Cell type-specific transcriptomic analyses of immunity in *Arabidopsis thaliana* roots

by

Charlotte Rich

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

School of Life Sciences

September 2018

Contents

List of Figures

List of Tables

Acknowledgments

Firstly, I would like to thank my supervisors Patrick Schäfer and Sascha Ott for providing limitless support and guidance throughout my PhD. Secondly, I would like to extend my appreciation to the the Schäfer and Ott lab groups for performing the wet-lab experiments that underpinned my work and for encouraging me on a daily basis. I would like to thank Elspeth Ransom and Jessica Finch for answering all of my stupid questions and drinking thousands of cups of tea with me. Finally, I would like to thank my family, and my partner Adam, for their years of unceasing support and tireless patience.

Declarations

This thesis is presented in accordance with the regulations for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work in this thesis has been undertaken by myself except where otherwise stated.

Abstract

Plant roots represent a complex organ consisting of different cell types with highly varied functions. Thus, the response of plant roots to environmental stresses, such as pathogen infection, requires the concerted action of many cell-types. Cell type-specific transcriptomic studies are essential to understand stress resistance signalling in such a complex organ.

In this thesis, the transcriptomic response to immunity elicitation is examined at the resolution of tissues and individual cell types in two large scale RNA-seq experiments. Firstly, Fluorescence-Activated Cell Sorting combined with RNA-seq was used to produce the first high-resolution gene expression atlas of plant root immunity. The resulting data set encompassed the transcriptomes of three root cell types which had been treated by two immunity elicitors. Differential gene expression analysis revealed that both immunity elicitors induced a largely cell-type specific response with a comparatively small set of genes differentially expressed in all three cell types. This strong specificity indicates that cell identity is a strong driver of the transcriptomic immune response.

Secondly, gene expression in root tips was analysed using the single cell technique Drop-seq. Clustering methods were used to identify cells from three developmental stages and multiple cell types, and the immune responses were characterised in these tissues.

In an effort to interpret and predict immunity network regulation in different cell types, a novel tool entitled the Paired Motif Enrichment Tool (PMET) was developed to investigate gene regulation by combinatorial transcription factor groups. The tool identifies enriched pairs of known regulatory motifs within immune-responsive gene sets and revealed that each cell type/immune response combination has a largely unique regulatory landscape. Furthermore, PMET has predicted new roles of transcription factors within immunity networks.

Abbreviations

BAM	binary alignment format
CDKs	cyclin dependent kinases
DE	differentially expressed
DEGs	differentially expressed genes
DGE	differential gene expression
FACS	fluorescence-activated cell sorting
HSPs	heat shock proteins
lncRNA	long non-coding RNA
LRC	lateral root cap
miRNA	microRNA
mtRNA	mitochondrial RNA
nUMI	number of Unique Molecular Identifiers
PCA	principal component analysis
PCR	polymerase chain reaction
PCs	principal components
ptRNA	plastid RNA
QC	quiescent centre
rRNA	ribosomal RNA

scRNA-seq	single cell RNA sequencing
STAMPs	single-cell transcriptomes attached to microparticles
t-SNE	t-Stochastic Neighbour Embedding
tRNA	transfer RNA
UMI	unique molecular identifiers

Chapter 1

Introduction

1.1 Plant stress and food security

Each year crop losses due to pathogen attack are devastating worldwide. To mitigate this, one of the most important aims of crop science is to enhance plant immunity to reduce crop losses. Global agricultural losses range between 20-40% as a result of pathogens, animals, and weeds (Savary et al., 2012). This alarming statistic highlights the importance of improving plant protection from diseases in order to meet growing demand for food of good quality as well as of sufficient quantity (Strange and Scott, 2005). However, improving plant disease management becomes more complex in light of environmental, economic and social concerns, as well as the ongoing reduction of natural resources due to industry and climate change (Strange and Scott, 2005; Smil, 2001; Brown, 2012).

One of the major sources of yield loss in crops is the phenomenon of immunity-induced root growth inhibition. This reduced root growth occurs when the plant immune system is triggered and growth is inhibited independent of any direct affect that a pathogen may have on the plant. When challenged with an avirulent form of powdery mildew, field grown barley suffered a 7% reduction in grain yield because immunity was induced despite the pathogen having no adverse affect on the crop (Smedegaard-Petersen and Tolstrup, 1985).

1.2 Primary root structure and development

Roots have evolved complex tissues comprising a diversity of cell types with different functions that govern overall root functionality and provide the necessary plasticity to cope with environmental stress. On a longitudinal axis, the primary root can

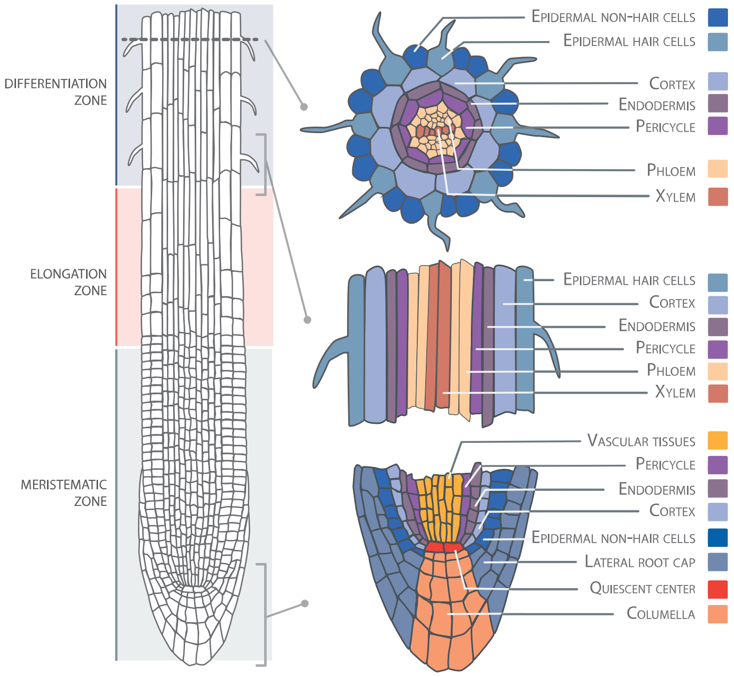


Figure 1.1: Longitudinal and transverse cross section of *Arabidopsis* root (Figure by Bouchè (2017))

be divided into three broad developmental zones: the meristematic zone at the tip of the root, the elongation zone above the meristem, and the differentiated zone which makes up the majority of the root (Figure 1.1, left). Across these zones, different cell types are organised in concentric layers of which epidermis cells form the outermost cell type (Figure 1.1, right). Epidermal cells can be further subdivided into trichoblast (root hair cells) and epidermal (non-root hair) cells. Root hairs are essential to increase water and nutrient (such as phosphorus and nitrogen compounds) uptake from the surrounding soil. The organisation of root hair and non-root hair cells is tightly controlled through patterning relative to the cortex cells beneath; if an epidermis cell is in contact with two cortex cells it will develop into a root hair otherwise it will remain a non-root hair cell. Continuing towards the centre of the root, beneath the cortex, there are endodermal cells which surround the pericycle and vascular tissues. The endodermis also forms a barrier called the Casparian strip to prevent unwanted solutes reaching the vascular tissue. Finally, the vascular tissue is composed of xylem and phloem cells which transport nutrients and water to and from the rest of the plant (Dolan et al., 1993).

Organisation of cell types is implemented by the stem cell niche in the root tip where cell fate is determined and cell types maintain their given identity throughout

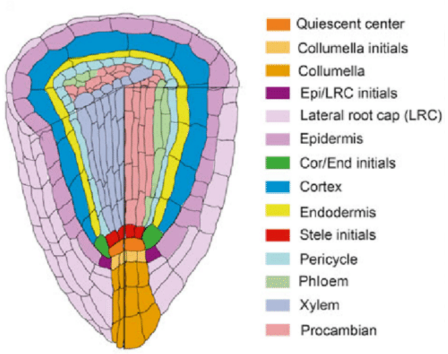


Figure 1.2: Root tip schematic reveals the organisation of 14 different cell types in the root tip (Figure from Duan et al. (2015))

their lifetime (Van den Berg et al., 1995; Sabatini et al., 2003; Wendrich et al., 2017). Under normal growth conditions cell fate is set after the first division from an initial cell based on cell position. Cells are produced from three sets of initial cells (Figure 1.2). The lateral root cap and columella are produced from the root cap initials, the cortex and endodermis (collectively known as ground tissue) originate from cortex/endodermis initials and the vasculature is produced from stele initial cells. These initial cells surround a group of two or three cells that form the quiescent centre (QC). The QC contains slowly dividing cells that replenish and maintain the initials (Dolan et al., 1993; Nawy et al., 2005).

Root growth is achieved by the combination of cell proliferation in the meristem and stem cell niche and through cell expansion in the elongation and differentiation zone. Within the meristem, growth is mediated through asymmetric cell division of stem cell initials to produce new meristematic cells whilst maintaining the stem cell niche. These new meristematic cells then divide symmetrically leading to cell proliferation between the stem cell niche and transition zone.. During the progression through the transition zone, the mitotic cell cycle is halted, halting cell division, and the cells begin to differentiate (Dolan et al., 1993; Nawy et al., 2005). The root meristem size is maintained by the antagonism of cytokinin and auxin at the boundary of the transition zone (Dello Ioio et al., 2008). The coordinated activity of cell division and expansion across developmental zones allows the establishment of a dynamic equilibrium between the dividing cells and those which differentiate, maintaining the size of the root meristem Dello Ioio et al. (2008).

Further up the root in the elongation zone, growth is generated through cell expansion. This switch from proliferation to expansion is accompanied in some plants, including *Arabidopsis*, with a switch from the mitotic cell cycle to the endocycle. The endocycle is a short-circuited version of the mitotic cell cycle, which stops short of cell division. It is the mechanism by which the genome is doubled within a cell, referred to as endoploidy (De Veylder et al., 2011; Breuer et al., 2014). Increased endoreduplication is correlated with increased cell size and is observed to occur up to three times in root cells (Sablowski and Carnier Dornelas, 2013; Bhosale et al., 2018). Unusually, compared to other organisms, endoreduplication in plants does not occur in all cells equally resulting in different sub-populations of cells with varied DNA content. This is particularly marked between different root cell files, which enter the endocycle at different points in development, and undergo endocycles at different rates (Bhosale et al., 2018). Based on the observation that cell files typically enter the endocycle prior to cell expansion, Bhosale et al. proposed that endoreduplication might prepare cells to cope with cell expansion by inducing cell wall modifications. They suggested that increasing the copy number of genes could help cells cope with the sudden increased demand for cell wall components.

1.3 Root growth inhibition

Root growth inhibition induced by activation of defence responses has been traditionally attributed to a plant simply reallocating resources away from development and growth into immunity. However recent research suggests that resource reallocation does not entirely explain growth inhibition. Firstly, a large study of many *Arabidopsis* accessions revealed than there was little correlation between nitrogen or carbon limitation and defence capability (Kleessen et al., 2014). Secondly, it was shown that immune elicitation in response to chitin (a fungal immune elicitor) does not affect growth, unlike treatment with bacterial elicitors flg22 and Ef-tu (Wan et al., 2008; Petutschnig et al., 2010). This suggests that root growth inhibition is instead the result of complex interactions between immunity and development.

Eichmann and Schäfer (2015) proposed an alternative model whereby interactions between immunity hormone networks and the cell cycle/endocycle mediate root growth inhibition. Specifically, they propose that the repression of gibberellic acid (GA) and increase of jasmonic acid (JA) signalling relieves the repression of DELLA proteins, which through a sequence of activation and repression of key cell cycle genes ultimately halts the cycle, inhibiting growth. Consistent with a strong link between stress and development, Bhosale et al. (2018) demonstrated that transcripts of genes that correlated strongly with endoploidy were good markers to predict the impact of stress on endocycling. Studying these molecular mechanisms within the complex structure of the root is essential to understand how the activation of immunity affects cell identity, and therefore growth and development.

1.4 Root cell type-specific transcriptomics

Within the complex structure of the root, each cell type is specialised making the root a perfect system to study cell type-specific responses (Benfey and Scheres (2000)). The study of gene expression has accelerated within the last few decades with the advent of microarrays and particularly RNA sequencing (RNA-seq). The RNA-seq protocol (utilised in Chapters 4 and 5 of this thesis) sequences all of the mRNA (and other populations of RNA if desired) in a sample, enabling the scientist to monitor changes in gene expression (amount of mRNA that is transcribed for a particular gene) over time, or determine differences between different groups or treatments. Recent advances have improved the resolution gene expression studies enabling the capture of the transcriptomes of a single tissue or cell.

The most widely applied method to study gene expression in individual tissues is to perform transcriptomic studies (either microarray or RNA-seq) on cells sorted using fluorescence-activated cell sorting (FACS) as developed by Birnbaum et al. (2003). This method uses fluorescent root cell type marker lines to isolate individual tissues prior to either microarray or RNA-seq. These transcriptomic studies are used to relate gene activity to cell fate and tissue specialisation. Birnbaum et al. (2003) produced the first gene expression map of the *Arabidopsis* root encompassing five cell types, and three developmental zones. This first study revealed that each cell type is characterised by a specific transcriptional identity. Furthermore, it revealed that regulation of hormones can be mapped to different cell types and developmental zones. This data set was later expanded to much higher resolution (Brady et al., 2007). Brady's extremely high resolution study aimed to understand the spatial and temporal control of transcriptional complexity in the root. They examined expression patterns in 14 non-overlapping cell types, across 15 developmental zones (cut along the longitudinal axis to contain approximately three cell layers per zone). GO term analyses were used to assign putative functions each cell type and highlighted the extent of cell specialisation. Testing for enrichment of cis-regulatory regions (CREs) also revealed the cell type-specificity of regulatory mechanisms. Identifying putative CREs is essential to understand cell type-specific regulation, as 5' upstream non-coding sequences control the major patterns of gene expression in the root (Lee et al., 2006).

The studies by Birnbaum et al. (2003) and Brady et al. (2007) revealed that regulation of cell identity in roots is complex under non-stressed conditions. However, roots in natural environments undergo constantly changing conditions. Therefore, stress responses must also be integrated into this complex system. Comparing the cell type-specific gene expression profiles generated by Birnbaum et al. (2003) to known stress responsive genes revealed that cell type-specific regulation of stress responses was highly likely (Ma and Bohnert, 2007). In order to realise the extent of specificity of stress responses, an array of FACS-based transcriptomics studies were performed in order to examine salt stress (Dinneny et al., 2008; Geng et al., 2013), iron deprivation (Dinneny et al., 2008) nitrogen depletion (Gifford et al., 2008) and varied pH levels (Iyer-Pascuzzi et al., 2011). Dinneny et al. (2008), Gifford et al. (2008) and Geng et al. (2013) showed that cell identity influences stress responses to iron deprivation, low nitrogen and high salinity, in a cell type- and developmental zone-specific manner. A meta-analysis of these studies by Iyer-Pascuzzi et al. (2011), revealed that in addition to these responses being cell type-specific, they were also treatment-specific as there was no universal stress response at cell type resolution. However, the study did show that some biological responses were consistent across multiple treatments. Iyer-Pascuzzi et al. (2011) also showed that cell identity markers remained highly expressed in all stress conditions, indicating that maintaining identity is a priority. Studies of cell type-specific transcriptomics based on FACS have also significantly advanced the knowledge of processes regulating root development (Birnbaum et al., 2003, 2005; Bargmann et al., 2013; Walker et al., 2017). Evidence provided in these studies have highlighted the functional individuality of cell types and the significance of a coordinated regulation of cell type-specific gene networks to master root development and secure overall root functionality (e.g. growth) under stress.

These previous studies all focussed on abiotic stresses, and to our knowledge, there has been no cell type resolution study of biotic stress responses in *Arabidopsis* roots. Based on these previous studies, it is expected that gene expression under immunity will also be regulated on a cell type-specific level. Indeed, initial evidence by Beck et al. (2014); Wyrsch et al. (2015) and Poncini et al. (2017) reveals that this is the case (see section 1.5.1). However, the gene networks underlying this cell type-specificity have yet to be elucidated. Studying both abiotic and biotic stress on a cell type-specific level is the key to understanding how roots control the stress response and balance the needs to protect and maintain normal cell function and identity.

1.5 Plant innate immunity

Unlike animals, the plant immune system is not composed of specialised immune cells. Instead, plant cells must be able to mount an immune response in addition to that cell's normal function. In order to recognise immune threats, plant cells possess a huge range of receptor kinases and receptor-like proteins (RLPs) at the plasma membrane to recognise pathogen-associated molecular patterns (PAMPs) from pathogens (or host-derived danger-associated molecular patterns (DAMPs)) which then activate an immune response (Figure 1.3). These receptors are referred to as pattern recognition receptors (PRRs). Receptor kinases typically possess an ectodomain potentially involved in ligand binding, a single transmembrane domain and an intracellular kinase domain. RLPs are similarly composed, except they lack a kinase domain. As such RLPs rely on the recruitment of co-factors to initiate downstream signalling (Zipfel et al., 2004; Fritz-Laylin et al., 2005; Couto and Zipfel, 2016).

Leucine-rich repeat (LRR)-containing PRRs such as FLAGELLIN SENSING 2 (FLS2), EF-TU RECEPTOR (EFR) and Pep receptors (PEPRs) preferentially bind proteins or peptides, such as bacterial flagellin or elongation factor Tu (EF-Tu), or endogenous AtPep peptides, respectively (Figure 1.3, Gómez-Gómez and Boller (2000); Zipfel et al. (2004, 2006); Yamaguchi et al. (2006)). These PRRs form heteromeric complexes with cofactors such as BRI1-ASSOCIATED RECEPTOR KINASE (BAK1) or SOMATIC EMBRYOGENESIS RECEPTOR KINASES (SERKs) to trigger downstream signalling (Chinchilla et al., 2007; Heese et al., 2007). By contrast, PRRs containing lysine motifs such as CHITIN ELICITOR RECEPTOR KINASE 1 (CERK1) and LysM-CONTAINING RECEPTOR-LIKE KINASE 5 (LYK5) form heteromeric complexes and bind carbohydrate-based ligands, such as fungal chitin or bacterial peptidoglycan. Finally, lectin-type PRRs bind extracellular ATP or bacterial lipopolysaccharides (LPS) (Ranf et al., 2015; Couto and Zipfel, 2016). These diverse receptors form the first layer of the plant immune system, referred to as pattern triggered immunity (PTI). PTI effectively repels most-non adapted pathogens due to the conserved nature and variety of the PAMPs recognised by PRRs. Perception of PAMPs by PTI receptors triggers signalling cascades that result in antimicrobial responses that limit pathogen infection, often at the expense of plant growth (Zipfel et al., 2004; Boller and Felix, 2009).

The next level of pathogen detection occurs within the cell. In order to get around plant PTI, many pathogens have evolved small 'effector' proteins that are secreted into cells and interact with cellular components in a multitude of ways

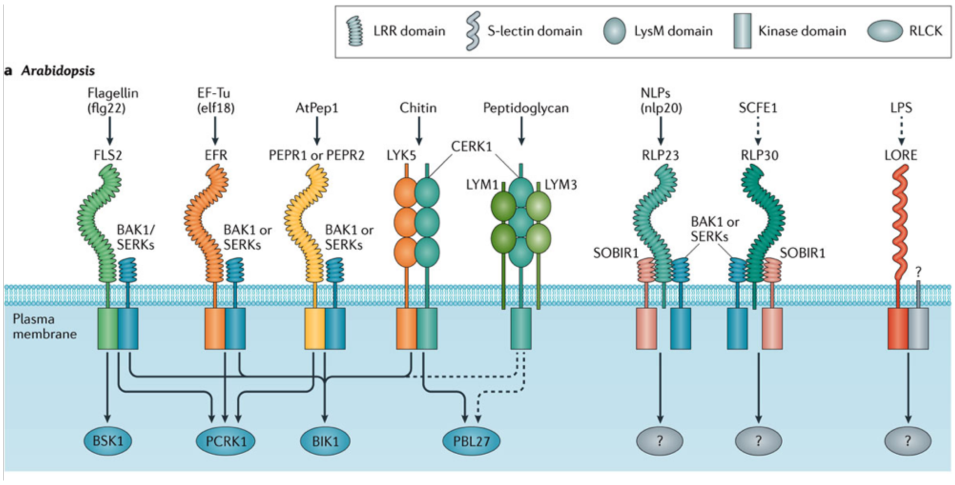


Figure 1.3: Diverse pattern recognition receptors respond to a variety of PAMPs from bacteria (flg22, Ef-Tu, LPS, and peptidoglycan) and fungi (chitin, sclerotinia culture filtrate elicitor1 (SCFE1) and Nin-like proteins (NLPs). Additional PRRs recognise endogenous DAMPs released in response to damage (AtPep1). Figure from Couto and Zipfel (2016).

to circumvent the immune response. In turn, plants evolved effector-triggered immunity (ETI), whereby these effectors are recognised by Nod-like receptor (NLR) proteins, and trigger the immune response. This co-evolutionary dynamic between pathogens and plants can be described using the zig-zag model of immune activation (Jones and Dangl (2006), Figure 1.4).

1.5.1 Flagellin perception in *Arabidopsis*

In the model organism *Arabidopsis thaliana*, defence against bacteria depends on perception of bacterial flagellin by the receptor FLS2 of which the active epitope is a 22-amino acid peptide called flg22 (Felix et al., 1999).

Studies on *Arabidopsis* leaves revealed that upon detection, FLS2 binds to the co-receptor BAK1 (Chinchilla et al., 2007) leading to an array of PTI responses including the rapid production of reactive oxygen species (ROS burst), Ca^{2+} signalling, MITOGEN-ACTIVATED PROTEIN KINASE (MAPK) phosphorylation and induction of immunity genes to stop pathogen infection (Felix et al., 1999; Gómez-Gómez et al., 1999; Asai et al., 2002; Zipfel et al., 2004; Chinchilla et al., 2007; Jeworutzki et al., 2010). MAPK signalling cascades activate a range of transcription factors including WRKY transcription factors (TFs) (Asai et al., 2002). In turn WRKYs act as both positive and negative regulators of the defence response network Pandey and Somssich (2009). PAMP recognition in leaves also triggers

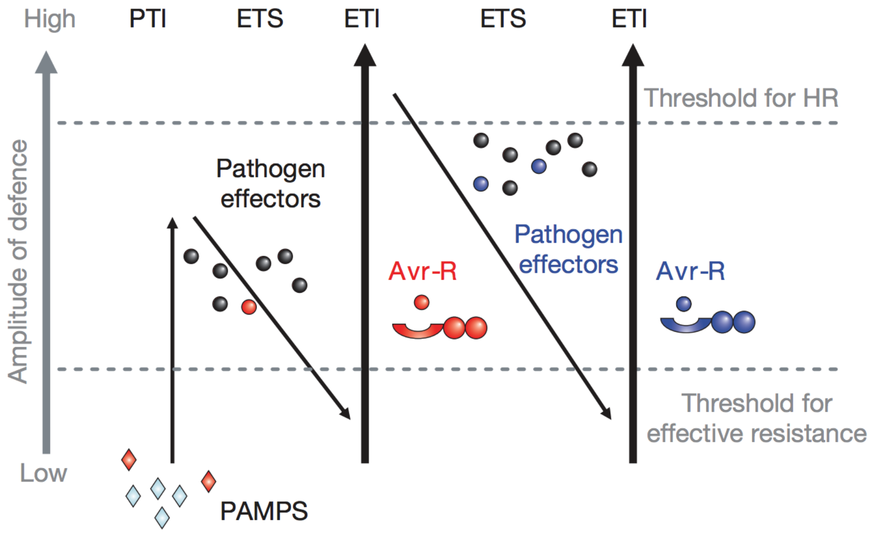


Figure 1.4: Zig-zag model of immune activation demonstrates the layers of immunity in plants, from Jones and Dangl (2006)

callose deposition which accumulates at the site of pathogen penetration and is believed to provide a physical barrier to pathogen attack (Aist and Bushnell, 1991). Overall, rapid recognition of flg22 by plants enhances the plant's ability to resist bacterial invasion (Zipfel et al., 2004). However, this enhanced resistance comes at a cost, as flg22 perception also leads to root growth inhibition (Gómez-Gómez et al., 1999; Jacobs et al., 2011).

The plant root constantly interacts with microbes in the rhizosphere. This constant interaction initially led to the belief that roots would not respond to flagellin. However, various studies have demonstrated that roots as a whole organ respond strongly to flg22, activating flg22-dependent downstream MAPK phosphorylation, the production of reactive oxygen species (ROS), the induction of defence marker genes, the production of anti-microbial compound camalexin, and callose deposition (Millet et al., 2010; Jacobs et al., 2011; Beck et al., 2014; Wyrsch et al., 2015). Millet et al. (2010) observed that the flg22 responsive genes were particularly active in the elongation zone, but PAMP-triggered callose deposition was observed along the whole root length. The complexity of flg22 responses in the root was further elucidated by studies of the expression patterns of *FLS2*, and signalling in different cell types.

Beck et al. (2014) showed the *FLS2* was expressed in a dynamic stress-responsive manner. In particular, under non-stressful conditions, *FLS2* expression is largely restricted to the stele and root cap, whereas this expression pattern is expanded outwards to encompass the entire root under different stress conditions including the perception of flg22. This dynamic response could be mediated by an internalisation mechanism (Robatzek et al., 2006). Furthermore, *FLS2* was shown to be expressed across all developmental zones, in a cell type-specific manner (Poncini et al., 2017).

In contrast to Beck et al. (2014) which examined the expression of *FLS2* under it's native promoter, Wyrsch et al. (2015) demonstrated the ability of individual root cell types to respond to flg22 using *FLS2* promoter fusions to cell type-specific promoters. Wyrsch et al. (2015) showed that when FLS2 is expressed in a cell type-specific manner, each cell type can mount a response to flagellin (demonstrated by a ROS burst and MAPK phosphorylation in the stele, pericycle, endodermis, and epidermis). These combined results demonstrate that FLS2 is both expressed (after induction by a stress) and can trigger both oxidative bursts and induce defence genes in all root cell types..

Wyrsch et al. (2015) also showed that the intensity of these responses varied between cell types in terms of the strength of ROS production, MAPK phosphorylation and defence gene activation implying that individual root cell types all have their own PAMP perception sensitivity and immune signalling competence. They suggested that these differences in PAMP perception contributed to "proper balance of defence responses according to the expected exposure to elicitors".

1.5.2 Danger-associated molecular patterns

In addition to PTI, plants can also activate immune responses using endogenous elicitors. These DAMPs such as Pep1 and it's homologs Pep2-7 (Bartels et al., 2013) are produced endogenously in response to damage or danger signals, including the perception of PAMPs. Pep1, a 23-amino acid peptide encoded by *PROPEP1*, signals through the plasma membrane receptors PEPR1 and PEPR2 (Yamaguchi et al., 2006; Krol et al., 2010), initiating MAPK and Ca^{2+} signalling, leading to the further induction of immunity genes, amplifying the plant's response to pathogens and other stresses (Qi et al., 2010; Bartels et al., 2013).

Pep1 has been suggested to act as an amplifier of defence responses as Pep-dependent signalling increases host resistance to bacterial and fungal pathogens, and also protect against herbivory (Huffaker et al., 2011, 2013; Tintor et al., 2013; Klauser et al., 2015). Pep1 is interpreted as a much stronger alarm signal than flg22, consistent with an amplification role (Poncini et al., 2017).

Pep1 signalling mechanisms overlap strongly with flg22 signalling. Like FLS2, the PEPRs interact with BAK1, and are likely stabilised by BOTRYTIS-

INDUCED KINASE 1 (BIK1) (Liu et al., 2013). Both flg22 and Pep1 initiate rapid increase of cytosolic Ca^{2+}, induce production of NO and ROS and signal through MAPK phosphorylation (Flury et al., 2013; Tintor et al., 2013). Above the cellular level, both treatments trigger callose deposition and inhibit growth. However, there are key differences in the ways by which flg22 and Pep1 activate these responses. For example, Pep1 has been shown to promote the influx of extracellular calcium, whereas flg22 triggers the release of Ca^{2+} from intracellular stores. Pep1 and flg22 also interact with different hormone pathways. Both peptides trigger the synthesis of ethylene (ET) in *Arabidopsis*, the two peptides increase the levels of antagonistic hormones SA and JA; flg22 perception elevates SA levels, whereas Pep1 slightly increases JA (Mishina and Zeier, 2007; Flury et al., 2013). Finally, Pep1 induces a much stronger root growth inhibition phenotype than PAMPs such as flg22 (Ma et al., 2014; Krol et al., 2010; Poncini et al., 2017). This could be consistent with the Peps' roles as amplifiers but also could implicate Peps in development.

Further evidence that Peps are involved in immunity include the fact that *atpepr1* and *atpepr2* knock-outs show a shorter root phenotype than the wild-type (Ma et al., 2014). Ma et al. states that this evidence suggests that these receptors could play positive roles in root growth, implying that biologically derived Pep1 should positively regulate root growth. However, root growth inhibition is observed at all concentrations of Pep1 treatment. They suggest it is possible that any root growth promotion by biologically derived Pep1 would occur at a cell type-specific level, based on the cell type-specific expression of different *PROPEP*s (Birnbaum et al., 2003; Brady et al., 2007). There have also been links made between *PROPEP*s and other developmental processes (reproduction and senescence, Yamaguchi et al. (2010); Gully et al. (2015)).

Poncini et al. (2017) demonstrated that, as with flg22, the whole root raises an immune response to Pep1. They used a luminol-based assay to measure the oxidative burst in response to flg22, chi7 (a synthetic peptide from chitin) and Pep1. The burst was stronger in response to flg22 and Pep1 than chi7. They also showed that PTI-associated MAPKs were phosphorylated (a marker of PTI signalling) in response to all three elicitors. Furthermore, analysis of knock-out mutants proved that these elicitors were signalling through the same receptors in roots as in leaves. Poncini et al. (2017) used fluorescent immunity-trigger marker genes to investigate the tissue specificity of elicitor reponses in the roots. In terms of Pep1, this revealed that an immune response is triggered across all developmental zones, however this response was not uniform. The Pep1 response was particularly strong in the transition zone and differentiated zone, and weaker in the meristem.

The root cap displayed the weakest response to Pep1. The marker gene expression also varied at the cell type level. For example the *MYB51* and *ZAT12* reporters were expressed more strongly in the stele compared to endodermis and cortex. The increased response in the stele could indicate that Pep1 effects transport of sugars, perhaps in order to redirect resources towards defence responses in the root. These differences indicate that Pep1 networks are likely to be regulated at the tissue and cell type levels.

1.6 Transcriptional regulation of gene expression

Transcription is the process by which mRNA is transcribed from DNA, prior to protein synthesis. The amount of mRNA that is transcribed, otherwise referred to as expression level, is largely controlled through regulation of transcription initiation, where RNA polymerases are recruited to transcribe mRNA. RNA polymerases are recruited to the region of DNA slightly upstream of the transcription start site (TSS). Upon recruitment, RNA polymerases move downstream and start to transcribe DNA into mRNA at the TSS. The activation of transcription relies on the recruitment of TFs which bind to the region upstream of the TSS called the promoter. Promoters can be several hundred kilobases long and contain a large number of TF binding sites.

TFs activate transcription via a range of processes including recruiting polymerases and other transcriptional activators to the promoter, or by altering chromatin structure to a more open state, making it amenable for transcription. Other TFs act as repressors of transcription. These repressors often bind elsewhere in the promoter and affect binding of activators indirectly, through mechanisms such as making activator binding less thermodynamically favourable through structural changes to the promoter (Ezer et al., 2014).

Many factors contribute to TF-DNA binding including nucleotide structure, the 3 dimensional structure of the DNA, the presence of cofactors, chromatin accessibility and other epigenetic markers (reviewed in Slattery et al. (2014)). These binding sites, otherwise known as motifs, can be represented in a variety of ways, one of the most common is the position specific motif matrix (PSSM) (Stormo, 2000). PSSMs encapsulate the nucleotide sequences of consensus binding sites and the relative conservation of individual bases. PSSMs are created by collating all the potential binding sequences for a single TF and expressing the conservation of each base in terms of the likelihood of observing that base at each position. The conservation of individual bases is quantified in terms of information content which can

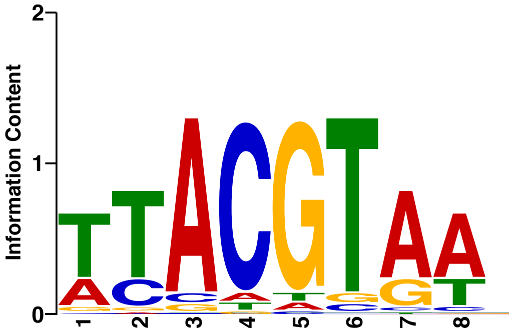


Figure 1.5: Example of a sequence logo visualisation of a motif

be visualised as sequence logos (Figure 1.5). The height of each base represents the conservation at that position. Whilst other models can include more information such as 3 dimensional structure, the PSSM is a simple and accurate representation of TF binding motif.

TF binding motifs in Arabidopsis have been identified using techniques such as DNA affinity purification sequencing (DAP-seq) and protein binding microarrays (PBMs) (Franco-Zorrilla et al., 2014; O'Malley et al., 2016). These studies revealed that binding sites for an individual TF can occur hundreds of times across a genome and that some TFs can recognise multiple motifs (Badis et al., 2009; Franco-Zorrilla et al., 2014). Furthermore, different TFs have varied intrinsic binding affinities for their cognate motifs.

Regulatory networks can take advantage of this variability in order to better control transcription. For example, high affinity binding sites can be utilised to maintain high concentrations of protein in one cell type, whereas low affinity binding sites in another cell type would results in low protein levels.

In addition to variability within motifs, regulatory networks can achieve highly specific regulation through the presence of multiple motifs. These can be homotypic (group of adjacent binding sites for the same TF) or heterotypic (multiple different TF binding sites). Homotypic clusters of motifs are found in bacterial and eukaryotic promoters, as well as in eukaryotic CREs. Ezer et al. reviewed three mechanisms by which homotypic clusters can affect gene expression, through direct or indirect cooperation, or through non-cooperative binding. Direct cooperation indicates that TFs bind as homodimers to closely spaced or even overlapping binding sites. Indirect cooperation suggests a model where two proteins do not interact but do influence each other's binding by, for example, stabilising each other's binding

(Wasson and Hartemink, 2009). Finally the non-cooperative model of homotypic clusters suggests that by simply having more binding sites, the chance of a single TF binding is increased and in fact multiple bindings could occur.

In addition to homotypic clusters, combinations of TFs (or heterotypic clusters) have been shown to represent a key principle in regulating gene networks in both animals and plants. Specific combinations of TFs can act through the same mechanisms as homotypic clusters, through direct contact and by indirectly enabling binding of the other TF. This represents another mechanism by which gene networks can achieve a high degree of specificity in signalling. In particular, combinatorial TFs can be used to fine-tune tissue- and cell-specific signalling (Halfon et al., 2000; Junion et al., 2012).

A genome-wide study of *Arabidopsis thaliana* revealed combinatorial TF motifs in a wide variety of biological processes including the cell cycle, light response and protein biosynthetic pathways (Vandepoele et al., 2006). Furthermore, TF combinatorics have been shown to perform an essential role in regulating gene networks in Arabidopsis including immunity and hormones (Michael et al., 2008; Achard et al., 2009; Van de Velde et al., 2014; Lewis et al., 2015). In a subsequent piece of work, Vandepoele highlighted the importance of integrating co-expression, gene ontology and motif data to enable understanding of gene networks, particularly referencing the identification cooperative elements within promoters (Vandepoele et al., 2009).

1.7 Organisation of this thesis

As demonstrated in this introduction, the regulation of immunity in *Arabidopsis thaliana* is incredibly complex, encompassing many levels of complexity. This thesis contains a chapter detailing on the development of a tool to analyse combinatorial TF binding sites and two chapters containing the results of two large experiments that examined the impact of immune elicitation on plant roots at the cell type and tissue level.

Following this introduction, Chapter 2 contains the materials and methods used to produce the work detailed in this thesis. The chapter contains summaries of the experimental methods and comprehensive descriptions of the data analysis methods used to analyse the experiments described in subsequent chapters. Chapter 3 details the development and implementation of a novel computational method designed to identify enriched pairs of motifs with sets of co-expressed promoters. This software, entitled the Paired Motif Enrichment Tool (PMET), uses a similar statistical approach to gene ontology (GO) enrichment analyses to predict potential regulators of co-expressed gene sets. Chapter 3 also contains a parameter sensitivity analysis, which examines the relative influence that various PMET parameters have on the statistical tests performed by the tool. This sensitivity analysis utilised gene sets that were identified later in Chapter 4.

Chapters 4 and 5 contain the results of two experiments that investigated the impact of immune elicitation on *Arabidopsis* roots at a cell type-specific resolution. In Chapter 4, RNA-seq analysis was used to discover cell type-specific immune responsive gene networks in three root cell types. Epidermis, cortex and pericycle cells that were treated with flg22, Pep1 or a mock treatment were isolated using FACS. Differential gene expression analysis was used to identify flg22 and Pep1 responsive genes in the epidermis, cortex and pericycle. Then potential cell type-specific regulatory networks were identified using PMET.

In Chapter 5, the new single cell RNA sequencing (scRNA-seq) method Drop-seq was to analyse the transcriptomes of mock- and flg22 treated root tips. Cell populations clustered using unsupervised methods were matched to specific cell types and developmental stages using a combination of known and novel marker genes. Finally, the flg22 response in the identified populations was analysed using differential gene expression analysis.

Chapter 2

Materials and Methods

2.1 Materials and methods for root cell type-specific RNA-seq

2.1.1 Plant growth and treatment

The following plant lines were utilised in this experiment:

- **Col-0** Wild-type *Arabidopsis thaliana* accession Columbia, I.D. N60000 obtained from the Nottingham A. thaliana Stock Centre (NASC, UK)
- ***pCORTEX:GFP*** *Arabidopsis thaliana* marker line expressing GREEN FLUORESCENT PROTEIN (GFP) fused to the promoter of *CORTEX*, a cortex specific marker gene.
- **E3754** *Arabidopsis thaliana* marker line that expresses GFP in pericycle cells adjacent to the xylem-pole.
- ***pGL2:GFP*** *Arabidopsis thaliana* marker line expressing GFP fused to the promoter of *GLABRA2 (GL2)*, an epidermis (non-root hair) specific marker gene.

Seeds of *pGL2:GFP*, *pCORTEX:GFP*, and E3754 lines were obtained from Miriam Gifford, University of Warwick, UK.

pGL2:GFP, *pCORTEX:GFP* and E3754 plants were sewn on vertical square Petri dishes containing *Arabidopsis thaliana* salt (ATS) medium (Lincoln et al., 1990) without sucrose and supplemented with 4.5 g L^{-1} Gelrite (Duchefa Biochemie), stratified for 1 day and then grown in a 22°C day/18°C night cycle (8 hour light) at

120 μmol m^{-2}s^{-1}. After 10 days, plants were treated on plates with 1 mL per plate of 1 μM solutions of flg22, Pep1 or H_2O as control for 1 hour prior to harvesting. Pep1 and flg22 peptides were used as described in Gómez-Gómez et al. (1999) and Krol et al. (2010). Three independent biological experiments were carried out for each marker line.

2.1.2 Separation of root cell types by FACS

Whole roots were cut into pieces and then incubated in protoplast solution for 1 hour. Protoplasts were filtered through 70 and then 40 μm cell strainers, centrifuged at 300 g for 3 minutes, resuspended in protoplast solution lacking cell wall-degrading enzymes and subjected to fluorescence-activated cell sorting (FACS). GFP-expressing protoplasts were collected by using BD Influx cell sorter (BD Biosciences), following previously published protocols (Birnbaum et al., 2003; Gifford et al., 2008; Grønlund et al., 2012). The cell sorter was equipped with a 100 μm nozzle and BD FACS-FlowTM (BD Biosciences) was used as sheath fluid. BDTM Accudrop Fluorescent Beads (BD Biosciences) were used prior to each experiment to optimize sorting settings. A pressure of 20 psi (sheath) and 21 – 21.5 psi (sample) was applied during experiments. Drop frequency was set to 39.2 kHz, and event rate was generally kept less than 4000 events per second. GFP-expressing protoplasts were identified using a 488 nm argon laser, plotting the outcome of a 580/30 bandpass filter versus a 530/40 bandpass filter to differentiate between green fluorescence and autofluorescence. Different cell populations were collected for microscopy in pre-experiments to determine the presence of GFP-expressing protoplasts. As previously published by Grønlund et al. (2012), these protoplasts were present in the high 530 nm / low 580 nm population. Sorting gates were set conservatively in following experiments based on these observations. For RNA-extraction, GFP-expressing protoplasts were sorted into Qiagen RLT lysis buffer containing 1% (v:v) β-mercaptoethanol, mixed, and immediately frozen at -80°°C. At least 10000 GFP-expressing protoplasts were sorted per experiment and treatment condition. Sorting times were kept below 25 minutes.

2.1.3 RNA isolation, RNA-seq library construction and RNA sequencing

Total RNA was extracted using the Qiagen RNeasy Plant Mini Kit including on-column DNase treatment with the Qiagen DNase kit. The 6000 Pico Kit (Agilent Technologies) was used to check quantity and quality of the RNA on a Bioanalyzer

2100 (Agilent Technologies). Preparation of amplified complementary DNA (cDNA) from total RNA and library construction were done with the Ovation® RNA-seq System V2 and Ovation® Ultralow Library Systems Kit (NuGEN Technologies), respectively, following standard protocols. 100 bp Paired end sequencing was carried out by the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics on an Illumina HiSeq2500 System.

2.1.4 Experimental replicate pooling

Cell-type specific protoplasts were processed in six FACS experiments, and sequenced in pools by replicate within cell types as can be seen in Table 2.1.

Table 2.1: RNA-seq sample names and processing pools.

Sample	Cell	Treatment	FACS	Pool	Replicate
WTCHG_129187_01	cortex	mock	1	1	1
WTCHG_129187_03	cortex	flg22	1	1	1
WTCHG_129187_05	cortex	Pep1	1	1	1
WTCHG_131167_01	pericycle	mock	1	2	1
WTCHG_131167_03	pericycle	flg22	1	2	1
WTCHG_131167_05	pericycle	Pep1	1	2	1
WTCHG_129189_01	cortex	mock	2	3	2
WTCHG_129189_03	cortex	flg22	2	3	2
WTCHG_129189_05	cortex	Pep1	2	3	2
WTCHG_125416_01	pericycle	mock	2	4	2
WTCHG_125416_03	pericycle	flg22	2	4	2
WTCHG_125416_05	pericycle	Pep1	2	4	2
WTCHG_129190_01	cortex	mock	3	5	3
WTCHG_129190_03	cortex	flg22	3	5	3
WTCHG_129187_07	cortex	Pep1	3	1	3
WTCHG_129190_05	pericycle	mock	3	5	3
WTCHG_129190_07	pericycle	flg22	3	5	3
WTCHG_129189_07	pericycle	Pep1	3	3	3
WTCHG_203594_01	epidermis	mock	7	6	1
WTCHG_203594_03	epidermis	flg22	7	6	1
WTCHG_203839_01	epidermis	Pep1	7	7	1
WTCHG_203594_05	epidermis	mock	8	6	2
WTCHG_203594_07	epidermis	flg22	8	6	2
WTCHG_203594_10	epidermis	Pep1	8	6	2
WTCHG_203839_04	epidermis	mock	9	7	3
WTCHG_203839_06	epidermis	flg22	9	7	3
WTCHG_203839_08	epidermis	Pep1	9	7	3

2.2 RNA-seq analysis methods

2.2.1 Databases

RNA-seq data was aligned to the *Arabidopsis thaliana* genome from Ensembl Release 39 containing the Araport11 genome annotation (FASTA and gene transfer format (GTF) files were downloaded from `https://plants.ensembl.org/info/website/ftp/index.html`, 2018-05-15). Illumina adaptor sequences were downloaded with the Trimmomatic tool v0.36, downloaded 2018-05-16).

The gene ontology (GO) database was downloaded using the R package `org.At.tairGO`. The motif database created by Franco-Zorrilla et al. (2014) was used for paired motif enrichment analysis (downloaded from `http://meme-suite.org/doc/download.html`, 2017-05-15). The database contains 113 motifs (in the form of letter-probability matrices) that characterize the target sequence specificity of 63 plant TFs from 25 families.

2.2.2 Trimming and alignment of raw sequences

Raw FASTQ files were assessed using FastQC (Andrews, 2010) version 0.11.7. FastQC returns sequence quality scores, quantifies adaptor content and identifies overrepresented sequences, indicating the quality of sequencing overall.

Sequences in FASTQ files were trimmed of adaptor sequences and low quality reads using Trimmomatic version 0.36 (Bolger et al., 2014). The default settings recommended for paired end data were used to remove Illumina TruSeq adaptors, low quality reads and reads without pairs.

The paired-end libraries (2 × 100 bp reads) were mapped to the *Arabidopsis* genome using the Spliced Transcripts Alignment to a Reference (STAR) Alignment Tool (Dobin et al., 2013). First, a genome dictionary was generated using the `genomeGenerate` function, using the genome FASTA and GTF files detailed above with the default settings except for `--sjdbOverhang 69`.

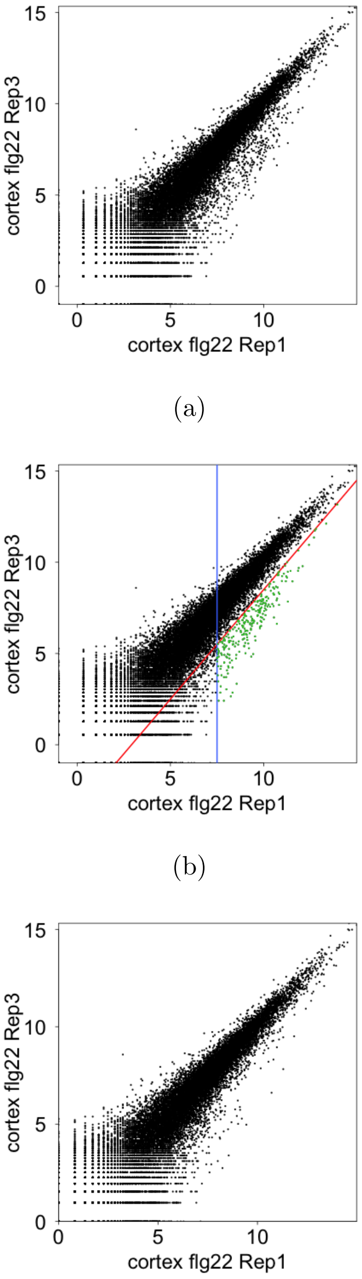
The alignment was then run with the following settings changed from the default: `--runThreadN 8`, `--outSAMtype BAM Unsorted`, `--limitOutSJcollapsed 2000000`, `--readFilesCommand zcat`. The quality of the alignment was assessed visually using the Integrative Genomics Viewer (IGV) (Robinson et al., 2011).

2.2.3 Read counting using HT-seq count

The reads mapping to exons were counted using LiBiNorm (an application analogous to HTSeq, Dyer et al. (2018)) using the `LiBiNorm count` command to produce read counts for each gene using the following parameters: `-r pos -i gene_id -s no -t exon -z -c`

2.2.4 Filtering artefacts

Read counts for each sample were normalised by size factor and $\log_2$-normalised counts plotted in pairs to identify batch effects (Figure 2.1). These artefacts could include extremely high read counts for non-protein coding genes such as ribosomal RNA (rRNA) in some samples, or 'side diagonals' caused by a large batch of genes being enriched in only one replicate, which can be seen to the right of the main grouping in Figure 2.1a. The side diagonal is isolated by plotting lines to capture these genes and using the equation for a line, genes forming the batch effect can be identified. In the example of Figure 2.1b, artefact genes are indicated in green, bounded by the blue line $x = 8$ and the red line $y = 1.2x - 4$. The artefact genes correspond to those that satisfy $(\log_2(\text{counts}_{\text{rep1}}) > 8)$ and $(\log_2(\text{counts}_{\text{rep1}}) \times 1.2 - \log_2(\text{counts}_{\text{rep3}}) > -4)$. This is repeated for each sample where there are obvious side diagonals. If the artefacts are the results of an obvious consistent source then they are filtered from all samples. The replicates are replotted to confirm artefact removal (Figure 2.1c). Organellar chromosomes, non-protein coding genes, and ribosomal proteins (listed in Appendix A) were filtered from all samples.



