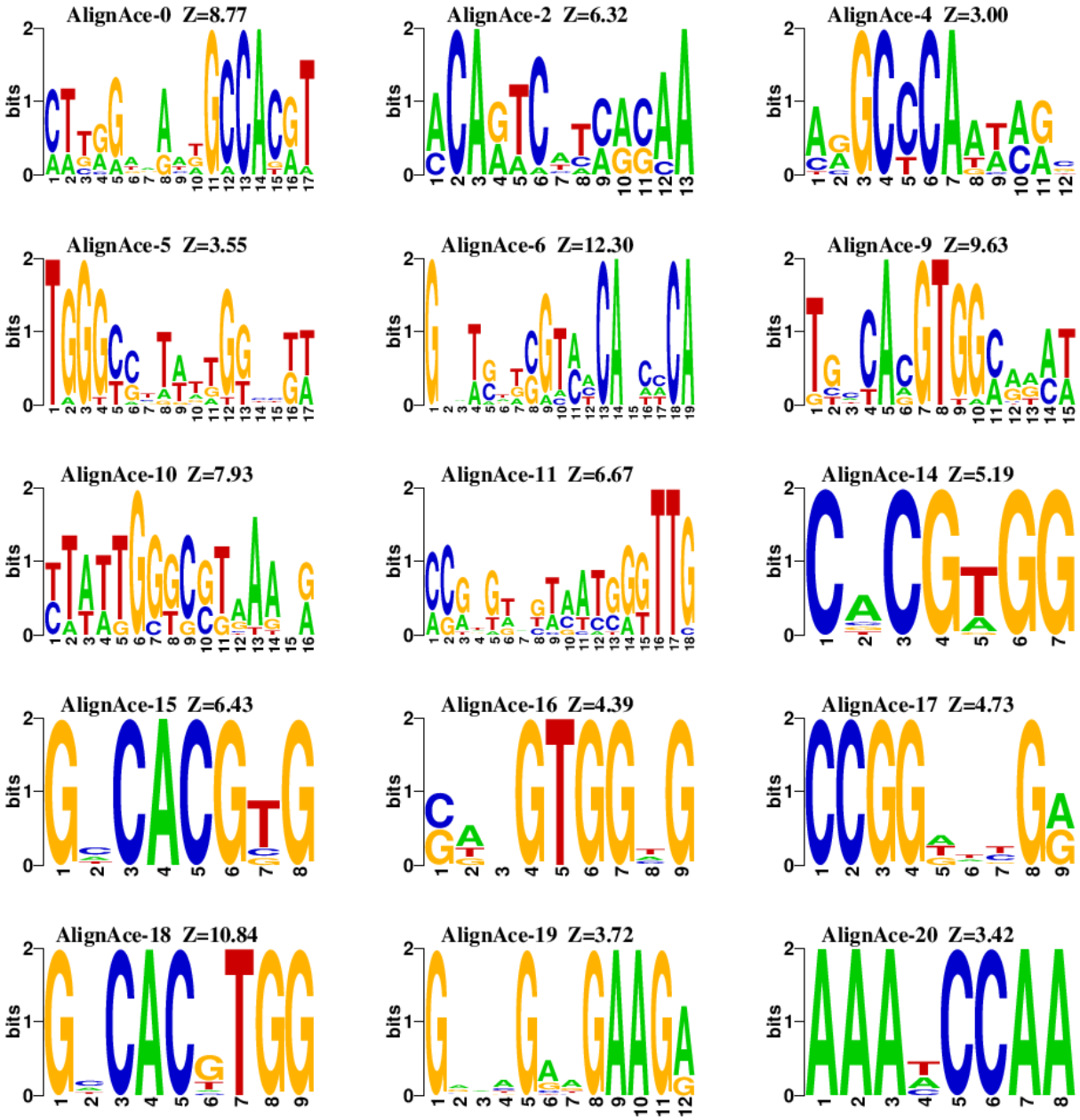



AlignAce-52 Z=4.58
AlignAce-53 Z=4.09
AlignAce-54 Z=3.66
AlignAce-55 Z=4.76
AlignAce-56 Z=4.87
AlignAce-57 Z=3.01
AlignAce-59 Z=5.28
AlignAce-60 Z=5.77
AlignAce-68 Z=5.33
AlignAce-72 Z=4.55
AlignAce-76 Z=6.91
MotifSampler-96 Z=3.18
MotifSampler-98 Z=4.75
MotifSampler-99 Z=3.85
MotifSampler-100 Z=3.99
Weeder-101 Z=3.63
Weeder-102 Z=6.60
Weeder-104 Z=3.32
bits

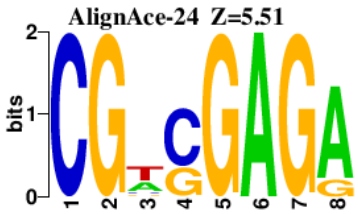
...plus 41 more.