(a)

(b)

(c)

Figure 2.1: Identifying and filtering artefacts. Normalised $\log_2$ read counts for two replicates are plotted against one another. (a) shows a replicate pair with a 'side diagonal' batch effect which is highlighted in green (b). If a consistent source of batch effects can be identified, then these genes can be filtered from the analysis as in (c).

2.2.5 Calculating fragments per kilobase of exon per million reads (FPKMs)

For individual gene plots, read counts were converted to fragments per kilobase of exon per million reads (FPKMs). FPKMs were calculated by diving the read count using the following equation:

$$FPKM_i = \frac{X_i}{l_i N} \times 10^9$$

where X_i is the read count for a particular gene i, l_i is the respective gene length and N is the total number of reads in that sample.

2.2.6 Principal component analysis for RNA-seq

The principal component analysis (PCA) was performed on size factor-normalised and regularised-log (rlog) transformed count data using `PlotPCA` from the `DEseq2` package (Love et al., 2014). Size factor normalisation is calculated by division of all values by a normalisation factor to equalise library sizes across all samples. The counts are transformed onto a $\log_2$ scale whilst minimising the differences between rows with low counts, thus minimising the impact of noise on the PCA.

2.2.7 Differential gene expression analysis

Deseq2 and EdgeR were used to calculate differentially expressed genes (DEGs) (Robinson et al., 2009; Love et al., 2014). The methods were compared by overlaying the expression of DEGs over the expression of non-DEGs, as in Figure 2.3. The best model fit was shown to be DESeq2, particularly for the pericycle data (Figure 2.2). EdgeR fitted the pericycle poorly characterising an excessive number of DEGs particularly at low read counts, whereas DESeq2 returned more reasonable results. As such DESeq2 was used to calculate DEGs throught this thesis.

Specifically, the DEGs were calculated using the `DESeq` function from the R package `DESeq2` (Love et al., 2014) using a paired replicate approach to account for batch effects. `DESeq` is a function which fits a generalised linear model (GLM) based on a negative binomial distribution. This model is used this to calculate differential gene expression between two conditions. The model was fitted using pairwise replicates. Differentially expressed genes were considered significant if the false discovery rate (FDR) was less than 0.05. To maximise the number of DEGs identified, `DESeq` as default applies independent filtering which excludes lowly expressed genes when calculating adjusted p-values.

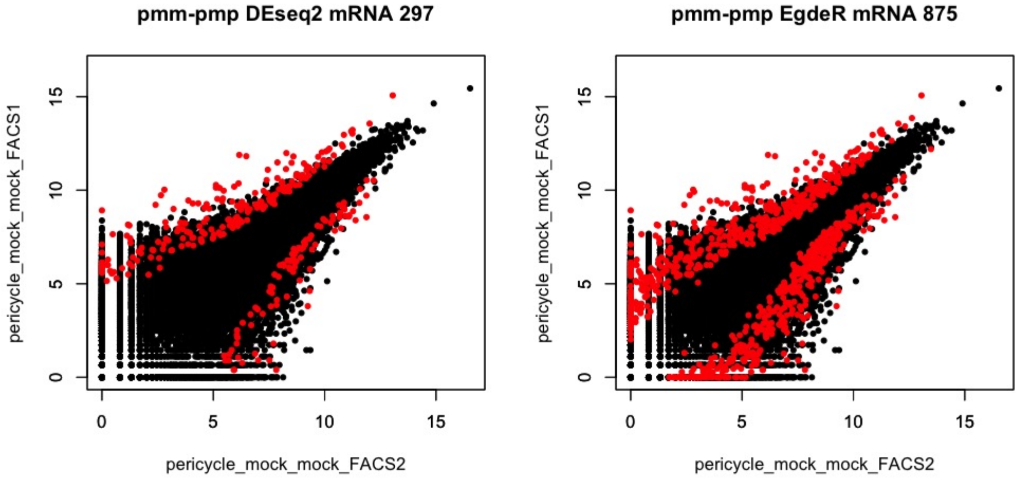


Figure 2.2: Comparison of methods to calculate differential gene expression. The expression of DEGs (red) overlaying non-DEGs (black) reveals the quality of differntial expression model fit for two tools; (a) DESeq2 and (b) edgeR.

Independent filtering maximises the number of FDR adjusted p-values returned by DESeq2 below a user-defined threshold (in this case $p < 0.05$). This is achieved by filtering out genes with little or no chance of being significant, filtering (based on average expression strength across all samples). Opting to remove count outliers means that genes with very wide variances, for example, a gene with very low read counts in all but one sample which expresses thousands of reads would be flagged for removal. Figure 2.3 shows that for the comparison cortex-mock vs. cortex-flg22, applying independent filtering increases the number of DEGs from 421 to 476, but removal of count outliers has no effect on differential gene expression, a pattern that is consistent across all comparisons. Based on these results, DEGs calculated using independent filtering were used in the subsequent analysis.

Immune responsive genes were defined as those responding in at least one cell type in response to flg22 and/or Pep1. They were calculated using pairwise comparisons between mock and treated samples in each cell type separately. Cell type-specific immune responsive genes were identified using Venn diagrams plotted using R packages gplots and VennDiagram (Chen and Boutros, 2011) and Java program EulerAPE for 3-set proportional Venn diagrams (Micallef and Rodgers, 2014) and gene subsets were extracted. Cell type- and treatment-specific genes were identified by pooling flg22 and Pep1 responsive genes and extracting the genes that responded to only one of the treatments. These genes were then split by cell-type expression using Venn diagrams.

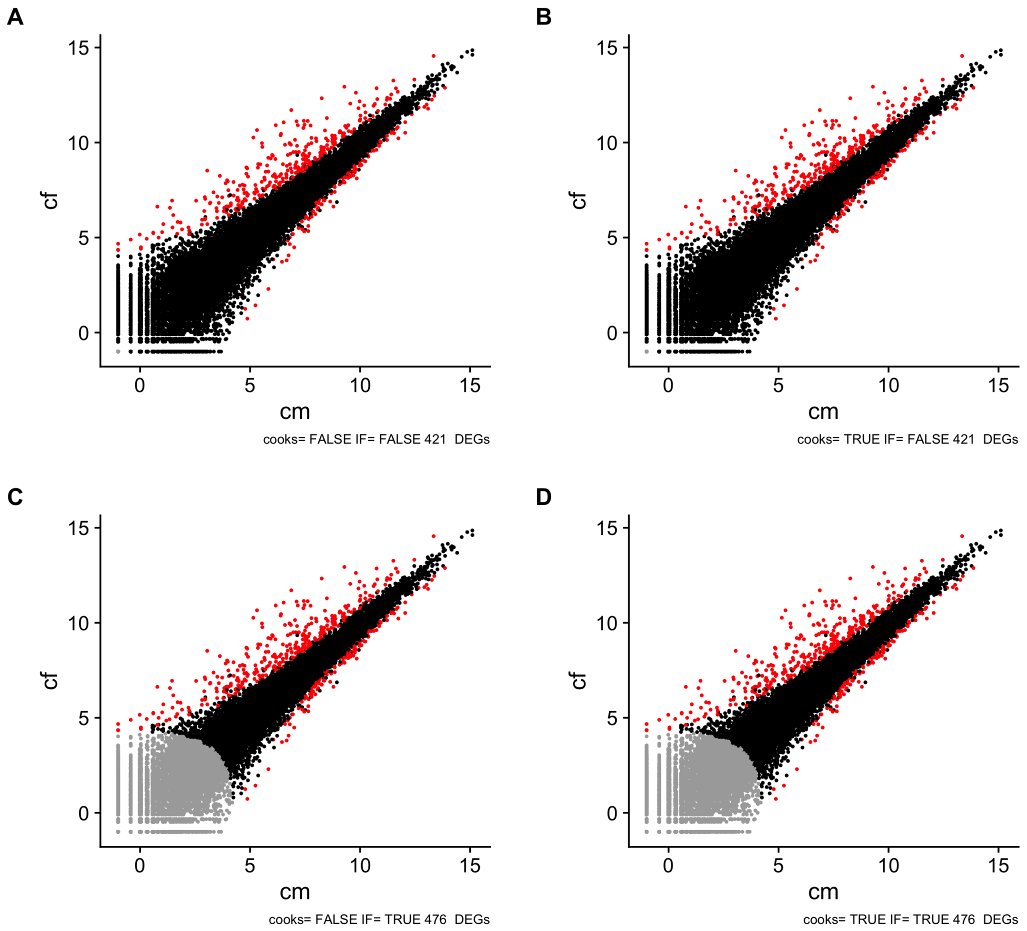


Figure 2.3: DEG replicate plots comparing the effect of applying independent filtering and/or count outlier removal by Cook's Distance for one representative comparison. Replicate plots show DEGs in red, non-DEGs in black and genes that have been filtered out in grey. Text beneath each plot indicates the parameters for the test and the number of DEGs calculated for that comparison.
(a) shows the effect of applying no independent filtering or count outlier removal.
(b) shows the effect of applying count outliers removal but no independent filtering.
(c) shows the effect of applying independent filtering but not count outliers removal.
(d) shows the effect of applying both independent filtering and removal of count outliers.

In order to characterise cell identity genes, all nuclear genes were categorised by the cell type they were most highly expressed in based on log-fold changes between cell types. Then using pairwise differential gene expression calculated for each possible pair of mock replicates, genes that were significantly higher ($p < 0.05$) in one cell relative to both other cell types were defined as cell identity genes for that cell type.

2.2.8 Gene ontology enrichment analysis

GO enrichment analysis was performed using the R package `GOStats` (Falcon and Gentleman, 2007). The `hyperGtest` function is used to test for over-representation of GO terms using a classical hypergeometric test. The resulting p-values are adjusted by FDR using the function `p.adjust`. GO term enrichment was visualised using bar plots and heat maps produced using the R package `ggplot` (Wickham, 2016).

2.2.9 Combinatorial motif analysis using the Paired Motif Enrichment Tool (PMET)

Promoter regions corresponding to 1,000 bp upstream from the transcription start site were extracted for all protein coding nuclear genes in the *Arabidopsis* genome. Mitochondrial or plastid chromosomes were excluded. For each motif and each promoter, the sequence was scanned for occurrences of the motif using FIMO (Grant et al., 2011) which assigns a probability score to each potential hit. In order to determine the number of hits to consider and to compute an overall score for motif presence in a promoter, we computed the geometric mean p of the top K FIMO probability scores for non-overlapping hits and computed the binomial probability of observing at least k hits of probability p in a 1 kb promoter. The value of k minimising the binomial probability was taken to indicate the most likely number of binding sites and as such k binding sites were used for subsequent analysis; k was restricted to values $1 \leq k \leq 5$. For each motif, the promoters were ordered by increasing binomial probability and the top $N = 5{,}000$ promoters were considered as containing the motif. The parameter N was chosen for high sensitivity (rather than specificity) as stringency is introduced when the pairing of motifs is considered. The binomial probability of the N^{th} promoter was recorded for each motif as a threshold. For each pair of motifs and for each promoter containing both motifs, overlaps of recorded motif hits were identified and the information content (IC) of the overlap (based on motifs) was calculated. If the IC of the overlap for either motif exceeded

4 (indicating highly conserved bases are part of the overlap), then these hits were removed and the binomial probability re-calculated for the remaining hits. If the re-calculated scores were still below the recorded motif-specific threshold, then the two motifs were considered as co-localised in the promoter. Finally, gene sets of interest were tested for enrichment of paired motifs using a pairwise hypergeometric test. Hypergeometric p-values were corrected for the number of motif pairs using local Bonferroni (calculating the correction for each gene set separately). Corrected p-values less than 0.05 were considered significant. For each comparison of results made between conditions, the gene sets tested were of equal size to make p-values comparable. To this end, gene set sizes were equalised by taking the top G genes from the larger gene set, where G is the size of the smaller gene set. For full details of PMET, see Chapter 3.

2.3 Materials and methods for Drop-seq

2.3.1 Plant growth and treatment

WOX5::GFP expressing plants were grown for nine days and treated with flg22 or with H_2O as a control (as in Section 2.1.1) at least three hours after the start of the daily light period. The *WOX5::GFP* marker line expresses *GFP* which has been fused to the promoter of *WUSCHEL-RELATED HOMEOBOX 5 (WOX5)*, a gene specifically expressed in the quiescent centre (QC).

2.3.2 Protoplasting root tips for Drop-seq

Two hours after treatment, 100-200 root meristems per treatment were harvested in protoplast solution on a microscope slide using a small hypodermic needle under a stereomicroscope. Two hours was chosen as a suitable time point that would capture the initial flg22 response, whilst being manageable within the constraints of the Drop-seq protocol. Roots were incubated in $\sim$500 μL enzyme solution for 2.5 hours with shaking at 110 rpm, and pipetting with a 100 μL pipette every 20 minutes. The protoplast suspension was passed through a 40 μm cell strainer to remove undigested clumps and other debris, followed through by another equal volume of enzyme solution, and the protoplasts in the resulting strained suspension counted using a Fuchs-Rosenthal Haemocytometer. The strained protoplast suspension was spun down at 300 rcf for 5 minutes to pellet the protoplasts, supernatant was removed and cells resuspended to a concentration of 300 cells μL^{-1} in cell buffer.

Table 2.2: Sequencing specifications for NextSeq

Read	Length (bp)	Description
Read 1	20	cell barcodes and UMIs
Read 2	50	cDNA
Read 1 Index	8	Barcodes to identify multiplexed samples

2.3.3 Drop-seq protocol summary

The protocol used to perform Drop-seq was the 'Online Drop-seq Protocol v3.1' from the McCaroll lab, (downloaded from `http://mccarrolllab.com/download/905/`, 2017-11). The machine used for droplet formation was the single-cell sequencing platform from Dolomite Bio.

In summary, cells from the protoplast suspension were encapsulated in droplets containing barcoded beads, tagging each mRNA molecule with unique molecular identifiers (UMI) and cell of origin barcodes. In general, each droplet containing both a bead and a cell captures ~11% of the mRNA from each cell, equivalent to ~20,000 transcripts. Following the machine run, droplets were broken and mRNA-bound microparticles were reverse transcribed into cDNAs to form a stable set of beads called single-cell transcriptomes attached to microparticles (STAMPs). STAMPs were treated with Exonuclease I to degrade excess bead primers that were not bound to RNA molecules. In order to prepare beads for sequencing, STAMPs were amplified using the polymerase chain reaction (PCR), the cDNA library was purified and the concentration was tested using a BioAnalyzer. Next, cDNA was tagmented using Nextera XT, purified and concentration was quantified again using a BioAnalyzer.

The samples were prepared for sequencing using the NextSeq 75 High Output kit and it's associated protocol. The sequencing specifications for NextSeq are shown in Table 2.2 The method is summarised in Figure 2.4, see Section 5.1 for a more comprehensive description of the Drop-seq protocol, and a larger figures.

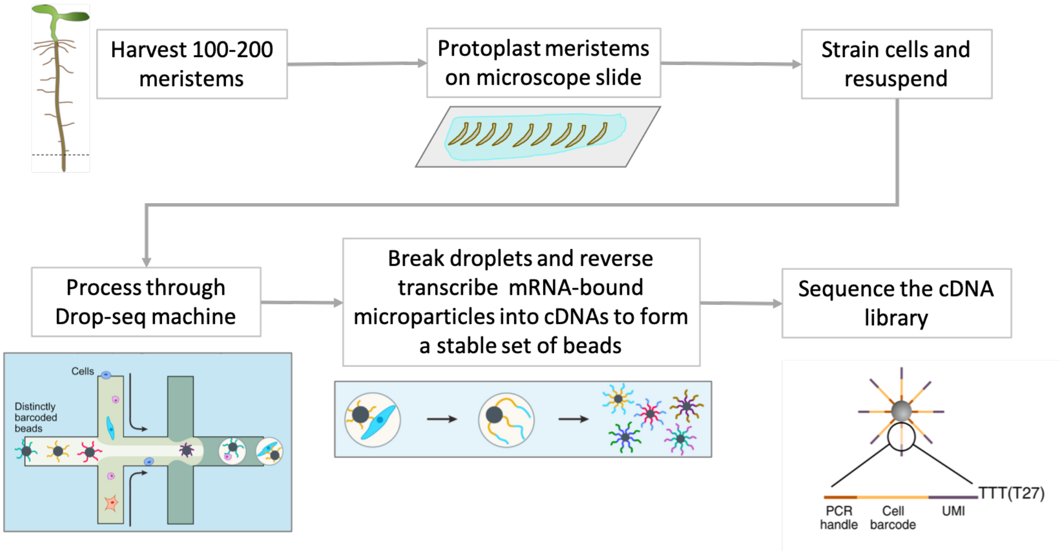

Harvest 100-200 meristems
Protoplast meristems on microscope slide
Strain cells and resuspend
Process through Drop-seq machine
Break droplets and reverse transcribe mRNA-bound microparticles into cDNAs to form a stable set of beads
Sequence the cDNA library
Cells
Distinctly barcoded beads
TTT(T27)
PCR handle
Cell barcode
UMI

Figure 2.4: Flow diagram demonstrating the application of the Drop-seq protocol to *Arabidopsis* root meristems. Seedling image from Sparks (2017), Drop-seq diagrams from Macosko et al. (2015).

2.4 Drop-seq analysis methods

The Drop-seq alignment protocol was performed following the ‘Drop-seq alignment cookbook’ v1.2 (Nemesh and McCarroll, 2016). This pipeline covers the steps alignment of FASTQ files to the genome and creating a differential gene expression (DGE) matrix containing read counts for each gene and cell as performed in Macosko et al. (2015).

2.4.1 Processing binary base call to compressed BAM files

Sequences were downloaded in pooled binary base call (BCL) format files. The files were demultiplexed into FASTQ files containing raw sequences from each sample using `bcl2fastq` using indexes detail in Table 2.3. FASTQ files contain uncompressed sequences and associated sequencing quality scores. `bcl2fastq` returns two FASTQ files each containing one of each paired read, per lane per samples. FASTQ files were merged into two paired files per sample using the unix `cat` command. These paired FASTQ files were merged into one binary alignment format (BAM) file using the `FastqToSam` program from Picard tools (Broad Institute, 2018). The BAM file is a compressed Sequence Alignment Format (SAM) file which contains sequence information, quality scores and have the capacity to store alignment positions and scores. However at this point our data is unmapped.

Lane	Sample_ID	Sample_Ref	Description	index	index2	Control	Recipe	Operator	Sample_Project
1	1_mockRep1	At	mockRep1	TAAGGCGA		FALSE	PE_indexing	PV	At_roots
1	2_flg22Rep1	At	flg22Rep1	CGTACTAG		FALSE	PE_indexing	PV	At_roots
1	3_mockRep2	At	mockRep2	AGGCAGAA		FALSE	PE_indexing	PV	At_roots
1	4_flg22Rep2	At	flg22Rep2	TCCTGAGC		FALSE	PE_indexing	PV	At_roots
1	5_mockRep3	At	mockRep3	GGACTCCT		FALSE	PE_indexing	PV	At_roots
1	6_flg22Rep3	At	flg22Rep3	TAGGCATG		FALSE	PE_indexing	PV	At_roots
2	1_mockRep1	At	mockRep1	TAAGGCGA		FALSE	PE_indexing	PV	At_roots
2	2_flg22Rep1	At	flg22Rep1	CGTACTAG		FALSE	PE_indexing	PV	At_roots
2	3_mockRep2	At	mockRep2	AGGCAGAA		FALSE	PE_indexing	PV	At_roots
2	4_flg22Rep2	At	flg22Rep2	TCCTGAGC		FALSE	PE_indexing	PV	At_roots
2	5_mockRep3	At	mockRep3	GGACTCCT		FALSE	PE_indexing	PV	At_roots
2	6_flg22Rep3	At	flg22Rep3	TAGGCATG		FALSE	PE_indexing	PV	At_roots
3	1_mockRep1	At	mockRep1	TAAGGCGA		FALSE	PE_indexing	PV	At_roots
3	2_flg22Rep1	At	flg22Rep1	CGTACTAG		FALSE	PE_indexing	PV	At_roots
3	3_mockRep2	At	mockRep2	AGGCAGAA		FALSE	PE_indexing	PV	At_roots
3	4_flg22Rep2	At	flg22Rep2	TCCTGAGC		FALSE	PE_indexing	PV	At_roots
3	5_mockRep3	At	mockRep3	GGACTCCT		FALSE	PE_indexing	PV	At_roots
3	6_flg22Rep3	At	flg22Rep3	TAGGCATG		FALSE	PE_indexing	PV	At_roots
4	1_mockRep1	At	mockRep1	TAAGGCGA		FALSE	PE_indexing	PV	At_roots
4	2_flg22Rep1	At	flg22Rep1	CGTACTAG		FALSE	PE_indexing	PV	At_roots
4	3_mockRep2	At	mockRep2	AGGCAGAA		FALSE	PE_indexing	PV	At_roots
4	4_flg22Rep2	At	flg22Rep2	TCCTGAGC		FALSE	PE_indexing	PV	At_roots
4	5_mockRep3	At	mockRep3	GGACTCCT		FALSE	PE_indexing	PV	At_roots
4	6_flg22Rep3	At	flg22Rep3	TAGGCATG		FALSE	PE_indexing	PV	At_roots

Table 2.3: Sample table containing indexes used to demultiplex bcl files into one FASTQ file per sample using `bcl2fastq`.

2.4.2 Databases

Drop-seq data was was aligned to the *Arabidopsis thaliana* genome from Ensembl Release 36 containing the Araport11 genome annotation (FASTA and GTF files were downloaded from `https://plants.ensembl.org/info/website/ftp/index.html`, 2017-08-17)

2.4.3 Creation of meta data for Drop-seq alignment

The Drop-seq alignment protocol was performed following the 'Drop-seq alignment cookbook' v1.2 (Nemesh and McCarroll, 2016). This pipeline covers the steps alignment of FASTQ files to the genome and creating a DGE matrix containing read counts for each gene and cell as performed in Macosko et al. (2015).

The Drop-seq alignment pipeline requires meta data in the form of the following input files:

- FASTA file: The reference genome sequences required by the aligner (downloaded from ensembl detailed in section 2.2.1).

- GTF file : Genomic features annotation file containing locations of genes, transcripts and exons (downloaded from ensembl, detailed in Section 2.2.1). Many other meta data files are derived from this file. The GTF file had to be edited in order to be compatible with the pipelines scripts that create meta data files. In particular, 'transcript_name' fields were filled by duplicating the 'transcript_id' field and any genes with missing empty 'gene_name' fields were filled by duplicating the 'gene_id' field. The *Arabidopsis* genome also contains genes with semi-colons and spaces in their genenames, which are used as column delimiters by the pipeline, as such these were replaced with hyphens. Care was taken in downstream analysis to check any genes containing punctuation were recorded in reports with the correct punctuation. An additional chromosome was added containing the gene sequence for GFP in order to detect reads from the *WOX5::GFP* fusion.

- Genome dictionary file: A file generated by Picard Tools `Create Sequence Dictionary` command.

- refFlat file: A file containing genomic locations of exons for each gene subsetted from the GTF file.

- Gene intervals file: Genes from the GTF file in interval list format.

- Exon intervals file: Interval list format file containing locations of exons from the GTF file.

- rRNA intervals file: Interval list format file containing locations of rRNA from the GTF file.

- Reduced GTF file: Human-readable GTF file.

2.4.4 Converting unmapped BAM to aligned and tagged BAM

The following steps of the Drop-seq alignment protocol were performed using tools were developed by the McCarroll lab as part of their 'Drop-seq tools' suite, unless otherwise specified:

1. **Tag cell barcodes.** `TagBamWithReadSequenceExtended` extracts bases encoding cell barcodes and creates BAM tags for those barcodes.

2. **Tag molecular barcodes.** `TagBamWithReadSequenceExtended` extracts bases encoding molecular barcodes and creates BAM tags. These tagging steps also tag low quality reads that contain at least 1 bp below a quality threshold (Phred score) of 10.

3. **Filtering reads** which have been tagged as low quality using `FilterBAM`.

4. **Trim 5' primer sequence.** `TrimStartingSequence` trims SMART Adapters that can occur at the 5' end of a read.

5. **Trim 3' polyA sequence.** `PolyATrimmer` searches for 6 contiguous A at the 3' end and removes these sequences.

6. **Convert SAM to FASTQ.** `SamToFastq` from Picard tools converts trimmed SAM files to FASTQ files, the required input by STAR.

7. **Alignment to genome.** STAR (Dobin et al., 2013) aligns the trimmed reads to the specified genome with settings as per defined in the Drop-seq alignment script with the addition of `--limitOutSJcollapsed 2000000` to ensure there is sufficient memory.

8. **Sort STAR alignment.** `SortSam` from Picard tools sort mapped BAM files in queryname order.

9. **Recover cell/molecular barcodes.** `MergeBamAlignment` from Picard Tools merges the STAR alignment-tagged BAM with the unmapped BAM that was

previously tagged with molecular/cell barcodes. This results in a mapped BAM file containing cell and molecular tagged reads.

10. **Add gene/exon and other annotation tags.** `TagReadWithGeneExon` adds a 'G' BAM tag onto reads that overlap with an exon of a gene. This tag later is used to produce the DGE matrix.

11. **Detecting bead synthesis errors.** Some batches of barcodes have been identified as containing high proportions of shared sequences, as they had not been synthesised correctly. `DetectBeadSynthesisErrors` identifies barcodes containing fixed bases (positions containing a high proportion of just one base). If this fixed base is at the end of the read, it is trimmed and if any of the trimmed sequences match they are merged. If a fixed base occurs in any other position reads with that barcodes are removed.

2.4.5 Generating the digital gene expression matrix

To count gene transcripts, the total number of UMI in each gene and within each cell are counted using `DigitalExpression` tool from the Drop-seq pipeline. If a pair of UMIs contain one matching base at the same position (Hamming distance = 1) these UMIs are merged. Then, the total number of unique UMI sequences are reported as the number of transcripts per gene, per cell. `DigitalExpression` also returns a summary of the digital gene expression matrix, containing the total number of genes and transcripts observed in each cell.

2.4.6 Cell selection

The efficiency of Drop-seq using the Dolomite Bio system is 20% as only a small proportion of droplets contain one bead and one cell. Most droplets contain no cells, or no beads, and some contain multiple cells (referred to as doublets). Droplets containing beads can also collect ambient RNA, which results in droplets that appear as cells with low read counts. Including doublets and ambient reads in the subsequent analysis can add noise which reduces the resolution and accuracy of the results. There are two suggested methods for determining the correct number of cells: BAM tag histograms and log nUMI-log barcode plots. `BAMTagHistogram` from the Drop-seq tools calculates the number of reads for all BAM tags in a BAM file associated with a tag such as the XC tag which indicates individual cells. These are then plotted as a cumulative distribution plot. The ideal plot will have a distinctive knee as shown in Figure 2.5 to indicate the ideal number of cells to select, however the plots from experimental data are rarely this well defined.

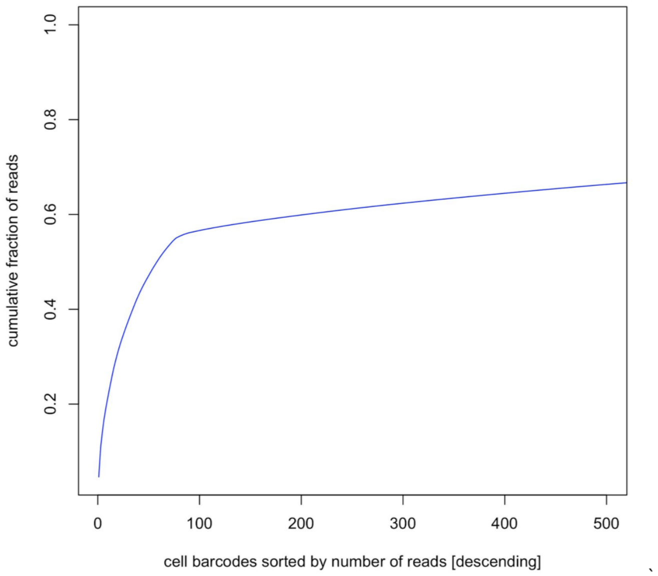


Figure 2.5: Ideal cumulative distribution plot of fraction of reads associated to BAM tags as calculated using `BAMTagHistogram`. The distinctive knee can be used to select the ideal number of cells to analyse further. Figure taken from the Drop-seq Alignment Cookbook v1.2

The alternative plot that can be used to determine the number of cells to select is the log transcripts-log barcodes plot. The number of transcripts (nUMI) is plotted on the y-axis and the barcodes ordered by decreasing UMI content on the x-axis, both log scaled, which should return a plot with two shoulders (Figure 2.6). The point at which the first shoulder starts to drop defines the point at which we are gaining no more information by adding new cells, and therefore defines our cell selection cut-off. The second shoulder indicates the drop off in information from barcodes associated with ambient RNA. In order to see all dips in the data you need to include data for all barcodes, which is obtained by running `DigitalExpression` for more cells than the dataset could contain (i.e. 1,000,000 cells).

If neither method produces a clear cut -off point, then an alternative approach would be to take only include the cells with a reasonable number of transcripts, e.g. select all cells with a minimum off 4000 transcripts, although this non data-driven method is more likely to result in loss of data.

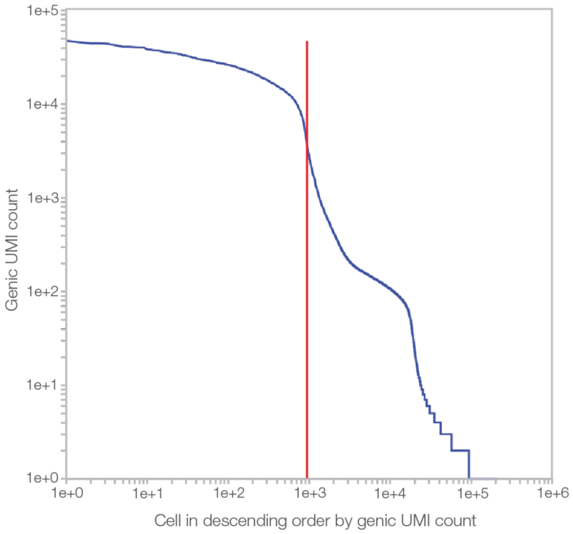


Figure 2.6: Log transcripts-log barcodes plot (blue line). The red line indicates the threshold for cell selection. Figure 4 from `https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/single-cell-rna-data-analysis-tech-note-1070-2017-001.pdf`

2.4.7 Filtering and quality control

`Seurat` is an R package (Butler et al., 2018) designed for QC, analysis, and exploration of single cell RNA sequencing (scRNA-seq) data. `Seurat` was used to filter the dataset, visualise clusters and assign identity to cell clusters. At this point, all samples were merged in order to maximise the resolution of the dataset.

`Seurat` was used to identify, interpret and, where appropriate, filter sources of heterogeneity from single cell transcriptomic measurements. Some sources of heterogeneity such as high concentration of mitochondrial RNA can result from the experimental procedure, so must be filtered out so as to not distort the subsequent results. Additionally, filtering cells with low information increases the performance of the downstream dimension reduction and clustering. For the same reason, genes that were only expressed in fewer than 3 cells were filtered out of the analysis. Retaining these lowly expressed genes would just contribute noise to the analysis.

Violin plots of nUMI, nGene, plastid and mitochondrial content were used to assess for potential contaminants. Based on these plots, cells containing >5% mitochondrial, >5% plastid reads, >50000 nUMI (indicating a potential doublet) or <200 unique genes were filtered out. Single cell datasets usually contain uninteresting sources of variation, such as technical noise, batch effects and biological sources of variation that are irrelevant to the experimental question. These can be regressed out to improve downstream analysis, using Seurat's `ScaleData` function. This function was used to constructs a linear model to predict gene expression, and used z-scores to scale the effects differences in gene expression driven by the number of detected molecules (nUMI), mitochondrial and plastid gene expression and sample identity (to minimise sample batch effects).

2.4.8 Principal component analysis

PCA was performed on the merged and filtered samples. Using the `Seurat` function `FindVariableGenes`, the most variable genes in the dataset were identified to focus on in the dimension reduction analysis. By focussing on the most variable genes, the computational complexity is reduced and the level of noise in the analysis is considerably reduced. The first two principal components (PCs) were visualised to confirm that batch effects between the samples were not the primary sources of variation in the dataset, and that no small cell populations were distant from others (indicating contamination) giving confidence to proceed with the t-Stochastic Neighbour Embedding (t-SNE) analysis. PCA is also a prerequisite to t-SNE.

t-SNE has a number of user-defined parameters which can affect the output. One of these parameters is the number of PCs to incorporate into the t-SNE, which in turn is affected by the number of variable genes used to perform the PCA. The `Seurat` package provides two methods to determine the best number of PCss; the elbow plot and visual inspection of gene expression in heat maps.

The elbow plot shows the standard deviation of each PC against that PC (Figure 5.12). The cut-off point for the number of significant PCs is the point at the 'elbow' where the standard deviation is no longer decreasing significantly. The second method investigates the expression of the top 20 most variable genes in the top 100 cells that represent each principal component. By examining these expression patterns for a range of PCs (plotted as heat maps), the point at which adding more PCs no longer adds significantly more information to the data set can be determined. The results of these methods are discussed in Chapter 5.

2.4.9 Visualisation using t-Stochastic Neighbourhood Embedding

The visualisation method t-SNE is used to reduce high-dimensional data to two dimensions with the aim of achieving an informative visualisation (Maaten and Hinton (2008), implemented in the R package `Seurat` by Butler et al. (2018)). The clustering of cells in the t-SNE plot can be used to determine meaningful groups of cells.

Read count data contains $n \times m$ dimensions where n is equal to the number of genes ($\sim 1.5 \times 10^4$) and m is equal to the number of cells ($\sim 2 \times 10^3$). Typically, a reduction from $\sim 1.5 \times 10^4$ to between 20 and 30 dimensions is calculated using PCA first. The resulting data is then utilised by t-SNE to further reduce the data to two dimensions.

One key difference between t-SNE and PCA is that PCA simultaneously maps the entire high-dimensional space to the low-dimensional space using linear algebra. In contrast, t-SNE finds low-dimensional coordinates that explain local structures rather than global structure. This means that similar data points cluster closely together, but the distances between different local structures or clusters are not necessarily informative. The t-SNE plot can be refined by altering a perplexity parameter which describes the average size of local clusters.

In order, to refine the analyses, t-SNE plots were produced using inputs from a range of included PCs and perplexity values, until a stable set of parameters were determined, where clustering was similar for multiple runs of t-SNE.

Cells were clustered based on the t-SNE coordinates using a shared nearest neighbour (SNN) modularity optimization based clustering algorithm, (implemented as the `FindClusters` function in the R package `Seurat` package). First k nearest neighbours were calculated from the SNN graph. A modularity function was then optimised to determine clusters (Waltman and Van Eck, 2013).

Identity was assigned to the clusters using two approaches. First, a supervised approach was used, whereby the expression of known cell type marker genes (detailed in Table 5.7) was compared between each cluster using violin plots and by overlaying the expression data on the t-SNE plot. In cases, where the marker genes were highly expressed in a small number of clusters, those genes were used to assign identity. Secondly, an unsupervised approach was used to identify novel marker genes for each cluster (as in Section 2.4.11).

2.4.10 Assigning cell cycle scores

Each cell was assigned to one of three cell cycle phases: G1, S or G2/M. For each cell, a score to define whether a cell was likely to be in a particular phase, was calculated based on the expression levels of S or G2/M phase associated genes. Cells that scored highly for the S or G2/M phase were assigned to that respective phase. The remaining low or ambiguous scoring cells were assigned to the G1 phase.

2.4.11 Identifying markers to identify cell subsets

Positive and negative markers of subsets of cells, relative to all other cells, were identified using differential gene expression (DGE) analysis based on the non-parametric Wilcoxon rank sum test (Wilcoxon, 1945). This DGE analysis was first used to determine whether clusters of potentially broken cells had a distinctive expression pattern (see Section 5.2.3). Identified marker genes were used to distinguish broken and whole cells. Secondly, novel marker genes that characterised the cell clusters identified using t-SNE. The known expression (based on data from Brady et al. (2007)), and biological function of these novel markers were analysed to assign identity to clusters where possible. Finally, flg22-responsive genes within different developmental zones were identified using DGE analysis

Chapter 3

Paired Motif Enrichment Tool Software Development

3.1 Introduction

The Paired Motif Enrichment Tool (PMET) was developed to identify pairs of transcription factors that are co-localised within the promoters of a given gene set and therefore could be acting together to regulate that gene set. The motivation for developing this tool was to identify different sets of TFs that could explain differing gene expression patterns that we observed in different cell-types in response to an immunity treatment (fully discussed in Chapter 4). Within this chapter is presented; the functionality of PMET, a brief summary of other available tools and a study to determine the optimum default parameters for the tool. PMET was developed from a framework for single motif analysis called HMT (Breeze et al., 2011).

3.2 Functionality and implementation of PMET

PMET is split into two algorithms; the first indexes motifs and promoters, and the second performs co-localisation enrichment tests. The first algorithm referred to as 'PMET index' scans all the promoter sequences in a given genome (or other provided sequences) for each motif individually. For each promoter-motif pair, the top K matches are extracted and a binomial score is used to calculate the number of 'true' motif matches, and the overall significance of that motif's presence in the promoter. This score is stored, and once all of the promoters have been scanned for a given motif, the top N promoters are retained, based on the best (lowest) binomial scores. The binomial score of the N^{th} promoter is retained as a binomial score threshold.

When all of the motifs from the provided database have been indexed, the paired enrichment testing can take place. For every possible pair of motifs, the top N sequences for each motif are assessed for overlaps and overlaps that exceed the information content (IC) threshold are removed from the analysis. The binomial score for that promoter is recalculated and if the new score exceeds the binomial score threshold for either motif, the promoter is filtered out of the analysis. Finally a hypergeometric test is performed to assess the enrichment of the motif pairs in target promoter set relative to the surviving promoters from the previous step.

PMET can be used to analyse paired enrichment in any genome with transcription start site (TSS) annotated to genes. The tool performs best when used to match motifs from organism specific databases to promoters in that organism but it can also be used to search for motifs from closely related organisms (e.g testing a generic plant motif database against a less studied plant species).

For each parameter, a default value is shown. These default values were chosen as biologically 'sensible' parameters which could be computed in reasonable time.

3.2.1 Current methods to investigate combinatorial motifs

The motivation for developing PMET was that none of the existing tools fitted our requirements. These requirements were: to account for multiple binding sites for individual motifs, to not constrain motif pairs based on the distance between motifs, to allow overlaps between motif pairs, and to not require chromatin accessibility or TF binding data.

In the early 2000s, an array of tools that identified motif pairs or composite motif modules within co-exprssed genes were developed. These tools built statistical models based on the input sequences of co-expressed genes and returned predicted motif modules. However, many of these modules such as 'Cister' (Frith et al., 2001), 'ModuleSearcher' (Aerts et al., 2003), 'ClusterBuster' (Frith et al., 2003), 'CisModule' (Zhou and Wong, 2004) and 'CMA' (Kel et al., 2006) required additional inputs such as expected number of modules, expected distance between motifs and expected distances between modules . These models all assume that cooperative TF binding sites are close together and therefore constrain their predictions based on distance. This assumption no longer fits with current models of TF cooperation. Whilst some cooperative TFs do bind to proximal sites, other cooperative TFs take advantage of DNA looping, which enables distal binding sites interact directly. TFs can also cooperate indirectly in cases where the first TF binding to the promoter makes the second TF binding site more thermodynamically favourable. The access to existing tools is also hindered by the lack of maintenance of older tools. In many cases, the tools published are no longer accessible or utilise databases that are out of date. Other tools that fit many, but not all, of the requirements of this analysis exist but have only been designed with mouse or human genomes in mind and are therefore not suitable, or able, to study *Arabidopsis* (Sharan et al., 2003; Meckbach et al., 2015).

Many recently developed tools have focussed on the integration multiple datasets, such as 'Chromia' (Won et al., 2010), 'CCAT' (Jiang and Singh, 2013) and 'TFcoop' (Vandel et al., 2017) which require combinations of ChIP-seq and or RNAi data or expression data. Similarly, 'Centipede' (Pique-Regi et al., 2011) and 'TFProb' (Lähdesmäki et al., 2008) rely on the integration of chromatin accessibility data and nucleosome positioning.

PMET combines some of the best features of existing tools; CMA uses a similar enrichment test to identify motif combinations that are specific to a co-regulated gene set. ModuleSearcher allows the motifs being tested to overlap. MSCAN combines p-values of multiple motif hits for a single motif to account for homotypic clusters, although utilises a different scoring method to PMET (Johansson et al., 2003).

Overall, most of the tools in the literature either rely on the user defining parameters that cannot easily discovered or utilise multiple data types which are not easily available to every experimentalist. The advantage of PMET is that the inputs are relatively simple and the parameters are designed to account more multiple mechanisms of TF cooperation, whilst making as few assumptions about the underlying structure of the data as possible.

3.3 PMET Part 1: Indexing

In the first stage of the PMET algorithm, individual motif instances are indexed within all promoters of the genome and the top promoters associated with each individual motif are stored for use in the enrichment test (Figure 3.1).

3.3.1 Genome sequence and motif database retrieval

PMET requires the following input files:

- genome sequences and index files in the form of `.fasta` and `.gff3` files

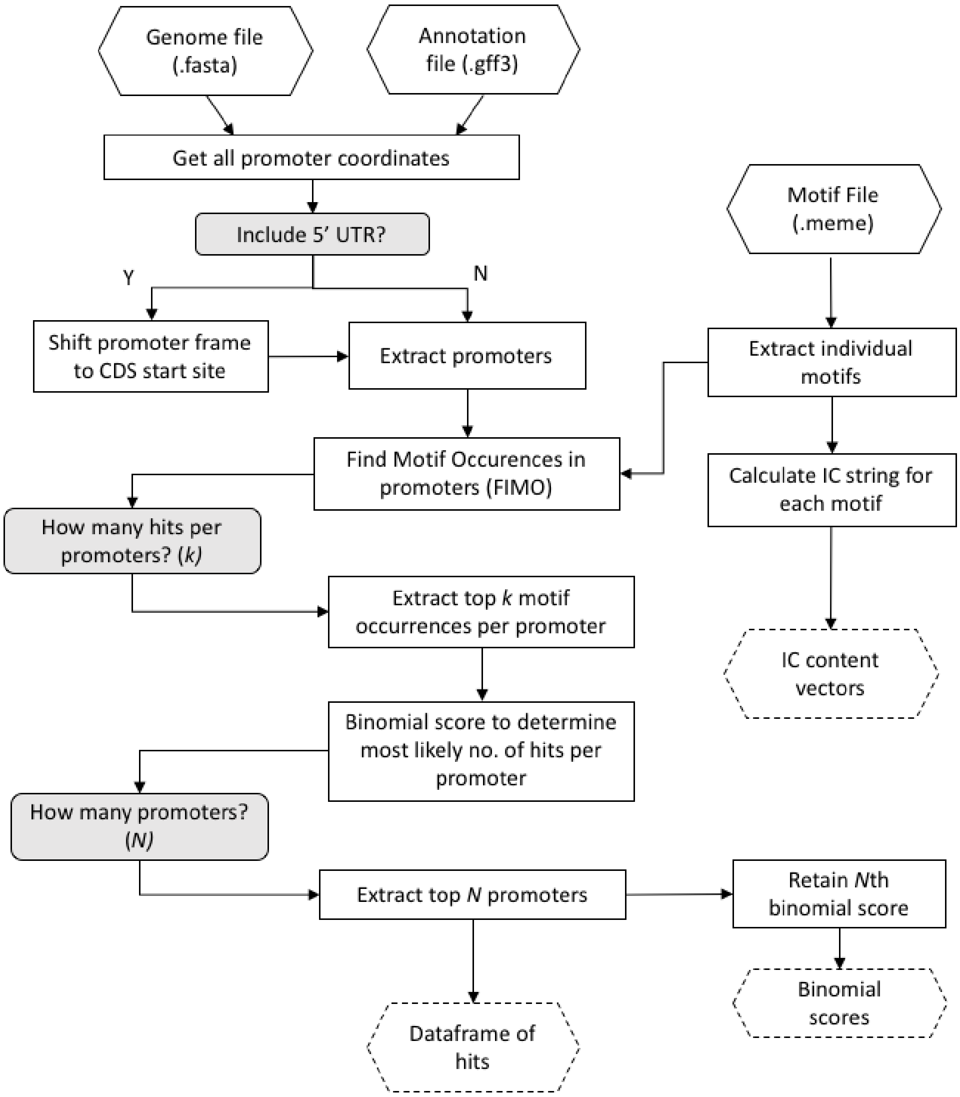

Genome file (.fasta)
Annotation file (.gff3)
Get all promoter coordinates
Motif File (.meme)
Include 5' UTR?
Y
N
Shift promoter frame to CDS start site
Extract promoters
Extract individual motifs
Find Motif Occurences in promoters (FIMO)
Calculate IC string for each motif
How many hits per promoters? (*k*)
Extract top *k* motif occurrences per promoter
IC content vectors
Binomial score to determine most likely no. of hits per promoter
How many promoters? (*N*)
Extract top *N* promoters
Retain *N*th binomial score
Binomial scores
Dataframe of hits

Figure 3.1: Work flow of Paired Motif Enrichment Tool Indexing. Hexagonal shapes with solid lines indicate data to be input by the user, hexagonal shapes with dashed outlines indicate data produced by the tool to be utilised later. Sharp-cornered rectangles indicate computational steps and grey rounded rectangles refer to user-defined parameters.

- a motif database in `.meme` format (described at `http://meme-suite.org/doc/meme-format.html?man_type=web`)
- a list of gene IDs to be tested for motif co-localisation

Alternatively, the user can provide custom FASTA sequences as described in Section 3.5.

3.3.2 Promoter extraction

PMET extracts n-bp promoters for all genes in the provided genome upstream of the TSS (where n is a user-defined parameter with a default of 1000 bp). If the user decides to include the 5' untranslated region (UTR), the promoter frame is shifted to start at the coding start site (CDS). These parameters were chosen based on a recent study of all *Arabidopsis* promoters revealed that motifs are distributed between -2000bp and +200bp from the TSS (Yu et al., 2016), with the majority of motifs in the first 1000bp.

3.3.3 Motif scanning

PMET uses the motif scanning tool FIMO (Find Individual Motif Occurences) from the MEME suite Grant et al. (2011). FIMO converts sequences provided in `.meme` format into log-odds position specific motif matrix (PSSM), independently scans each sequence for motifs, and reports a log-odds score and p-value to quantify the quality of all potential matches.

3.3.4 Binomial test for multiple motif instances occurring in the promoter

The tool allows for multiple matches for each individual motif in order to account for the prevalence of 'homotypic' clusters of motifs in eukaryotic genomes. FIMO returns all potential matches, and the top K most significant matches based on p-values are extracted for each promoter-motif pair. K, the maximum number of such matches is a user-defined parameter, with a current default value of $K = 5$. A binomial score, initially used in HMT and adapted for PMET, predicts the most likely number of 'true' TF binding sites up to a maximum of K in each promoter.

The binomial score $P(k)$ calculates the probability of at most k matches occurring in a promoter of length n for k between 1 and K: $P(k)$ is defined by

$$P(k) = \sum_{l=0}^{k} \binom{n}{l} p^l (1-p)^{n-l}$$

If $k = 1$, then the probability of that match occurring is the p-value calculated by FIMO (p). If $k > 1$, then the probability of k matches occurring is calculated as the geometric mean of the p-values associated with those k matches. The value of k that returns the minimal binomial score is then taken as the 'true' number of motif matches. The original implementation of the binomial score used the Hommel method (Vovk, 2012) to combine p-values but Hommel was weighted too weakly towards single matches such that the addition of any weak matches always resulted in a stronger score. The binomial score combined with the geometric mean more accurately models homotypic clusters.

Figure 3.2 describes three promoters where the binomial score returns different numbers of most likely k in a promoter of length n (in this case 1000bp) up to a maximum, K, of 4. In each case the binomial score is calculated K times, and the lowest scoring of these tests indicates the number of motifs least likely to occur randomly, and therefore most likely number of 'true' motif matches.

In the first promoter (Figure 3.2a), FIMO has returned four potential motif matches with p-values; 1×10^{-6}, 0.01, 0.01, and 0.01. The binomial score will be calculated four times.

1. The first binomial test calculated that the likelihood of the single lowest scoring (most significant) motif match occurring in a promoter of 1000bp is 0.002.

2. The second test determined that the likelihood of the combination of the two most significant motif matches, (defined as the geometric mean of the two lowest p-values) occurring in the promoter is 0.02.

3. The likelihood of the three most significant motif matches occurring was 0.07

4. The likelihood of all four occurring was 0.14.

In this example, the lowest scoring binomial test was the first one; accordingly the most likely number of 'true' motif matches is one. This single match was retained for the subsequent analysis, and the rest were discarded. The FIMO p-values in Figure 3.2a agree with this conclusion as the first motif match has a much lower P-value than the other three potential motif matches (highlighted in the figure). The lowest binomial score from these tests (corresponding to one motif match) was also stored by PMET and used to rank all the promoters tested based on how strong their respective motif matches are, and then extract the top N.

In the second promoter (Figure 3.2b), FIMO returned four potential motif matches with p-values; 1×10^{-6}, 1×10^{-4}, 0.01, and 0.01. In this case the four binomial tests calculated scores of 0.002, 0.0002, 0.06 and 0.14 for 1 to 4 hits,

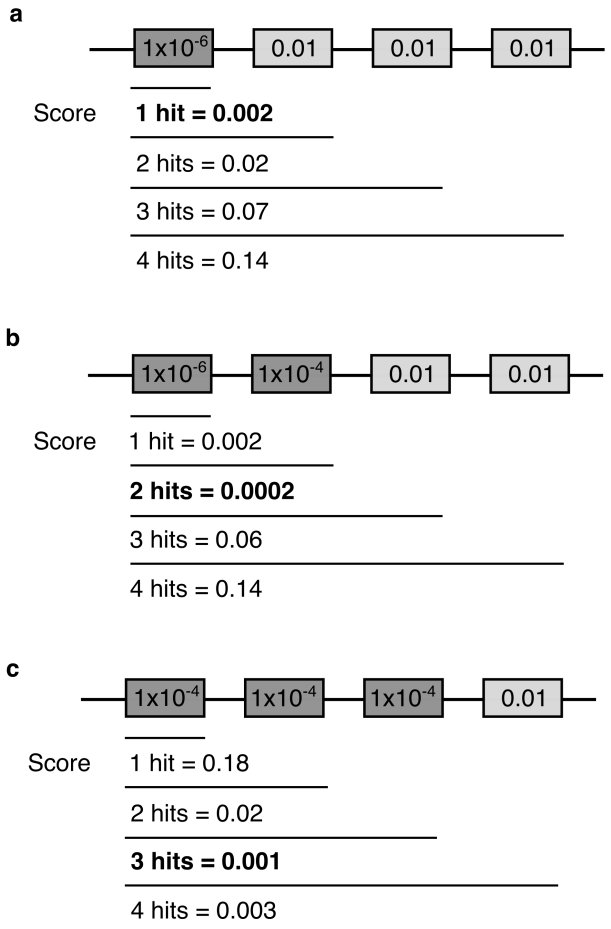


Figure 3.2: Binomial scoring of promoters. Each figure contains a representation of motifs in a promoter as boxes on a line. The FIMO p-values are ordered by significance (shown inside boxes ordered from left to right), and the binomial scores are calculated for 1, 2, 3 or 4 matches, shown below the promoter. (a-c) demonstrate three potential scenarios where the binomial score returns different numbers of most likely k in promoters up to a maximum, K, of 4.
(a) An example with one highly significant match, and three higher scoring matches. The binomial score returns the smallest value for one motif match (shown in bold).
(b) An example where the combination of two matches returns in the smallest binomial score.
(c) An example where the combination of three weaker matches returns the smallest binomial score.

respectively. The lowest scoring test was the second, indicating that the number of matches most likely to occur in this promoter is two. These two motifs would be retained by PMET and the score of 0.0002 used in promoter rankings.

In the third promoter (Figure 3.2c), FIMO returned four potential motif matches with p-values; 1×10^{-4}, 1×10^{-4}, 1×10^{-4}, and 0.01. The lowest binomial score corresponds to the third iteration, and therefore the most likely number of motif matches in this promoter is three. The final rankings of these three examples, based on lowest binomial score would be: promoter 2, promoter 3, promoter 1.

Whilst optimising PMET to work effectively for larger genomes, the following rearranged version of the binomial score was implemented to decrease computation time.

$$\begin{aligned}
P(k) =& \sum_{l=0}^{k} \binom{n}{l} p^l (1-p)^{n-l} \\
=& \prod_{i=1}^{l} \Big(\frac{n+1-i}{n}\Big) p^l (1-p)^{n-l} \\
=& \sum_{l=0}^{k} \exp\left(\Big(\sum_{i=1}^{l} \log(n-i+1) - \log(l)) \Big) \times \Big(l \cdot \log(p) \Big) + \Big((n-l) \cdot \log(1-p) \Big) \right)
\end{aligned}$$

3.3.5 Extracting top N promoters

For each motif, the top N promoters with the lowest binomial scores are retained to be tested for paired motif enrichment. N is a user-defined parameter with a default value of 5000. This value returns significant hits for all co-regulated gene sets tested so far, whilst limiting computation time to a few hours. The binomial scores of the N^{th} promoters for each motif are also retained to be used as a threshold following overlap removal. The novel approach of utilising the top N promoters rather than applying a threshold has the benefit that each motif is treated in an unbiased manner.

3.3.6 Calculating information content of motifs

Finally, the level of conservation for each base position in a motif logo is quantified using the IC:

$$IC_i = 2 + \sum_b f_{b,i} \log_2 f_{b,i}$$

Information content is calculated from the relative probabilities of observing each base at a single position. If all bases have an equal probability of being observed at a

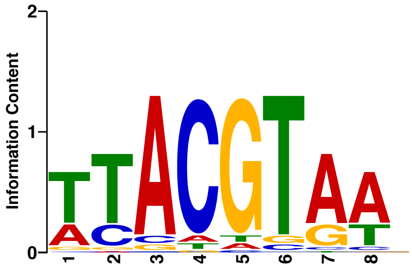


Figure 3.3: Information content in ANAC55_2 motif from (Franco-Zorrilla et al., 2014). The conservation of each base at each position is quantified in terms of IC where a fully conserved base will have an IC of 2.

single position, then the information content is very low, whereas if the chance of one particular base is high relative to all the other bases (a highly conserved base) then the information content is high. This can visualised as a sequence logo (Figure 3.3). The information content per position for each motif is stored by PMET in order to quantify overlaps between potential co-localised motifs during the enrichment test (Section 3.4.1).

3.4 PMET Part 2: Enrichment testing

The purpose of the PMET enrichment test is to quantify how likely a co-localisation between two motifs is to occur in a test promoter set relative to co-localisation across the genome. For every possible pair of motifs, promoters that contain both are identified. Within each such promoter, instances of motifs are checked for overlaps. Any promoters containing overlaps that exceed a threshold are removed. Next, an enrichment test quantifies the significance of the size of the overlap between the target promoter set and the set of promoters containing both motifs. This comparison ensures that the paired motifs identified by the tool are specific to the promoter set that the user wants to test. Finally, a multiple testing correction is applied to all the tests, and the results are returned (Figure 3.4).

The computation time of the co-localisation tests is directly proportional to $N^2/2 - N$ where N is the number of motifs. This means that the size of motif database has a considerable impact on computation time, and is an important consideration before running PMET.

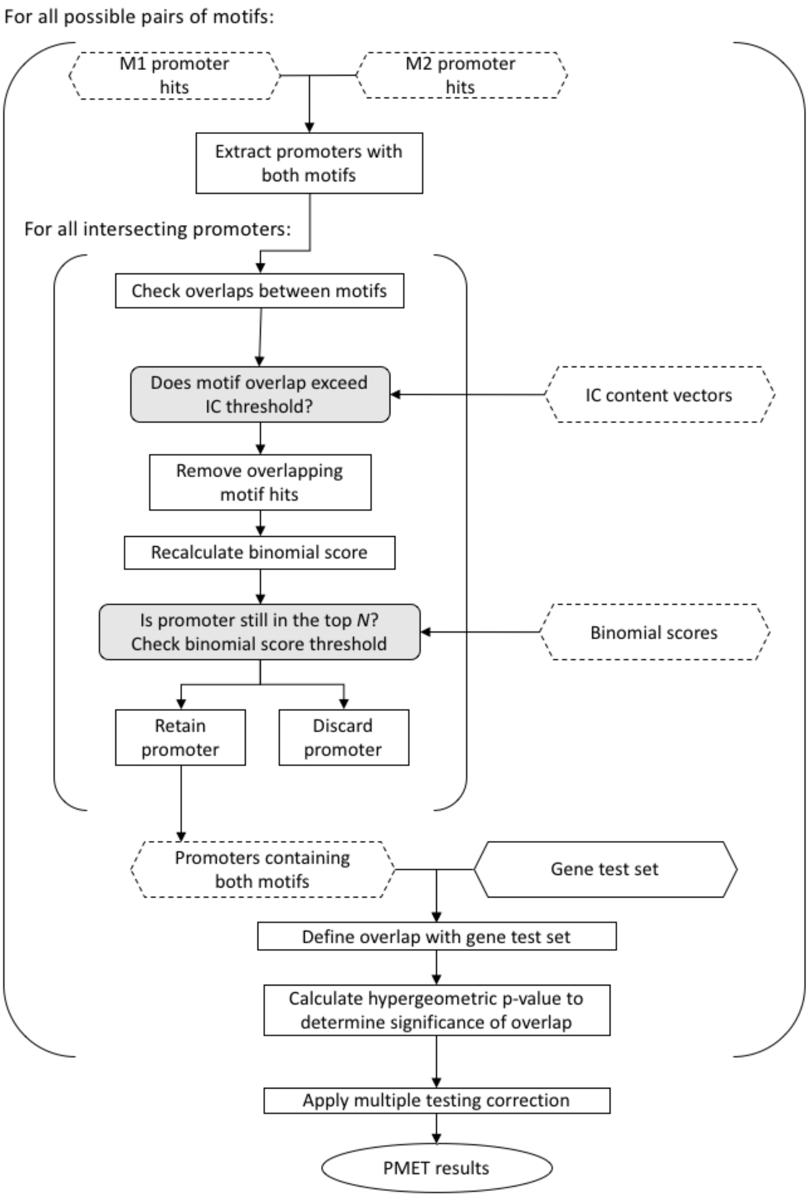

For all possible pairs of motifs:
M1 promoter hits
M2 promoter hits
Extract promoters with both motifs
For all intersecting promoters:
Check overlaps between motifs
Does motif overlap exceed IC threshold?
IC content vectors
Remove overlapping motif hits
Recalculate binomial score
Is promoter still in the top N? Check binomial score threshold
Binomial scores
Retain promoter
Discard promoter
Promoters containing both motifs
Gene test set
Define overlap with gene test set
Calculate hypergeometric p-value to determine significance of overlap
Apply multiple testing correction
PMET results

Figure 3.4: Workflow of PMET Enrichment Testing. Hexagonal shapes with solid lines indicate data to be input by the user, hexagonal shapes with dashed outlines indicate data produced by the tool to be utilised later. Sharp-cornered rectangles indicate computational steps and grey rounded rectangles refer to user-defined parameters or steps that require input from the indexing algorithm.

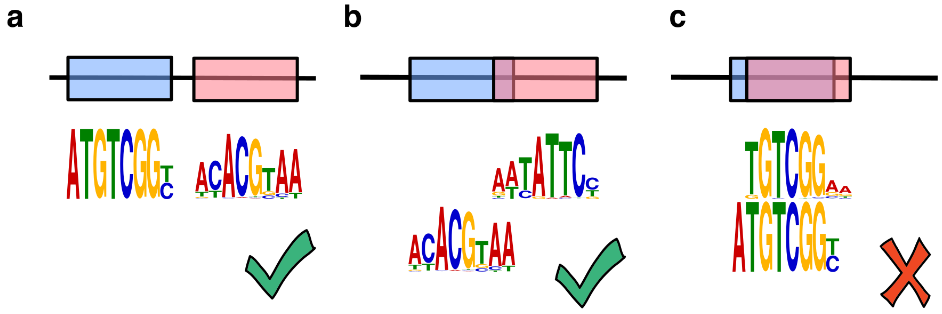


Figure 3.5: Removal of overlapping motifs from PMET analysis. (a) shows two co-localised motifs that do not overlap, (b) shows two motifs that overlap slightly across 2 less-conserved bases. The sum of IC for those two bases is less than 4 for both motifs, so this paired is retained. (c) shows two highly overlapping motifs where the sum of IC for the overlapping bases exceeds 4 for at least one of the motifs and therefore this matching pair would be rejected and removed from the PMET index.

3.4.1 Overlap checking

Unlike single motif enrichment tests which only assess whether or not a motif is present, multiple-motif enrichment tests must consider the amount of overlap motifs must be considered. If the motif positions were not considered (and therefore any overlaps were allowed), any two identical or highly similar motifs would be errantly determined by PMET to be highly co-localised. This would, in turn, make any co-localisations between distinct motifs appear insignificant by comparison. The obvious alternative is to remove any motif instances which overlap. However, recent work on combinatorial TF binding sites (Rodríguez-Martínez et al., 2017) indicates that two binding sites can overlap and bind TF pairs. The solution implemented in PMET is to assess the matches for each motif pair within a promoter individually, and apply a threshold whereby a small amount of overlap is permitted (Figure 3.5b), whereas a large overlap is rejected (Figure 3.5c).

For every pair of motifs, the promoters that contain both motifs are identified. Within each promoter, any overlaps between motif matches are identified based on motif positions. When an overlap is identified, PMET calculates the sum of the IC for the overlapping positions (which had been stored during the indexing, Section 3.3.6). If the total IC of the overlap exceeds the specified threshold in either motif, then that motif-pair is removed (Figure 3.5). The default IC threshold for PMET is 4, corresponding to two fully-conserved base pairs. In practice, an overlap of 4 corresponds to two motifs overlapping by more than two base pairs, as motif positions are rarely fully conserved in flanking bases. Observing an overlap of IC 4 or larger implies that either highly similar motifs have been matched to the same positions, or that these binding sites could no longer function, as steric hindrance would prevent two TFs binding cooperatively.

After the motif matches that exceed the IC threshold have been removed, the binomial test for motif occupancy is recalculated. If the new binomial score exceeds the previously calculated binomial score threshold for either motif, then that promoter is excluded from the analysis. The surviving promoters are used to calculate the hypergeometric p-value for paired motif enrichment.

3.4.2 Hypergeometric test

The pairwise hypergeometric test calculates statistical significance of the overlap between the set of promoters which contain both motifs and target promoter set (significance of the size of the orange set compared to red and yellow sets, respectively (Figure 3.6), relative to number of promoters in the genome. Computation of the log-scale p-value for the pairwise hypergeometric (Fisher's exact) test is based on the efficient function proposed by Meng et al. (2009).

3.4.3 Multiple testing correction (MTC)

PMET provides three multiple testing corrections (MTCs) to correct the hypergeometric p-values; Benjamini-Hochberg, Bonferroni and global Bonferroni (Bonferroni, 1936; Hochberg, 1988; Benjamini and Hochberg, 1995).

- The Benjamini-Hochberg correction ranks the p-values in order of significance and multiplies each p-value by rank number.
- The Bonferroni correction is applied by multiplying all hypergeometric p-values by the number tests performed within each gene cluster.
- The global Bonferroni multiplies hypergeometric p-values by the number of tests performed across all gene clusters.

3.4.4 Results and visualisation of enriched motif pairs

The data that PMET returns to the user includes corrected and uncorrected p-values for each cluster and motif pair combination, the associated set sizes, and the

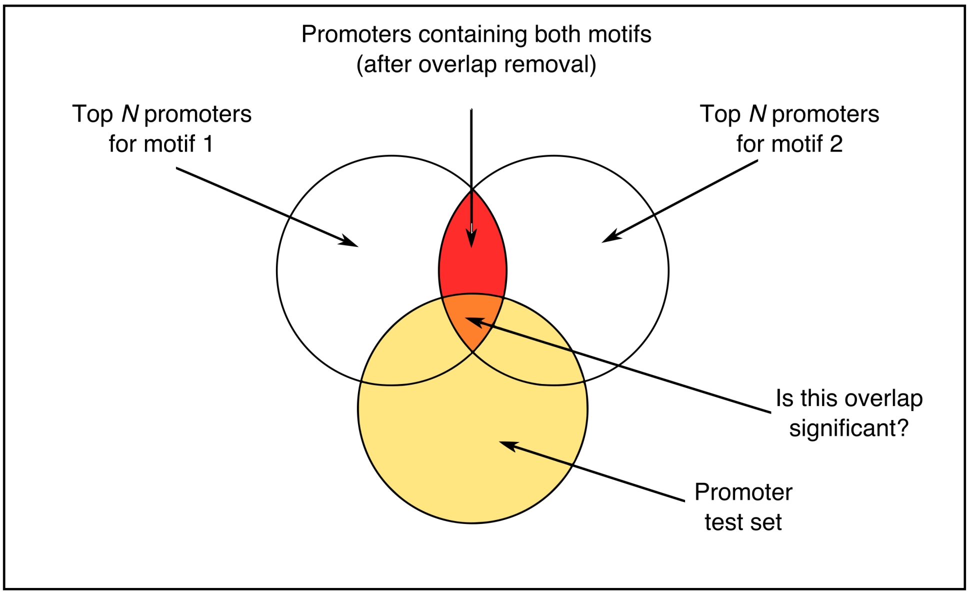


Figure 3.6: Hypergeometric testing diagram. PMET tests whether the size of the overlap (orange set) between the set of promoters containing both motifs (red set) and the target promoters (yellow set) is significant.

list of genes that overlaps between the target promoters and the set of promoters that contain both motifs. PMET's standard output also includes a set of heat maps for each promoter test set that contains a positive hit. Significant pairs of motifs are coloured according to the p-value (insignificant p-values are shown as white) as shown in Figure 3.7.

3.4.5 Parallel computations

The tool was initially designed for *Arabidopsis thaliana* data sets but was subsequently developed to work for any genome. In order to process the larger genomes, the code was extensively profiled and optimised to increase calculation speeds. As genome size increases, the number of promoters to test increases. As detailed in Section 3.3.4, the binomial score in particular was optimised to calculate more efficiently since this is calculated K times for each motif-promoter pair. To further increase computation speeds, parallel programming was utilised to index motifs in promoters in the PMET index algorithm (utilising the unix package `parallel`).

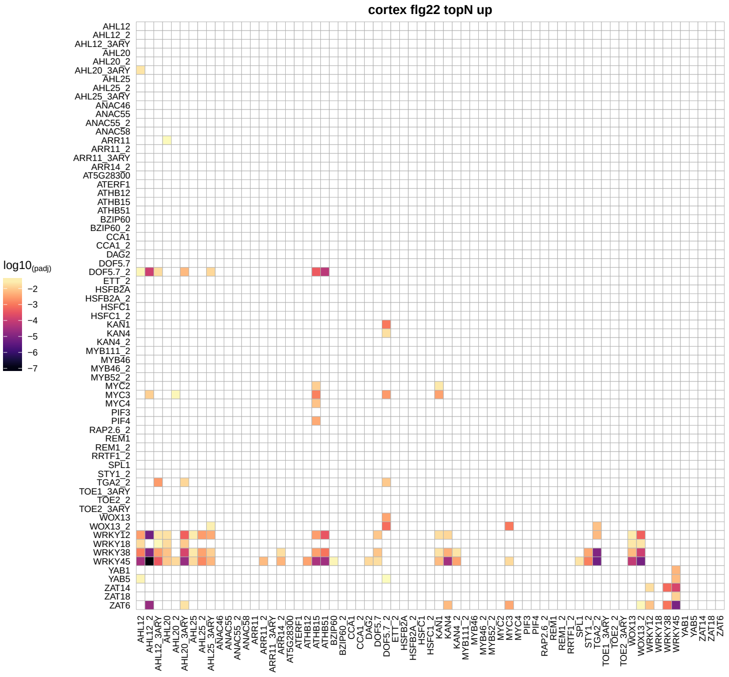


Figure 3.7: Example PMET visualisation. A matrix of all motifs is plotted as a heatmap and significant pairs are coloured on a log-scale according to p-value (more significant p-values indicated by darker colours, insignificant p-values false coloured in white).

3.5 Application of PMET to ATAC-seq datasets

PMET can also be used to investigate motif co-localisation within co-regulated promoters identified in ATAC-seq or similar studies of chromatin accessibility. In this case, the user identifies regions of chromatin that have opened or closed in response to a treatment and compares them to all the open chromatin regions identified in the genome (referred to as the universe). To utilise PMET, the user must provide sequences of all the open chromatin regions in the universe (instead of the genome FASTA and `.gff3` file). These sequences are used instead of promoters so the PMET index algorithm starts from the motif scanning stage. The tool utilises the FASTA sequence headers as 'promoter names' so the user must provide a text file containing FASTA headers of the co-regulated sequences as the target set.

3.6 Parameter sensitivity analysis of PMET

A parameter sensitivity analysis was used to explore the optimal values for each parameter. Investigating the variation in the number of significant hits returned by PMET across a range of values for each parameter revealed the relative importance and the range of optimal values to be recommended to the user. Performing a sensitivity analysis also revealed any parameter sets under which the validity of the statistical test, or the biological results would not hold.

The following parameters were tested in the sensitivity analysis:

- maximum motif matches permitted per promoter (K)
- promoter length
- number of promoters tested per motif (N)
- maximum allowed overlap between motif pairs (measured in IC content)

This was performed by testing a range of values for one or two parameters at a time whilst fixing the remaining parameters at their default value (shown in Table 3.1). All analyses were performed on the *Arabidopsis thaliana* genome and the motif database created by Franco-Zorrilla et al. (2014). Within each analysis, six sets of immune-responsive genes (containing between 128 and 365 genes) and 15 random gene sets (containing between 250 and 750 genes) were analysed for the enrichment of paired motifs. The immunity genes are composed of two sets of flg22 up-regulated gene sets specifically expressed in either epidermis or cortex root cells and four sets of Pep1 responsive genes, split into up- and down-regulated genes that

Table 3.1: Default parameters for PMET and ranges of values tested in the parameter sensitivity analysis. The multiple testing correction (MTC) and p-value threshold (p_{adj}) were consistent across all tests.

Parameter	Default value	Range of values tested
K	5	1-9
Promoter length	1000	100-5000
N	5000	100-27000
IC	4	0-14
MTC	Bonferroni	not tested
p_{adj}	0.05	not tested

are specifically expressed in either epidermis or cortex root cells. These genes were identified and described in Chapter 4, Section 4.2.6.

3.6.1 The effect of varying K

Firstly, the number of significantly enriched motifs pairs (referred to as 'hits') was compared for different values of K between 1 and 12.

In general, there was a dramatic increase in the number of significant hits for $K = 2$ (accepting 1 or 2 motif matches per promoter, based on the binomial score) compared to $K = 1$ (extracting a single motif hit per promoter). Further gains were made by increasing K up to values of 7 (Figure 3.8). The dramatic increase between $K = 1$ and $K = 2$ was particularly evident in epidermis Pep1 up-regulated gene sets, (Figures 3.8e-f). For epidermis Pep1 up-regulated genes the number of significant hits that were identified increased from 3 to 14 as the K increased from 1 to 2. Similarly, over the same K change, the number of significant hits in epidermis Pep1 down-regulated genes increased from 1 to 14. Cortex Pep1 up-regulated genes were the exception to this pattern as the number of significant motif pairs decreased from 5 to 3 over the same K change. However, the number of significant hits peaked strongly at $K = 5$ and the overall pattern is consistent with the other tests. Against random genes, there was at most one significant hit returned from each test, suggesting that increasing K does not increase the number of false positives. The value of K that corresponded to the highest number of significant hits varied by gene set, suggesting that the optimal parameter is dependent on the gene set being tested. However, values of K between 3 and 7 typically returned the most significant results, and therefore this is the range that this study recommends one chooses K from.

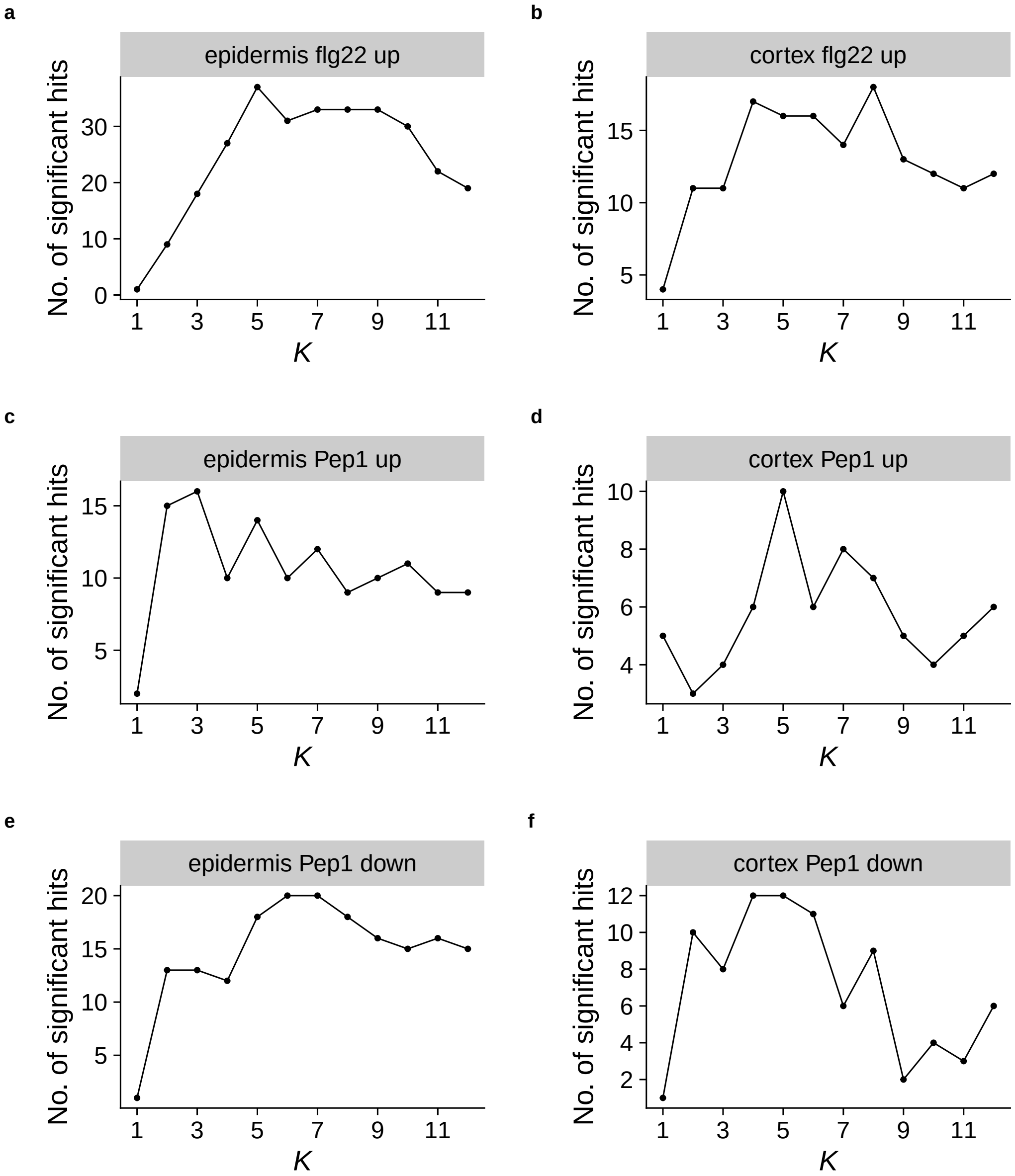


Figure 3.8: Effects in number of significantly increased enriched motif pairs as K changes for six lists of DEGs: (a-b) flg22 up- and (c-d) Pep1 up- and (e-f) down-regulated genes in the epidermis and cortex, respectively. Additional PMET parameters listed in Table 3.1.

3.6.2 The optimum promoter length is different for up- and down-regulated genes

Promoter length is one of the most important parameters in the study of gene promoters. In general, promoters are ill-defined as there is no known promoter start signal (Korkuć et al., 2013). Generally in promoter studies, an arbitrary length between 200 and 2000 bp upstream of the TSS is selected to represent the promoter. Varying the promoter length parameter used by PMET between 200 and 5000 bp revealed that there was no optimal promoter length to capture the most significant hits across all gene sets. Instead, the analysis revealed a distinct difference between up- and down-regulated genes. For up-regulated genes (Figure 3.9a, b, c and e), the number of significant hits identified by PMET peaked between 500 and 1500 bp and returned few or no significant hits at the shortest and longest promoter lengths tested. For the down-regulated genes (Figure 3.9c and f) the number of significant hits was very low in short promoters and increased as promoter length increased from 1000 to 3000 bp, peaking at 3000 bp and 2500 bp for cortex and epidermis Pep1 down-regulated genes respectively. Testing random gene sets returned a maximum of one weakly significant hit with no correlation to promoter length. This indicates that the tool does not return many false positives at any promoter length. Overall, for these gene sets the recommended default value of 1000 bp would be appropriate to capture a reasonable picture of immunity gene regulation, although some resolution may be lost when testing down-regulated genes.

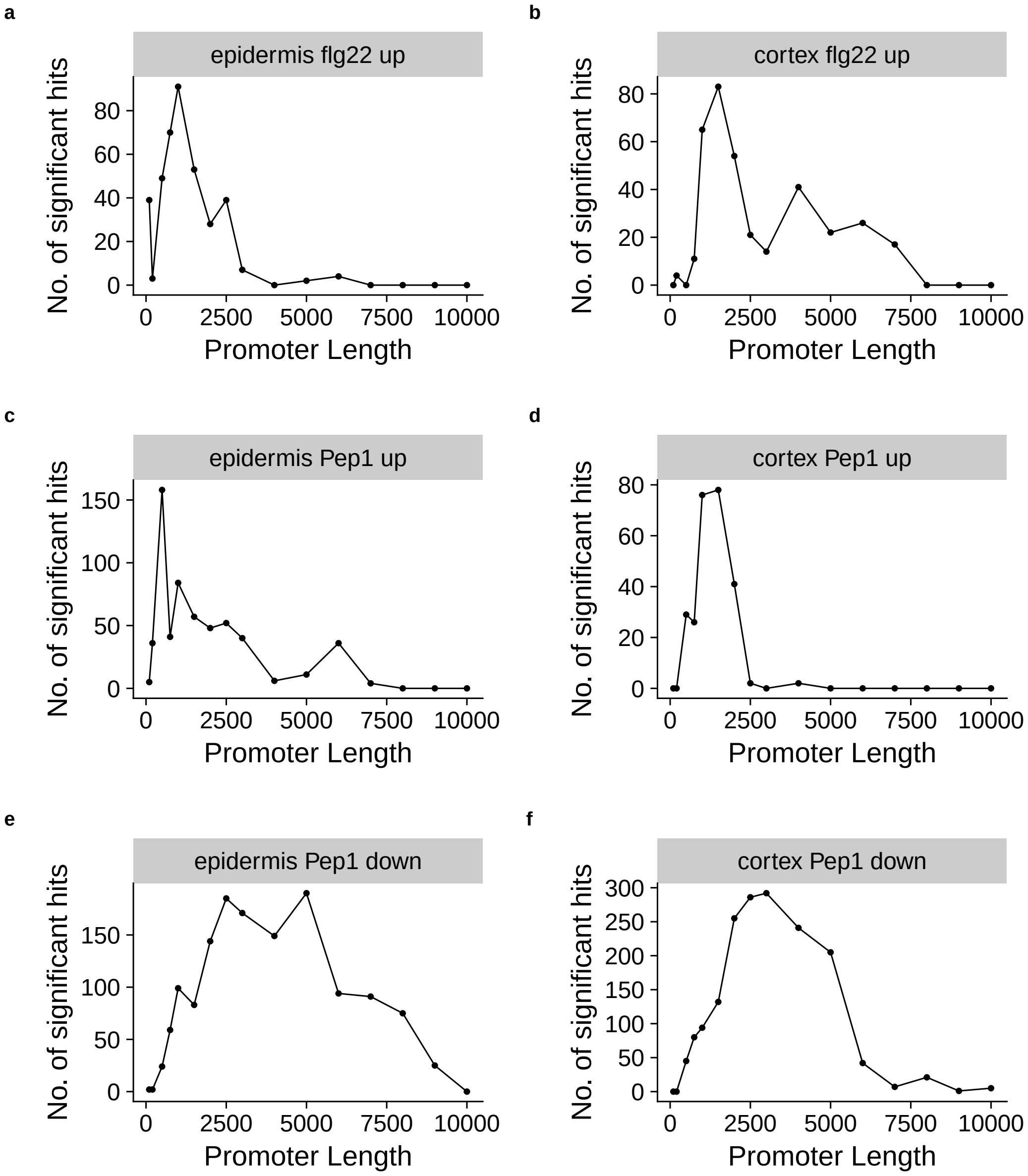


Figure 3.9: Effect on significant hit count under changing promoter length for six lists of DEGs: (a-b) flg22 up- and (c-d) Pep1 up- and (e-f) down-regulated genes in the epidermis and cortex, respectively. Additional PMET parameters listed in Table 3.1.

3.6.3 The relationship between K and promoter length

When considering the parameters for a given run of PMET, the choice of promoter length is intrinsically linked to the choice of K, as in longer promoters there is more space for homotypic motif clusters than in shorter promoters. In order to investigate how this relationship is reflected in the PMET results, the number of significant hits was assessed over varied values of both parameters. Promoter length was varied between 200 and 5000 bp, and K took values between 1 and 9. The highest numbers of significant hits in up-regulated genes were observed for short promoters and low K, whereas the opposite was observed in down-regulated genes (Figure 3.10). The highest number of significant hits in up-regulated gene sets occurred when promoter length was shorter than 2000 bp and K was less than 7. However, in down-regulated genes, the highest number of significant hits occurred when promoter length was more than 2000 bp and K was more than 7.

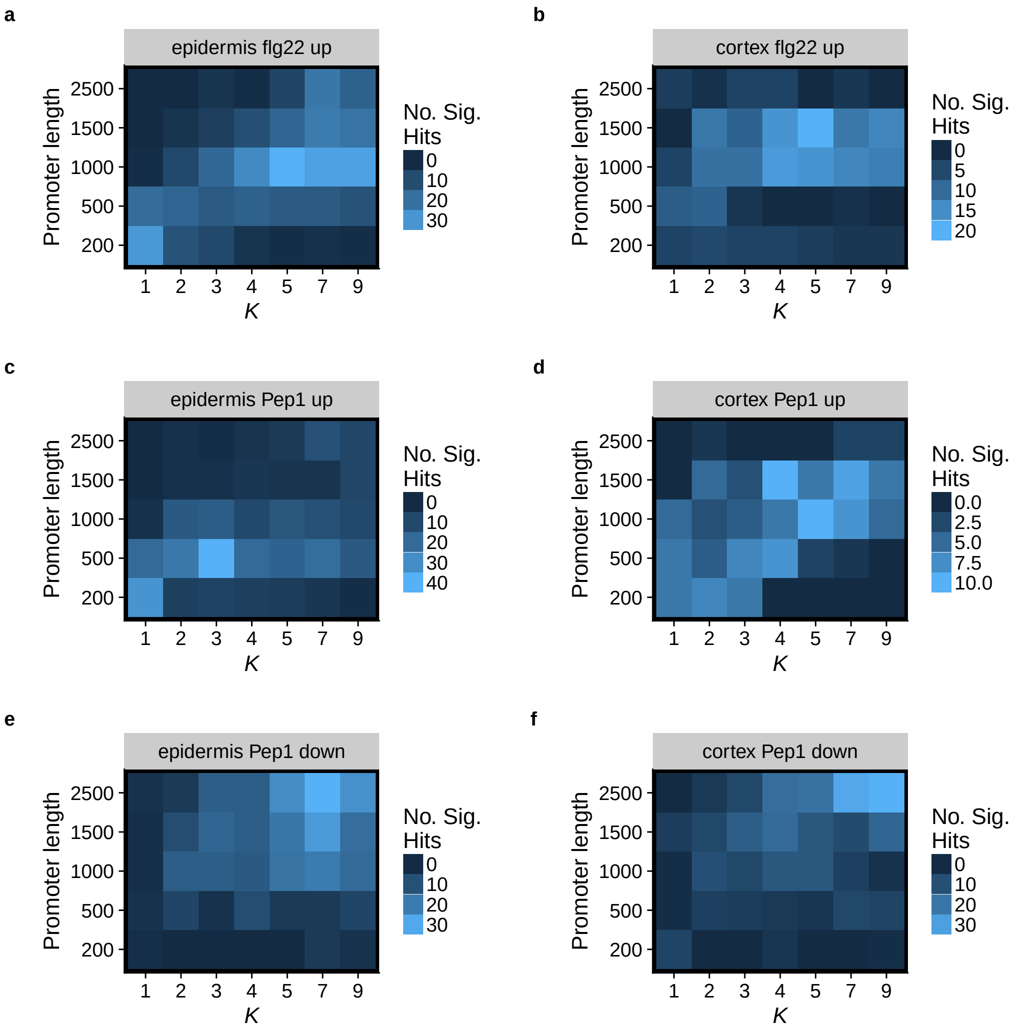


Figure 3.10: Combined effect of varying K and promoter length for six lists of DEGs: (a-b) flg22 up- and (c-d) Pep1 up- and (e-f) down-regulated genes in the epidermis and cortex, respectively. Additional PMET parameters listed in Table 3.1.

3.6.4 Varying *N*

One of the key features of PMET is that it casts a very wide net to capture a large number of potential motif-containing promoters. The algorithm achieves this by first accepting the N most significant promoters that contain each individual motif (lowest scoring promoters according to binomial score), referred to as the top N promoters. The default value for N is 5000. However, Figure 3.11 reveals that the value of N that returns the most significant hits ($p < 0.05$), is higher than 5000 in all six sets of immunity genes. In fact, for the tested genes sets the N that returns the most significant hits is between 15000 and 20000. Furthermore, testing random gene sets returned between 0 and 1 significant hits, uncorrelated with N, indicating that the significant hits observed at high values of N are not false positives.

Changing the value of N will usually change the size of the overlap between the test set and the top N promoters (the observed overlap) and therefore change the p-value corresponding to that overlap. Therefore, for each value of N the minimum number of genes required to obtain a p-value of 0.05 will be different. This number of genes is referred to as the *critical value*. The following example consider the enrichment of a motif pair in a hypothetical gene set of 100 genes, against a genome of 10000 genes.

- If $N = 1000$, then in order to reach a significance level of $p < 0.05$, a minimum of 5 genes need to overlap between the test set and the top N. Therefore the critical value is 5.

- If $N = 5000$, to reach a significance level of $p < 0.05$, a minimum of 42 genes need to overlap between the test set and the top N. Therefore the critical value is 42.

Over a range of N, the value of N that returns the most significant p-value corresponds to the N with the largest difference between the observed overlap and the critical value, referred to as the *optimal* N. The optimal N is different for each test set and motif-pair combination. Figure 3.12 shows that for the majority of motifs, the optimal N is between 15000 and 20000, however some motif pairs have a much lower optimal N.

It is important to note that by taking larger values of N, more motif-pairs become enriched because the standard for significant enrichment is lower. Figure 3.13 shows that as N increases the binomial score threshold increases. Allowing higher binomial scores means that promoters containing weaker motif hits will pass the threshold. However, the method of ranking motifs required for the binomial

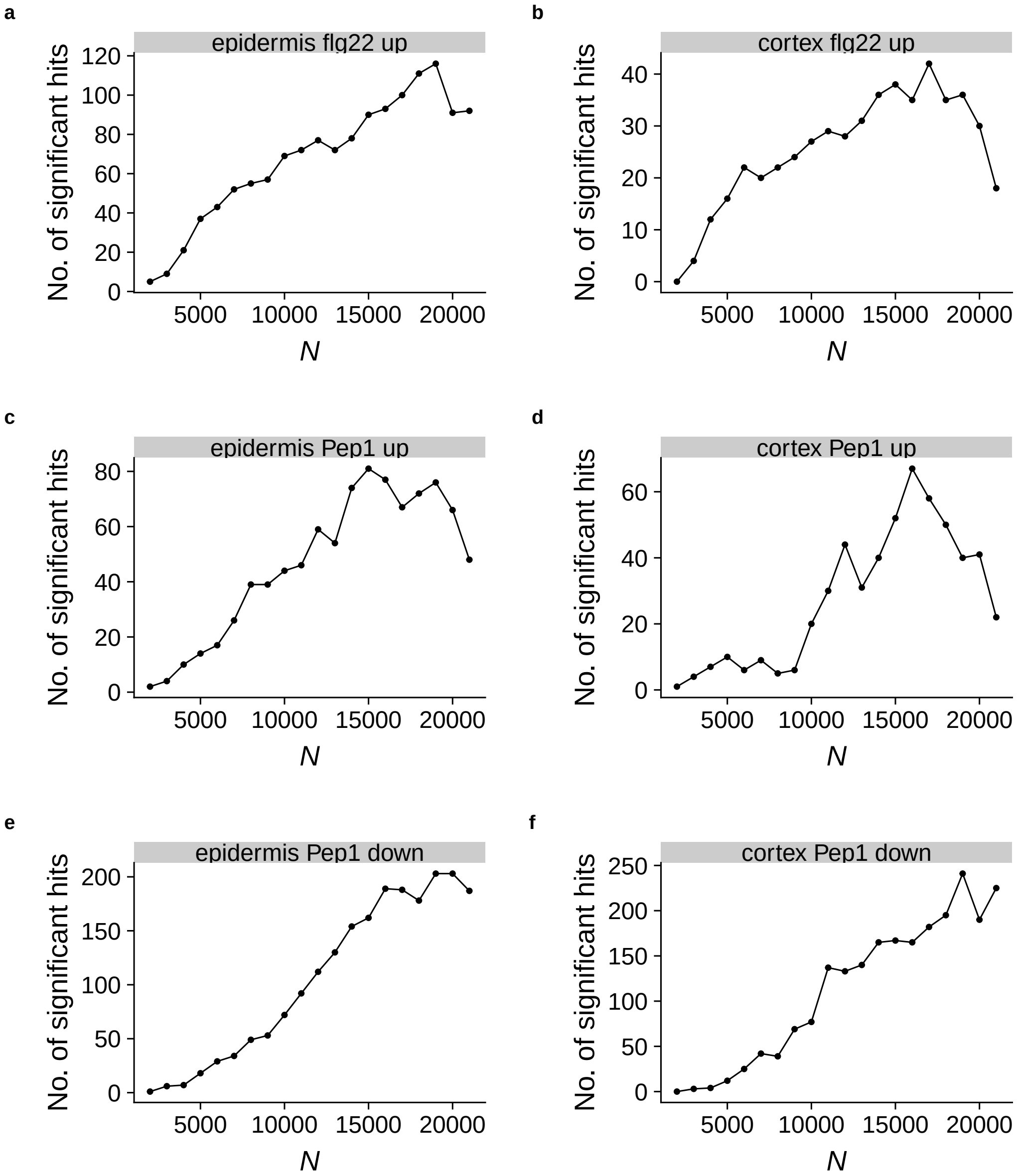


Figure 3.11: Effect on significant hit count under changing N for six lists of DEGs: This is shown for (a-b) flg22 up- and (c-d) Pep1 up- and (e-f) down-regulated genes in the epidermis and cortex, respectively. Additional PMET parameters listed in Table 3.1.

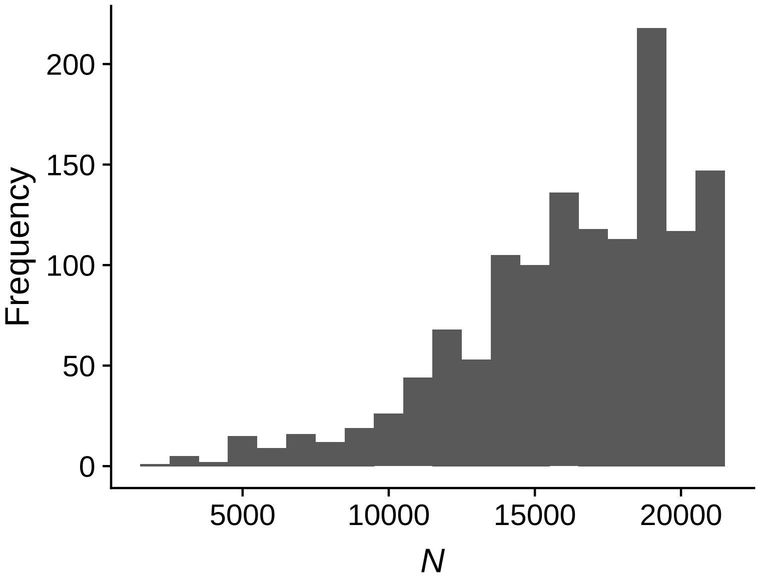


Figure 3.12: Optimal N values for different motif pairs. For each motif-pair, the value of N that returns the most significant p-value is referred to as the *optimal* N. The optimal N is different for each test set and motif-pair combination and varies between 100 and 20000, peaking at 19000.

scoring continues to be effective as the binomial scores do not reach 1 even for large values of N.

Overall, this analysis reveals that the best N to return the highest number of significant hits is between 15000 and 20000. However, using higher values of N includes weaker scoring promoters and therefore may not represent the true biology. Furthermore, increasing N requires considerably more computation time. The user must strike an appropriate balance between these considerations. This computation time restriction led to a choice of $N = 5000$ for all analyses shown in Chapter 4.

3.6.5 Impact of changing IC

One of the key differences between PMET and other published motif tools is the way that PMET manages overlaps between motif pairs. The user defines a maximum overlap allowed between a pair of motifs in terms of IC.