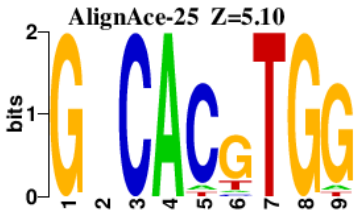
C.2 Cortex

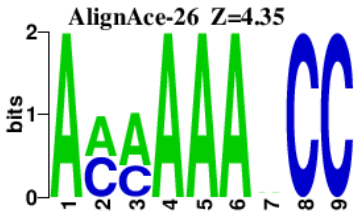
Table C.2: Cortex-specific motif sequence logos.

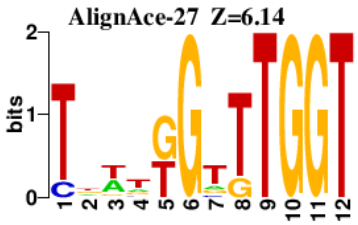



AlignAce-21 Z=3.05
bits

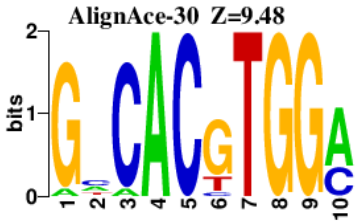

AlignAce-24 Z=5.51
bits

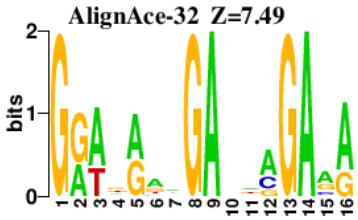

AlignAce-25 Z=5.10
bits


AlignAce-26 Z=4.35
bits

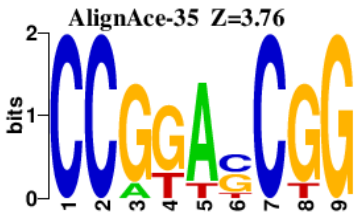

AlignAce-27 Z=6.14
bits

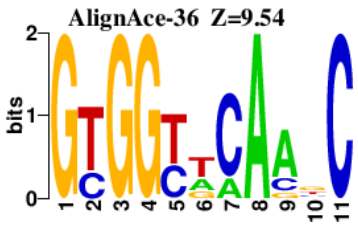

AlignAce-28 Z=5.38
bits

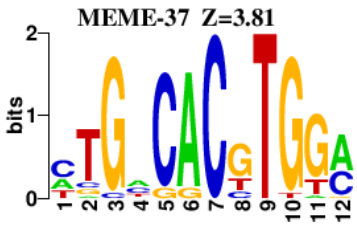

AlignAce-30 Z=9.48
bits

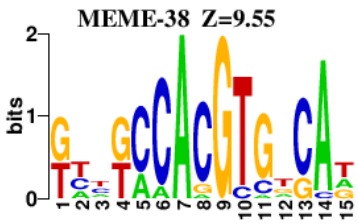

AlignAce-32 Z=7.49
bits


AlignAce-33 Z=4.81
bits

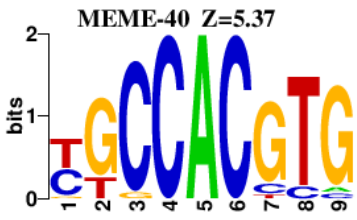

AlignAce-35 Z=3.76
bits

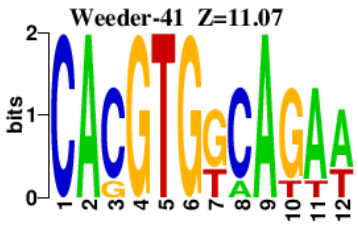

AlignAce-36 Z=9.54
bits


MEME-37 Z=3.81
bits


MEME-38 Z=9.55
bits

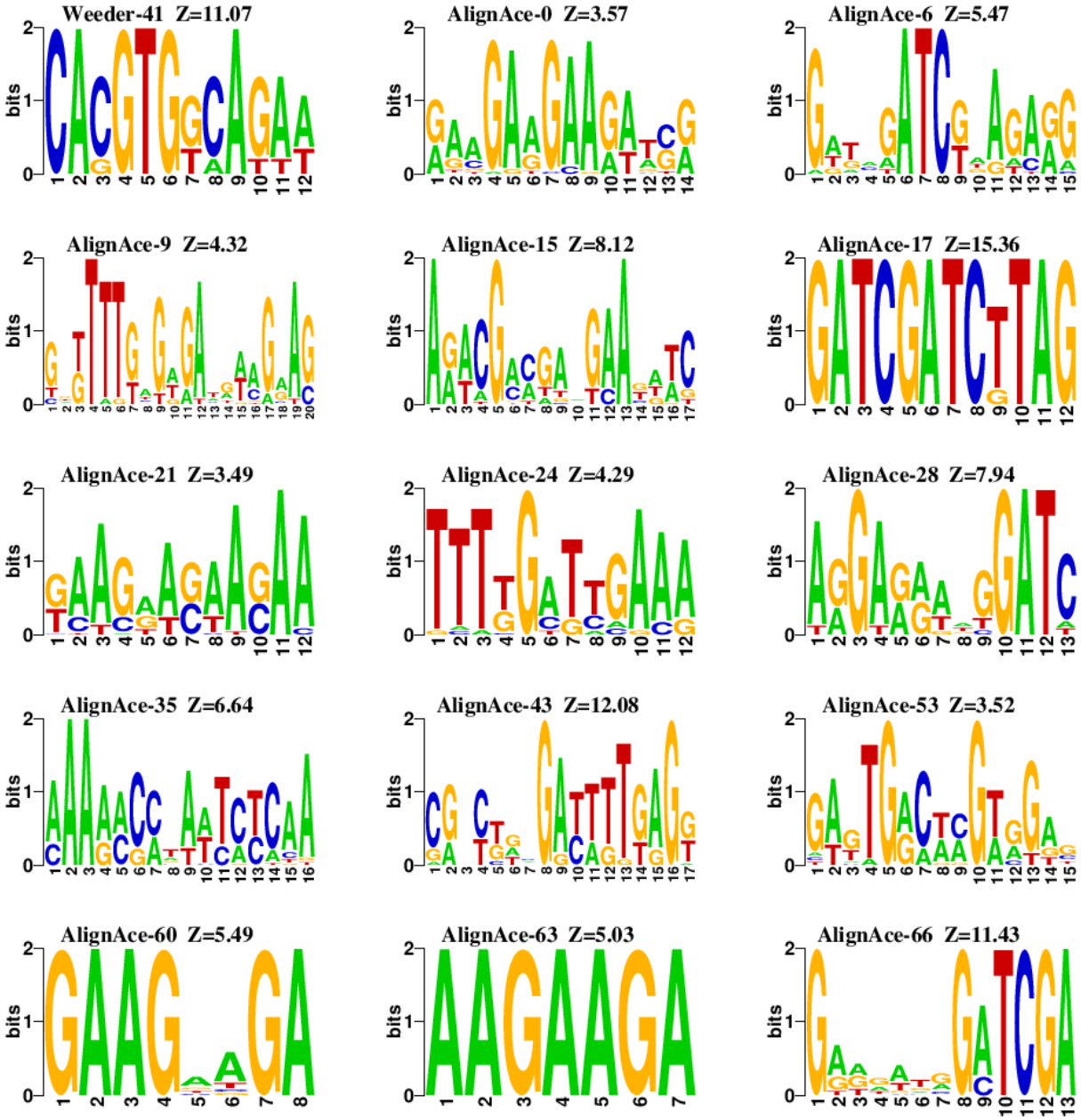

MEME-39 Z=5.25
bits

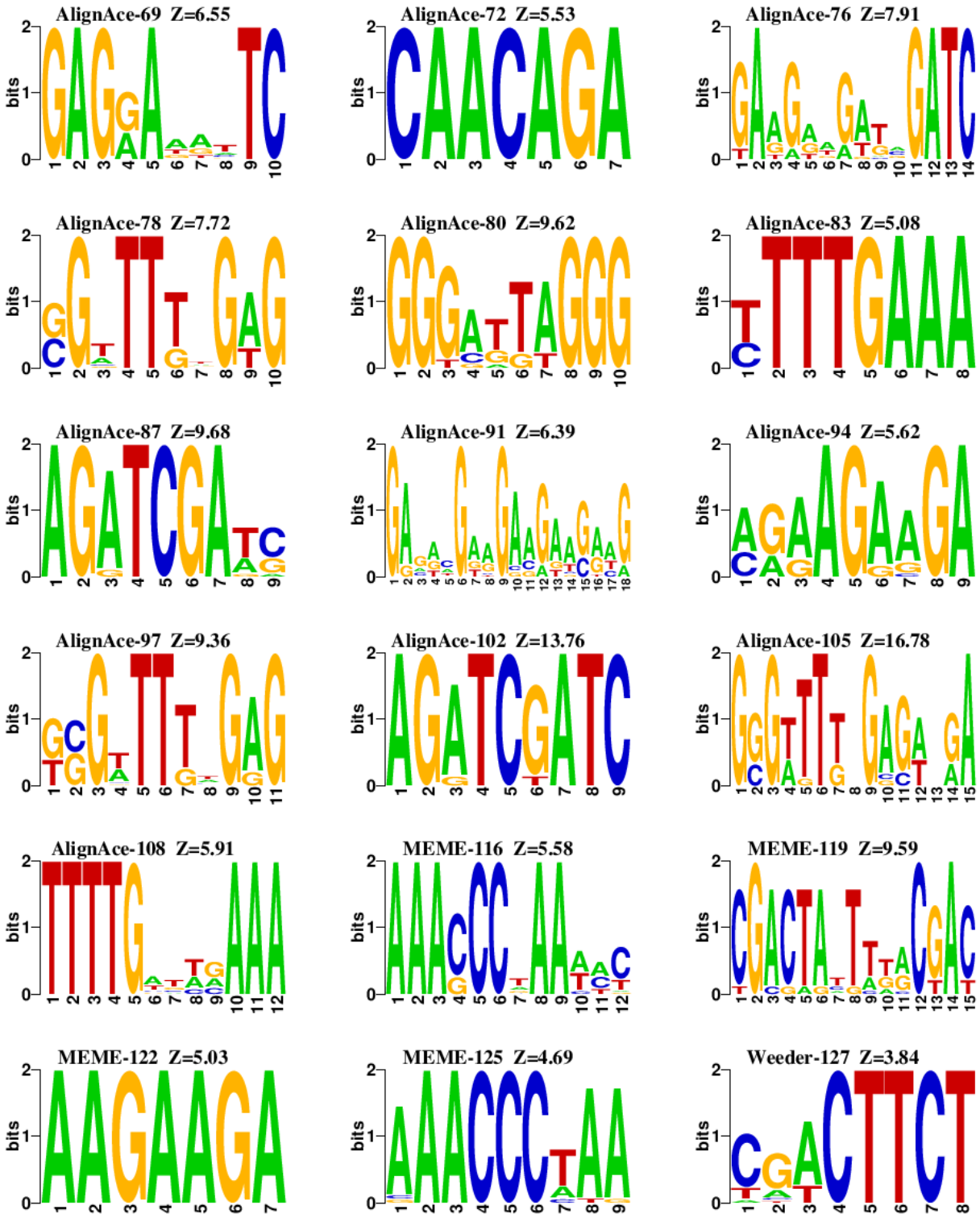