For up-regulated genes, the general trend is that as the maximum allowed overlap is increased, the number of significant hits is also increased (Figure 3.14a, b, d and f). This is particularly evident in the epidermis flg22 up-regulated genes, as the number of significant hits increases from 20 to 47 as the allowed overlap is increased from 0 to 14 IC. This trend is not observed in down-regulated genes

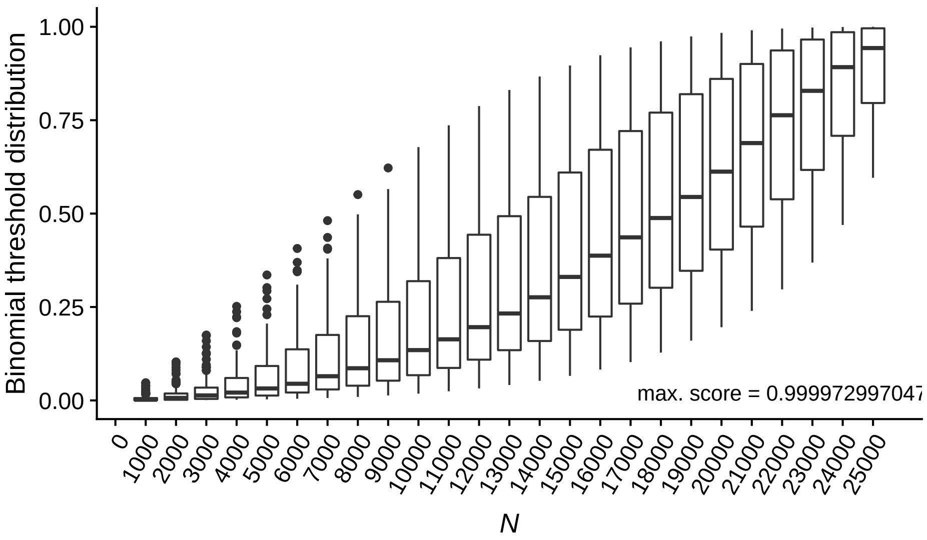


Figure 3.13: The binomial score thresholds increase as N increases. As part of the PMET index algorithm, each motif is assigned a binomial threshold defined as the binomial score for the N^{th} lowest-scoring promoter (according to binomial score). Plotting the distribution of binomial thresholds for all motifs over different values of N reveals that the median of these scores increases as N increases, and the variance widens.

(Figure 3.14c and f). In the epidermis, the number of significant hits varies between 17 and 19 across all overlap options, so the difference in signal is negligible and likely just noise. However, in the cortex, the difference between the minimum and maximum number of significant hits is more substantial, so cannot be explained as noise. For cortex Pep1 down-regulated genes, the number of significant hits is equally high at low and high IC, and drops for middle values of IC.

In all tests, the maximum number of significant hits plateaus as maximal overlap is increased. This plateau could equate to highly similar motifs entirely overlapping one another, as the mean of total IC content across all motifs in the database is 10.2, which is where the start of the plateau is seen in most of the gene sets. This is confirmed by the data in Figure 3.15, which shows that at high maximal overlap (high IC) the new hits being identified are highly similar motifs. Therefore, the IC threshold should be chosen to be between two and six. This is because such a choice returns the most significant hits that are not highly similar.

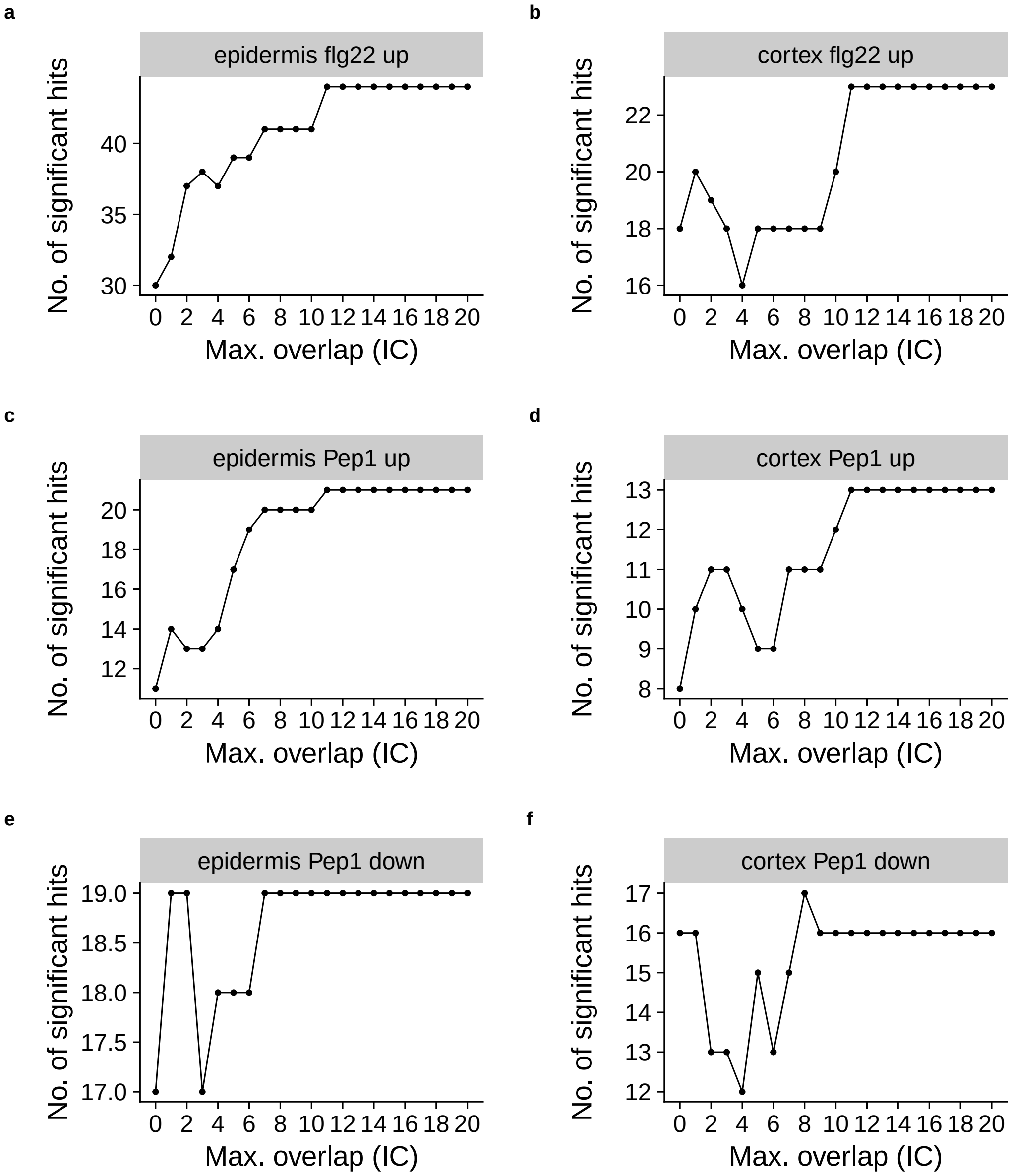


Figure 3.14: Increasing the maximum overlap threshold between motifs increases the total number of significantly enriched paired motifs in (a-b) flg22 up- and (c-d) Pep1 up- and (e-f) down-regulated genes in the epidermis and cortex, respectively. All other PMET parameters are left as default, listed in Table 3.1

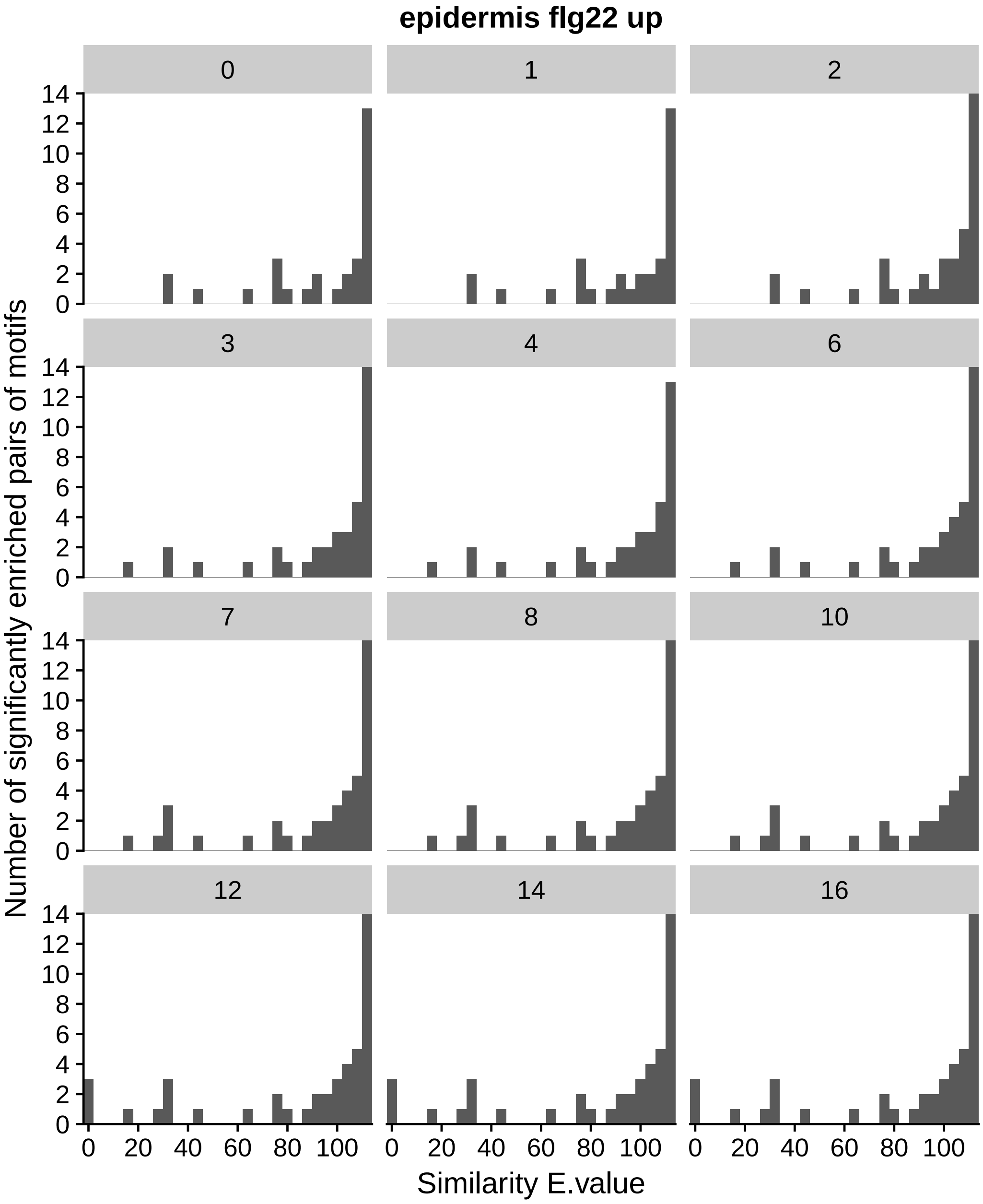


Figure 3.15: Comparison of the similarity between motif pairs relative to the number of promoters they jointly occur in, subject to a varying maximum overlap threshold (IC, indicated in grey title boxes). Each point represents a motif pair. Similarity scores (E-values, defined between 0 and 100) were calculated such that a low score (near 0) indicates high similarity, and high score (near 100) indicates low similarity. Increasing the maximum overlap threshold (in terms of IC) results in more highly similar motifs being detected in promoters. (Data used: epidermis flg22 up-regulated genes). All other PMET parameters set to default values as in Table 3.1.

3.7 Discussion

PMET was developed as a tool that can identify potential regulators of co-expressed genes in any eukaryotic genome. The development of such tools is essential for researchers to start to unpick the complexity of gene regulation in eukaryotes. Eukaryotic gene regulation by TFs is complex on multiple levels. Firstly, within each potential binding site, different bases are conserved to different extents, and therefore TFs can bind to multiple slightly different sequences, likely with differing efficiency. The variability within these sites enable regulatory systems to better control transcription. For example, differences in binding efficiencies could enable one TF to control the expression of several genes at different expression levels (Stormo, 2000). Secondly, and possibly as a result of these less conserved binding sites, some TFs rely on multiple potential binding motifs within a promoter in order to work effectively (Ezer et al., 2014). Thirdly, often multiple TFs must act on a gene promoter in order to activate or repress an individual gene (Pilpel et al., 2001; Wasson and Hartemink, 2009). These features combined with spatial and temporal expression of TFs, enable gene networks to be finely controlled as well as robust. However, the overall complexity makes researching how gene networks are regulated a huge undertaking.

The multiple levels of complexity involved in this regulation means that no developed method is able to capture the complete picture. However, the combination of a variety of methods within PMET allows it to account for a higher level of complexity than many other motif tools. PMET is able to consider multiple binding sites for each TFs, and scores promoters based on the quality of motif multiple matches to each hit, before considering combinatorial regulation by motif pairs. The parameter sensitivity analysis revealed that using a value of K (the number of individual motif matches per promoter) that is more than 1 makes a large difference to the number of significant hits. This insight highlights the importance of considering multiple homotypic motifs as well as pairs of heterotypic motifs when analysing combinatorial control of gene regulation by TFs. There are many mechanisms whereby individual transcription factors utilise multiple binding sites within one promoter (Ezer et al., 2014) and extensive evidence that combinatorial regulation can enhance the specificity of signalling (Zhang et al., 2012; Suryamohan and Halfon, 2015), but little cohesion between the two ideas. As far as we are aware, PMET is unique in its ability to combine information from homotypic and heterotypic clusters, and as such can return high resolution results, based on a variety of gene regulation models.

The fact that there there are no easy to define optimal parameters for PMET, and instead they are different for each tested gene set reflects the complexity of gene regulation. PMET could be indirectly reflecting the variety of mechanisms by which gene networks are regulated on a treatment and cell type-specific level.

3.7.1 PMET sensitivity analyses reveal biological insights.

An unexpected result of the PMET sensitivity analysis was the observation that up- and down- regulated immunity genes respond differently as parameters were changed. In particular, the differences between the optimal values of promoter length and K between up- and down-regulated genes were striking. These patterns suggest that these gene sets might favour different mechanisms of signalling. However, it is important to note that these differences cannot be generalised as specifically distinguishing between up- and down-regulation mechanisms. PMET data does not explicitly reveal whether any of the identified TFs are actively regulating a particular gene set, it merely shows that these TFs are likely to regulate that gene set in some way. The following four simplified scenarios demonstrate why these results must be interpreted carefully.

- Scenario 1: A gene set that is induced or repressed based on the presence or absence of a single TF or TF-pair. In this case, PMET would return the same motifs whether this gene set were being tested as an up- or down-regulated gene set.

- Scenario 2: A gene set that is induced by one TF and repressed by another, possibly through a mechanism of direct competition. In this case, PMET would identify that the motifs for both of these transcription factors are 'co-localised' despite the fact they act in opposition to one another.

- Scenario 3: A set of genes is regulated by TFs that toggle between activator and repressor functions based on environmental signals. In this case, PMET would identify the motifs associated to those TFs, and the mechanism of action could be implied by the direction of differential expression, i.e. if a motif was found enriched in a set of up-regulated genes, the mechanism of the TF is likely to be activation or de-repression. However, it would be hard to identify these TFs as able to toggle between activation and repression using PMET unless two experiments returned the same sets of genes under different experimental conditions.

- Scenario 4: A set of genes is up-regulated by one consistent set of TFs, but a variety of TFs and TF-pairs repress different subsets of these genes. In this case, PMET would identify the motifs corresponding to the TFs that are upregulating the genes as significantly co-localised but not identify the motifs associated with down-regulation, as the enrichment signal would be too weak.

Only in scenario 4 would PMET correctly identify a specific difference between the mechanisms of induction and repression. That being said, genes regulated by mechanisms akin to scenario 4are not unlikely and could potentially be inferred by examining the motif pairs identified by PMET. If the associated TFs to these motif-pairs have been previously shown to act specifically as either repressors or activators, then the PMET results could be shown to be consistent with the experiment that identified the co-regulated gene set. If scenario 4 holds in this case, the pattern of down-regulated genes preferring longer promoters could be an indication of long-range indirect transcriptional repression through mechanisms such as chromatin remodelling (Payankaulam et al., 2010). The extent to which PMET is able to capture differences in signalling mechanisms could only be determined by performing a much larger study of genes known to be co-regulated, preferably in a variety of biological systems.

The most difficult parameter for a PMET user to optimise would be the value of N (the number of promoters which have been indexed for individual motif occupancy in which to test for motif pairs). In this case, the optimal N is defined as the value that returns the most significant hits, which may not be the best N to choose. It might not be the best choice because as N is increased weaker promoters are included in the analysis. However, as shown in Figure 3.9, there is a peak N after which the number of significant hits decreases again, indicating that the binomial score is sufficiently rigorous to rank promoters correctly even at high values of N.

The fact that individual motifs have different optimal values of N must also be considered. In fact, for other motif databases the distribution of optimal Ns may be different. In practice, it would be advisable to perform a preliminary analysis on a small subset of motifs in order to determine the optimal N for that database. The user could either use a random subset of motifs, or choose specific motifs that are particularly relevant to their experiment. Having tested a subset of motifs, the user can then make an informed decision to pick a value of N that is returning significant hits, whilst maintaining a reasonable computation time.

The sensitivity analysis also revealed that increasing the maximum overlap permitted between heterotypic binding sites does not have a large impact on the analysis. One would assume that when the equivalent of a whole motif is allowed to overlap (IC $>$ 10), the number of significant hits would increase dramatically as highly similar motifs assigned to the same positions would be identified as co-localised. Figures 3.14 and 3.15 revealed that in fact, increasing IC to the equivalent of a whole motif only introduces a few extra significant hits. This minimal effect can be explained by the nature of the hypergeometric test: if 2000 promoters were identified as containing both motifs (in a total universe of 20000 promoters), then to obtain a p-value of 0.05 (before MTC), the target set would have to overlap by at least 29 promoters, whereas if 400 promoters were identified as containing both motifs, then to obtain a p-value of 0.05, the target set would have to overlap only 16 promoters. Therefore, is it more likely that one would observe significant hits against smaller 'both-motif' promoter sets. This means that if a very high number of promoters are identified as containing both motifs either because they are highly similar or they are frequently co-localised across the whole genome, then observing a significant hit is much less likely. Conversely, if a pair of motifs co-localises less frequently, a smaller overlap with the target set is required to return a significant hit, but that positive hit would be more specific to the target gene set. The fact that we do not observe a large spike at high IC confirms that the hypergeometric test only identifies motif-pairs that are specific to the gene set being tested, and thus is returning only biologically relevant results.

3.7.2 Validation of PMET results

The top priority following the development of a tool such as PMET is experimental validation of the results. This is essential to test whether predictions of enriched motif-pairs are biologically relevant to the regulation of the target genes. PMET infers transcription factor interaction and activity purely through presence or absence of motif binding sites, but cannot predict whether these sites are actually bound in the context of a particular cell type and/or environmental signal. There are a variety of experimental techniques that could be utilised to validate these results. Firstly, the ability of predicted binding sites to bind the transcription factor in question could be tested *in vitro* using yeast 1-hybrid (Y1H) (Bass et al., 2016), or *in vivo* within particular promoters using chromatin-immunoprecipitation (ChIP)-polymerase chain reaction (PCR) (Mukhopadhyay et al., 2008). The binding of a TF can also be assessed across the whole genome using ChIP-seq (Mundade et al., 2014). In particular, ChIP-seq results could be used to test whether the enrichment of a particular TFs are specific to the gene sets predicted by PMET. The action of a TF on a particular pathway could also be examined using knock-out lines.

PMET does not explicitly predict the interaction of TF pairs, however in many cases pairs of TFs may interact directly in order to co-regulate gene expression. Potential interactions between pairs of TFs can be checked against existing databases such as String (Szklarczyk et al., 2014) or BioGRID (Stark et al., 2006). Alternatively, one could use yeast 2-hybrid (Y2H) to experimentally screen TF interactions *in vitro* and confirm the interactions *in vivo* using techniques such as co-immunoprecipitation (CoIP) or bimolecular fluorescence complementation (BiFC), (reviewed in Xing et al. (2016)).

Additionally, RNA-seq data can be used to test whether the expression of genes coding for TFs is consistent with the observed gene set. However, this can be of limited value, as firstly, transcription factor genes can be expressed at very low levels making them hard to detect and gene expression data does no account for any protein trafficking between cells. This was shown to be the case in *Arabidopsis* roots by Brady et al. (2011) in their study of gene regulatory networks (GRN) in root stele tissue. This study used microarray data to map TF expression and the expression of their targets. They showed that TFs and their targets often showed 'expression domain overlap' (more than zero expression in the same cell type). However, there was limited co-expression with respect to expression levels, i.e. high levels of TF expression did not correlate with high level expression of their targets and vice versa.

Finally, the role of the identified motifs as binding sites in specific promoters could be investigated using quantitative PCR, or reporter constructs. More specifically, the binding sites for one or both of the TFs could be mutated to make the sites non-functional. If the TFs are both required for strong expression of targets, then the expression read out in the mutant promoters would be reduced, or absent. This approach could also be used to assess whether these motifs are required to maintain spatial, temporal or signal-response specificity. For example, promoter constructs fused to fluorescent tags could be used to identify changes in cell type specific gene expression. If a particular motif is required to maintain cell-type specificity then mutation in that site would result in the fluorescent reporter being expressed non-specifically, or in a different domain. Synthetic promoters containing just paired motifs linked by random sequences could also be used to determine whether these motifs are sufficient to induce or repress gene expression.

In addition to the interpreting the biological results, the binomial score for motif occupancy would benefit from validation in order to further improve the mathematical model PMET is based on. The binomial score uses the geometric mean to combine the p-values associated with multiple binding sites into one combination score and the likelihood of these combined p-values appearing by chance in a promoter of that length is calculated. This score is based on the assumption that strong multiple binding sites can result in greater increases of gene expression (Ezer et al., 2014). However, this is based on sequence prediction alone, and doesn't account for whether a TF actually binds to those sites or how specific the binding is. ChIP-seq studies of a few TFs from a variety of TF families would be the ideal method to assess the reality of the binding landscape in order to confirm and further develop this model of motif occupancy.

3.7.3 Opportunities for future development.

One of the shortcomings of PMET is that it does not return motif locations or the number of binding sites that are identified within each promoter. This means that one cannot identify how many motif-pairs physically overlap in a given analysis, and therefore the relative importance of allowing these small overlaps cannot be determined. PMET could be further developed to return structural information about the promoters such as motif positions, scores, and visualisations of overlaps between motifs. This would enable the user to better interpret the datasets beyond the enrichment tests. The user would be able to determine if one or both motifs exist as homotypic clusters, and whether the TFs that bind these motifs are likely to interact with one another directly. In addition to revealing the relationships between motif pairs, location information could also reveal structural information about the promoters, for example whether the target promoters contain A/T or G/C enriched regions.

In addition to developing the output of PMET to be user friendly, comparative genomics could be used to enhance the specificity of PMET's predictions. For example, conservation of non-coding regions between multiple species or cultivars. Conservation is considered to be a 'reliable pointer to essential regulatory elements' and is used as an alternative approach to identify motifs in tools such as APPLES (Baxter et al., 2012). If in addition to being enriched in the dataset, the motifs identified by PMET are also found to be conserved in promoters of other plant genomes, they are especially important and worth further study. A PMET study of *Arabidopsis* promoters could test for conservation between different cultivars using data from the 1001 genomes project (Cao et al., 2011), or between different plant genomes.

In the current implementation of PMET, the computational time restricts which parameter sets can be used. For example, a PMET run with N of 20000 takes multiple days to compute, whereas N of 5000 only takes hours. Increases in motif database size and genome size also increase the computation time. In order to make the tool more attractive to potential users, the computation time must be reduced, perhaps by implementation of further parallelisation or more efficient management of data storage.

Currently the top N is a binary test, either a promoter is in the top N or it isn't. However, each promoter has a quality score associated to it. Future development of PMET's mathematical models could potentially incorporate weighting of promoters based on binomial scores. This could potentially improve the specificity of PMET's biological predictions. In it's current form, the binomial scoring can be used to highlight the most significant promoters as ones that merit further investigation.

Overall, PMET represents a solid framework that can be used to identify highly specific regulatory mechanisms, in order to predict how complex gene networks are maintained. In the sensitivity analyses in this Chapter and in the results of Chapter 4, it has been shown that highly specific biological insights can be made using PMET to examine the differences between between cell type- and treatment specific gene networks.

Chapter 4

Cell type-specific transcriptomic studies of immunity in *Arabidopsis thaliana* roots

4.1 Introduction

Motivated by recent findings suggesting distinct competences of different root cell types in launching PTI (Wyrsch et al., 2015), this chapter analyses the contribution of different root cell types to PTI activation by examining cell type-specific transcriptomic responses. Here I describe the transcriptional networks of three Arabidopsis root cell types; epidermis, cortex and pericycle, following treatment by flg22 and Pep1. *Arabidopsis.* The subsequent analysis shows that different immunity gene networks are activated in the three cell types and, hence, these cell types contribute differently to overall root pattern triggered immunity (PTI). The data was further used to examine the interplay between cell identity and cell type-specific immunity networks, and discuss how plant roots are able to use cell type-specificity to secure root integrity under conditions of environmental stress. Combinatorial TF motif analyses using PMET were used to predict potential regulators of cell type-specific immunity networks, and explain how cell type-specificity is maintained.

Findings presented in this chapter have been submitted to Plant Cell as a manuscript entitled, 'Cell type identity determines transcriptomic immune responses in *Arabidopsis thaliana* roots'. The treatment, FACS and RNA-seq of the main experiment was performed by Ruth Eichmann and Marco Reitz. Sequencing was performed at the Welcome Trust Centre for Human Genetics. All bioinformatic analysis was performed as part of my PhD studies.

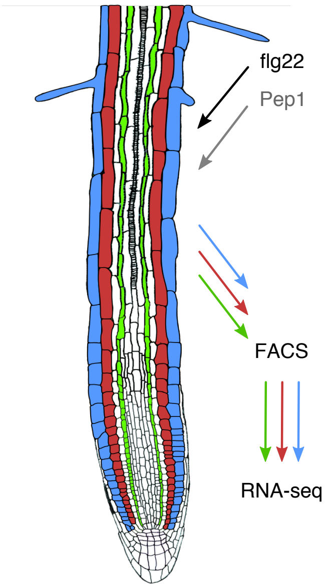


Figure 4.1: Experimental design applied to 10 day old *Arabidopsis* roots. Whole roots were treated with immunity elicitors flg22 or Pep1 (or water as a mock). Epidermis, cortex and pericycle cells were then extracted using fluorescence-activated cell sorting (FACS) and these cell populations were processed by RNA-seq in order to perform cell type-specific gene expression analysis.

4.2 Results

Plants expressing fluorescent cell type markers for epidermis, cortex or pericycle cells were treated with either flg22 or Pep1. Then fluorescent cells were isolated using FACS, and the transcriptome for each cell type was sequenced by RNA-seq (Figure 4.1). The sequenced reads were aligned to the *Arabidopsis thaliana* genome, and differential gene expression analysis was performed to identify genes that respond to immunity activation on a cell type-specific level. Finally, potential regulatory factors were identified using combinatorial motif analysis tool PMET.

4.2.1 Quality control and alignment of RNA-seq data

A total of 716.5 million 100bp paired reads were generated by RNA-seq, with an average of 26.5 million reads per sample. Poor quality reads were trimmed from the dataset, and the quality of the remaining reads was investigated (per sample) using FastQC (Andrews, 2010). FastQC assess the sample quality using a range of criteria including quality of sequencing, adapter content and sequence duplication levels. For each criteria FastQC returns a pass, fail or warning result, and a visualisation of that result to enable the user to understand the source of any warning or fail results.

FastQC analysis revealed that all samples in this dataset passed the criteria for 'adapter content', 'per base nitrogen content', 'per base sequence quality' indicating that in general the sequences are of good quality. The only 'failures' identified by FastQC were for 'sequence duplication levels' in 18 epidermis samples.

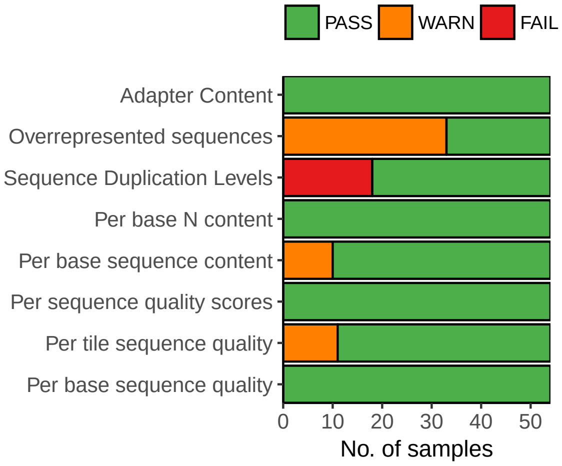


Figure 4.2: Summary of FastQC results for trimmed RNA-seq samples. The number of samples that pass, fail or get labelled as warning are shown for 8 FastQC metrics.

This result indicates that within these samples there is likely to be some kind of enrichment bias, such as PCR over-amplification of reads (Figure 4.2). Further investigation revealed that, on average only 41% of the sequences in these samples are unique, compared to an average of 88% sequences from the cortex and pericycle. The epidermis samples were processed separately to the cortex and pericycle samples, which suggests this is a batch effect from the experimental protocol for those samples caused over amplification of some reads. Parekh et al. (2016) states that duplicated sequences can result from sampling and fragmentation bias as well as PCR preferential amplification of individual reads, however "removal of duplicated sequences improve neither the accuracy nor precision and can actually worsen the power and the False Discovery Rate (FDR) for differential gene expression". As a result, duplicated sequences are retained in subsequent analysis, but further tests for preferential amplification were performed.

33 samples (including the 18 epidermis samples detailed above) were also flagged with a warning for overrepresented sequences. This module issues a warning if any individual sequences represent more than 0.1% of the total library. Inspection of these samples reveals that those flagged with warnings included samples from all three cell types and from various sequencing pools and replicates, indicating that this was not a batch effect but a consistent pattern. One of the overrepresented sequences that occurred in these samples was 'TGTTTGATTTTTTTTTTTTTTTTTTTTTT TTTTTTTTTTTTTTTTTTTTT' indicating these duplicated reads correspond to polyA tails attached to an adaptor which would be unlikely to align to the genome, and therefore will be filtered out of the analysis during alignment.

Eleven samples were flagged with a warning for 'per tile sequence quality' indicating that within the indicating that some of the flow cells tiles were of low quality at some read positions. Visual inspection of the 'per tile sequence quality' for representative sample WTCHG_129180_01 reveals 3 distinct tiles with low quality bases (Figure 4.3) but overall the quality of sequencing per tile is high and that these lower scoring tiles will not affect the subsequent analysis..

Ten samples all sourced from the pericycle material failed the 'Per base sequence content' criterion. In a random library, the proportions of the four bases are expected to approximately equal, however if the libraries contain overrepresented or Kmer sequences at the end of reads then there is more likely to be a bias towards specific bases. In our dataset there is a consistent pattern of biases towards certain bases in the first 8-9 bases of the sequences, (Figure 4.4, although this is observed in most samples, not just those marked with a warning. This feature occurs when some of the hexamers used for random priming of the library production are favoured over others rather than priming with equal efficiency. These consistent biases are very similar across all libraries and did not effect the subsequent alignment.

Overall, performing quality control of the RNA-seq samples using FastQC revealed that the data was of sufficiently high quality to enable us to be confident in the subsequent analysis.

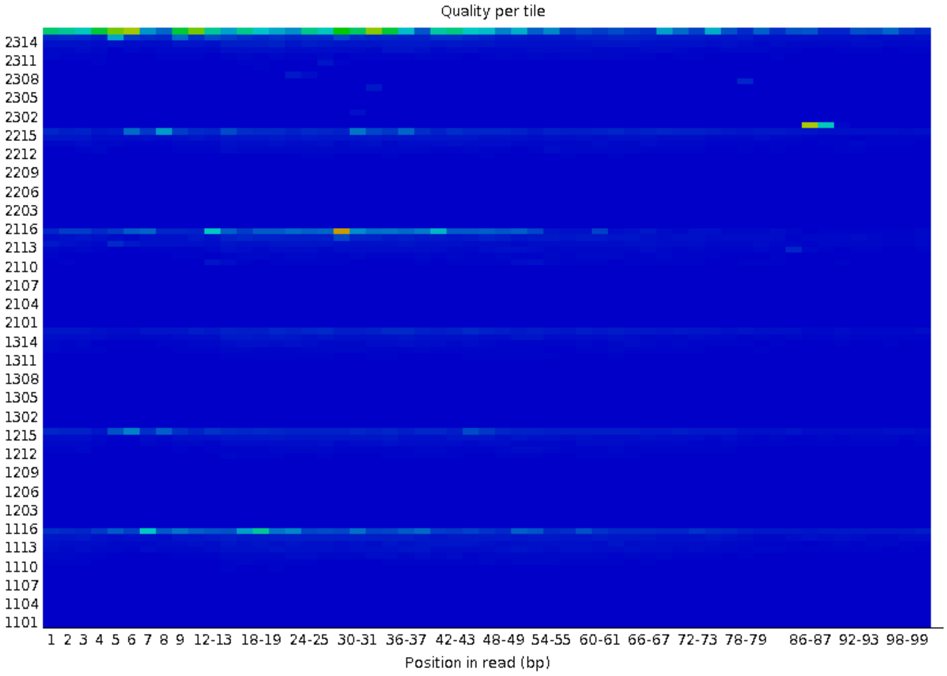


Figure 4.3: Quality per tile of sequencing flow cells for sample WTCHG_129189_01_R2. The plot shows the deviation from the average quality for each tile, per bp of sequence (plot produced using FastQC). The heat map is plotted on a cold to hot scale, with cold colours being positions where the quality was equal to or above average for that base in the run. Hotter colours indicate that a tile had on average worse quality than other tiles for that base.

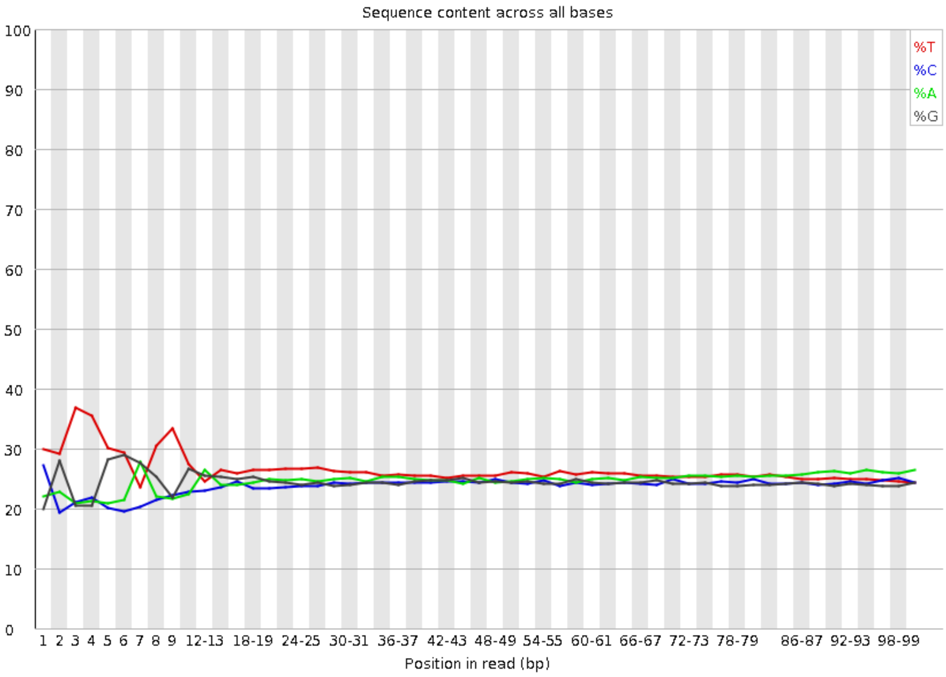


Figure 4.4: FastQC per base sequence content plot shows the proportion of each base position in a sample. The plot shows that there is bias towards specific bases occurs in the first 8-9 bases of sequences in representatives sample WTCHG_125416_03_R2 (plot produced using FastQC).

4.2.2 Inspecting alignment quality

After the initial quality control checks, approximately 373 million reads (an average of 13.8 million read pairs per sample) were uniquely mapped to the *Arabidopsis* genome. On average across all samples, 57.9% of all reads mapped to unique positions in the genome, whereas 34.1% of reads mapped to multiple positions and only 7.9% reads did not map to the genome (Tables 4.1 and 4.2).

Since some potential batch effects had been identified by cell type, the alignment statistics were grouped and a statistical summary of each metric was plotted as a box plot (Figure 4.5). This plot revealed that the pericycle had the widest variance in total read counts, whereas the epidermis samples were much more consistent in size. The epidermis samples contained on average fewer reads than the other two cell types, but the proportion of those reads that were multi-mapped was much smaller. In contrast, the mean read number of uniquely mapped and multi-mapped reads were very similar in the pericycle sample indicating a high proportion of these reads did not map to unique positions. By proportion, 70.9% of epidermis reads uniquely mapped to the genome, whereas only 52.8% of cortex and 50.0%

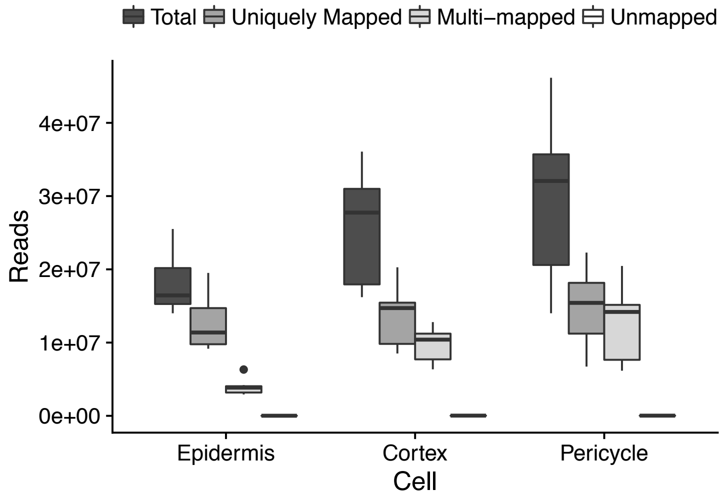


Figure 4.5: Alignment statistics summarised by cell type. Box plots show the summary statistics of read numbers from Table 4.1 grouped by cell type.

of pericycle reads mapped uniquely. Of these aligned reads, 271 million aligned to exons in the *Arabidopsis* genome (an average of 10 million per sample, Table 4.3). The majority of reads that did not align to exons are likely to result of contaminant DNA in the RNA sample or unspliced RNA. Some reads could also be the result of unnanotated exons, or splicing events.

Replicate quality across samples was assessed by plotting the read counts of replicates against each other. Ideally two replicates from the same condition should correlate closely together such as between the epidermis flg22 replicates shown in Figure 4.6a. Inspecting the correlations between replicates revealed potential data artefacts in some replicates. For example in the cortex there were 'side peaks' below the main distribution (Figure 4.6b), and in some epidermis samples there were 'side peaks' above the main distribution and the distribution did not collapse into a tight peak at high read counts. The genes contributing to these side peaks were assessed using the method described in Materials and Methods and revealed to be highly enriched in genes coding for ribosomal proteins indicating preferential cDNA synthesis of ribosomal protein mRNAs and rRNAs prior to RNA-seq. These artefacts increase noise within the dataset in turn reducing the accuracy of differential gene expression analysis. As a result, we restricted the analysis to nuclear-encoded protein coding genes and filtered reads that corresponded to ribosomal protein mRNAs (listed in Appendix A) from all samples. The reduction of side peaks following filtering is clear in the replicate plots in both the cortex (Figure 4.7a) and epidermis (Figure 4.7b). The pericycle replicates were consistently noisier than the other two

Table 4.1: Alignment statistics for mapped reads. Table shows the total number of reads per sample and the number of reads that mapped uniquely to the *Arabidopsis thaliana* genome and the proportion relative to total number of sequenced reads for each sample.

Sample	Total Reads	Uniquely Mapped Reads	% Uniquely Mapped Reads
WTCHG_125416_01	3.49×10^{7}	1.68×10^{7}	48.2
WTCHG_125416_03	3.21×10^{7}	1.54×10^{7}	48.1
WTCHG_125416_05	2.88×10^{7}	1.46×10^{7}	50.6
WTCHG_129187_01	3.10×10^{7}	1.54×10^{7}	49.8
WTCHG_129187_03	2.86×10^{7}	1.55×10^{7}	54.0
WTCHG_129187_05	3.27×10^{7}	1.88×10^{7}	57.3
WTCHG_129187_07	1.77×10^{7}	9.82×10^{6}	55.5
WTCHG_129189_01	1.79×10^{7}	8.53×10^{6}	47.5
WTCHG_129189_03	2.67×10^{7}	1.32×10^{7}	49.4
WTCHG_129189_05	3.61×10^{7}	2.03×10^{7}	56.2
WTCHG_129189_07	1.73×10^{7}	8.62×10^{6}	49.8
WTCHG_129190_01	1.62×10^{7}	8.51×10^{6}	52.5
WTCHG_129190_03	2.77×10^{7}	1.47×10^{7}	53.0
WTCHG_129190_05	1.40×10^{7}	6.72×10^{6}	48.0
WTCHG_129190_07	4.62×10^{7}	2.23×10^{7}	48.3
WTCHG_131167_01	3.62×10^{7}	1.82×10^{7}	50.1
WTCHG_131167_03	3.57×10^{7}	1.87×10^{7}	52.4
WTCHG_131167_05	2.06×10^{7}	1.12×10^{7}	54.5
WTCHG_203594_01	2.02×10^{7}	1.47×10^{7}	72.9
WTCHG_203594_03	1.53×10^{7}	1.11×10^{7}	73.0
WTCHG_203594_05	1.40×10^{7}	9.77×10^{6}	69.8
WTCHG_203594_07	1.44×10^{7}	9.75×10^{6}	67.7
WTCHG_203594_10	2.25×10^{7}	1.70×10^{7}	75.3
WTCHG_203839_01	1.53×10^{7}	1.14×10^{7}	74.1
WTCHG_203839_04	2.55×10^{7}	1.95×10^{7}	76.5
WTCHG_203839_06	1.64×10^{7}	9.16×10^{6}	55.7
WTCHG_203839_08	1.85×10^{7}	1.35×10^{7}	73.1

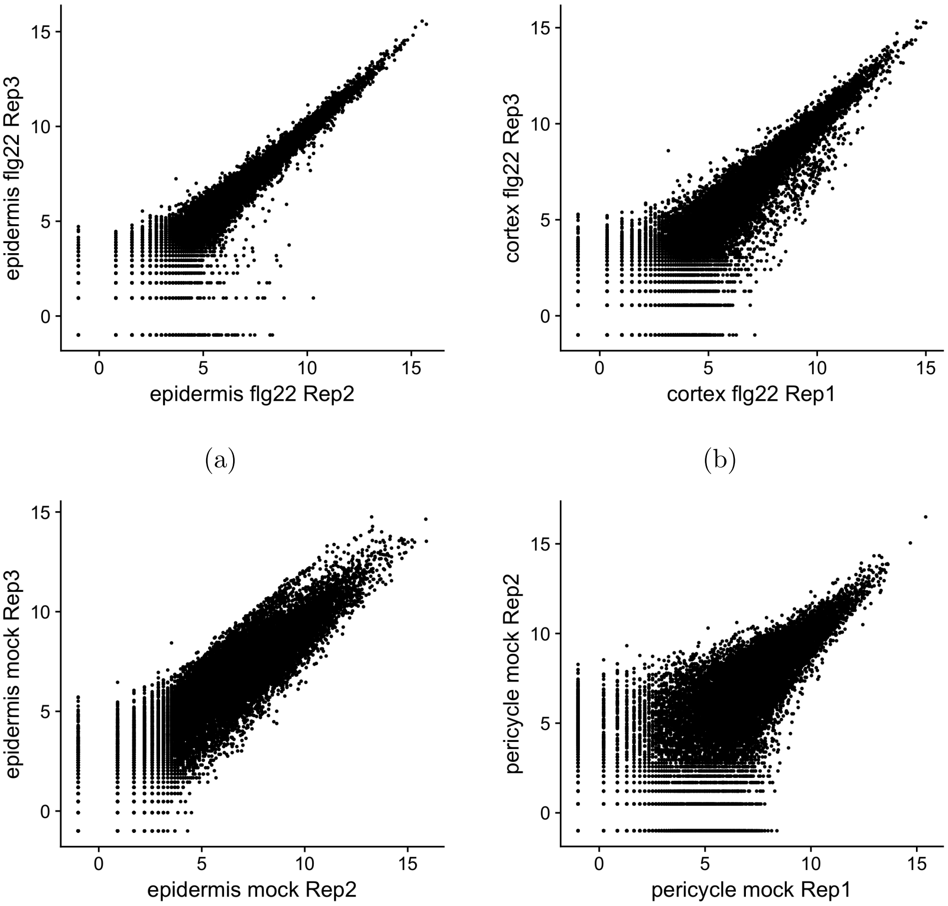
Table 4.2: Alignment statistics for unmapped or multi-mapped reads. Table shows the total number of reads that either multi-mapped and or were unmapped to the *Arabidopsis thaliana* genome and proportions relative to total number of sequenced reads are shown for each sample.

Sample	Total Reads	Multi-mapping reads	% Multi-mapping reads	Unmapped reads	% Unmapped reads
WTCHG_125416_01	3.49×10^{7}	1.55×10^{7}	44.3	25861	7.42
WTCHG_125416_03	3.21×10^{7}	1.42×10^{7}	44.2	24561	7.66
WTCHG_125416_05	2.88×10^{7}	1.18×10^{7}	41.0	24177	8.40
WTCHG_129187_01	3.10×10^{7}	1.27×10^{7}	41.1	28106	9.07
WTCHG_129187_03	2.86×10^{7}	1.05×10^{7}	36.5	27190	9.50
WTCHG_129187_05	3.27×10^{7}	1.04×10^{7}	31.7	35944	11.00
WTCHG_129187_07	1.77×10^{7}	6.48×10^{6}	36.6	1.4×10^{4}	7.89
WTCHG_129189_01	1.79×10^{7}	7.71×10^{6}	43.0	17067	9.51
WTCHG_129189_03	2.67×10^{7}	1.12×10^{7}	42.1	22916	8.59
WTCHG_129189_05	3.61×10^{7}	1.28×10^{7}	35.5	29939	8.30
WTCHG_129189_07	1.73×10^{7}	7.34×10^{6}	42.4	13554	7.83
WTCHG_129190_01	1.62×10^{7}	6.35×10^{6}	39.2	13399	8.27
WTCHG_129190_03	2.77×10^{7}	1.04×10^{7}	37.5	26387	9.51
WTCHG_129190_05	1.40×10^{7}	6.16×10^{6}	44.0	11232	8.02
WTCHG_129190_07	4.62×10^{7}	2.05×10^{7}	44.3	3.4×10^{4}	7.37
WTCHG_131167_01	3.62×10^{7}	1.51×10^{7}	41.8	29275	8.08
WTCHG_131167_03	3.57×10^{7}	1.43×10^{7}	40.0	27169	7.61
WTCHG_131167_05	2.06×10^{7}	7.65×10^{6}	37.1	17303	8.40
WTCHG_203594_01	2.02×10^{7}	4.04×10^{6}	20	14220	7.05
WTCHG_203594_03	1.53×10^{7}	3.00×10^{6}	19.7	11223	7.35
WTCHG_203594_05	1.40×10^{7}	3.18×10^{6}	22.7	10487	7.49
WTCHG_203594_07	1.44×10^{7}	3.6×10^{6}	25.0	10515	7.30
WTCHG_203594_10	2.25×10^{7}	3.93×10^{6}	17.5	16205	7.20
WTCHG_203839_01	1.53×10^{7}	2.92×10^{6}	19.0	10463	6.82
WTCHG_203839_04	2.55×10^{7}	4.19×10^{6}	16.4	18087	7.09
WTCHG_203839_06	1.64×10^{7}	6.32×10^{6}	38.4	9669	5.88
WTCHG_203839_08	1.85×10^{7}	3.85×10^{6}	20.8	11114	6.02

Table 4.3: Alignment to exons as calculated by LiBiNorm. Table shows the number of aligned reads that uniquely mapped to exons and the relative proportion to total aligned reads, The remaining reads are split into three categories: 'no feature' (did not map to an exon), 'ambiguous' (ambiguously mapped, e.g. did not map completely to an exon) or 'alignment not unique' (mapped to multiple exons).

Sample	Unique mapped to exon	Unique mapped to exon (%)	No feature	Ambiguous	Alignment not unique
WTCHG_125416_01	1.06×10^{7}	28.0	4.07×10^{6}	2.14×10^{6}	3.24×10^{7}
WTCHG_125416_03	9.68×10^{6}	27.6	3.74×10^{6}	2.00×10^{6}	2.97×10^{7}
WTCHG_125416_05	9.87×10^{6}	31.5	3.12×10^{6}	1.56×10^{6}	2.53×10^{7}
WTCHG_129187_01	1.00×10^{7}	30.3	3.09×10^{6}	2.33×10^{6}	2.65×10^{7}
WTCHG_129187_03	1.11×10^{7}	36.1	2.48×10^{6}	1.90×10^{6}	2.16×10^{7}
WTCHG_129187_05	1.41×10^{7}	40.3	2.61×10^{6}	2.01×10^{6}	2.17×10^{7}
WTCHG_129187_07	6.85×10^{6}	36.5	1.78×10^{6}	1.19×10^{6}	1.36×10^{7}
WTCHG_129189_01	5.38×10^{6}	27.1	1.92×10^{6}	1.24×10^{6}	1.61×10^{7}
WTCHG_129189_03	8.78×10^{6}	29.7	2.70×10^{6}	1.69×10^{6}	2.33×10^{7}
WTCHG_129189_05	1.50×10^{7}	37.4	3.22×10^{6}	2.08×10^{6}	2.68×10^{7}
WTCHG_129189_07	5.45×10^{6}	29.0	2.13×10^{6}	1.04×10^{6}	1.55×10^{7}
WTCHG_129190_01	5.74×10^{6}	33.6	1.72×10^{6}	1.05×10^{6}	1.33×10^{7}
WTCHG_129190_03	9.97×10^{6}	34.1	2.90×10^{6}	1.84×10^{6}	2.18×10^{7}
WTCHG_129190_05	4.17×10^{6}	27.9	1.76×10^{6}	7.94×10^{5}	1.30×10^{7}
WTCHG_129190_07	1.33×10^{7}	27.1	6.15×10^{6}	2.82×10^{6}	4.37×10^{7}
WTCHG_131167_01	1.18×10^{7}	30.2	3.58×10^{6}	2.77×10^{6}	3.21×10^{7}
WTCHG_131167_03	1.32×10^{7}	34.1	3.23×10^{6}	2.29×10^{6}	2.98×10^{7}
WTCHG_131167_05	8.02×10^{6}	35.9	1.86×10^{6}	1.35×10^{6}	1.60×10^{7}
WTCHG_203594_01	1.27×10^{7}	58.1	1.30×10^{6}	7.27×10^{5}	8.42×10^{6}
WTCHG_203594_03	9.53×10^{6}	57.4	1.06×10^{6}	5.46×10^{5}	6.28×10^{6}
WTCHG_203594_05	8.13×10^{6}	53.7	1.09×10^{6}	5.58×10^{5}	6.76×10^{6}
WTCHG_203594_07	7.97×10^{6}	50.8	1.18×10^{6}	6.07×10^{5}	7.63×10^{6}
WTCHG_203594_10	1.47×10^{7}	60.8	1.34×10^{6}	8.85×10^{5}	8.56×10^{6}
WTCHG_203839_01	9.87×10^{6}	59.1	1.01×10^{6}	4.93×10^{5}	6.09×10^{6}
WTCHG_203839_04	1.72×10^{7}	62.3	1.51×10^{6}	8.09×10^{5}	8.81×10^{6}
WTCHG_203839_06	6.45×10^{6}	35.6	1.77×10^{6}	9.33×10^{5}	1.30×10^{7}
WTCHG_203839_08	1.15×10^{7}	57.9	1.28×10^{6}	7.04×10^{5}	8.17×10^{6}

cell types, with a much wider variance of read counts between replicates (Figure 4.6d), however there are no obvious artefacts between replicates in the dataset.



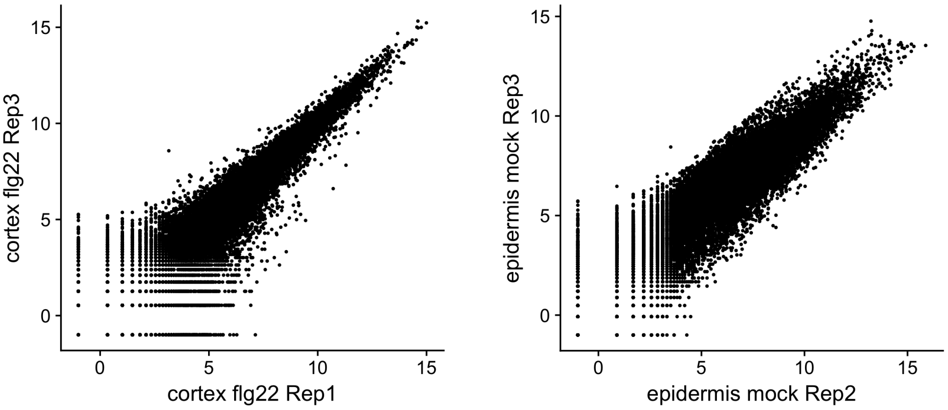
(a)

(b)

(c)

(d)

Figure 4.6: Unfiltered replicate plots. The normalised $\log_2$ read counts for aligned reads to each exon from two replicates were compared to examine the correlation between replicates. The comparisons reveal (a) the ideal distribution between epidermis flg22 replicates 2 and 3, (b) 'side diagonals' below the main distribution between cortex flg22 replicates 1 and 3, and (c) above for epidermis mock replicates 2 and 3 and (d) a less strongly correlated pair of replicates for pericycle mock replicates 1 and 2.



(a) (b)

Figure 4.7: Filtered replicate plots. The normalised $\log_2$ read counts for aligned reads to each exon from two replicates were compared to examine the correlation between replicates. Filtering of data artefacts results in a reduction of ‘side diagonals’ between (a) cortex flg22 replicate 1 and 3 and (b) epidermis mock replicates 2 and 3.

4.2.3 Principal component analysis

The principal component analysis (PCA) of the RNA-seq samples revealed the majority (approx. 83%) of the variation within the dataset was contained within the first three principal components (Figure 4.8). Plotting PC1 and PC2 against one another revealed close clustering of samples by cell type indicating that cell identity was the principal source of variation, (Figure 4.9, left) accounting for 78% of the variation (62% and 16% from PCs 1 and 2 respectively). Consistent with this, cell type marker genes were highly expressed in the respective cell type populations (Figure 4.10). The samples are distinctly separated by immune response in PC3 (5% variation, Figure 4.9b) indicating that flg22 and Pep1 must induce different transcriptional responses compared to each other and the mock treatment.

4.2.4 Differential gene expression analysis of cell type-specific immune responsive genes.

Based on the separation of flg22 and Pep1 in the PCA (Figure 4.9), differential gene expression analysis was performed to identify the genes contributing to the differing immune responses. Replicate plots which overlaid the differentially expressed genes were plotted to check the quality of the `DESeq2` differential gene expression (DGE) results (Figure 4.11). In each of the plots the red dots indicating differentially expressed genes (DEGs) are clustered at the edge of the read count distribution, indicating a good model fit.

A total of 3276 unique genes were differentially expressed in response to one or both elicitors in at least one cell type. Consistent with a recent study (Poncini et al., 2017), Pep1 treatment elicited the most DEGs (3082), whereas many fewer DEGs (884) were elicited by flg22 treatment. Poncini et al. used a variety of PTI assays, including ROS burst, MAPK phosphorylation, and qPCR analysis of defence genes, to demonstrated that Pep1 elicits a stronger immune response than flg22 or fungal chitin in whole *Arabidopsis* roots.

A stronger Pep1 response than flg22 was observed in the DGE results in all three cell types. In the epidermis 601 genes were DE (569 up- and 32 down-regulated) in response to flg22, in the cortex 476 genes were DE (344 up and 132 down) and in the pericycle 98 genes were DE (74 up and 24 down). By contrast Pep1 elicited 1700 DEGs in the epidermis (964 up and 736 down), 2,187 in the cortex (1148 up and 1039 down) and 528 in the pericycle (335 up and 193 down, Figure 4.12). Across all three cell types, flg22 induced vastly more genes than it represses whereas Pep1 induces and represses genes in equal measure, and Pep1 consistently

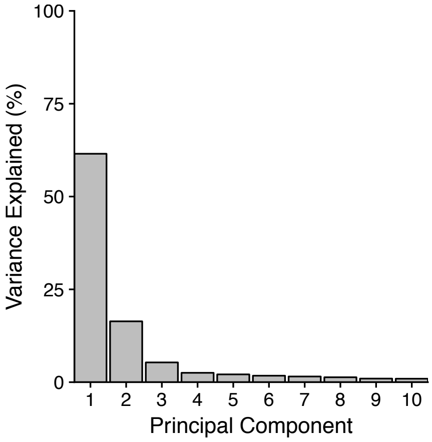


Figure 4.8: Scree plot showing the percentage of variance explained by the top ten principal components.

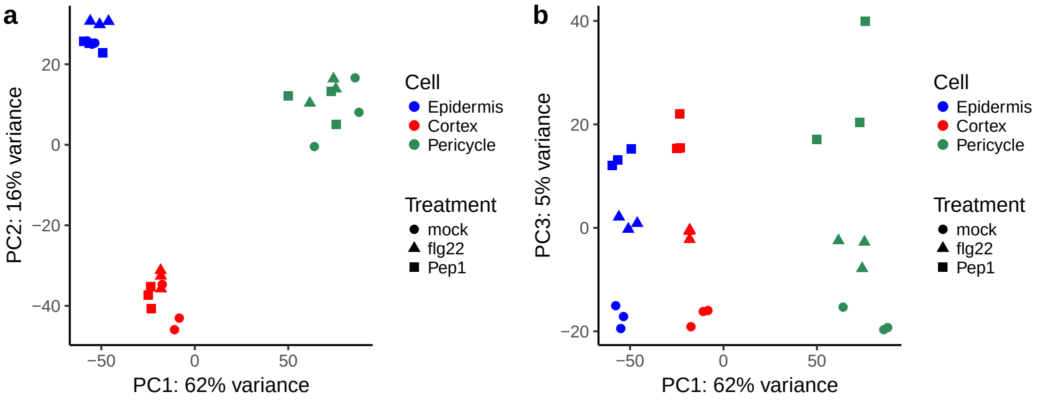
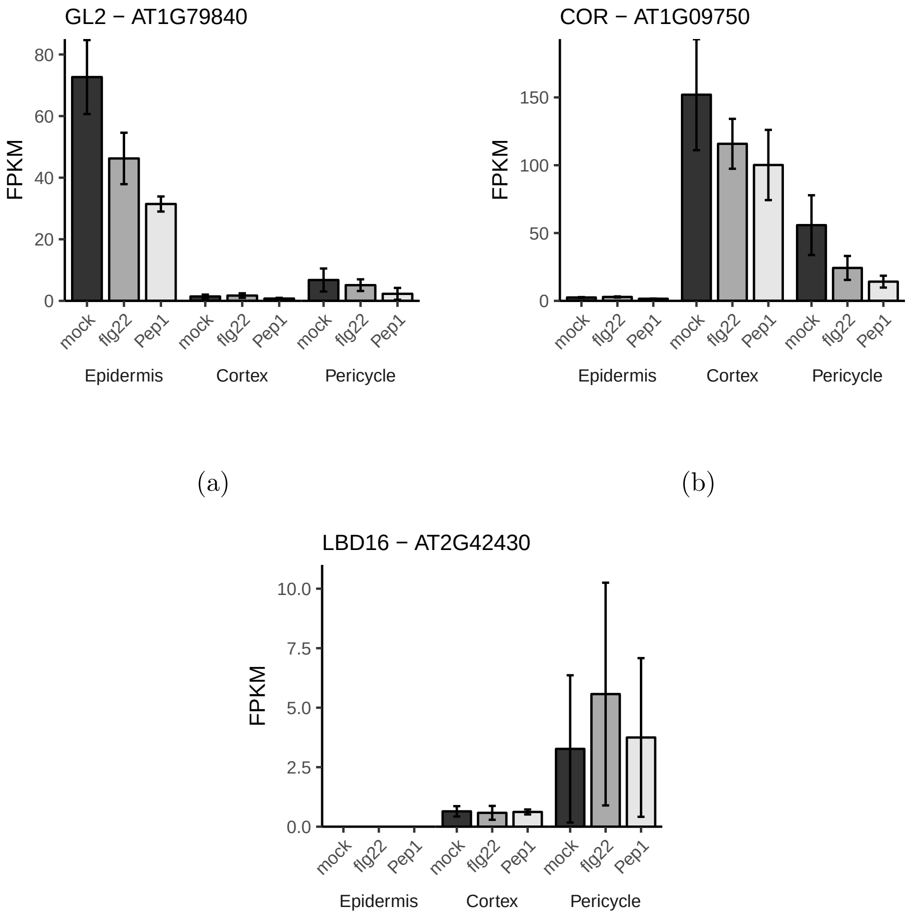


Figure 4.9: PCA of RNA-seq samples.
(a) Plotting PC1 vs. PC2 reveals that samples cluster strongly according to cell type; epidermis (blue), cortex (red) and pericycle (green).
(b) Plotting PC1 vs. PC3 reveals separation between treatments; Pep1 (square points), flg22 (triangular points) and mock (circles). The % variance that each PC represents is indicated on the plot axes.

up-regulated more genes than flg22. The differences in number of DEGs and the difference in expression pattern support the hypothesis that Pep1 activates different transcriptional networks to flg22, despite evidence that they act through overlapping pathways.

There are substantially fewer DEGs between treatments in the pericycle compared to the other two cell types. The pericycle samples have much higher variance between replicates, which in turn reduces the clarity of differences between genes, reducing the number of significant DEGs. This high variance could be the result of technical variability, due to the challenges of sorting smaller cells, or a lower RNA content in smaller cells (Poulíèková et al., 2014). Alternatively the larger variation between replicates may be a reflection of greater heterogeneity within pericycle cell populations compared to cortex and epidermis. This heterogeneity is implied in Brady et al. (2007) which separated the pericycle into three distinct populations.



(a)

(b)

(c)

Figure 4.10: Log2 FPKM expression of cell type markers.
(a) Log2 FPKM expression of (a) epidermis marker *GLABRA2 (GL2)*, (b) a cortex marker *CORTEX (COR)*, and (c) a pericycle marker *LOB DOMAIN-CONTAINING PROTEIN 16 (LBD16)* in all three cell types after mock, flg22 or Pep1 treatment.

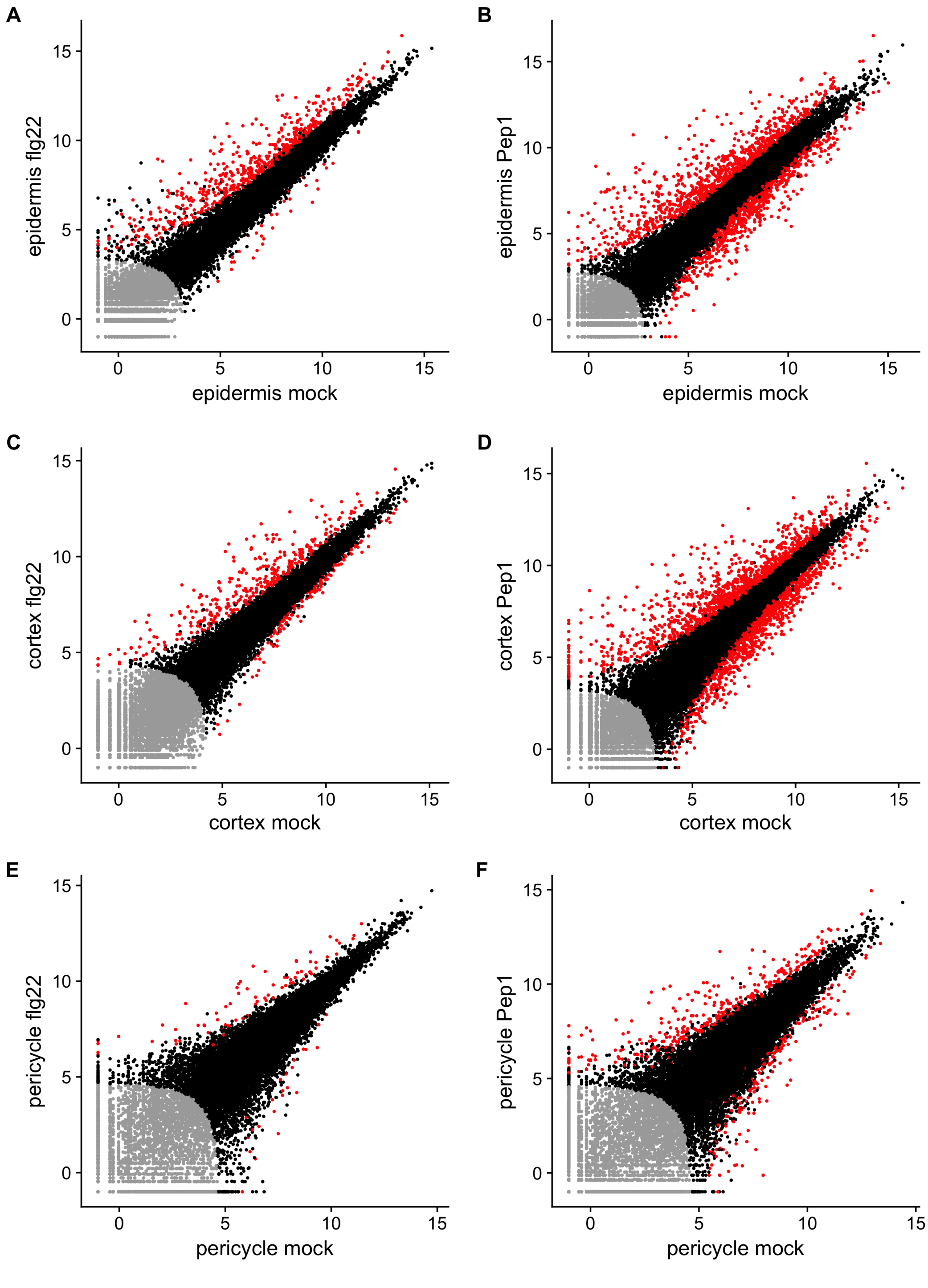


Figure 4.11: DEG replicate plots to assess the model fit of DGE analysis. The mean read counts for mock replicates against flg22 (a, c, e) or Pep1 (b, d, f) replicates are plotted in black for each cell type. The read counts for DEGs are overlaid in red and read that have been filtered out are shown in grey.

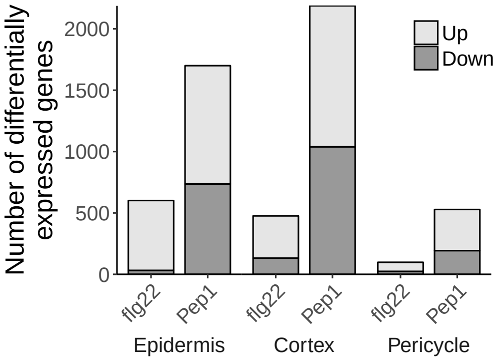


Figure 4.12: Numbers of differentially expressed genes in response to flg22 and Pep1 in different cell types.

4.2.5 Effect of protoplasting on differential gene expression

Birnbaum et al. (2003) used microarrays to quantify the extent that rapid protoplasting treatment followed by cell sorting using FACS affects global gene expression. They tested this by comparing the expression profiles of roots that had been treated with untreated roots and concluded that the technique did not induce major changes in global gene expression, as the treated samples correlated strongly with the untreated samples. However, protoplasting was shown to activate transcription of a few hundred transcripts including some stress response genes. It is possible that some protoplasting-induced genes were not identified by Birnbaum et al. as microarray studies are limited to the gene features on the microarray chip, a limitation that does not apply to RNA-seq. In this experiment, mock and immune-elicitor treated samples were processed identically to negate any effect that protoplasting might have on the comparative analyses. To confirm this, the effect of protoplasting was quantified by comparing compiled lists of all DEGs across the three cell types to a published list of protoplasting-induced genes (Birnbaum et al., 2003). For example, following flg22 treatment, 884 genes were differentially expressed across the three cell types, 43 (or 4.9%) of these genes have previously been shown to be induced by protoplasting. After Pep1 treatment, 3.6% (110/3082) DEGs were in both lists. The very small overlap between lists of genes induced by protoplasting and DEGs in response to immunity treatments confirmed that sample processing has not affected differential gene expression analysis.

There is the potential for studies that use protoplasts to be biased by developmental zone, as younger cells are more readily protoplasted as during early development the cell wall are not lignified or suberised. This bias is hard to quantify within the context of bulk RNA-seq, as the expression of developmental associated genes is averaged out across all cells. However, testing for the presence genes associated to specific developmental stages, e.g. Casparian strip development, could be used to confirm the presence of some cells at each developmental stage. Theoretically, the simplest way to avoid this bias is to harvest cells by developmental zone in addition to cell type, however this would result in a more complex, expensive and time intensive experimental design.

4.2.6 Immune responses in *Arabidopsis thaliana* roots show clear cell type differences

Considering the two treatments separately, 74% (644 genes) of all flg22-responsive were expressed in a single cell type and 25% were expressed in more than one cell type (Figure 4.13a). In terms of Pep1-responsive genes, 66% (2018 genes) were expressed in a single cell type (Figure 4.13a) and 35% in more than one cell type. For this test, any overlaps between flg22 and Pep1 were not considered, and as such some genes appear in both flg22 and Pep1-responsive gene lists. It is striking that in both the epidermis and cortex, more than half of the genes responding to a treatment were expressed in cell type-specific manner indicating that flg22 and Pep1 are activating different gene networks in each of the three cell types. These cell type-specific networks are more prominent in the cortex and epidermis than in the pericycle.

A gene ontology (GO) term enrichment analysis was used to explore whether the observed cell type-specificity in the epidermis and the cortex of flg22 and Pep1-responsive gene networks reflects specific functions. In order to make enrichment p-values directly comparable between analyses, gene set sizes were equalised by the top most significant (false discovery rate (FDR) adjusted p-value) 128, 365 and 337 for flg22 up-, Pep1 up- and Pep1 down-regulated DEGs, respectively, based on Figure 4.13. Equalising the gene sets is beneficial to the analysis as the only changed parameter between statistical significance test is the change in overlap size between the test set and the GO term gene set. This means that the p-values can be used to quantify differences between tests. If different size gene sets are used then two parameters in the significance test change (overlap and gene set size) and therefore the enrichment values should not be used to quantify the differences in enrichment. In this case p-values should be used as a binary metric, i.e. is it significant or not. Since fewer DEGs were identified to be down-regulated by flg22 and overall in the pericycle, individual GO term analyses were performed for sets between 50 and 100 genes and gene function was investigated directly for sets smaller than 50 genes .

Firstly, the up-regulated genes following both flg22 and Pep1 treatments in both the epidermis and cortex were strongly enriched in many immunity-associated terms such as “immune system process” and “regulation of defence response” (Figure 4.14a-d). These terms are enriched in both the epidermis and cortex despite the lack of overlap in the genes tested indicating that both cell types are immune-responsive via either temporally separated induction or represent immunity sub-networks.

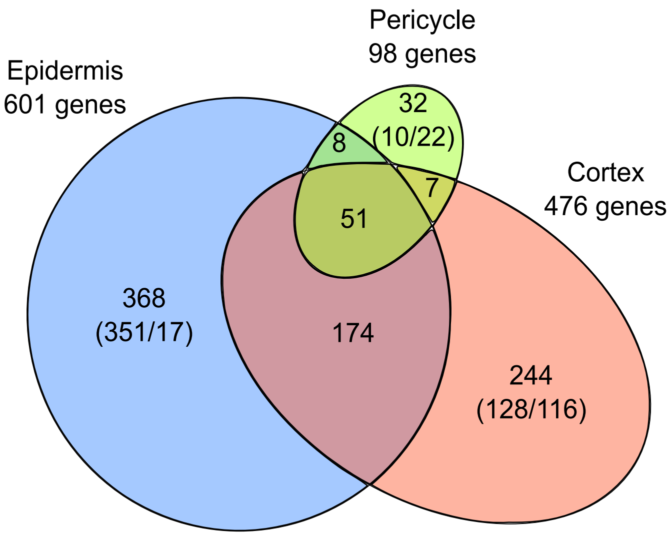
In total, 24% (31/128) of epidermis-specific flg22-responsive genes were associated with immunity and defence terms annotated to genes such as *NDR1/HIN1-LIKE PROTEIN 10 (NHL10)*, *MITOGEN-ACTIVATED PROTEIN KINASE 5 (MPK5)* and *CHITINASE IV (AtCHITIV)*, and in the cortex, this proportion was 22% (28/128) of the cortex-specific DEGs including *CHITIN ELICITOR RECEPTOR KINASE 1 (CERK1)*, *WALL ASSOCIATED KINASE-LIKE 2 (WALK2)*, *WRKY17*, *WRKY27* and *WRKY51.* Pep1 induced similar proportions of cell type-specific genes associated with immunity and defence; 18% of DEGs in the epidermis (65/365) including *WRKY72* and *PROPEP3*, and 19% in the cortex (71/365) including *BAK1-INTERACTING RECEPTOR-LIKE KINASE 1 (BIR1)*, *WRKY22* and *ARABIDOPSIS NAC DOMAIN-CONTAINING PROTEIN 019 (ANAC019).*

In addition to the strong cell type-specific immune responses observed, GO enrichment analysis revealed functional specificity in both the epidermis and the cortex. "Oligopeptide transport" and "organic acid transport" were specifically enriched in the cortex flg22- and Pep1-specific DEGs ($p < 10^{-5}$ and $p < 10^{-9}$), associated to genes such as the sugar transporter *PROBABLE POLYOL TRANSPORTER 6 (PLT6)*, *NITRATE TRANSPORTER 1.8 (NRT1.8)* (flg22 response, Figure 4.14a-b) and the sulphate transporter *SEEDLING LETHAL 1 (SEL1)* (Pep1 response, Figure 4.14c-d). The role of peptide synthesis and transport has been highlighted in leaf studies that demonstrated that antimicrobial synthesis and delivery was highly important to ensure effective PTI (Kwon et al., 2008; Bednarek et al., 2009; Li et al., 2009; Nekrasov et al., 2009; Saijo et al., 2009).

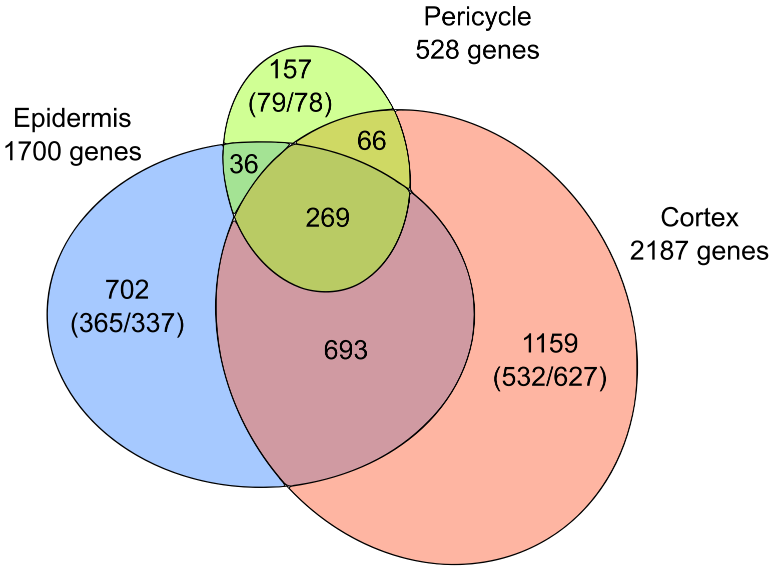
In the epidermis, up-regulated genes did not reveal a unique function as these genes were dominated by immunity-associated genes. However, the Pep1-repressed genes in the epidermis were more strongly enriched in development-associated GO terms such as "root morphogenesis","root system development" and "post-embryonic root development" (Figure 4.14e-f) than in other tested gene sets ($p < 10^{-5}$ in the epidermis Pep1 down-regulated versus $p < 10^{-2}$ in the cortex,"root morphogenesis" term). This enrichment was associated with expansins such as *EXPANSIN 14 (EXP14)*, auxin-associated gene *AUXIN-INDUCED IN ROOT CULTURES 1 (AIR1)* and development-associated transcription factor *KANADI 4 (KAN4).* The enrichment of growth associated terms in epidermis down-regulated DEGs, suggests that this cell type is either the main driver of root growth inhibition, or perhaps more likely the first cell type to slow growth. As the outermost cell type, a reduction in growth is likely to then constrain growth of the inner cell types. In contrast, the cortex down-regulated genes are uniquely enriched in terms associated with the repression of secondary metabolism and brassinosteroid biosynthesis.

The DEG response to flg22 in the pericycle was limited to 32 (10 up- and 22 down-regulated) largely uncharacterised genes with putative functions in protein modification and ion exchange. The Pep1 response in the pericycle was stronger, inducing 79 genes and repressing 78. The top most significant GO terms associated with the up-regulated response were RNA methylation, macromolecule modification and protein targeting (Figure 4.15a). "Anion transport" and "nitrate transport" were weakly enriched in the Pep1-repressed genes (Figure 4.15b). Whilst these terms represent small groups of genes, they indicate that the pericycle might be playing a supportive role in PTI, similar to the cortex. Two expansins were also identified in the pericycle Pep1 down-regulated DEGs: *EXPANSIN 15 (EXP15)* and *EXPANSIN B3 (EXPB3)*, indicating that the pericycle perhaps also undergoing structural modifications. These combined results show that both flg22 and Pep1 are activating gene networks with distinct functions in all three cell types.

In addition to cell type-specific patterns of GO terms, we also identified treatment-specific differences, particularly in the Pep1 response in both the epidermis and the cortex. For example, Pep1 additionally appears to impact brassinosteroid signalling with "brassinosteroid biosynthetic process" enriched in both cortex and epidermis Pep1-repressed gene sets.



(a)



(b)

Figure 4.13: Cell type-specific gene expression. (a) Venn diagram showing cell type-specific networks responding to flg22, (b) Venn diagram showing cell type-specific networks responding to Pep1. Numbers in brackets indicate up- and down-regulated genes.

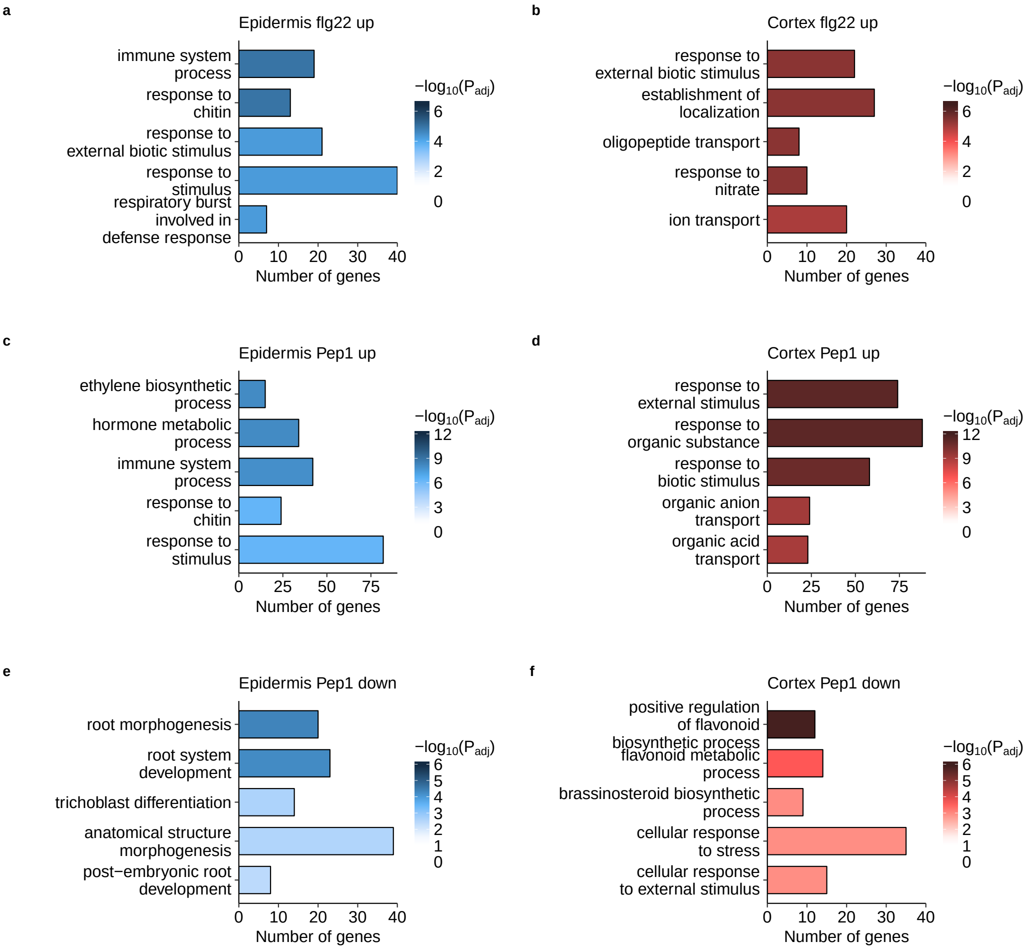


Figure 4.14: Significantly enriched GO terms in the epidermis and cortex.
(a) Top five most significantly enriched GO terms in genes up-regulated after flg22 treatment in the epidermis but not DE after flg22 treatment in the cortex or pericycle,
(b) Top five most significantly enriched GO terms in genes up-regulated after flg22 treatment in the cortex but not DE after flg22 treatment in the epidermis or pericycle,
(c,d,e,f) Top five most significantly enriched GO terms in exclusively up- or down-regulated genes after Pep1 treatment in the epidermis and cortex.

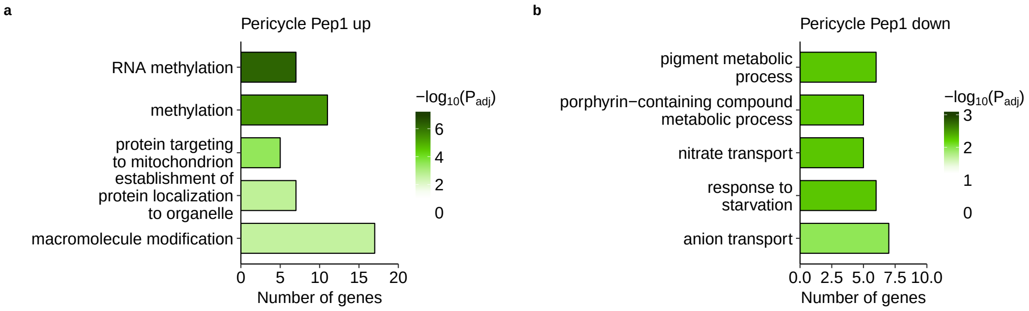


Figure 4.15: Significantly enriched GO terms in the pericycle. Top five GO terms based on all (a) up- and (b) down-regulated by Pep1 in the pericycle but not the epidermis or cortex.

4.2.7 Cell- and treatment-specific genes

Prior studies have shown flg22 and Pep1 to signal through overlapping pathways, activating the same MAPK pathways. Furthermore, there are strong overlaps in the observed phenotypic responses to flg22 and Pep1 including ROS bursts and root growth inhibition. The GO term enrichment analysis showed that the flg22 and Pep1 responses were both strongly enriched in immunity and defence terms. Next, the immune responsive genes were split into sets based on the cell and treatment responsive expression in order to determine the extent of gene network overlap, and whether the cell type-specific gene sets observed above were also treatment specific (Figure 4.16).

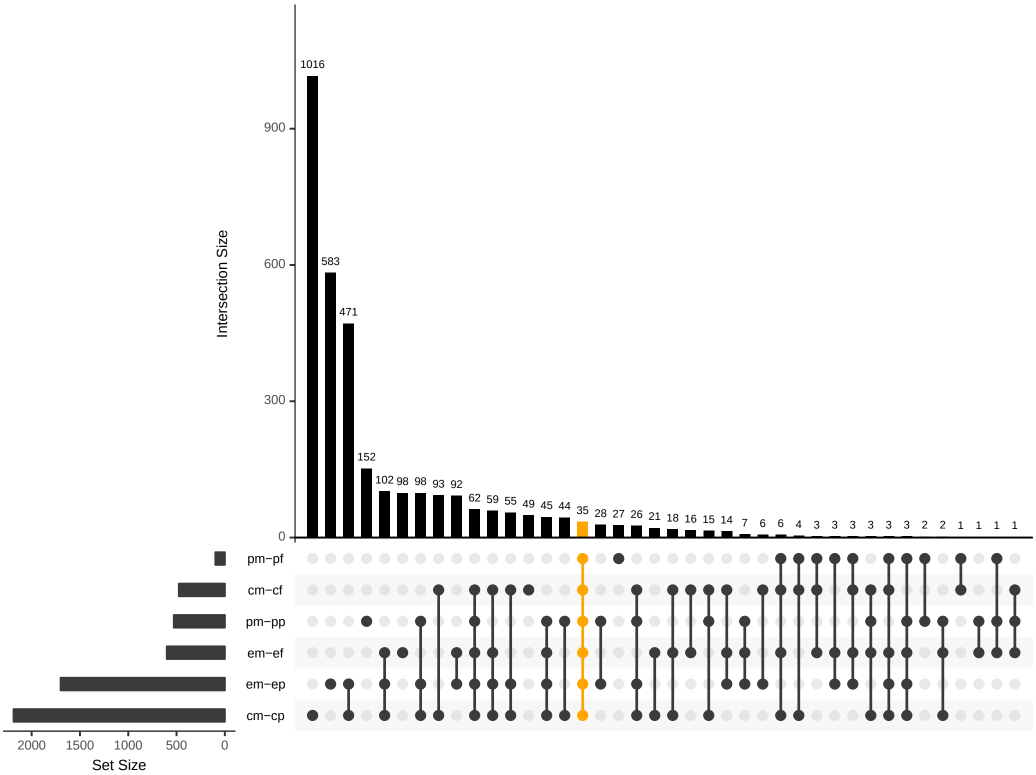


Figure 4.16: Intersections between sets of immune-responsive DEGs in the three cell types, visualised using the 'UpSet' technique from the UpSetR package (Conway et al., 2017). Overlaps between samples (indicated as cm-cf equals cortex mock-flg22, em-ep equals epidermis mock-Pep1 etc.) are shown as a matrix of dots below a vertical bar chart. Vertical bars indicate the size of the gene set specific to the overlaps indicated below. The horizontal bar chart shows the relative sizes of the input samples, with pm-pf (pericycle mock-flg22 responsive) containing only 98 genes vs. cm-cp (cortex mock-Pep1) containing 2187 genes. The intersection between all 6 gene sets is indicated in orange.

35 genes formed the “core” PTI response; expressed across all cell types in response to both immune elicitors (highlighted in orange, Figure 4.16). Within this core set we identified genes encoding three *GLUTATHIONE S-TRANSFERASEs (GSTF) 6,7* and *12*, three O-methyltransferase family proteins; *INDOLE GLUCOSINOLATE METHYLTRANSFERASEs (IGMT) 2, 3* and *4*, three peroxidases including *PEROXIDASE (PER) 4* and *5* and *PRX71* and various chitinases. These gene families are particularly associated with oxidative stress in plants, having been shown to respond to both biotic and abiotic stress (Kawano, 2003).

4.2.8 The flg22 response is largely encompassed by the Pep1 response

Figure 4.16 shows that many of the flg22 responsive genes also respond to Pep1, whereas the Pep1 response is more treatment-specific. The extent of this overlap is revealed in Figure 4.17 which shows that after aggregating cell type responses for each treatment approximately 78% (690 of 884 DEGs) of flg22-responsive genes were also regulated by Pep1 and only 194 genes were specifically responsive to flg22. Figure 4.17 also shows that the majority of genes that respond to one treatment, are also cell type-specifically expressed. 92% (174 DEGs) of the flg22-specific DEGs are only expressed in one cell type with 98, 49 and 27 DEGs showing epidermis, cortex and pericycle-specific expression, respectively. For Pep1, 78% (2392 of 3082 DEGs) of the DEGs showed Pep1-specific expression out of which 73% (1751 DEGs) were expressed in only one root cell type with 583, 1016 and 152 DEGs displaying epidermis, cortex and pericycle-specific expression, respectively.

The majority of DEGs that responded to flg22 in the epidermis and cortex were induced whereas in the pericycle they were largely suppressed (Table 4.4). GO term analysis was performed for each cell type and flg22-specific gene sets and revealed that the flg22-specific epidermis expressed genes was enriched in GO terms such as “regulation of proton transport” and “response to nitrate” and much less strongly enriched in immunity and defence associated terms compared to the enrichment for all flg22 responsive genes specific to the epidermis. In the cortex the flg22-specific genes are enriched in the GO term “oligopeptide transport” and “amide transport”, which matches the most strongly enriched GO terms from all the flg22-responsive genes specific to the cortex. This indicates that the enrichment of these peptide transport terms is a flg22-specific response in the cortex.

In response to Pep1, the cell type-specific DEGs showed similar levels of up- and down-regulation across all three cell types (Table 4.4). Consistent with the GO term enrichment of all Pep1-responsive genes, the Pep1-specific epidermis expressed genes were enriched in growth and hormone terms, particularly "ethylene biosynthetic process" and "hormone metabolic process". In turn, the Pep1-specific cortex expressed genes were more strongly enriched in broad defence terms such as "response to biotic stimulus" encompassing genes such as *WRKY18*, *BIR1*, and *MYELOBLASTOSIS 122 (MYB122).* These data suggest firstly that the Pep1 response shows strong cell type-specificity and that epidermis and cortex DEGs are indicative of contributing different functions to the Pep1 induced PTI. Moreover, these data indicate that, compared to Pep1, flg22 elicits a weaker response in root cells and appears to be largely encompassed by the Pep1 response.

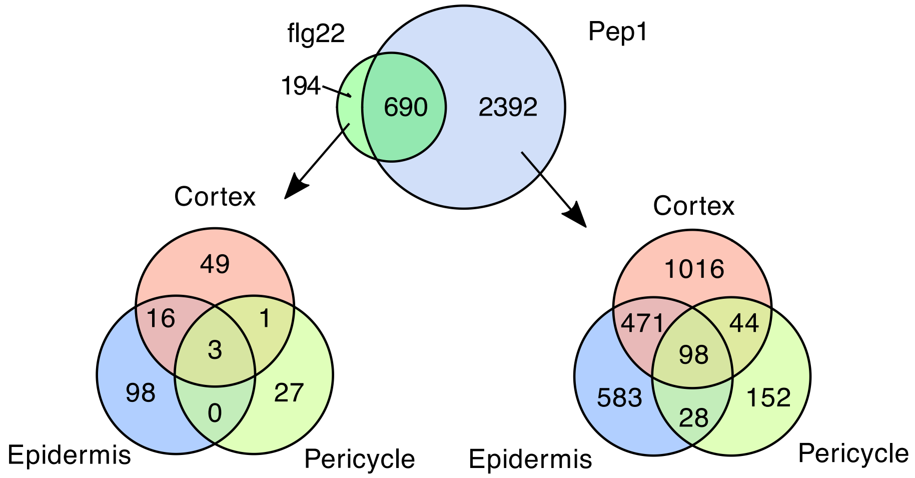


Figure 4.17: Differences between responses to flg22 and Pep1 in each cell type.
Upper Venn diagram shows the overlap of flg22- and Pep1-responsive gene sets, aggregated across cell types.
Lower Venn diagrams show the split by cell types for genes responding to flg22 in at least one cell type and not responding to Pep1 in any cell type (left) and vice versa (right).

Table 4.4: Numbers of DEG in cell type and treatment specific gene sets. Total, up- and down-regulated gene numbers are shown for flg22 and Pep1 treatments in epidermis, cortex and pericycle cells.

Treatment	Cell type	No. of DEG	Up	Down
flg22	epidermis	98	98	0
flg22	cortex	49	40	9
flg22	pericycle	27	10	17
Pep1	epidermis	583	259	324
Pep1	cortex	1016	445	571
Pep1	pericycle	152	77	75

4.2.9 Cell identity is unchanged by immune stress

Considering the substantial transcriptional changes in individual cell types upon immune elicitor treatments, we asked if immunity would affect the cell type identity. For each root cell type, identity is determined at the stem cell niche (Nakajima et al., 2001; Sabatini et al., 2003; Aida et al., 2004; Sarkar et al., 2007; Stahl et al., 2009). The maintenance of cell type identity, as defined by cell type-specific housekeeping functions, is of outstanding importance for overall root integrity and functionality (e.g. root growth), especially under stress (Iyer-Pascuzzi et al., 2011; Geng et al., 2013). Interestingly, immunity is known to halt root growth suggesting that housekeeping functions are over-ridden during root PTI (Gómez-Gómez et al., 1999; Jacobs et al., 2011). We therefore wanted to know if immunity affects cell type identity and with it the core of root tissue integrity and functionality. Cell type identity transcriptomes of unchallenged epidermis, cortex and pericycle cells were defined as described in Section 2.2.7. 950 genes were identified as specifically enriched in the epidermis, 512 in the cortex and 1055 in pericycle. These enriched datasets were confirmed to highly overlap ($p < 10^{-6}$, hypergeometric test) with published cell identity gene sets (Bargmann et al., 2013). Distinct GO terms were associated with each set of identity genes denoting the different functions of the three cell types. The epidermis is enriched in cell division-associated GO terms (e.g. "mitotic cytokinesis") (Figure 4.18a), the cortex governs processes related to protein metabolism such as "Golgi organisation" and "calcium transport" (Figure 4.18b) and the pericycle is enriched in terms such as "S-glycoside biosynthetic process" and "glucosinolate biosynthetic process" indicative of secondary metabolism in addition to being strongly enriched in "response to stimulus" (Figure 4.18c).