MEME-40 Z=5.37
bits


Weeder-41 Z=11.07
bits

C.3 Endodermis

Table C.3: Endodermal-specific motif sequence logos.




AlignAce-69 Z=6.55
AlignAce-72 Z=5.53
AlignAce-76 Z=7.91
AlignAce-78 Z=7.72
AlignAce-80 Z=9.62
AlignAce-83 Z=5.08
AlignAce-87 Z=9.68
AlignAce-91 Z=6.39
AlignAce-94 Z=5.62
AlignAce-97 Z=9.36
AlignAce-102 Z=13.76
AlignAce-105 Z=16.78
AlignAce-108 Z=5.91
MEME-116 Z=5.58
MEME-119 Z=9.59
MEME-122 Z=5.03
MEME-125 Z=4.69
Weeder-127 Z=3.84
bits

...plus 55 more

Curriculum Vitae

Name:	Keegan Leckie
Post-Secondary Education and Degrees:	University of Western Ontario London, Ontario Canada 2015 - 2017 M.Sc.
	University of Western Ontario London, Ontario Canada 2010 - 2015 B.Sc.
Related Work Experience:	Teaching Assistant The University of Western Ontario 2015 - 2017
	Honors Thesis Program University of Western Ontario 2014-2015
	Laboratory Assistant, Science Internship Agriculture and Agri-Food Canada, London On. 2013-1214

Scientific Posters:

Leckie K, Austin RS. Transcriptional Regulation of Cell-type specific expression in the Arabidopsis Root. 2nd Symposium on Synthetic Biology, London, Canada. July

Bergin F, Leckie K, Croft M, Austin RS. Cell-type specific chromatin dynamics in the Arabidopsis Root. 25th International Conference on Arabidopsis Research (ICAR 2014) Poster # 296, Vancouver, Canada. July 28-Aug1, 2014.

Oral Presentation:

Transcriptional control of cell-type specific expression in the Arabidopsis root. Agriculture and Agri-food Canada, London, Ontario Canada. Jan. 2017.

Efficient mapping of Agrobacterium T-DNA insertions in Arabidopsis using whole-genome resequencing data. University of Western Ontario, London, Ontario Canada. April. 2015.

2013-2014 Science Internship Research Assistant of Agriculture and Agri-Food Canada. University of Western Ontario, London, Ontario Canada. Sept. 2014