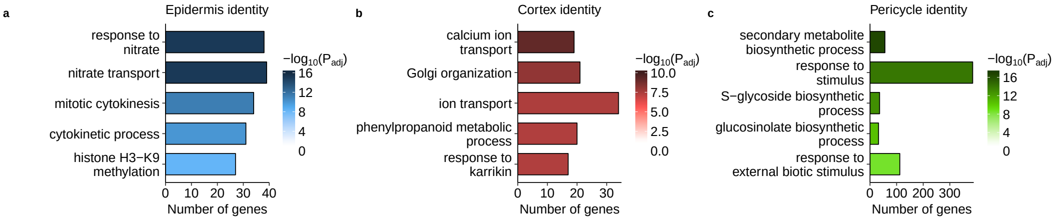
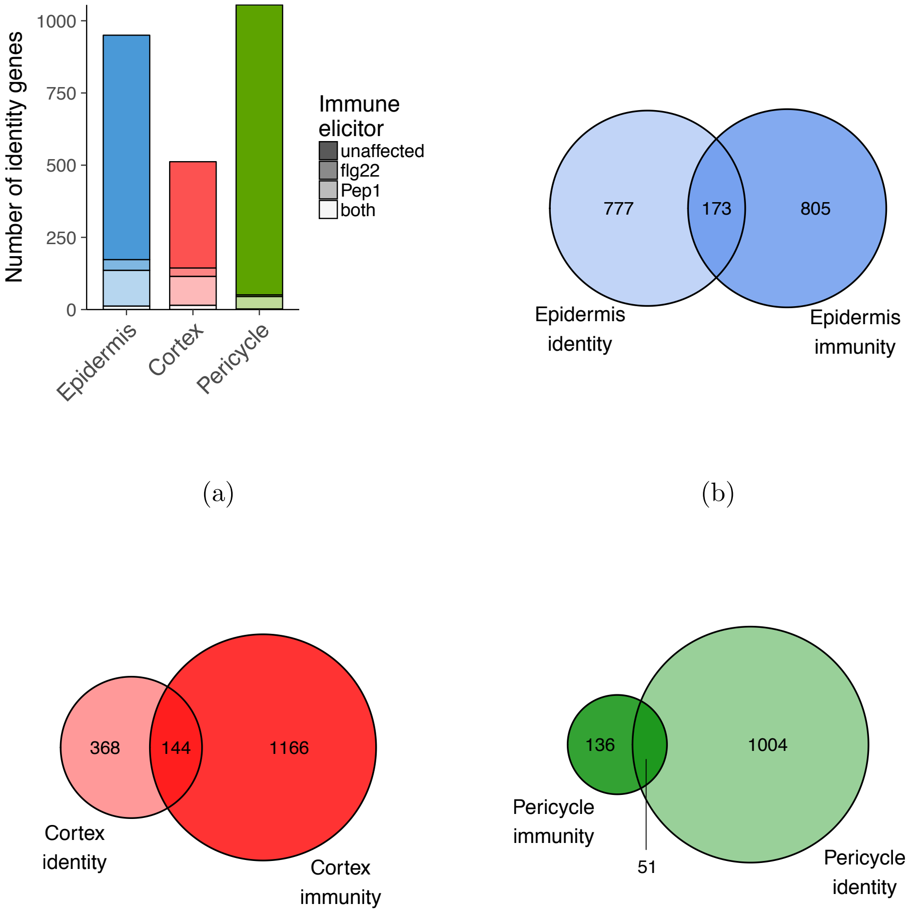


Figure 4.18: Significantly enriched GO terms in cell identity gene sets Top five GO terms enriched in the top 1699 identity genes in the (a) epidermis, (b) cortex and (c) pericycle.

In order to answer the question of whether immunity affects cell type identity, we quantified the overlap between immunity genes and cell identity genes. We found that the regulation of only 18% of epidermis, 28% of cortex and 5% of pericycle-associated identity genes, respectively, was affected by either or both immune elicitor(s) (Figure 4.19a). On the other hand, the identity genes make up a larger proportion of the cell type-specific PTI response (to flg22 or Pep1) in each cell type. 28%, 21% and 32% of epidermis, cortex and pericycle PTI DEGs, respectively, are represented in the cell type identity gene sets (Figure 4.19b-d). These findings are consistent with earlier studies, where cell identity was found to be unaffected by abiotic stress (Dinneny et al., 2008; Iyer-Pascuzzi et al., 2011; Geng et al., 2013).



(a)

(b)

(c)

(d)

Figure 4.19: Number of identity genes affected by either flg22 and/or Pep1. (a) Stacked bar plot indicating identity DEGs responsive to flg22, Pep1, bot or neither treatment. (b-d) Venn diagrams showing the overlap between identity genes and genes DE in response to flg22 or Pep1 in only one cell type (termed immunity genes) in the (b) epidermis, (c) cortex and (d) pericycle

4.2.10 PMET reveals developmental motifs are enriched in cell identity genes

In order to generate comparable p-values, the identity gene sets were equalised by taking all 512 cortex-enriched genes and the top 512 cell type enriched genes from the epidermis and the pericycle. There were a large number of highly significant enrichment of a number of motif pairs in epidermis- and cortex-specific genes, but fewer and weaker signals for pericycle-specific genes (highlighted motifs in Figures 4.20,4.21 and 4.22 and full results shown in Appendix B.1-B.3). Enriched motifs were identified in the promoters of 472 epidermis (92%), 442 cortex (86%) and 461 pericycle (90%) cell identity genes. The pattern of enriched motif combinations was highly distinctive in all three cell types with a large number of motif pairs highly significant in one cell type and not statistically significant in the others.

Within the promoters of epidermis identity gene motifs three different WRKY TFs (WRKY12/38/45) were found to pair with a wide range of motifs (Figure 4.20). In particular, WRKYs were enriched with motifs for AT-HOOK MOTIF CONTAINING NUCLEAR LOCALIZEDs (AHL) TFs (AHL12/20/25, $p < 10^{-7}$, Bonferroni-corrected p-value corresponding to enrichment score of AHL12_2 and WRKY45 motif pairs) ARABIDOPSIS NAC DOMAIN-CONTAINING PROTEINs (ANACs) ($p < 10^{-6}$ or ANAC_55/WRKY45)and ARABIDOPSIS THALIANA HOMEOBOX (ATHB) TFs ($p < 10^{-10}$, Bonferroni-corrected p-value corresponding to enrichment score of ATHB51 and WRKY45 motif pairs). ATHB and/or AHL co-localised with WRKY in 185 of 512 epidermis identity genes (36%). WRKYs represent a large family of Arabidopsis TFs (> 70 members) with known regulatory functions in plant innate immunity, abiotic stress and developmental processes (Pandey and Somssich, 2009; Rushton et al., 2010). In turn, AHL TFs have distinct roles in growth and development (Matsushita et al., 2007; Hur et al., 2015) and ATHBs take key roles in adapting growth according to environmental conditions such as shade avoidance or photocontrol (Prigge et al., 2005). ANAC motifs co-localised strongly with AHL in particular ANAC55_2 with AHL_2 and AHL_3ARY (both $p < 10^{-8}$). This is another example of development and stress TF motifs interacting as ANACs, and particularly ANAC55 have been shown to be involved in stress responses (Hickman et al., 2013).

In both the epidermis and cortex, gene promoters showed enriched pairing of MYC and PHYTOCHROME-INTERACTING FACTOR (PIF) TF binding motifs with nine AHL and two ATHB TF binding motifs (Figure 4.20 and 4.21). MYCs were also paired with ATHB12 in the cortex identity genes, although more weakly than the pairs of ATHB15 and ATHB51 with MYCs ($p < 10^{-2}$ for

ATHB12/MYC2 vs. $p < 10^{-7}$ for ATHB51/MYC2). In total, 80 and 95 genes were associated with ATHB/MYC pairs in the epidermis and cortex, respectively. These are further examples of the pairing of stress and development-related TFs as MYC2-4 are well known integrators of plant immune signalling (Fernández-Calvo et al., 2011; Schweizer et al., 2013), whereas PIFs and ATHBs take key roles in adapting growth according to environmental conditions such as shade avoidance or photocontrol (Prigge et al., 2005; Leivar and Quail, 2011).

In addition, gene promoters in the cortex were specifically enriched in pairs of MYB and ATHB motifs (Figure 4.21), accounting for 176 out of 512 genes. MYB46 and MYB111 have been reported to be involved in flavonol glycoside metabolism (Stracke et al., 2010) and secondary cell wall synthesis, respectively (Zhong et al., 2007; Ko et al., 2009).

Fewer strong paired motifs were enriched in the pericycle identity genes, although paired motifs were identified in the same proportion of pericycle identity as epidermis or cortex. The two most prominent pairs were between AHL and DAG2 motifs, and between AHL and the ZAT6 motif (Figure 4.22), observed in 21% (108/512) promoters . Neither TF family is well-studied, but DAG2 has been shown to act during seed germination (Santopolo et al., 2015) and ZATs appear to have a general role in mediating abiotic stress tolerance (e.g. cold, drought) (Yin et al., 2017). Both are expressed in the pericycle and vasculature and not in the outer root tissues (Brady et al., 2007), suggesting they may have additional roles in the root.

Overall, the paired motif analyses detected AHL motifs co-localising with a variety of motifs from different TF families. PMET detected co-occurring motif pairs with a predominant gene network regulatory role of WRKY, MYC and AHL combinations in the epidermis, AHL and ATHB, MYB and MYC combinations in the cortex and AHL with DAG2 and ZAT6 in the pericycle.

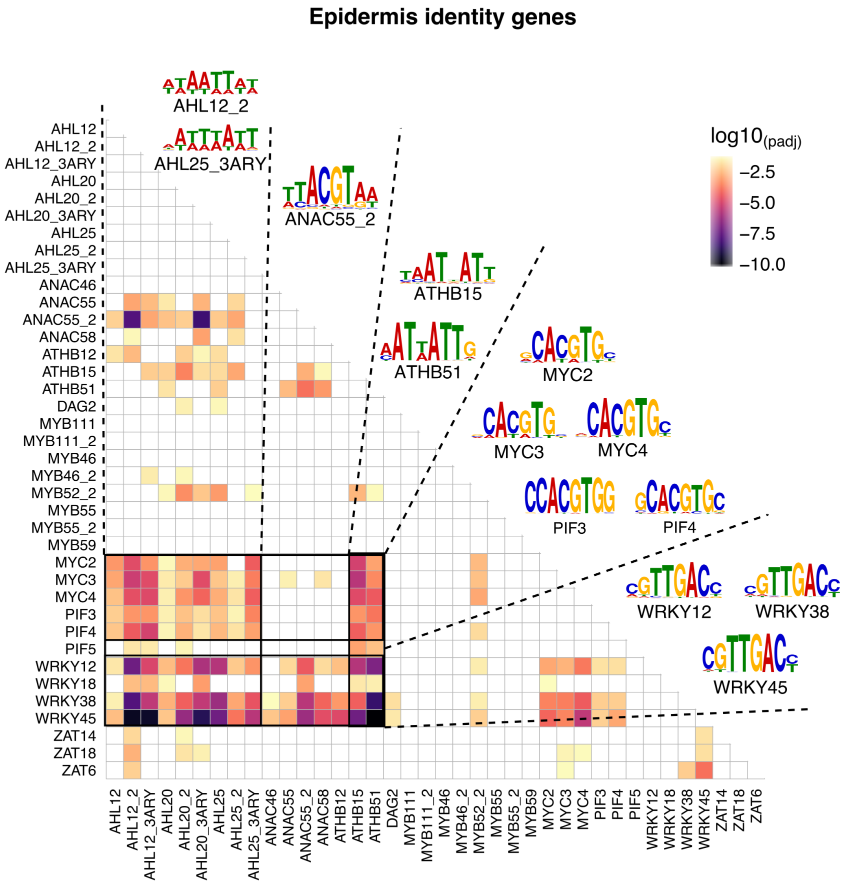


Figure 4.20: Paired motifs in epidermis identity genes. Heat map highlights the top paired motifs enriched in the 1000bp promoter upstream of the transcription start site (TSS) plus the 5' untranslated region (UTR) region in the top 512 epidermis identity genes in the epidermis alongside sequence logo representations of selected motifs. Colour indicates significance of association for values of $p \leq 0.05$ on a log10 scale ($\log_{10}(0.05) = -1.3$), $p > 0.05$ are coloured as white.

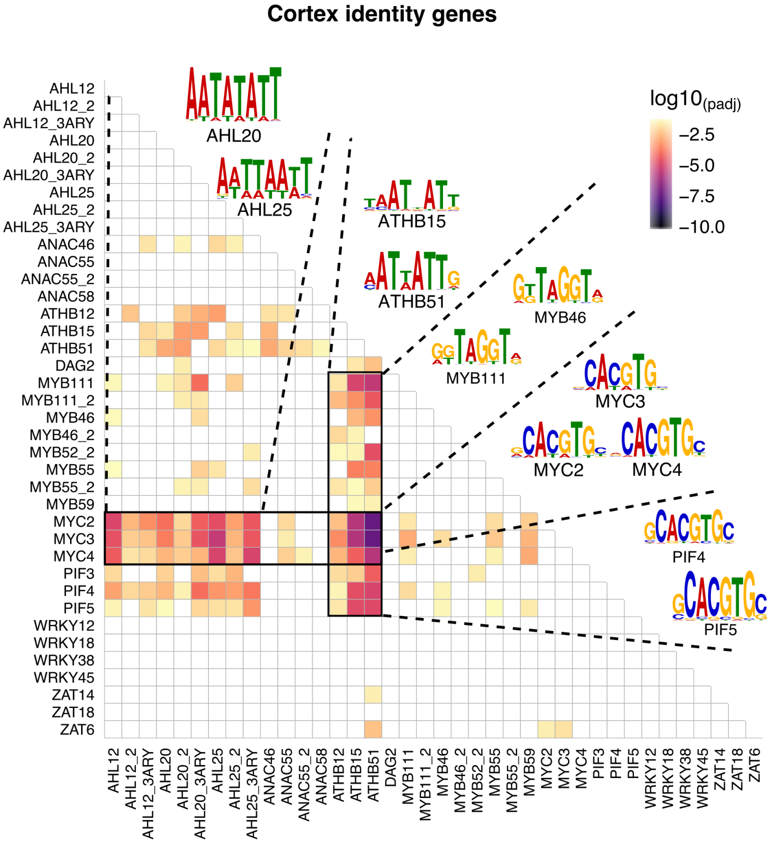


Figure 4.21: Paired motifs in cortex identity genes. Heat map highlights the top paired motifs enriched in the 1000bp promoter upstream of the TSS plus the 5' UTR region in the top 512 cortex identity genes in the epidermis alongside sequence logo representations of selected motifs. Colour indicates significance of association for values of $p \leq 0.05$ on a log10 scale ($\log_{10}(0.05) = -1.3$), $p > 0.05$ are coloured as white.

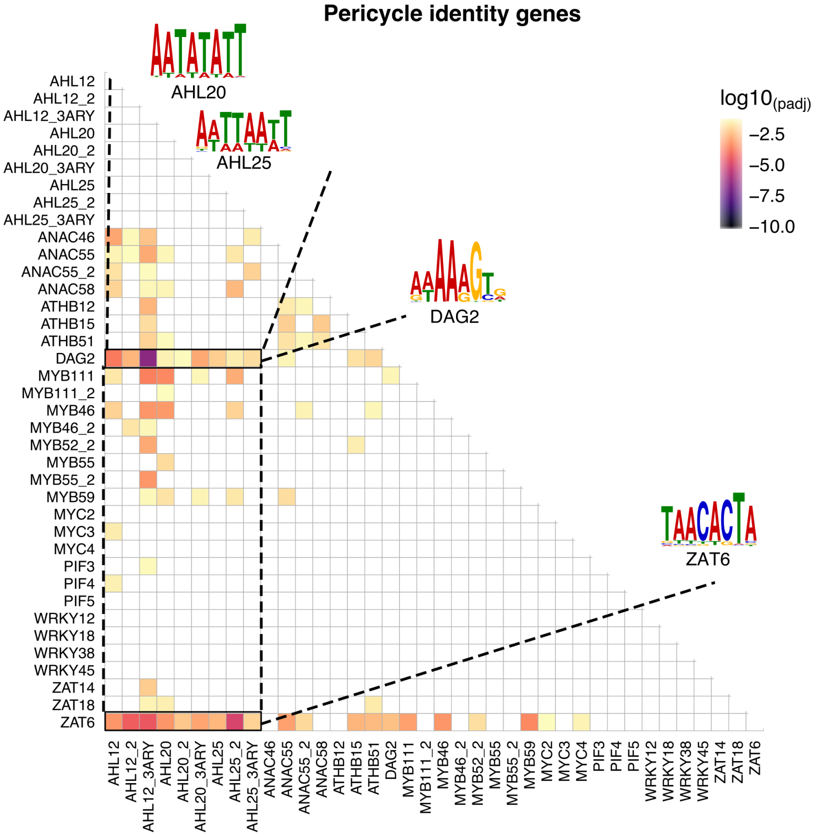


Figure 4.22: Paired motifs in pericycle identity genes. Heat map highlights the top paired motifs enriched in the 1000bp promoter upstream of the TSS plus the 5' UTR region in the top 512 pericycle identity genes in the epidermis alongside sequence logo representations of selected motifs. Colour indicates significance of association for values of $p \leq 0.05$ on a log10 scale ($\log_{10}(0.05) = -1.3$), $p > 0.05$ are coloured as white.

4.2.11 WRKYs cooperate with developmental TFs to regulate cell type-specificity of flg22 immunity

PMET analyses were conducted in order to understand if the cell type-specific rewiring of gene networks upon flg22 is reflected by a distinct enrichment in promoter motif pairs in the epidermis and cortex. In order to make the enrichment scores across flg22-treated cell types directly comparable, gene set sizes were equalised by taking all of the cortex-specifically up-regulated genes (128 genes, from Figure 4.13a) and the top 128 most significantly up-regulated, epidermis-specific genes. The analysis revealed enrichments of highly specific motif pairs in the promoters of 116 (out of 128) epidermis and 110 (out of 128) cortex genes up-regulated by flg22 (Appendix B.4-B.5). Overall, the promoters of flg22-responsive DEGs were particularly enriched in WRKY motifs. In both the epidermis and the cortex AHL motifs are co-localised with WRKYs, however the co-localisations are more strongly enriched in the epidermis ($p < 10^{-6}$ vs. $p < 0.01$, for AHL25_2/WRKY38 in the epidermis and cortex, respectively). This striking feature is also observed in epidermis identity genes, suggesting that AHL/WRKY interactions may explain overlaps between cell type-specific immunity and identity networks in the epidermis. In addition, to these overlapping pairs, there were also unique pairings with WRKY in both the epidermis and the cortex. In the epidermis WRKYs specifically paired with ANACs (ANAC55, ANAC55_2, ANAC58), DOF5.7, REM1/REM1_2 and ZAT6 (Appendix Figure B.4), whereas in the cortex WRKYs specifically paired with ETT_2, HSFB2A, HSFB2A_2, HSFC1, MYB111_2, MYB46 and MYB46_2 (Appendix Figure B.5). In the epidermis, WRKY motifs also showed particularly enriched pairing with KAN1, KAN4 and KAN4_2 motifs ($p < 10^{-5}$; Bonferroni-corrected p-value corresponding to enrichment score of KAN4 and WRKY38 motif pairs), which bind the KANADI family of TFs (Figure 4.23a). KANADIs have been shown to act as a negative regulator in embryo development (McAbee et al., 2006), root development (Hawker and Bowman, 2004) and vascular tissue formation (Ilegems et al., 2010).

Next, the expression of these TFs was tested to see if it matched the paired motif enrichment patterns. Consistent with the stronger enrichment of paired motifs, KAN2 and KAN4 were significantly DE in the epidermis and not in either of cortex and pericycle, consistent with the gene expression database ePlant (Waese et al., 2017). KAN/WRKY motif pairs were predominant and present in the promoters of 35 of the 128 flg22-induced epidermis genes and this gene set included core immune signalling receptor-like kinases from the CYSTEINE-RICH RECEPTOR-LIKE KINASE (CRK) and WALL ASSOCIATED KINASE-LIKE (WAKL) families. The combination of motif enrichment and expression patterns suggests that the

KAN/WRKY paired motifs are an epidermis-specific flg22 signalling mechanism. In the cortex, this enrichment of WRKY/KAN motif pairs was weaker and only partially seen ($p < 0.005$, Bonferroni-corrected p-value corresponding to enrichment score of KAN4 and WRKY38 motif pairs). In the epidermis flg22-up regulated genes, WRKYs also uniquely paired with YAB1 motif, an AT-rich motif similar to AHL. The YAB TF has been linked to development in the shoot apical meristem, but has no defined function in the root (Bowman, 2000). The cortex flg22-induced enriched paired motifs largely overlapped with the epidermis, however the cortex was uniquely enriched in ATHB15/51 pairs with WRKY12/18/38.

Comparing the results for the flg22 and cell identity gene sets revealed strong overlaps in motif pair enrichment, including AHL/WRKY pairs in the epidermis identity and both epidermis and cortex flg22-induced gene sets.

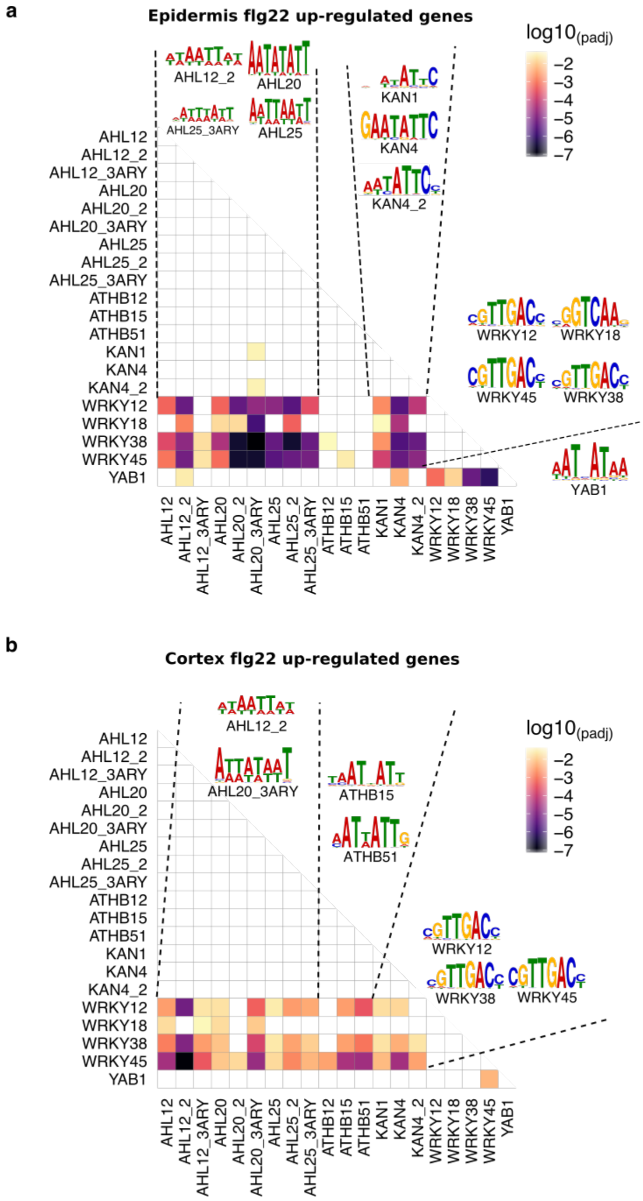


Figure 4.23: Paired motifs in flg22-induced genes in the epidermis and cortex. Heat map highlights the top paired motifs enriched in the 1000bp promoter upstream of the TSS plus the 5' UTR region n the top 128 flg22 responsive genes specifically expressed in (a) epidermis and (b) cortex cells alongside sequence logo representations of selected motifs. Colour indicates significance of association for values of $p \leq 0.05$ on a log10 scale ($\log_{10}(0.05) = -1.3$), $p > 0.05$ are coloured as white.

4.2.12 Context dependent linkage of immunity and cell identity networks

Having observed clear overlaps between the identity genes and the flg22 response, paired motifs in Pep1 up- and down-regulated genes were investigated to see was consistent across multiple immune response pathways. Furthermore, identification of paired motifs that are unique to the Pep1 response may begin to explain how the non-overlapping flg22 and Pep1 networks are regulated.

The up-regulated and down-regulated Pep1 response were investigated separately, and again the gene set sizes were equalised. All epidermis-specifically Pep1-induced and Pep1-suppressed genes were tested (365 and 337 genes, respectively; Figure 4.13b) as well as the 365 and 337 most significantly Pep1-induced and -suppressed, cortex-specific genes, respectively. The analyses identified paired motif enrichment in the promoters of 212 and 219 out of 365 epidermis- or cortex-induced genes, respectively, and 191 and 168 out of 337 epidermis- or cortex-suppressed genes, respectively (Figures 4.24). Consistent with the flg22 response, Pep1 induced genes are enriched in WRKY TF motifs paired with a wide variety of motifs including AHLs, ANACs, ATHBs, KANs, MYBs, WOX13 and YABs (Appendix Figure B.6). In particular, Pep1-induced genes in the epidermis KAN4 and YAB1 showed enriched pairing with WRKY12 and 18 and 45. KAN/WRKY pairs were found in 62 out of 365 tested promoters ($p < 10^{-8}$ for KAN4/WRKY45 pair; Figure 4.24a) and 44 genes were associated to YAB1/WRKY enrichment. In the cortex, pairs associated with Pep1 up-regulated genes deviated more strongly from the flg22 response. The most dominant observed pairs of MYC (MYCs 2 and 3) with WRKY TFs (WRKY38 and 45), a pairing that was found in 47 out of 365 promoters of cortex-specifically Pep1-induced genes (Figure 4.24b).

Pep1-suppressed genes were not enriched in WRKYs. Instead, AHLs (AHL20-_2/ AHL25/ AHL25_3ARY) to pair with MYC2/3/4 in Pep1-suppressed genes in the epidermis (found in 70 out of 337 promoters; Figure 4.24c). This suggests that MYCs act as repressor or activators dependent on context. In the cortex, enriched pairing of MYB TF motifs (MYB111/MYB111_2/MYB46) was detected particularly with ATHBs (ATHB12/15/51), found in 47 promoters of 337 cortex-specifically Pep1-suppressed DEGs (Figure 4.24d). Interestingly, MYC TF motifs were enriched in promoters of Pep1-induced genes in the cortex (paired with WRKYs) and Pep1-suppressed genes in the epidermis (paired with AHLs). These findings indicate the efficiency of paired motif enrichment analysis and its potential in reinterpreting or confirming previous gene expression studies in roots. MYC TFs have been implicated in Pep1-mediated signalling in particular as Peps specifically induce the

MYC2-dependent branch of jasmonic acid (JA)-responsive signalling (Bartels and Boller, 2015). Additionally, MYC2 has been shown to act as both an activator and a repressor in JA-mediated gene expression (Dombrecht et al., 2007).

Comparing cell type-specific identity and immunity networks (using flg22 or Pep1-induced genes) based on our paired motif enrichment analyses, we observed significant patterns (Figure 4.25). In the epidermis, WRKY12/18/38/45 and AHL12/20/25 connect identity with immunity networks by pairing with KAN4 and ANAC46/55/58 (dominating epidermis immunity networks) and with ATHB51 (dominating epidermis identity networks), respectively. In turn, in the cortex, MYC2/3/4, PIF3/4/5, AHL12/20/25 and ATHB12/15/51 tie both networks by pairing with WRKY12/38/45 and ZAT6/14 (dominating cortex immunity networks) and with MYB46/52/55/111 and YAB5 (dominating cortex identity networks).

Our paired motif enrichment analyses suggest that cell identity integrates with and determines root immunity by specific TF pairs in individual cell types. In summary, the patterns of motif pairing that are statistically linked with our cell type-resolved expression data provide distinctiveness across cell types, across treatments, and in comparisons of gene induction vs. gene suppression. Moreover, it allows the identification of potential regulatory mechanisms to connect cell identity with immunity networks.

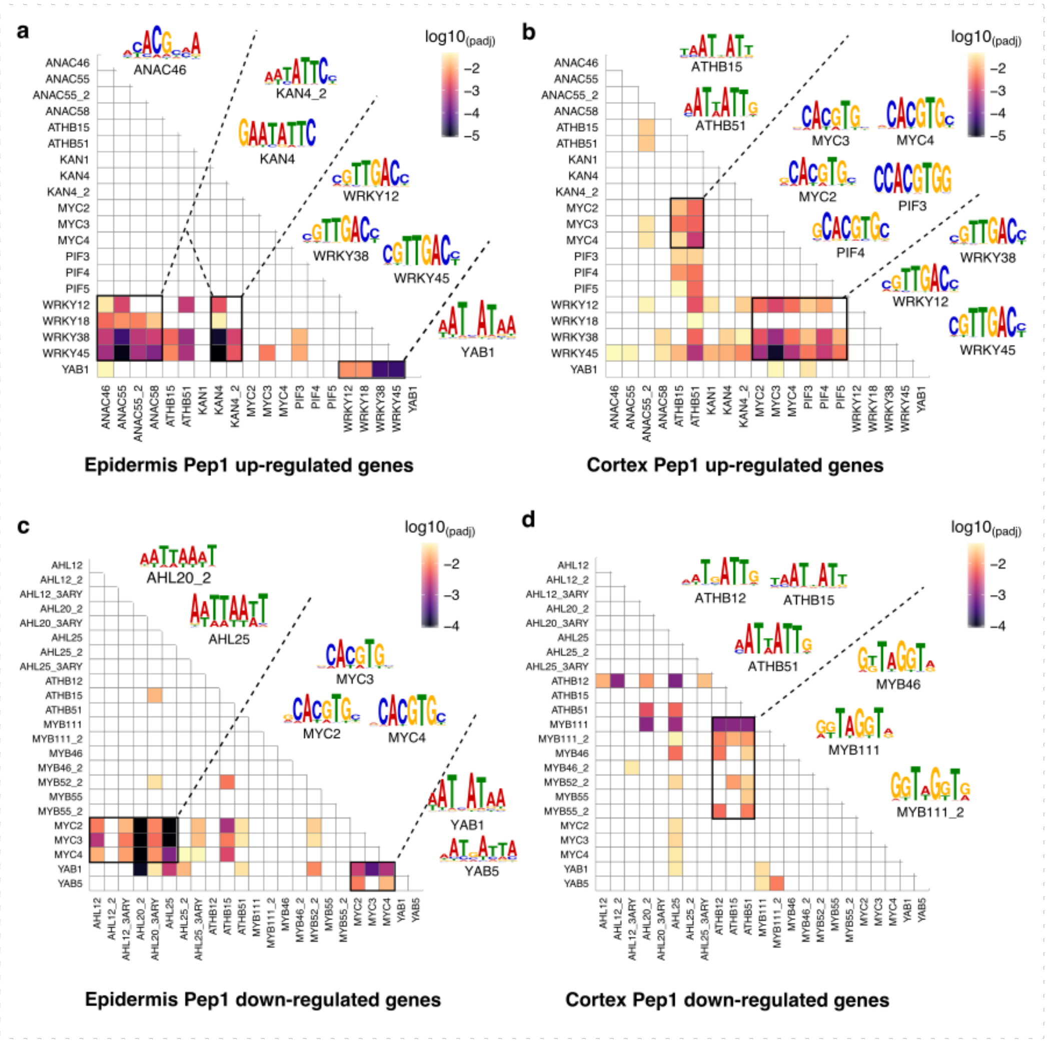


Figure 4.24: Paired motifs in Pep1 up- and down-regulated genes in the epidermis and cortex. Heat map highlights the top paired motifs enriched in the 1000bp promoter upstream of the TSS plus the 5' UTR region in (a,b) the top 365 Pep1-induced genes specifically expressed in (a) epidermis and (b) cortex cells alongside sequence logo representations of selected motifs. Heats maps (c) and (d) shows the enriched paired motifs in the top 337 Pep1 down-regulated genes in the (c) epidermis and (d) cortex. Colour indicates significance of association for values of $p \leq 0.05$ on a log10 scale ($\log_{10}(0.05) = -1.3$), $p > 0.05$ are coloured as white.

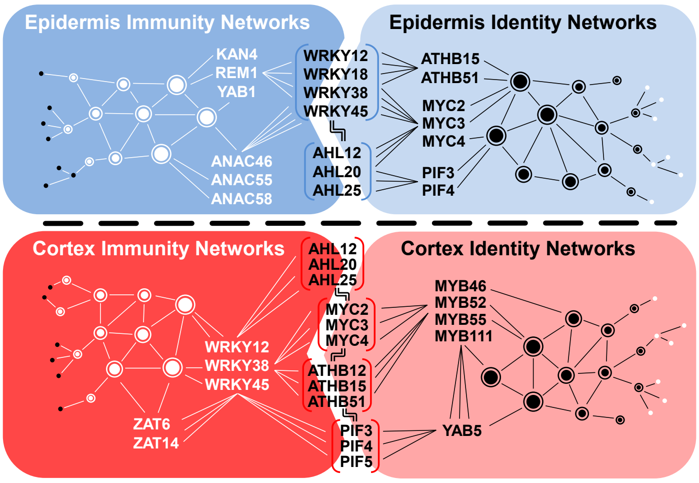


Figure 4.25: Model for the connection between cell type-specific identity and immunity networks. Connection of epidermis and cortex-specific cell identity networks with immunity networks via transcriptions factors (bracketed) that are part of respective immunity and identity networks.

4.3 Discussion

4.3.1 Pep1 and flg22 signalling networks are highly cell-type specific

Many studies have shown that different root cell types respond differently to abiotic stress (Dinneny et al., 2008; Iyer-Pascuzzi et al., 2011; Geng et al., 2013; Gifford et al., 2013). In this chapter, the immune response was shown to be similarly cell type-specific. RNA-seq analysis was used to examine the immune responsiveness of three root cell types as a first step to decipher the coordination of immunity in a complex tissue. In particular, the immune response in epidermis and cortex cells, which build the outer frontier to the rhizosphere, and pericycle cells, as outer frontier of the inner root vasculature, were analysed.

The analyses reveal that epidermis and cortex are highly immune responsive and all three cell types respond very differently to the immune elicitors flg22 and Pep1. By comparing the responses to flg22 and Pep1, specific gene networks were determined for each of the three root cell types. Both elicitors activated different gene sets in each cell type which demonstrates the remarkable complexity of immunity in roots.

Although these networks may require a higher degree of coordination (e.g. numerous regulatory and signalling proteins), maintaining cell type-specific networks may add the robustness and flexibility needed for a root system to adapt to constantly changing environmental stimuli. Consistent with this, recent studies suggested qualitative differences in the immune competences of different root cell types in response to different immune elicitors (Beck et al., 2014; Wyrsch et al., 2015; Poncini et al., 2017). Wyrsch et al. (2015) reported the perception of flg22 in isolated root systems and subsequent activation of PTI markers such as ROS production and MAP kinase activation. Furthermore, they observed flg22 responsiveness in individual cell types across different root development zones by expressing the flg22 receptor FLS2 in a cell type- and root development-specific manner. By analysing defined PTI responses (e.g. MAPK phosphorylation and immunity marker gene expression) their study suggested different contributions of root cell types to PTI.

Consistent with Poncini et al. (2017), our study showed that Pep1 is a stronger elicitor of immune signalling than flg22. Poncini et al. (2017) hypothesised that this difference occurs as Pep1 is 'interpreted as a stronger alarm signal by the root when compared to flagellin (flg22) or chitin because these pathogen-associated molecular pattern (PAMP)s are an abundant component of the rhizosphere'. Our results additionally showed that the flg22 immune response is largely encompassed by the Pep1 responsive network. If Pep1 represents a 'stronger alarm signal', the Pep1-signalling network may encompass responses to multiple attacks including from bacteria, explaining the large overlap between Pep1 and flg22-immune responsive genes. This is further supported by other Pep1 studies that have shown Pep1 signalling helps protect the root from multiple threats including increasing host resistance to bacterial or fungal pathogens (necrotrophic and biotrophic) and offering some protection against herbivores (Huffaker et al., 2011, 2013; Tintor et al., 2013).

4.3.2 Integration of stress and cell type identity in roots

Studies with *Arabidopsis* roots exposed to abiotic stress implicated cell type specificity in stress integration. Irrespective of the nature of abiotic stress (e.g. salt stress, iron starvation, nitrogen depletion) each cell type responded differently and in a highly coordinated manner to maintain root functionality under stress (Iyer-Pascuzzi et al., 2011). Furthermore, Geng et al. (2013) observed a transient root growth inhibition phenotype under salt stress that coincided with the cell type-specific rewiring of hormone signalling to reconfigure root growth-regulating networks. Such plasticity in root growth and development appears to be fundamental as it has also been observed under nitrogen depletion and was shown to drive lateral root development (Gifford et al., 2008; Walker et al., 2017). The integration of stress and growth signalling further underlines the hierarchy of root (cell) function with the maintenance of growth and development kept as top priorities under fluctuating environments. This 're-wiring' could be mediated by post-translational modification of transcription factors or epigenetic changes such as histone modification. Compromising cell identity would jeopardise root tissue function and hence, plant fitness and survival. Accordingly, the data in this chapter demonstrates that root cell types keep their identity under biotic stress and that housekeeping and stress-responsive gene networks co-exist in each root cell type. This observed overlap could be of importance to overall root plasticity (Figure 4.25). In line with previous studies on abiotic stress integration (Dinneny et al., 2008; Geng et al., 2013), the data in this chapter suggests that cell identity networks underpin cell type-specific immune responses. Linking immunity to cell identity networks could guarantee cell type-specific regulation of immune responses according to the functional competence of each cell type, leading to exquisite cell specificity in response to signal perception. Such a co-regulatory model would likely be applicable to all environmental stresses.

4.3.3 Differential immune responses across developmental zones

The work in this chapter does not address whether the immune response is variable across different developmental zones. As discussed in Section 4.2.5, there was likely bias of protoplasting towards younger tissue, as the cell walls are less tough. Equally, as large sections of roots were harvested, a large number of mature cells would have been harvested. The landscape of the mature root is complex. Across the elongation and differentiation zones cells undergo endoreplication, with the result that different cells have varied DNA content (Bhosale et al., 2018). This is associated with an increase in metabolic activity and a global increase in transcription (Bourdon et al., 2012; Pirrello et al., 2018). Since the stage of entry(in development) and total number of endocycles varies between cell types (Bhosale et al., 2018), this would likely have an impact on gene expression of an cell-type specific level. Within this dataset, the extent of endocycling at low resolution could be investigated in terms of the differential expression of endocycling-associated genes. However, in the absence of developmental stage data, it would be of limited added value to the dataset. This lack of resolution is unfortunate since the extent and rate of endocycling is likely to play a role in plant immunity. Bhosale et al. (2018) showed that they could predict the *Arabidopsis* root endoreplication response to various abiotic stresses demonstrating that endoploidy and stress adaptation are inherently linked. The application of single-cell RNA-seq (as described in Chapter 5) to *Arabidopsis* roots could be used to address these problems.

4.3.4 Combinatorial motifs integrate stress and identity networks

High resolution combinatorial promoter analyses can provide new and important insights into how immunity and cell identity networks are coordinated. PMET identified enriched paired motifs associated to the majority of genes (up to 90%) in each network, thus providing strong evidence to explain the differences in gene regulation between cell types and treatments. Importantly, such a combinatorial analysis allowed us to consider cooperative binding of different TFs, which can enhance the flexibility of gene regulation under changing environments (Van de Velde et al., 2014). Pairing of TF motifs differed strongly between cell types suggesting that different TF families act together to create the highly cell type-specific networks. The observation of differential enrichment for specific family members suggests that these networks can be further tuned by using specific combinations of TF family members.

Throughout the PMET analyses, a striking pattern occurred in the pairing of promoter motifs for known stress-regulatory TF families with motifs for developmental TFs in each cell type-specific gene network. Moreover, certain TF combinations prevailed in specific cell types in a treatment-specific manner. For instance, WRKY TFs might have a more prominent function in regulating epidermis-specific networks together with specific developmental TFs, with WRKY/MYC and WRKY/ATHB paired motifs regulating cell identity networks and KAN/WRKY and ANAC/WRKY paired motifs regulating cell immunity networks.

In turn, cortex function relies on MYC, ATHB, and PIF TFs as central regulators that combine with WRKY TFs to regulate cortex-specific immunity and with AHLs to regulate cortex identity-specific gene networks. It will be interesting to explore in future studies to what extent this combination of stress and developmental TFs contribute to stress integration and how this relates to growth regulation under immunity. Furthermore, PMET analysis identified motif co-localisations patterns that distinguished elicitor-specific networks in each cell type. These differences between PTI elicitors might reflect the life strategies of pathogens. In plants, flg22-induced PTI evolved to defend specifically against bacteria whereas Pep1 is an endogenous PTI elicitor that is activated by different hormones and, hence, might trigger the full array of all immune responses against a larger variety of pathogens.

The limited difference between flg22 and Pep1 DEGs was also reflected in the motif enrichment. In the epidermis, both flg22 and Pep1 responsive DEGs were enriched in KAN/WRKY and YAB/WRKY paired motifs. In the cortex WRKY and ATHB motifs were paired in both responses. This further emphasises the idea that Pep1 induces signalling through the same pathways as flg22, and therefore induces the same genes.The additional enrichment of ANAC, MYC and PIF motifs in Pep1-responsive gene sets reveals the potential mechanism by which the additional observed DEGs are regulated.

Tight coordination is required to maintain the complexity of cell type-specific immunity networks. The PMET results strongly suggest that TF combinations mediate this complexity through cell type-specific gene regulation. This study provides the first insight into cell type-specific immunity networks, and combinatorial TF motifs associated with specific responses and cell types. In some cases PMET reveals that different combinations of motif pairs implicates individual TFs in strikingly different networks. For example, MYC motifs in Pep1 responses appear to have context-dependent associations. For promoters of genes specifically up-regulated in cortex cells MYC motifs, and in particular the MYC4 motif, were found to pair with WRKYs. In contrast MYCs were found to be paired with AHL20/25 in promoters of down-regulated genes in the epidermis and MYC2 and AHLs were the most strongly enriched. Neither up-regulation in epidermis nor down-regulation in cortex were found to be linked to either of these motif combinations. Therefore PMET analyses have the potential to detect the regulation of disparate networks by different TF family members acting with a variety of partners in different cell types to effect contrasting transcriptional outputs. These findings highlight how context dependency of regulatory function may reduce perceived redundancy among regulatory factors that recognise highly similar sites if investigated in isolation.

Chapter 5

Dropseq analysis of *Arabidopsis thaliana* root tips

5.1 Introduction

The observation, in Chapter 4, that cell identity is prioritised over a cell's response to immunity leads to the following questions: how is cell identity established in the root tip, and does immune activation affect the ability to establish cell identity.

Fluorescent cell type marker genes have been shown to be expressed in cells very early in development, and the same set of transcription factors governs identity and proliferation of the stem cells as well as the fates of daughter cells (Moreno-Risueno et al., 2015) implying that identity is established early in development. However, the meristem has also been shown to have the capacity to regenerate a functional root tip after the QC and initial cells have been excised (Efroni et al., 2016). This regenerative ability only occurs within the meristem as these cells can revert to a pluripotent state, suggesting that despite cell fate being defined after the first division, cell identity is not as strongly established in the meristem. The second major interest of studying root tip transcriptomics is to further understand the mechanism of root growth inhibition in response to immune activation. Root growth is enacted via two processes, cell division in the meristem and cell elongation in the elongation zone. Preliminary data produced from our lab implicates the meristem in root growth inhibition, as meristem length is reduced after immune activation by flg22.

High throughput single cell RNA sequencing (scRNA-seq) is highly suited for the examination of root tips. Unlike in mature roots, fluorescence-activated cell sorting (FACS) approaches are less suited to root tip studies firstly due to the limited

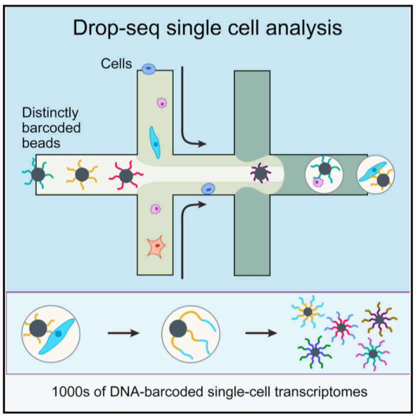

Drop-seq single cell analysis
Cells
Distinctly barcoded beads
1000s of DNA-barcoded single-cell transcriptomes

Figure 5.1: Drop-seq microfluidics schematic. Drop-seq uses a microfluidics system to combine cells and beads inside droplets. Droplets contain lysis buffer so that upon capture cells are lysed to release mRNA which binds to poly(T) tails on the primer beads. (Figure shows graphical abstract from Macosko et al. (2015))

availability of appropriate fluorescent markers and the difficulties in processing cell types that only form a very small proportion of the root tip such as the QC. Using scRNA-seq circumvents these challenges by capturing a snapshot of all cell types in a tissue without the need for any cell type markers. Examining the effect of flg22 on different cell types in the root could reveal if flg22 is differentially activating different cells based on identity or developmental stage, and identify pathways that are targeted by flg22 in the root tip.

Drop-seq is a single cell sequencing method that encapsulates individual cells in droplets and utilises a unique barcode system attached to microparticle beads in order to identify both the cell of origin and a unique molecular identifiers (UMI) for each transcript. The machine uses microfluidics to combine one microparticle bead and one cell into one droplet made up of water based lysis buffer kept separate from other droplets using oil. Once encapsulated in a droplet, cells are immediately lysed releasing RNA that binds to primers on the bead surface (Figure 5.1). Each bead contains a polymerase chain reaction (PCR) handle directly bound to the bead, and then attached to each PCR handle is a 12 bp cell barcode unique to that bead, 4^8 different 8 bp UMIs and finally an oligo-dT sequence is synthesised to the 3' end of all the oligos (Figure 5.2). Bead-primer-RNA complexes are referred to as single-cell transcriptomes attached to microparticles (STAMPs). Finally the droplets are lysed and the RNA is reverse transcribed into cDNA, then sequenced.

Once the transcriptomes have been sequenced, dimension reduction and clustering methods are used to group cells by transcriptome similarity. Cell identity is

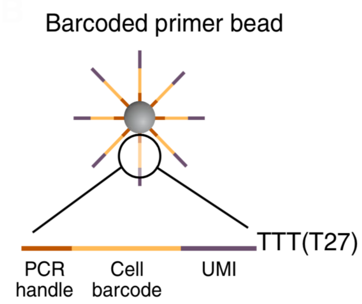


Figure 5.2: Drop-seq beads schematic. The surface of each bead is coated with oligonucleotide sequences containing a PCR handle, a cell barcode which is unique to that bead, a UMI to identify individual mRNAs, and a poly(T) tail that will bind the poly(A) tail of mRNAs in the droplet. Figure is taken Macosko et al. (2015)

assigned to clusters based on the expression of known markers and gene expression profiles. Prior studies utilising Drop-seq in mammalian tissues have been used to identify novel cell types and novel markers of known cell types but there are no published Drop-seq experiments on plant tissue.

In this chapter, I present a novel Drop-seq experiment that produced single cell transcriptomic data for root meristems untreated and treated by flg22. Root meristems (cut at the base of the first elongating cells) were harvested and protoplasted and single cells were processed through Drop-seq and sequenced (three replicates per condition). I optimised the analysis pipeline for *Arabidopsis thaliana* and discuss the challenges of assigning cell identity to developing tissue. At the end of the chapter I show the preliminary results of the differential gene expression (DGE) analysis between flg22 and mock treated root meristems, and suggest further analyses and experiments to explore the topics introduced above.

Experimental design was performed by Jessica Finch and myself with useful input from Ruth Eichmann. The plant growth, treatment, harvesting and protoplasting was performed by Jessica Finch. Processing the cells by Drop-seq and preparation for sequencing was performed by Emma Lucas. Sequencing was performed by the Genomics Facility, School of Life Sciences at the University of Warwick. All bioinformatic analysis was performed as part of my PhD studies.

5.2 Results

Following sequencing and alignment of reads, the Drop-seq analytical pipeline considers any bead that has captured any RNA to be a cell. Unfortunately, the microfluidics process can produce beads that contain incomplete transcriptomes, referred to as 'low quality barcodes'. Low quality barcodes can occur through a range of processes including contamination by non-endogenous mRNAs from cell free RNA admixed with cells in the input solution. If an empty droplet picks up some of this cell free mRNA, then the resulting sequences reveal an almost empty cell, referred to as ambient barcodes. This contamination source has been shown to be present in even the most ideal data sets (Zheng et al., 2017).

Additionally low quality barcodes can occur when a barcode binds to transcripts from multiple cells (referred to as doublets) or when captured cells are broken or killed resulting in partial transcriptomes binding to beads. In this experiment, damaged protoplasts and cells damaged during harvesting can release RNA from organelles and the cytosol which are then captured and sequenced. The first step of the analytical pipeline is to filter out all low quality barcodes from the dataset. There is no definitive method to identify low quality barcodes, so quality filtering is based on a range of criteria, including UMI content, mitochondrial RNA (mtRNA) content, and the relationship between the number of genes and the number of UMIs (equivalent to reads in bulk RNAseq) detected in the cell (Macosko et al., 2015; Ilicic et al., 2016; Butler et al., 2018). Additionally, in this plant-based experiment, we must also consider plastid RNA (ptRNA) content.

5.2.1 Determining the number of cells captured by Drop-seq protocol

Two approaches are recommended for the determining the threshold between real cells and ambient barcodes in a dataset, the binary alignment format (BAM) tag histogram and the log-log plot, described in Section 2.4.6. Unlike the ideal BAM tag histogram plots (see Materials and Methods, Figure 2.5), none of the samples processed had a clear elbow in the histograms in Figure 5.3. The clearest elbow was observed in flg22 replicate 1 around 500 cells. In all the samples the amount of information starts to level out past 1500 cell barcodes, indicating that adding more cells to the analysis is not adding new information, and would likely just add noise to the dataset.

The ideal log-log plot shows two distinctive shoulders, the first of which delimits the break between 'real' cells and damaged, broken or ambient cells. In our samples, the log-log plot for flg22 replicate one displays the clearest pattern of two shoulders (Figure 5.4), from which we can infer the most likely number of real cells is approximately 400 cells indicated by the vertical dashed line in Figure 5.5b), corresponding to cells containing at least 2000 UMIs. Shoulders can also be

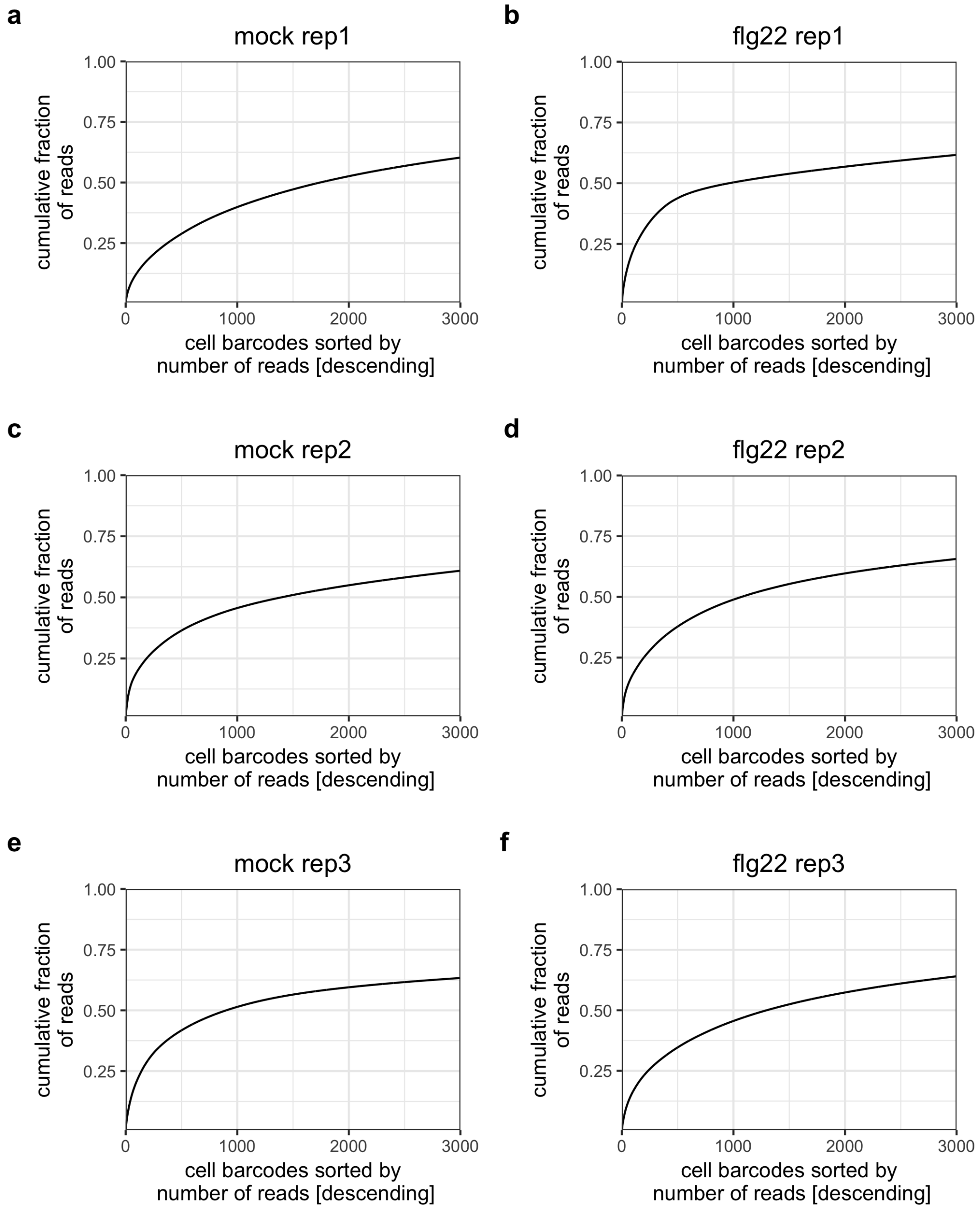


Figure 5.3: Cumulative fraction of reads per cell barcode. Histograms show the cumulative fraction of reads (UMIs) per cell barcode sorted by decreasing number of reads for six Drop-seq samples, (a,c,e) mock replicates 1-3, (b,d,f) flg22 replicates 1-3.

discerned in Figures 5.5c) and e for mock replicates two and three.

However as this break is unclear for the other samples, 2000 nUMIs was used as the minimum threshold to dictate which cells were taken forward through the analysis corresponding to 3022 cells (between 375 and 658 cells per sample (see Table 5.1)). The samples were later merged to ensure that sufficient cells were used in clustering analyses.

The process of removing the low UMI cells removes a large proportion of ambient barcodes, however this step should not be seen as removing all of the ambient barcodes as it is possible than ambient barcodes could have a higher UMI count. Since the signature of an ambient barcode is unknown, applying this threshold and exploring the remaining cells for potential contaminants is the best current solution to ensure that only the best quality cells are retained. These methods are based on fairly arbitrary cut-offs that are unclear in many datasets. The development of mathematical or machine learning models to more accurately define which barcodes correspond to real cells will improved the precision and sensitivity of these decisions.

Table 5.1: Number of single cells per sample containing more than 2000 UMIs, prior to quality control and filtering potential contaminants.

Sample	Cells containing >2000 UMIs
mock rep 1	494
mock rep 2	537
mock rep 3	658
flg22 rep 1	375
flg22 rep 2	553
flg22 rep 3	405

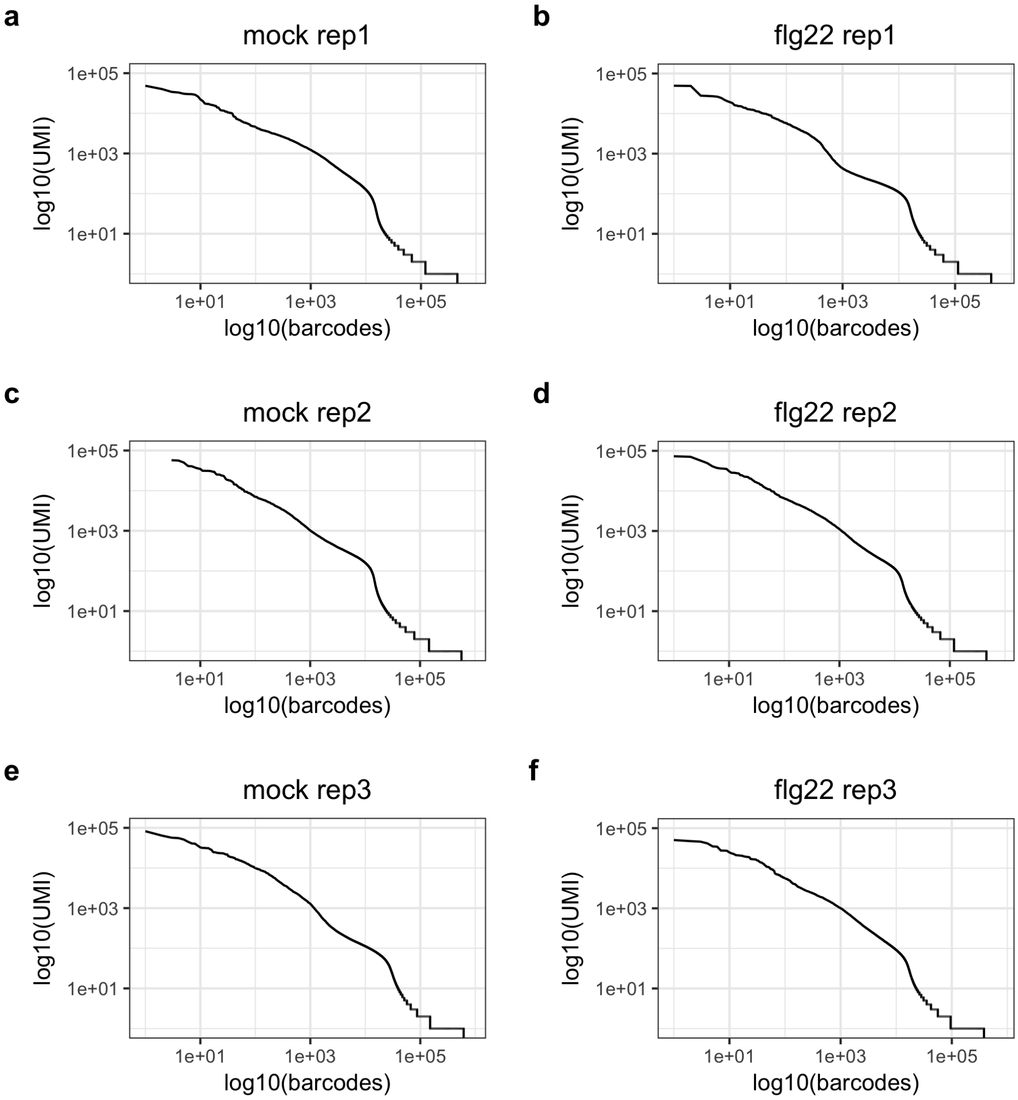


Figure 5.4: Log10 number of Unique Molecular Identifiers (nUMI) vs log10 barcode plots. The number of UMIs per barcode (in order of decreasing UMI content) is plotted on a log-log scale to determine the most likely number of real cells in each sample, (a,c,e) mock replicates 1-3, (b,d,f) flg22 replicates 1-3.

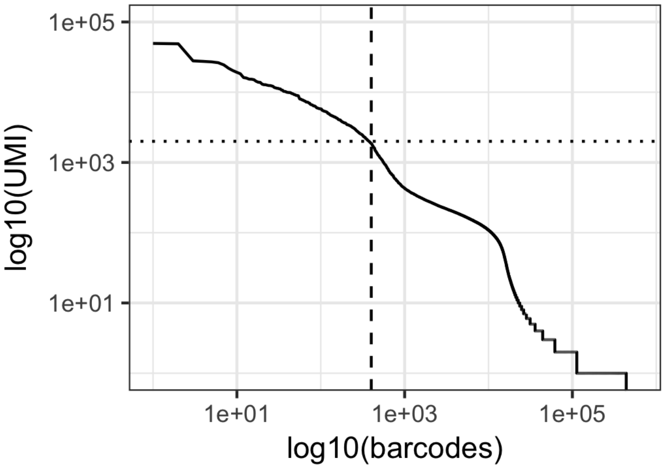


Figure 5.5: Log-log plots of nUMI vs barcode for flg22 replicate 1.The number of UMIs per barcode (in order of decreasing UMI content) is plotted on a log-log scale to determine the most likely number of real cells in each sample. The shoulder at approximately 400 cells, indicated by the vertical dashed line, indicates the likely number of 'real' cells in this sample. This corresponds to cells containing at least 2000 UMIs, indicated by the horizontal dotted line.

5.2.2 Doublet detection

The distribution of the UMIs in each sample can reveal the presence of doublets (barcode binds to transcripts from multiple cells, Figure 5.6 and Table 5.2). 75% of cells in the data set contain fewer than 7500 UMIs with the median varying across the samples between 2918 and 3810. Consistent with the best log-log plots, flg22 replicate 1 and mock replicates 2 and 3 have higher nUMI, indicating that these samples contain higher quality cells. There is a small proportion of cells that contain more than 50000 reads in mock replicates 2 and 3, and flg22 replicate 2. It is plausible that these cells are in fact doublets, and the high transcript level is the result of capturing mRNA from multiple cells. Macosko et al. (2015) observed between 0 and ∼10 % doublet concentration dependent on the concentration of cells, determined using microscopy and multi-species experiments. Within the context of this single-species experiment, there is no experimental protocol available to quantify doublets so a conservative threshold of 100000 reads was used to remove potential doublets on the basis that individual cells are unlikely to contain this many unique reads compared to the average distribution. Whether or not these cells are in fact doublets, these cells will contain radically different expression profiles from the rest of the dataset, which would make the cells difficult for algorithms to cluster with other cells in the downstream analysis. The assumption that doublets will inevitably

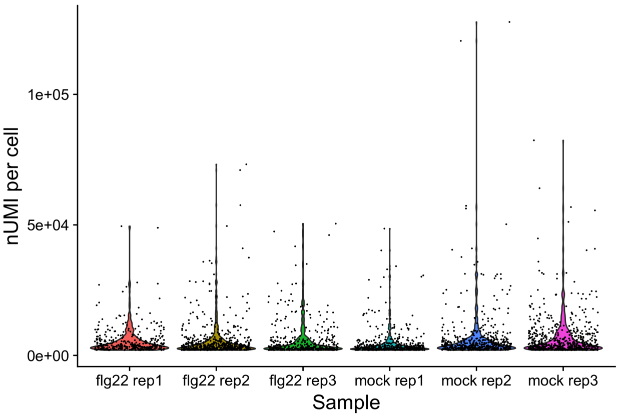


Figure 5.6: Distribution of nUMI per cell in 6 Drop-seq samples for all cells that contain more than 2000 UMIs. Dots indicate individual cells, and violin plots visualise distribution of the read numbers.

have higher UMI content than singlets is weak, because the expression profiles are different in different cells. It remains possible that these cells with very high expression levels are indeed single cells with transcript profiles that are outliers compared to regular cells. Additionally, the dataset is likely to contain more doublets that are composed of multiple cells that both have low expression levels. However these are challenging to detect within the current pipeline.

Table 5.2: Statistical summary per sample of nUMI distribution for all cells containing more than 2000 UMIs.

Sample	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
mock rep 1	2008	2384	2918	4373	3916	48583
mock rep 2	2007	2702	3810	6495	5980	127744
mock rep 3	2007	2559	3759	6530	7414	82401
flg22 rep 1	2010	2690	3685	5416	5868	49502
flg22 rep 2	2011	2488	3417	5659	5342	73208
flg22 rep 3	2005	2434	3069	5640	5517	50481

5.2.3 Detecting broken cells and contaminants

The harvesting, protoplasting and the Drop-seq protocol could create broken or dead cells in each sample. These broken cells which are passed through the machine will be low quality noisier cells that can be detrimental to downstream analysis. Broken cells in the scRNA-seq datasets have been associated with lower read counts overall and higher proportions of reads associated with the mitochondria (Ilicic et al., 2016).

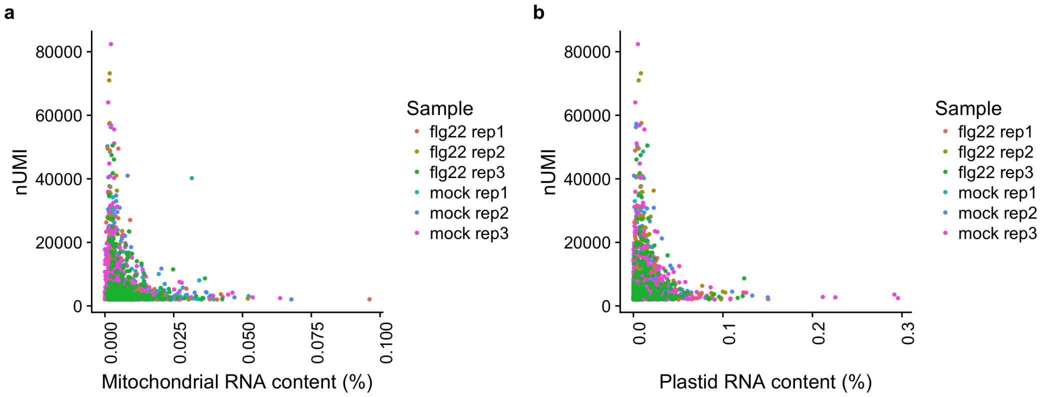


Figure 5.7: Expression of mitochondrial and plastid RNA. Scatter plots show the relationship between nUMI and percentage (a) mitochondrial RNA or (b) plastid RNA across 6 samples.

It is likely that broken cells in single cell datasets generated from plants would also be enriched in plastid reads, although this has not been currently published.

Figure 5.7a shows that very few cells contain high percentages of mitochondrial RNA and that the cells containing high proportions of mitochondrial reads also have very low expression levels. There are more cells in the data set with high plastid content (Figure 5.7b) but these are also observed in cells with very low expression levels. This pattern is consistent with Ilicic et al.'s (2016) description of broken cells, and as such these cells are to be removed from the analysis. The recommended threshold for removing cells with high proportions of mitochondrial reads is 5% (Butler et al., 2018), so cells containing more than 5% mitochondrial or plastid reads were removed from the dataset. This correponded to the removal of between 4 and 123 cells from each sample, with between 264 and 635 cells remaining in each sample (Table 5.3).

The final stage in the data quality processing pipeline is to inspect the relationship between the number of genes detected per cell (nGene) and nUMI. Figure 5.8 shows that all 6 samples show a positive correlation between number of genes expressed in each cell relative to UMIs, indicating that, as expected, cells containing high UMIs are also expressing high numbers of detected genes. However, in flg22 replicates 2 and 3, and mock replicates 1 and 2 (Figure 5.8b-e), a subset of cells with low nGene and higher than expected UMI is observed offset from the distribution. The black line on these plots indicates the separation between 'offset' cells and cells following the expected distribution. The offset cells explain the lower median nUMI values observed in these samples, and potentially why these samples performed poorly in the log-log plots. In order to investigate the origin of these offset cells, first we looked at the experimental design to establish whether an experimental detail such as time between protoplasting and being processed by the machine could explain the unusual expression profile of these cells. However there was no pattern between variation within the experimental procedure with the samples containing more offset cells.

In Chapter 4, some bulk RNA-seq replicates contained high ribosomal RNA (rRNA) content (despite the use of poly-T tails to isolate mRNA). Based on this, the non protein coding RNA content was assessed to see if the offset cells resulted from the capture of non-protein coding RNAs. However, the expression of potential contaminants such as non-coding RNAs including transfer RNA (tRNA), microRNA (miRNA) and rRNA, was extremely low ($< 0.08\%$) indicating that the poly-T tail have successfully prevented the capture of non polyadenylated RNAs. A small number of cells contained high proportions of long non-coding RNA (lncRNA), which can be polyadenylated (Kashi et al., 2016). Six cells contained $> 10\%$ lncRNA content, including one that contained 73% lncRNA) and were contained within the subset of offset cells. However this was insufficient to be considered causative. There was also no correlation between mitochondrial or plastid RNA content in these offset cells. The lack of organellar RNA in these offset cells casts some doubt on them being broken cells captured by the Drop-seq machine, as it is expected that organelles are more likely to be captured as cells.

In order to characterise the offset cells further, DGE analysis was performed comparing the offset cells in each sample relative to the main distribution of cells (described in Materials and Methods Section 2.4.11) which revealed that three genes - *tRNA (ADENINE(34)) DEAMINASE (TADA)*, *CHAPERONIN CONTAINING TCP1 8 (CCT8)* and lncRNA *AT3G09745* - are definitive markers for the offset cells across all samples. Figure 5.9 shows the expression of these three markers in flg22 replicate 3 which is representative of all the data. *TADA* encodes a nuclear encoded-deaminase, which localisaes to the chloroplast, that deaminates adenosines to inosines in tRNA-Arg(ACG) (Delannoy et al., 2009). *CCT8* encodes a T-complex protein 1 subunit theta (TCP1θ), a molecular chaperone involved in ATP hydrolysis (Fichtenbauer et al., 2012). *AT3G09745* is a lncRNA of no defined function. *TADA* and *CCT8* are both expressed in the meristem broadly across multiple cell types, (Brady et al. (2007), expression visualised on ePlant Waese et al. (2017)). Based on this information, it is hard to state definitively whether these cells are intact or broken cells: the low expression and the fact these cells are only observed in some replicates indicates that these are likely broken cells, but the consistent high expression of marker genes is unlikely to be observed in true broken cells that are

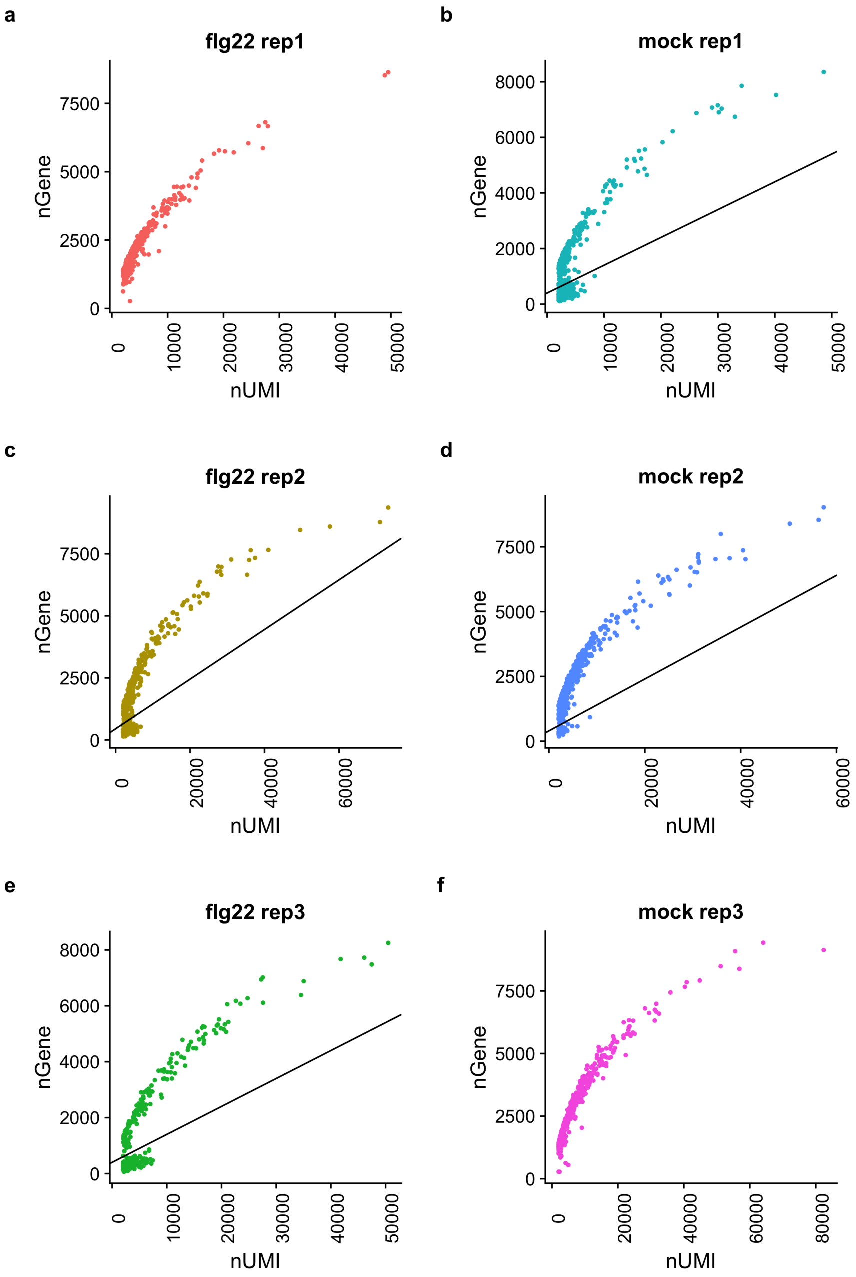


Figure 5.8: nGene vs nUMI scatter plots. The relationship between number of genes detected (nGene) and number of UMIs (nUMI) is shown as a scatter plot for each sample in a-f.

(b-e) In flg22 replicates 2 and 3 and in mock replicates 1 and 2, a black line indicates the separation between cells that followed the expected distribution above the line and ‘offset’ cells that expressed a higher number of genes relative to nUMI than expected.

produced from a variety of cell features. It seems unusual that these two unrelated genes which are expressed in different cellular compartments can distinguish these potential artefact cells. However, given ambiguity of the cell's provenance and in order to focus the analysis on the most clear data, these cells were filtered out from the dataset based on *TADA* expression. The overall cell numbers numbers before filtering were comparable and after removing offset cells the number of cells in mock replicate 1 and flg22 replicates 2 and 3 was reduced by $> 50\%$ (Table 5.3). The median expression in all of the samples is increased by filtering (Table 5.4).

Table 5.3: Number of cells remaining per sample after filtering out cells containing fewer than 2000 nUMIs, more than 5% mitochondrial, plastid or lncRNA reads, and 'offset' cells.

Sample	Minimum 2000 nUMI threshold	Mitochondrial, plastid and lncRNA filter	Offset cells filter
mock rep1	494	490	203
flg22 rep1	375	364	363
mock rep2	537	522	441
flg22 rep2	553	537	259
mock rep3	658	635	629
flg22 rep3	405	392	145

Table 5.4: Median nUMI per sample before and after filtering out low quality cells. The median nUMI in every sample increases after filtering out cells containing more than 5% mitochondrial, plastid or lncRNA reads, and 'offset' cells

Sample	Median nUMI before offset filter	Median nUMI after offset filter
mock rep 1	2918	3540
mock rep 2	3810	4118
mock rep 3	3759	3833
flg22 rep 1	3685	3733
flg22 rep 2	3417	4737
flg22 rep 3	3069	6682

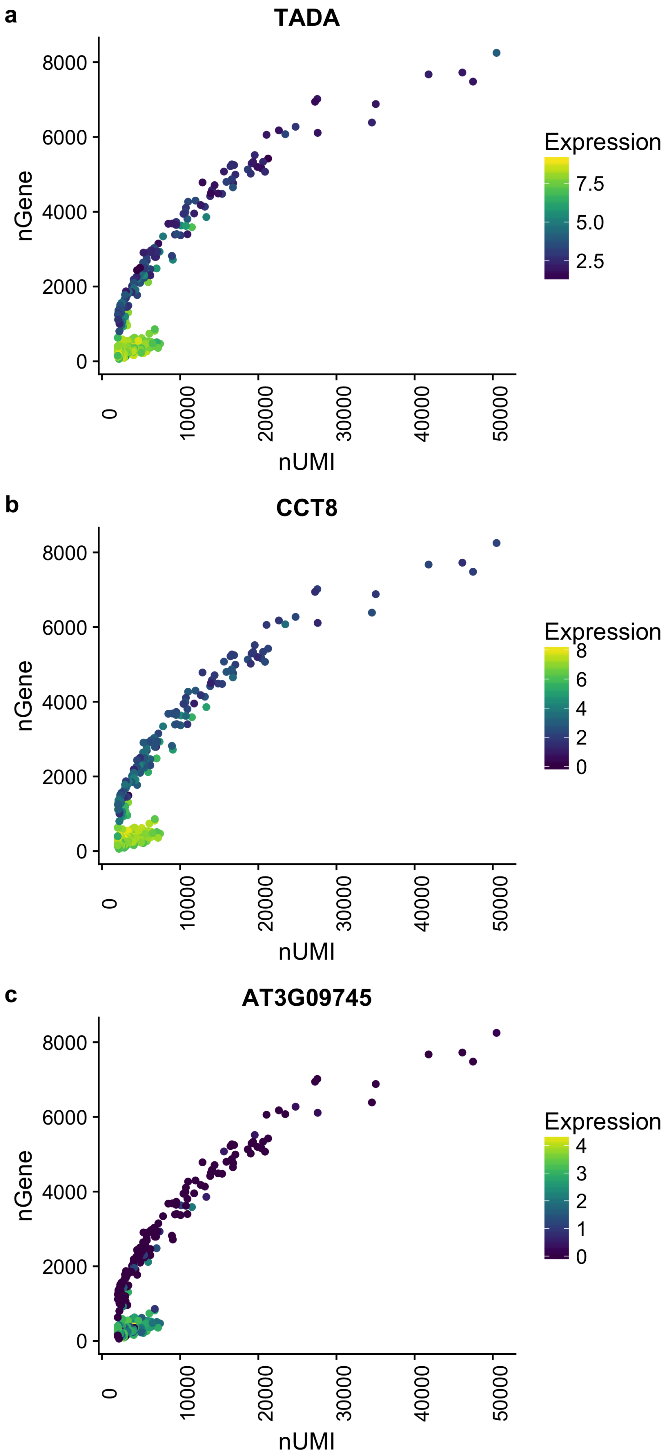


Figure 5.9: Genes enriched in cells deviating from the expected distribution. The expression of three genes: (a) *TADA*, (b) *CCT8* and (c) *AT3G09745* is enriched in 'offset' cells visualised by colour on the distribution of nGene vs nUMI content per cell for flg22 replicate 3.

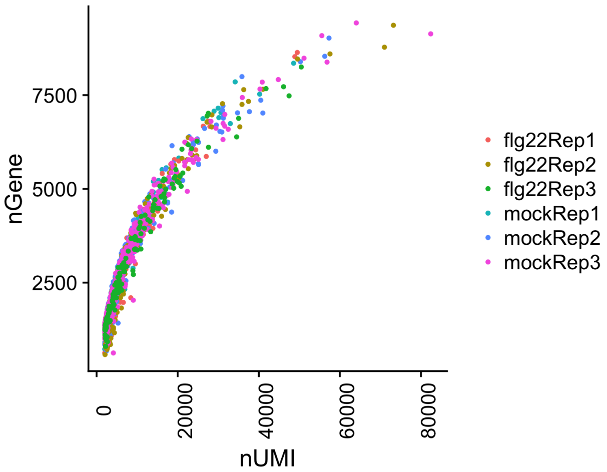


Figure 5.10: nUMI vs nGene distribution for all samples after filtering out cells containing fewer than 2000 nUMI, more than 5% mitochondrial, plastid or lncRNA reads, and ‘offset’ cells.

5.2.4 Dimension reduction

The median nUMI of the highest quality cells reveals how sparse Drop-seq data is. The *Arabdopsis thaliana* genome has $\sim$ 27000 annotated genes, so if all genes were expressed uniformly only 15% of the genes would be covered by 1 nUMI. Drop-seq provides a surface snapshot of the single cell transcriptomes, rather than a high resolution analysis. This snapshot was investigated using the dimension reduction techniques principal component analysis (PCA) and t-Stochastic Neighbour Embedding (t-SNE) which were used to group the cells and assign identity and function. scRNAseq data is very high dimensional, so in order to extract biological information from this level of complexity, dimension reduction tools can be used to reduce the complexity of the dataset in order to visualise the underlying structure of the data.

PCA projects data into a reduced number of independent dimensions. These dimensions capture the highest variance possible, preserving short- and long-range distances between data points. However unlike a bulk RNAseq analysis, PCA is not the ideal tool to distinguish differences between cells types and treatments, as PCA is restricted to linear dimensions and assumes that the dataset is normally distributed (often not the case for scRNAseq data). In a scRNAseq analysis, PCA is ideally suited to identify batch effects and outliers, and order the dimensions by variance (Andrews and Hemberg, 2017). PCA is an essential precursor to t-SNE, as t-SNE is performed using only 10-30 dimensional data.

5.2.5 Assessing batch effects using PCA

PCA reveals that in the first two dimensions there is no obvious pattern to the greatest sources of variance in the dataset (Figure 5.11a). Between replicates there is a shift observed between replicates 2 and 3 (Figure 5.11b). This batch effect is likely caused by variation in number of cells between replicates. In a perfect experiment we would expect no batch effects as plant root structure is consistent across all roots, so the proportions of cells are fixed, and since large numbers of root tips were harvested for each replicate, we would expect the cell proportions to average out. However when considering the number of cells in each sample (between 145 and 600) vs the number of cells harvested (at least an order of magnitude more), it seems likely that the cells processed would not necessarily represent a full sample of all sources of variation. Furthermore the disparity between the large number of cell in mock replicate 3 (629 cells) vs the small number of cells (145) in mock replicate 1, further increases the likelihood that mock replicate 3 has captured a wider variety of cells. Overall the differences between the replicates are mild, and some cells from all three replicates are spread across the plot, so no further corrections need to be made.

There are also no apparent differences between mock and flg22 treated cells on the PCA plot (Figure 5.11c) which is expected, given that in most scRNAseq data sets cell type differences are the greatest source of variance and the immune response has already been shown to be a weaker source of variance than cell type differences in cell type-specific RNAseq (Chapter 4).

5.2.6 Parameter optimisation of t-SNE dimension reduction and clustering

Single cell RNA-seq datasets contain far more complex data structures, compared to bulk RNAseq data, usually consisting of many globular clusters of different sizes and variance arranged in complex patterns in sample space, which cannot be visualised using PCA. t-SNE is a stochastic algorithm designed to visualise large high-dimensional datasets into 2 or 3 dimensions developed by Maaten and Hinton (2008). Unlike PCA, t-SNE does not preserve the structure of the entire dataset, instead t-SNE only preserves local structure ignoring long range distances between data points, specifically projecting data into isolated areas. These groups can then be defined as clusters calculated based on the t-SNE structure. In other words, t-SNE can be used to assign cells to clusters which often correspond to cell types. It is important to remember when interpreting t-SNE plots that the whole structure

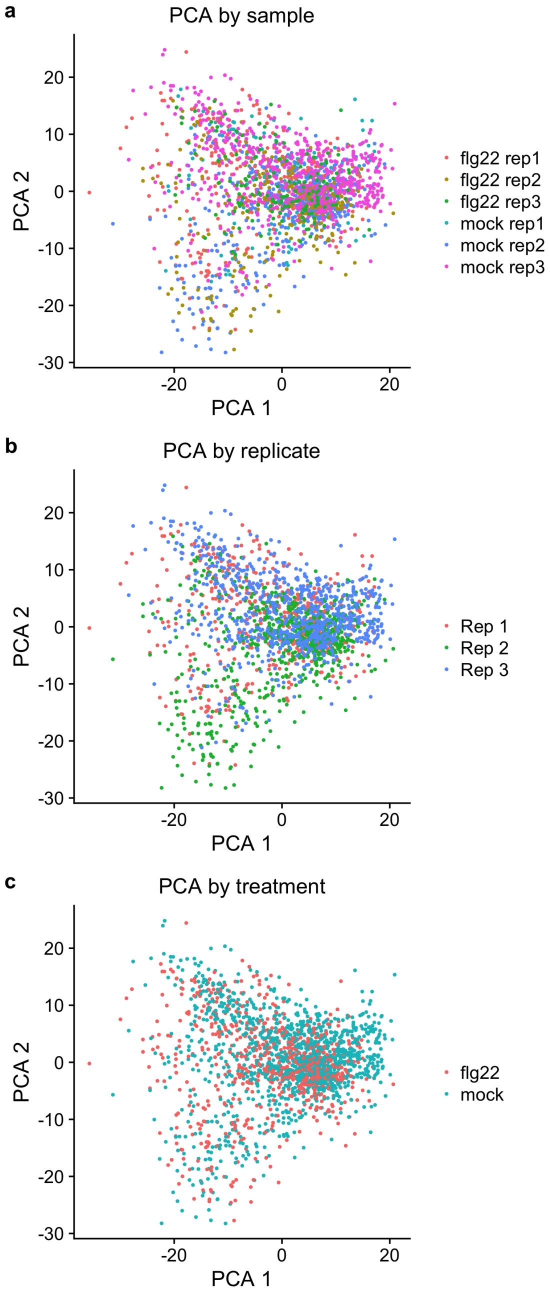


Figure 5.11: PCA plot of all merged samples comparing PC dimensions 1 and 2. In (a) cells are coloured by sample origin and in (b) cells are coloured by replicate, and in (c) cells are coloured by treatment.

of the dataset is not represented, and long distances between clusters may be meaningless. The major drawback of t-SNE is its sensitivity to various parameters such as the perplexity, defined as the number of 'close' neighbours each cell has. Setting the perplexity too high or too low will result in poor clustering of cells.

In scRNAseq, PCA can be used in conjunction with t-SNE. First, the top PCs are calculated to capture the majority of the variance in the dataset and this data is fed into t-SNE for projection into 2 dimensions. The number of PCs used is another parameter that requires optimising for the t-SNE plot. A third parameter that affects t-SNE is the number of variable genes used to calculate the principal component analysis. Using only the most variable genes can improve the PCA and thus improve the clustering on the t-SNE plot, while too few genes results in a loss of important information.

The problem with t-SNE's sensitivity to these parameters is that there are currently no automated optimisation techniques that can predict the best parameter sets. There are some metrics to determine the number of principal components (PCs) that encompasses the majority of the variation of the data, but no well-developed methods for automatically optimising the combined parameter sets. Figure 5.12 shows that the standard deviation contained in PCs decreases slowly after 20 PCs indicating that the majority of the variance in the dataset is contained within the first 21 PCs. Investigating the expression in most variable genes and cells suggested that the first 30 PCs could contain meaningful variance.

In order to ascertain the optimum set of parameters to represent the data, a wide of range of parameter sets (shown in Table 5.5) were used to calculate a multitude of t-SNE plots. These plots were inspected to determine the parameter ranges where the apparent number of clusters, and overall shape had stabilised. The plot was deemed to be stabilised when the resulting t-SNE plot structure was similar after calculating t-SNE using the same parameters multiple times using different random 'seeds' or start points. The final optimised parameters for the t-SNE plots for the data set are shown in Table 5.6, for the plot shown in Figure 5.14. Interestingly, having inspected t-SNEs produced from a wide range of PCs (4-20), the clearest clustering was obtained from 8-10 PCs, indicating that cluster differences are defined by many fewer dimensions than predicted by the elbow plot and apparent variance in gene expression across PCs.

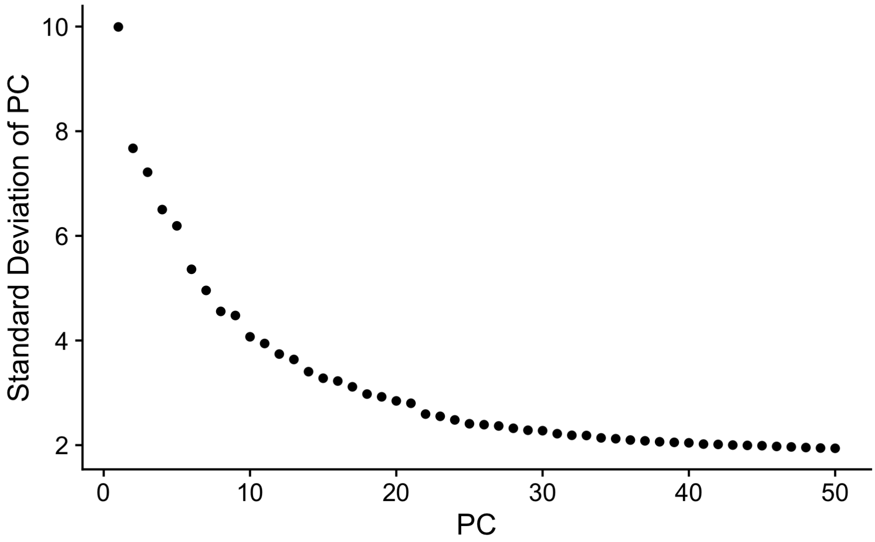


Figure 5.12: Principal component elbow plot. The standard deviation per PC reveals that as the dimensions of PCA get higher the standard deviation and therefore the information contained within the PC decreases. The 'elbow' where the plot gradient flattens, at approximately 20 PCs, indicates the number of PCs that contain the majority of the variance in the dataset.

Table 5.5: Range of t-SNE parameters tested in order to identify the optimal settings to display the variation in the dataset into clear clusters. All possible combinations of parameters were tested.

Parameter	Minimum value	Maximum value
Number of variable genes	500	3000
PCs included	4	30
Perplexity	20	40

Table 5.6: Optimal parameters for t-SNE identified using parameter testing.

Parameter	Optimum parameter value
Number of variable genes	2308
PCs included	9
Perplexity	35

Using the optimised parameters, a t-SNE plot that split up the cells into 11 distinct clusters was produced (Figure 5.14). The data was sorted into 11 clusters based on the shape of the t-SNE plot. Clusters 0-2 contain large numbers of cells which are loosely packed. Cluster 3-10 was comprised of smaller populations of cells around the edge of the plot. These clusters are more tightly packed and distinct than clusters 0-2, indicating they consist of cells with highly similar transcriptional profiles.

As with the PCA, potential batch effects were assessed by plotting the t-SNE coloured by replicate, treatment and sample (Figure 5.13a-c). Figure 5.13a reveals that replicates 2 and 3 contribute unequally to clusters on the t-SNE plot, as replicate 3 is overrepresented in the top right and replicate 2 is overrepresented in the bottom left of the plot. In all clusters, there are at least a few cells from more than one replicate, indicating that there are no entirely unique sources of variation originating from one replicate. Consistent with the PCA, there is no strong distinction between mock and flg22 treatment on cells apparent from the t-SNE plot. This further indicates that treatment differences are not driving the primary variation in the dataset, instead clusters are based on the fundamental differences between cell transcriptomes.

5.2.7 Supervised approach to identifying cell type identity in t-SNE clusters.

In Chapter 4 and in various published scRNA-seq datasets, the greatest driver of variation between clusters is cell type identity. Based on these findings, the expression of known cell type marker genes was examined in relation to the clusters (Table

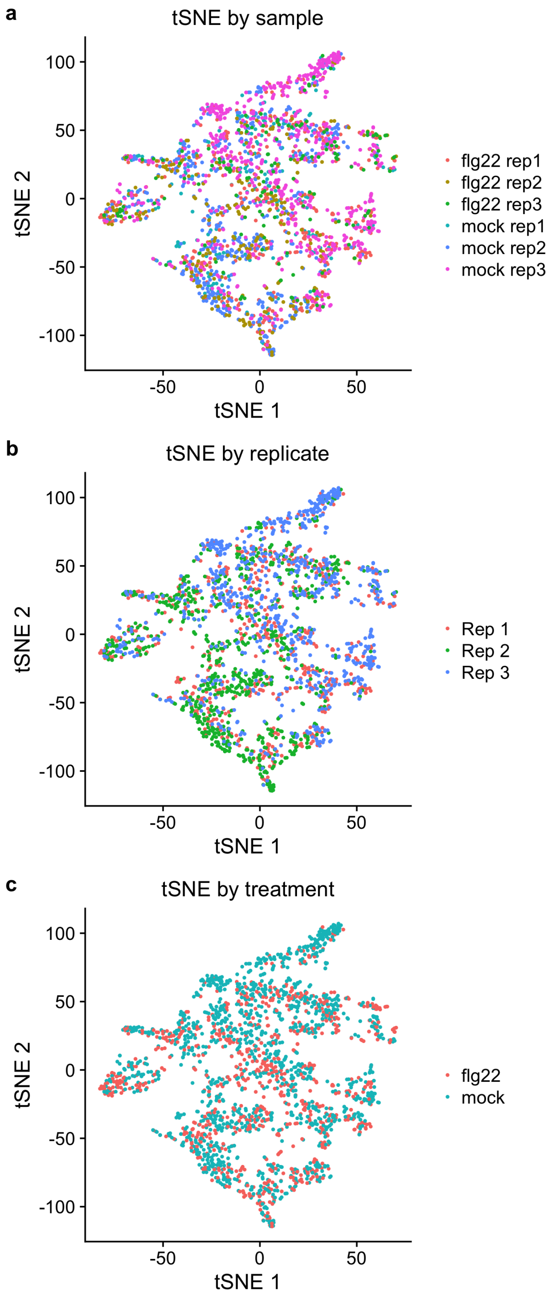


Figure 5.13: Two dimensional t-SNE plot of all merged samples. In (a) cells are coloured by sample origin and in (b) cells are coloured by replicate, and in (c) cells are coloured by treatment.

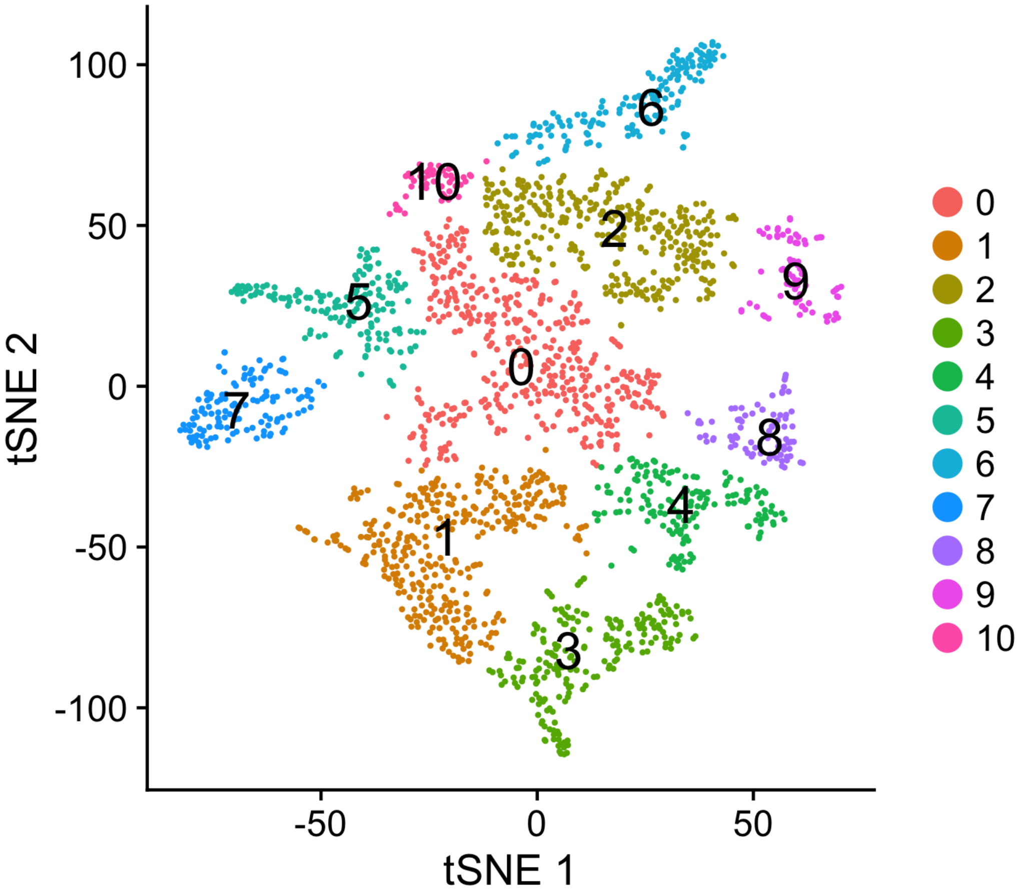


Figure 5.14: Eleven clusters were calculated using two dimensional t-SNE plot of all merged samples.

5.7, Figure 5.15). The expression of all the markers listed in Table 5.7 were checked against Brady et al. (2007)'s developmental root microarray datasets to confirm expression in the meristem.

The most strongly expressed marker was *ROOT CAP PROTEIN-2LIKE PROTEIN MDK4.20, AT5G52370 (MDK4.20)*, a root cap (encompassing columella and lateral root cap (LRC) cells) expressed marker (Lilley et al., 2011). The violin plot in Figure 5.16a shows that *MDK4.20* is highly expressed in cluster 3, but is also expressed in adjacent clusters 1 and 4. The feature plot in Figure 5.16b which shows the expression of *MDK4.20* overlaid on the optimised t-SNE plot shows that cells expressing high levels of *MDK4.20* cluster closely at the bottom of the plot, indicating these cells are likely to be root cap cells. The rest of the cell type-specific markers are less defined by borders between clusters. *AUXIN TRANSPORTER PROTEIN 1 (AUX1)*, a more specific marker to the LRC only (Marchant et al., 2002; Brady et al., 2007) is also expressed in cluster 3 although it is more strongly expressed in clusters 1 and 9. This could indicate that cluster 1 contains lateral root cells, whereas cluster 3 contains both LRC and columella cells.

Consistent with *AUX1* expression, *PIN-FORMED 2 (PIN2)* and *WEREWOLF (WER)* (Lee and Schiefelbein, 1999) markers that are expressed both in lateral root cap and epidermis cells are most strongly associated to cluster 1. *WER* is also expressed broadly across clusters 0, 1 and 3 (Figure 5.17a-b), with the strongest expression in the cluster 1 (LRC).

Epidermis specific markers *GLABRA2 (GL2)* and *CAPRICE (CPC)* (which are not expressed in the lateral root cap) are enriched most strongly in cluster 9. Looking at the feature plots in Figure 5.17c and e shows that although *GL2* and *CPC* are most strongly expressed in that cluster, it is not a definitive marker for the cluster, like *MDK4.20.* The cells not expressing the epidermis markers are either cells from a different lineage or epidermis cells, with lower expression of the *GL2* or *CPC* such that it was not captured by the Drop-seq beads. The structure of the cluster is less tightly packed than cluster 3 suggesting that these cells are less strongly defined as one identity. Additionally, cells expressing *GL2* and *CPC* and a fourth epidermis marker *P-GLYCOPROTEIN 4 (PGP4)* appear widely across the clusters, indicating that potential epidermis cells are not restricted to clusters 1 and 9.

A large proportion of cell type markers expressed in cortex, endodermis and stele tissues are expressed most highly in cluster 5. For example, *SCARECROW (SCR)*, *SHORTROOT (SHR)* and *PIN-FORMED 3 (PIN3)* are all most strongly expressed in cluster 5, but similarly to the epidermis markers, the expression of these markers is not restricted to one cluster. Given the number of cells in cluster 5 it would be very surprising if it contained all the cortex, endodermis and stele cells. These unclear expression patterns indicate that an alternative source of variation between the cells is superseding cell type identity and making it impossible to assign definitive identity to the majority of cells.

Finally, the expression specific quiescent centre (QC) markers *WUSCHEL-RELATED HOMEOBOX 5 (WOX5)*, *AGAMOUS-LIKE 42 (AGL42)* (Nawy et al., 2005) and QC expressed markers *PISTILLATA (PI)* (which is also expressed in young cortex and endodermis cells) and *PERIANTHIA (PAN)* (which is also expressed in columella cells) were examined to try to identify potential QC cells. Given the number of QC cells relative to the total number of cells in the meristem, it would be expected that these cells would make up a very small proportion of the captured cells, but we would expect them to have a distinct transcriptional identity, as QC cells are slowly dividing relative to the surrounding cells and undertake a different function to the initial cells surrounding them. The canonical QC marker *WOX5* was only detected in 2 cells across the dataset. However the QC marker *AGL42*, and QC/columella marker *PAN* were detected in cells in clusters 2 and 5. Figure 5.18 shows that *AGL42* is weakly expressed and the cells expressing it do not cluster closely into a defined cluster. The expression pattern of *PI* more closely matches the other endodermis markers and does not distinguish the QC. As such we cannot identify definitive QC cells within the dataset.

5.2.8 Unsupervised identification of cluster marker genes

Due to the limited success of assigning cluster identity using known marker genes, an unsupervised approach was used to identify the top markers for each of the clusters. This alternative approach enabled us to further understand the variation underlying the clustering. Markers were identified based on positive differential gene expression between one cluster and all other clusters. In total, between 11 and 1000 potential markers were identified ($p < 0.05$) per cluster, and the top ten for each cluster are visualised in Figure 5.19. All of the clusters have distinctive expression patterns, although the expression of some markers are overlapping in multiple clusters; for example *TUBULIN BETA-5 CHAIN (TUBB5)*, which codes for a tubulin component, is expressed strongly in both clusters 6 and 9 and *GLUTATHIONE S-TRANSFERASE U5 (GSTU5)*, a gene associated with redox) is expressed in both clusters 4 and 8. The analysis identified between 1 and 500 unique (differentially expressed in only one cluster) positive markers per cluster. Clusters 3-9, corresponding to the clusters on the edges of the t-SNE plot, are represented by strongly expressed

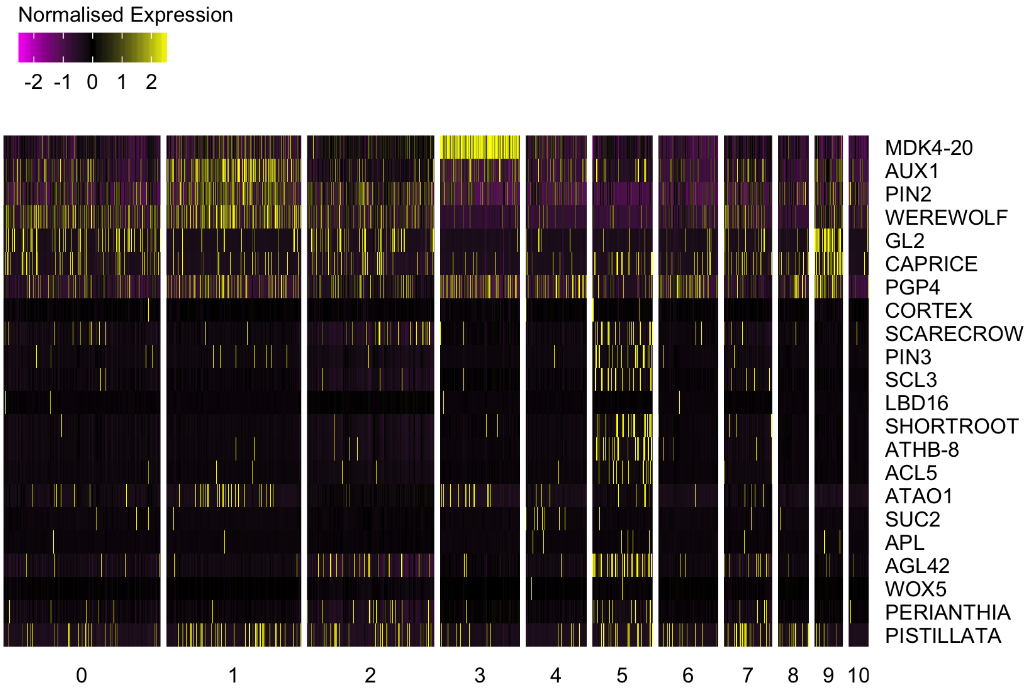


Figure 5.15: Root cell type marker gene expression in cell clusters. The expression in every cell (one vertical line represents one cell) of known cell type marker genes (rows, genes detailed in Table 5.7) grouped by clusters from Figure 5.14.

Table 5.7: Root cell type markers. Table details the observed expression pattern of cell type-specific markers, and a relevant citation.

Gene ATG code	Gene name	Cell type expressed	Published
AT5G54370	MDK4-20	root cap	Lilley et al. (2011)
AT2G38120	AUX1	lateral root cap	Marchant et al. (1999)
AT5G57090	PIN2	lateral root cap, epidermis	Müller et al. (1998)
AT5G14750	WEREWOLF	lateral root cap, epidermis	Lee and Schiefelbein (1999)
AT1G79840	GL2	epidermis	Lin et al. (2015)
AT2G46410	CAPRICE	epidermis	Wada et al. (1997)
AT2G47000	PGP4	epidermis	Wyrsch et al. (2015)
AT1G09750	CORTEX	cortex	Dinneny et al. (2008)
AT3G54220	SCARECROW	endodermis	Malamy and Benfey (1997)
AT1G70940	PIN3	stele, endodermis,cortex	Birnbaum et al. (2003)
AT1G50420	SCL3	stele, cortex, endodermis	Birnbaum et al. (2003)
AT2G42430	LBD16	pericycle	Wyrsch et al. (2015)
AT4G37650	SHORTROOT	stele,	Helariutta et al. (2000)
AT4G32880	ATHB-8	stele, pericycle	Bargmann et al. (2013)
AT5G19530	ACL5	stele	Birnbaum et al. (2003)
AT4G14940	ATAO1	stele	Ghuge et al. (2015)
AT1G22710	SUC2	phloem	Gottwald et al. (2000)
AT1G79430	APL	phloem	Lee et al. (2006); Bonke et al. (2003)
AT5G62165	AGL42	QC	Nawy et al. (2005)
AT3G11260	WOX5	QC	Sarkar et al. (2007)
AT1G68640	PERIANTHIA	QC and columella	
AT5G28770	PISTILLATA	QC and young cortex and endodermis	

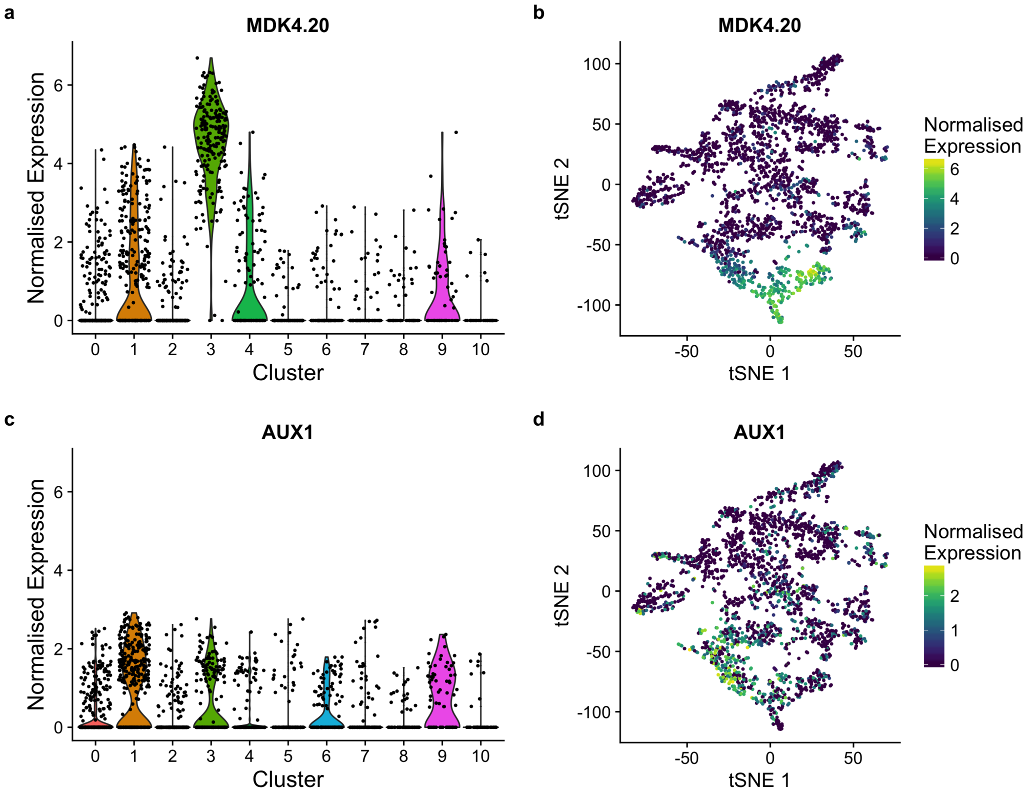


Figure 5.16: Expression of root cap markers.
(a-b) Root cap marker *MDK4.20* is very strongly expressed in cluster 3 and also expressed in clusters 1, 4 and 9 (a), corresponding to the bottom of the t-SNE plot (b).
(c-d) Lateral root cap-specific marker *AUX1*is strongly expressed in cluster 1 and further expressed in clusters 3, 6 and 9 (c), visualised on the t-SNE plot in (d).

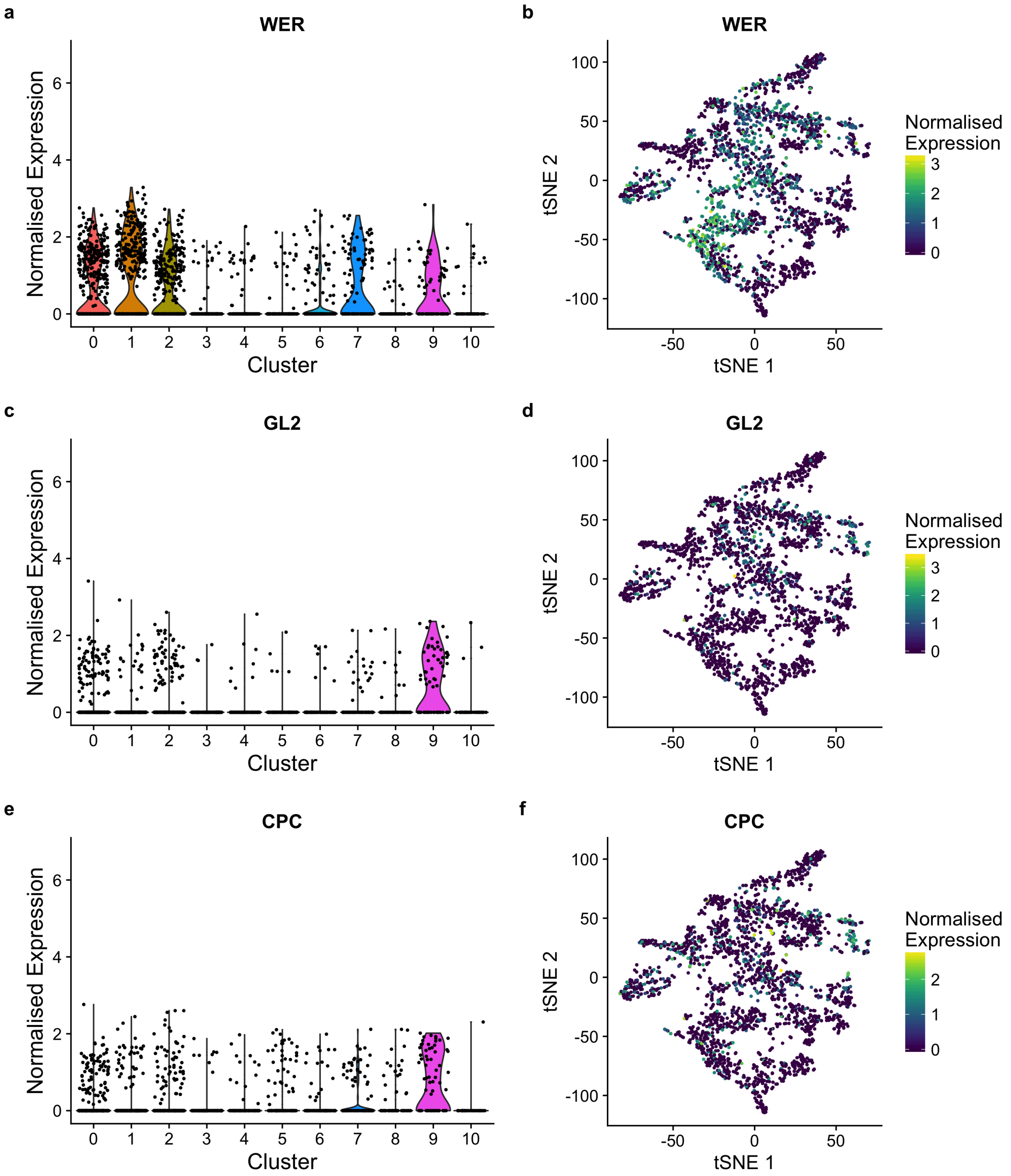


Figure 5.17: Expression of epidermis-specific markers.
(a-b) Epidermis marker *WER* is expressed in clusters 0, 1, 2, 7 and 9 (a), corresponding to a wide band of expression across the center of the t-SNE plot (b).
(c-d) Epidermis marker *GL2* is expressed most strongly in cluster 9 (c), a distinct cluster on the edge of the t-SNE plot in (d).
(e-f) Epidermis marker *CPC* is expressed most strongly in cluster 9 (e), matching the expression of *GL2* on the t-SNE plot in (f).

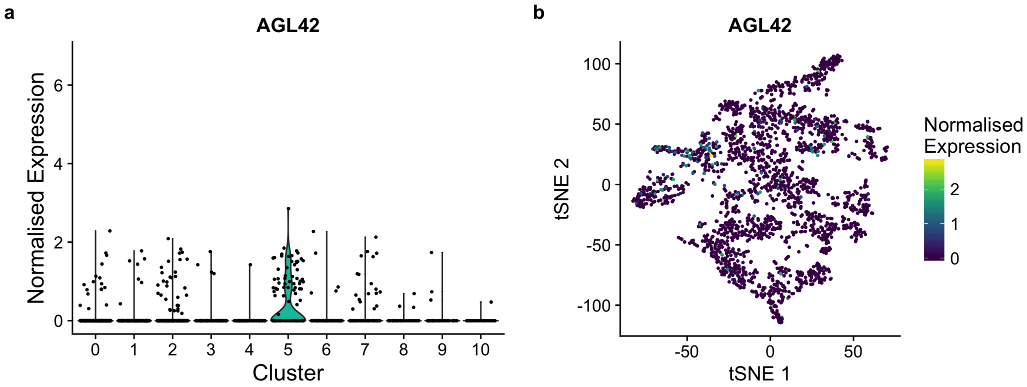


Figure 5.18: Expression of epidermis-specific markers.
(a-b) QC marker *AGL42* is most strongly expressed in cluster 5 and more weakly expressed in clusters 2 and 7 (a), visualised on the t-SNE plot (b).

markers. In contrast, clusters 0, 1, 10 and particularly 2 are characterised by much more lowly expressed markers.

Clusters 1 and 3 have been assigned potential identity based on Figure 5.15; *MDK4.20* was identified independently as a top marker for cluster 3. The top novel marker genes included *AT3G19430* and *AT4G27400* whose corresponding proteins contain a root cap specific domain (Hundertmark and Hincha, 2008). The gene expression was also confirmed to be root cap specific in Brady et al. (2007). In total 148 potential root cap specific markers genes were identified (cluster 3). *AT5G60520*, a marker gene for cluster 1, also contains a root cap-specific domain and is expressed in the lateral root cap (Brady et al., 2007).

In Figure 5.15, cluster 5 is most strongly enriched in genes that are known to be expressed in vasculature and in the cortex and endodermis cells layers. However the expression pattern is much more uneven than the markers for clusters 3 and 9. The top novel markers identified for cluster 5 include a polar auxin transporter *LIKE AUXIN RESISTANT 2 (LAX2)*, a transcription factor *BASIC HELIX-LOOP-HELIX 144 (bHLH144)* and a co-activator *GRF1-INTERACTING FACTOR 1 (GIF1)*. *LAX2* expression has been detected in young vascular tissue, the QC and in columella cells (Péret et al., 2012). *bHLH144* has been shown to be a potential interactor with *LONESOME HIGHWAY (LHW)* in the regulation of root vascular initial populations (Ohashi-Ito and Bergmann, 2007). *GIF1* is expressed in both cortex and phloem companion cells in the meristem (Brady et al., 2007). It is involved in cell proliferation in leaves (Kim and Kende, 2004), but has no defined role in roots. The combined evidence of the known marker genes and novel markers suggests that this cluster contains cells from the young vasculature, but that the

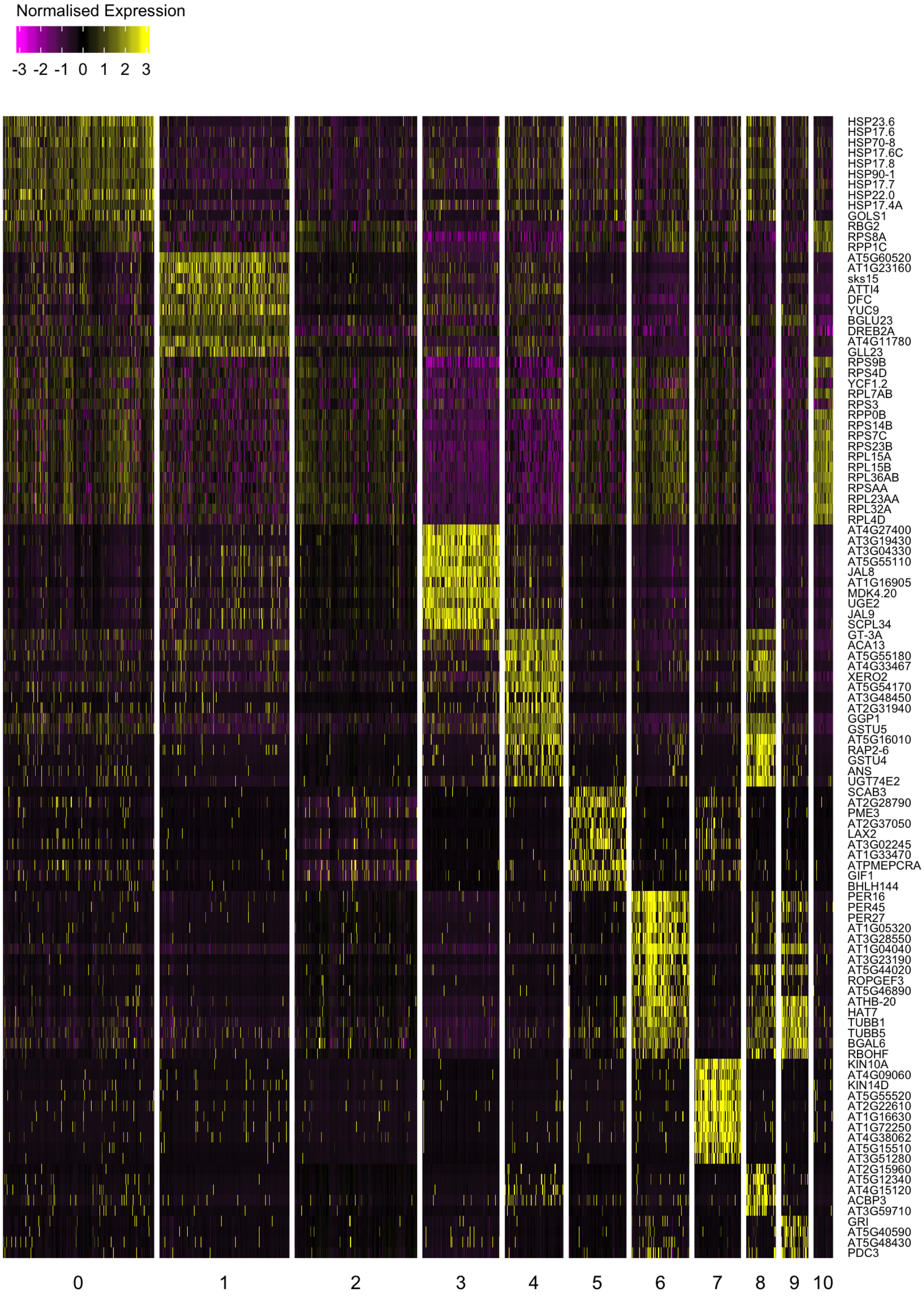


Figure 5.19: Expression of top cluster marker genes. Heatmap shows the normalised expression of the top markers (rows) for each cluster in each cell (one cell per vertical line) separated by cluster (split by grid).

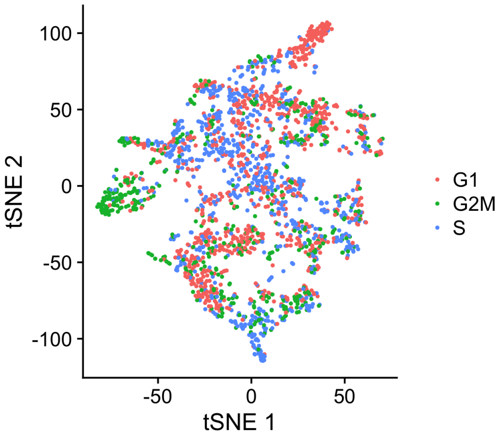


Figure 5.20: Two dimensional t-SNE plot of all merged samples coloured by cell cycle phase.

gene expression patterns are insufficient evidence to define the clusters as specifically vasculature.

Clusters 6 and 7 are clearly distinct in Figure 5.14 suggesting that these cells have gene expression patterns that strongly differ from the other cells in the data set. However the investigation of known cell type markers (Figure 5.15) does not suggest that these cells belong to a specific known cell type. The unsupervised identification of novel marker genes revealed that cluster 7 is dominated by cell cycle associated genes including *KINESIN-LIKE PROTEIN 10A (KIN10A)* and *KIN14D* in the top 10 markers. In total 253 potential markers were identified for this cluster ($p < 0.05$) including 9 Cyclins, 2 cyclin dependent kinases (CDKs), 2 CDK regulatory subunits and 20 kinesin-like proteins. Assigning cell cycle phase to each cell reveals that cluster 7 includes genes that are actively dividing cells in G2/M phase (Figure 5.20.

The top 3 markers defining cluster 6 are peroxidases: *PER16*, *PER27* and *PER45* which have a role in cell wall loosening (Francoz et al., 2015; Dunand et al., 2007) during elongation in roots. Investigating the known expression patterns (Brady et al., 2007) revealed that peroxidases are expressed in cells on the border between the meristem and the elongation zone (data for *PER16*, shown in Figure 5.22e and f). This suggests that cluster 6 represents a population of cells that have begun elongating.

Finally, looking at clusters characterised by weakly expressed markers: cluster 0 is strongly enriched in heat shock proteins (HSPs), which although characterised as being highly responsive to a variety of stresses, in tissues from both above and below ground in plants (Swindell et al., 2007), have also been linked to regulation of steroid hormone receptors (Kregel, 2002) and involved in root development (Petti et al., 2014).

The list of markers for clusters 1 & 2 are enriched in genes coding for ribosomal proteins. Ribosomal proteins were observed to be an artefact in Chapter 4, so the potential for this observed enrichment to have resulted as an artefact from individual samples was investigated. Figure 5.21 reveals that percentage content of ribosomal protein genes in individual cells varied from 0% to 25-30% across all samples. The distribution of these ribosomal protein percentages within each sample were bimodal, where some cells have a high proportion of genes coding for ribosomal proteins (and therefore a high number of ribosomes) and others have a very low proportion. This is particularly pronounced in flg22 replicate 1 and mock replicates 2 and 3. Based on the consistency of variation across all the samples, ribosomal protein content is unlikely to be an artefact in this dataset. One possible explanation for this bimodal expression is that some cells in the meristem might require more ribosomes in order to facilitate cell division, through increased protein production. Alternatively, ribosomal proteins could play a regulatory role in root development. In *Arabidopsis* leaves, ribosomal proteins have been shown to mediate cell proliferation and cell expansion (Fujikura et al., 2009), in some cases undertaking different functions based on stress status (Ferreyra et al., 2010). Examining the known expression of ribosome markers identified in Figure 5.19 revealed that the bulk of the meristem above the QC is enriched in ribosome protein genes relative to the root cap and elongation zones Brady et al. (2007), shown for *60S ACIDIC RIBOSOMAL PROTEIN P0-2 (RPP0B)* in Figures 5.22b and d .

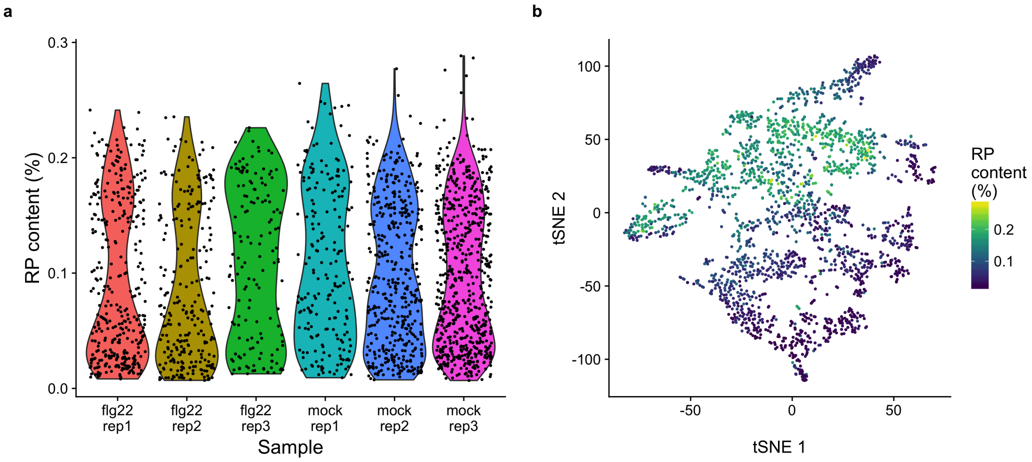


Figure 5.21: Percentage content of genes coding for ribosomal proteins (RP content) per cell.
(a) The distribution of RP content in cells is distinctly bimodal within each sample.
(b) The cells with high RP content are clustered together in the t-SNE plot.

Combining the observations of root cap markers, ribosomal protein genes and peroxidases, reveals that the development of the root can be mapped strongly to the t-SNE plot (Figure 5.22). Clusters 1 and 3 correspond to the root cap, clusters 0, 2, 4, 5, 7, 8, 9, 10 correspond to the cells in the meristem and cluster 6 corresponding to cells that are beginning to elongate. The relative proportions of cells within these clusters make sense within the context of the experiment, as root meristems were harvested with the cuts made at the border between the meristem and the elongation zone.

Cell cycle phase data supports this developmental axis across the t-SNE plot (Figure 5.20). Cells were scored based on their likelihood to be in S or G2M phase, and the remaining cells were assigned as G1 cells. This is a limitation of the cell cycle assignment method. Cells assigned as G1 in fact incorporates cells in gap phase, as well as non-cycling and endocycling cells. The cells identified as elongating are all assigned to G1 phase. The up-regulation of *SIAMESE* in cluster 6 suggests that these cells are in fact endocycling. This adds to the evidence that cluster 6 contains cells that are starting (or about to start) to elongate. By contrast, the cells within the meristem represent a variety of cell cycle stages, consistent with the cycling expression associated with cell division.

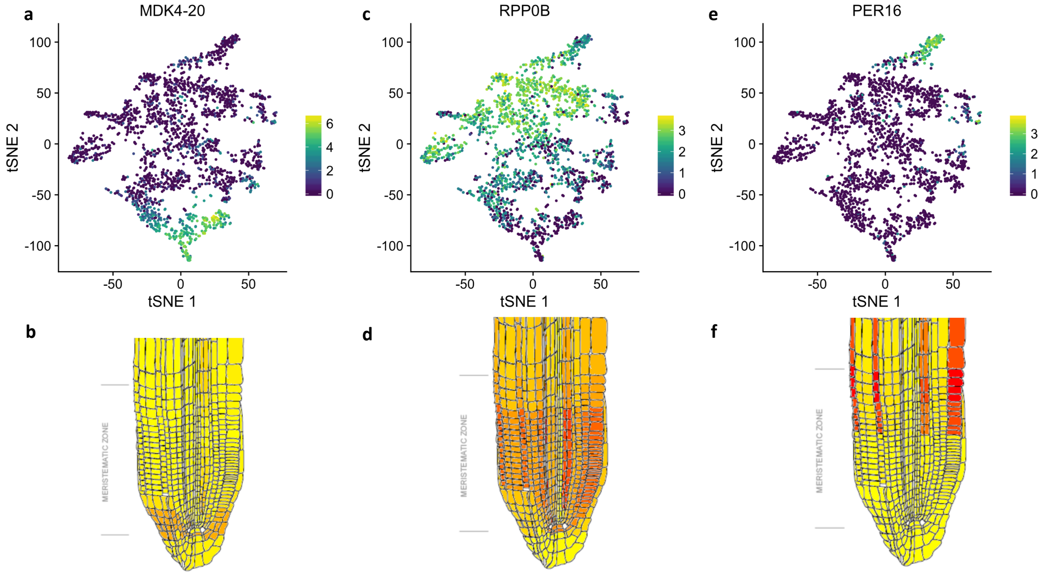


Figure 5.22: T-SNE clusters can be mapped to developmental stage.
(a-b) Root cap marker *MDK4.20* expression on the t-SNE plot (a) and visualised on a root tip using expression data from Brady et al. (2007).
(c-d) Ribosomal protein *RPP0B* expression on the t-SNE plot (c) and visualised on a root tip (d)
(e-f) *PEROXIDASE 16 (PER16)* expression on the t-SNE plot (e) and on a root tip (f)

5.2.9 Differential expression between flg22 and mock treated roots

DGE analysis compared flg22 and mock treated cells for each of the three developmental stages. Cells were compared within the meristem (defined as clusters 0, 2, 4, 5, 7, 8, 9, and 10), root cap (clusters 1 and 3) and early elongation zone (cluster 6). Ten differentially expressed genes (DEGs) were found in the meristem (Table 5.8), eight in the root cap and five in the elongating cells. Consistent with the flg22 response observed in cell type-specific RNAseq of mature root cell types, flg22 has only induced genes in the root tip, and did not significantly repress any genes.

HSPs were differentially expressed in both the meristem and the root cap. Specifically *HEAT SHOCK FACTOR 7A (HSFA7A)* was differentially expressed (DE) in both the root cap and the meristem, *HEAT SHOCK PROTEIN 17.6A (HSP17.6A)* was DE in the meristem only, and *HEAT SHOCK PROTEIN 17.6C (HSP17.6C)*, *HEAT SHOCK PROTEIN 70-5 (HSP70-5)*, and HSP associated gene *HSP ORGANIZING PROTEIN 3 (HOP3)* were DE in the root cap. HSPs are known to be involved in the cellular response to various forms of stress besides heat and transcriptional profiling of HSPs revealed that they represent an interaction point between multiple stress response pathways including the immune response. Furthermore profiling revealed that when responding to stressed other than heat or osmotic stress, HSPs exhibited family or tissue specific expression (Swindell et al., 2007). HOP3 has been shown to play an essential role during endoplasmic reticulum (ER) stress in plants particularly during immune response (Fernández-Bautista et al., 2017). *AT3G09070*, a wound responsive protein (Cheng et al., 2017) was also DE in both the meristem and root cap. The DEGs induced in the elongating cells did not overlap with the meristem or root cap DEGs and included probable WRKY transcription factor *WRKY47*.

Table 5.8: flg22 induced genes in the meristem

Gene symbol	Description	Avg. logFC	P value (adj)
AT2G35382	snoRNA	0.774	2.73×10^{-12}
AT2G43140	Transcription factor bHLH129	0.31	9.86×10^{-7}
AT3G15450	AT3g15450/MJK13_11	0.322	2.75×10^{-5}
ATCTH	Zinc finger CCCH domain-containing protein 23	0.251	6.59×10^{-4}
AT1G25275	AT1G25275 protein	0.46	5.97×10^{-3}
AT1G08643		0.326	7.37×10^{-3}
AT3G07090	PPPDE putative thiol peptidase family protein	0.308	1.69×10^{-2}
STY46	Serine/threonine-protein kinase STY46	0.275	2.81×10^{-2}
HSFA7A	Heat stress transcription factor A-7a	0.322	4.53×10^{-2}
HSP17.6A	17.6 kDa class I heat shock protein 1	0.325	4.87×10^{-2}

Table 5.9: flg22 induced genes in the root cap

Gene symbol	Description	Avg. logFC	P value (adj)
HSFA7A	Heat stress transcription factor A-7a	0.559	2.62×10^{-4}
HOP3	Hsp70-Hsp90 organizing protein 3	0.751	4.39×10^{-4}
HSP70-5	Hsp70b	0.926	3.46×10^{-3}
AT3G07090	PPPDE putative thiol peptidase family protein	0.499	6.11×10^{-3}
BAG6	BAG family molecular chaperone regulator 6	1.09	7.58×10^{-3}
CLPB1	Chaperone protein ClpB1	0.589	1.66×10^{-2}
HSP17.6C	17.6 kDa class I heat shock protein 3	1.17	2.2×10^{-2}
UGT85A1	UDP-glycosyltransferase 85A1	0.578	3.12×10^{-2}

Table 5.10: flg22 induced genes in the early elongation zone

Gene symbol	Description	Avg. logFC	P value (adj)
AT5G66050	Putative uncharacterized protein At5g66050	0.816	3.62×10^{-3}
WRKY47	Probable WRKY transcription factor 47	0.271	4.49×10^{-3}
AT5G01760	ENTH/VHS/GAT family protein	0.527	6.38×10^{-3}
PSBP1	Oxygen-evolving enhancer protein 2-1, chloroplastic	0.296	8.73×10^{-3}
SNOR105	SNOR105 (SMALL NUCLEOLAR RNA 105); snoRNA	0.341	2.61×10^{-2}

5.3 Discussion

The *Arabidopsis thaliana* root tip is a complex structure composed of many different cell types. Studying these cell types on a cell type-specific level has been limited by the availability of fluorescent marker genes and the difficulties in obtaining sufficient tissue to perform FACS-based transcriptomics (Efroni and Birnbaum, 2016). Drop-seq represents a great opportunity to understand the plant root at the level of single cells, in order to tackle complex questions relating to development, cell cycle and immunity triggered root growth inhibition. In this chapter, a large scale Drop-seq dataset for mock- or flg22- treated root meristems were analysed. This is the first Drop-seq dataset obtained for root tips, and is among the first Drop-seq experiments to be performed on plants (Shulse et al., 2018). In order to analyse this novel dataset, the existing pipelines were adapted to the *Arabidopsis* genome and as the analysis progressed, challenges unique to performing Drop-seq in plants were identified, in addition to the general challenges of scRNAseq. Cell populations corresponding to three developmental zones and two specific cell types were identified. Finally, DEG analysis identified flg22-induced genes responding across the three developmental zones.

5.3.1 The challenges of utilising Drop-seq in plants

The primary challenge in Drop-seq is establishing the number of captured cells. In particular, this involves counting the number of doublets, and identifying broken cells. The offset cells in Figures 5.8 and 5.9 are particularly representative of this problem. As described in Section 5.2.3, the low expression levels could indicate that these are broken cells, but the identification of highly consistent marker genes suggests that they could be real cells. If this is the case then a great deal of information is likely to have been filtered out of the dataset (e.g. $> 50\%$ of cells in flg22 replicate 3).

As Drop-seq becomes a more popular method to perform scRNAseq, new techniques are being developed to identify these artefacts. Ilicic et al. (2016) utilised microscopy of the microfludics device in the Drop-seq machine to determine which cells were broken, empty, or were doublets from a mammalian tissue dataset, and performed gene expression analysis to determine the footprint of low quality cells. They discovered that mtRNA and genes associated with the GO term 'membrane' were up-regulated in broken cells. A similar experiment needs to be performed in plant tissue as the transcriptional footprint of broken cells from plants will be different from animal cells due to fundamental differences such as the presence of the ptRNA (from plastids in roots or chloroplasts in aerial plant tissues). If we assume that these offset cells are broken, the expression pattern of broken cells might even be tissue specific as the marker genes *TADA* and *CCT8* are expressed strongly in the root meristem (Brady et al., 2007). A secondary advantage to performing microscopy alongside Drop-seq is that the number of expected cells would become known, and the pipeline wouldn't rely on predictive methods to discover the most likely number of cells in the dataset. However such an experiment would be very time consuming and difficult to automate.

In addition to studying broken cells, Ilicic et al. (2016) also examined the footprint of empty cells, but again a plant specific experiment needs to be performed. It could be argued that low expression is not sufficiently discriminative to distinguish broken cells. In flg22 replicate 3 more than 50% of the cells were filtered as broken cells whereas very few broken cells were detected mock replicate 3. Given that there was no discernible difference between the experimental procedure in the replicates which contained offset cells and those that did not, it seems unlikely that the proportion of broken cells should vary that much. Cells with high mtRNA or ptRNA are filtered from the dataset based on a standard threshold of 5% (Figure 5.7). They are filtered out on the basis that they may be broken cells, and even if they are not the high organellar RNA content is likely to increase noise in the dataset and reduce quality of clustering. Very few cells in the dataset contain more than 5% mtDNA content whereas many more cells contain more than 5% ptRNA. If the plastid concentration naturally varies more than mitochondrial content then the 5% threshold may not be optimal for plant datasets. This threshold should be tested to ensure that valuable information is not lost in the analyses.

The biases associated with Drop-seq data collection are not fully known. In theory, Drop-seq is unbiased in cell type collection and sequencing as all cells are treated equally. However, Drop-seq data only reveals the strongest transcriptional signals and so cells with higher RNA content are much more likely to be captured and retained through filtering, than those with low RNA content. This is particularly relevant to our study of root tips as the RNA content in epidermis cells is known to be higher than QC cells. Brennecke et al. (2013) compared the RNA content of *WOX5*-expressing QC cells and *GL2*-expressing epidermis cells (extracted using laser microdissection) and showed that on average *GL2* expressing cells contained 60pg of RNA whereas QC cells only contained 10pg RNA. In discussing the difficulties in general scRNAseq methods, they stated that the amount of biological RNA captured can differ widely due to technical differences in the efficiency of cell lysis and biological differences in cell size and total RNA content of each cell. The Dolomite Bio Drop-seq machine utilised in this experiment operates at a 5% cell capture efficiency. As the efficiency of this technology improves a greater proportion of cells will be retained and the effect of these potential biases in cell capture will be reduced.

5.3.2 Assigning cell identity

The aim of this analysis was to perform cell type-specific transcriptomics but assigning cell-type identity (or cell fate) has proven tricky. With the exception of root cap marker *MDK4.20*, cell type markers are lowly expressed widely across clusters. In general, the low resolution of Drop-seq data makes it highly likely that the expression of any given cell type-specific marker will be absent or low. Additionally, Efroni et al. (2015) observed low-level sporadic expression of known marker genes for cells of a different identity to the cell being tested, which could not be attributed to technical artefacts. These challenges mean that the expression of multiple cell type-specific transcripts must correlate strongly in order to be confident of cell identity. In contrast to Efroni et al.'s 2015 approach, this analysis has used an unsupervised approach to cluster cells and then assign identity to clusters, rather than using a supervised assignment of cell identity. This unsupervised approach has revealed that developmental gradients more strongly define cells' relationships with each other, and that despite cell fate being fixed after the first division from a stem cell, cell lineages are not distinct from each other. Instead other sources of variation such as cell cycle stage, ribosomal protein content and developmental gradients define cell differences.

Cell cycle phase is a strong source of variation within meristem cells (clusters 0, 2, 4, 5, 7, 8, 9, and 10 in Figure 5.14, coloured by cell cycle phase in Figure 5.20). In particular, cluster 7 which is distinctly separated from the other meristem cells is dominated by cells in G2M phases indicating that they are actively dividing. The documentation for the Seurat package (Butler et al., 2018) suggests that in some cases cell identity that has been masked by the effects of cell cycle heterogeneity can be recovered if the effects of cell cycle genes are regressed out of the dataset. However, in other cases and particularly in datasets containing a mix of stem and differentiated cells, regressing out the cell cycle can be detrimental to assigning cell identity (Butler et al., 2018). In this case Butler et al. suggest regressing out the difference between the G2M and S phase scores. In this analysis, cells have been assigned to a three cell cycle phases based on the expression of periodic genes (see Methods 2.4.10) but the effect of these genes has not been regressed out. One of the next steps in the analysis is to attempt to regress out cell cycle effects and in doing so perhaps reveal a different data structure to the meristem cells, hopefully defining more distinct clusters.

Cell identity could also be examined in the context of development by clustering subsets of cells from the same developmental stage. By removing or reducing that source of variance, cell identity may become clearer. If this is true, then the establishment of developmental gradients could also be related to cell type and the question of how different cell types contribute to developmental gradients could be tackled.

5.3.3 Developmental gradients in the root meristem

Rather than relying on cell clustering, a diffusion mapping approach could also be used to answer questions about how cell identity is established. By ordering cells into developmental 'pseudotime', the expression of genes can be modelled as identity is established. Wendrich et al. (2017) proposed that cells in the *Arabidopsis* root meristem gradually transition from stem cell to differentiated cells. Their study utilised three fluorescent marker genes which were expressed in different developmental gradients to perform expression studies of root meristems (but did not have single cell resolution). They observed that often genes were expressed in two opposing gradients, characterised as development and differentiation, and they related these gradual changes in expression to changes in protein accumulation and cellular properties. The presence of strong transcriptional gradients across development in our dataset is consistent with their observations. Wendrich et al.'s study relied on fluorescent markers to extract tissue according to gradient. This prevents the collection of cell type-specific information, and biases the experiment towards those cells expressing the markers (which was not consistent across the root cell lineages). Defining the root development based on gradients of marker expression is also sensitive to fluctuations in fluorescence. Drop-seq avoids these shortcomings as all root meristem cells are collected and the developmental gradients are established based on the whole transcriptome rather than a few markers.

Efroni et al. (2016)'s studies on root regeneration following QC excision suggests that young cells can change their identity if the root is damaged and reform a new meristem from partially differentiated cells. This ability is restricted to young meristem cells, suggesting as cells develop and begin elongating cell identity becomes permanently fixed. Since each initial cell type produces more than one differentiated cell type, identity cannot be established immediately, as the gene and protein networks that define identity are unlikely to be created instantaneously. However, whilst cell identity is not fixed at this point, Efroni et al. does observe a reduction of cell identity in stele cells (based on marker gene expression) in response to root tip removal, implying that final cell fate remains flexible even after cell identity has begun to be established.

5.3.4 Immunity in root meristems

Differential expression analysis in response to flg22 revealed DEGs in the meristem, root cap and early elongating cells. However, the small number of cells in the dataset limits the usefulness of DEG analysis.

Clusters containing small numbers of cells are likely to be more heterogenous than large clusters, as across a large number of cells the gene expression profile will be smoother. This higher degree of heterogeneity in small datasets decreases the resolution of DGE analysis. However, even in a larger dataset, the complete response would be impossible to detect using Drop-seq. Despite this limitation, the response of key genes can be investigated using this dataset. As the clustering of cells is developed, the resolution of flg22 DGE analysis is likely to improve. If this is the case, then the dataset could be used to investigate how flg22 responsive genes responds in different cell types. Rather than inspecting differentially expressed genes within clusters, differential expression could be examined in the context of development. For example, a pseudotime series analysis could be applied to cells which compares the expression profiles of flg22 treated cells across the diffusion map, against untreated cells. This could detect genes with divergent gene expression patterns in flg22 responsive cells, which may relate to development. This could either be performed by comparing cells within one diffusion map containing all data, or by comparing diffusion maps created separately for treated and untreated cells using newly published tools like 'cellAlign' which can compare expression dynamics within and between single-cell trajectories (Alpert et al., 2018).

This experiment could be expanded to incorporate different time points after flg22 treatment, in order to model the responses in the context of both time and developmental stage, revealing the key changes that induce the immune response and root growth inhibition. Furthermore increasing the range of developmental tissue tested could reveal which tissues have the strongest response to flg22. Similarly, stronger treatments such as Pep1, that have a more dramatic effect on root growth could be investigating using Drop-seq.

Chapter 6

Discussion and Future Work

6.1 Conclusions

Plant roots are comprised of numerous cell types, which work in a concerted manner to maintain homeostasis, growth and respond to the environment. Various studies of abiotic stress in *Arabidopsis* roots revealed that different cell types are characterised by cell type-specific gene networks. These networks are managed at the transcriptional level by combinations of transcription factors working together to achieve a high degree of specificity. The work in this thesis reveals that immune signalling is also characterised by cell type-specificity. As was observed in multiple abiotic stress studies (Iyer-Pascuzzi et al., 2011), these specific transcriptomic responses are strongly aligned to cell identity. This is revealed by the significant overlap between immune-responsive and identity genes, and the enrichment of promoter elements between these gene sets. The development of PMET as a method to investigate enriched promoter elements in cell type-specific gene networks proved to be very effective. The tool identified cell type-specific motif-pairs in each gene set tested, revealing that immunity networks often rely on the pairing of a known developmental transcription factor and a transcription factor associated with stress response in order to elicit a cell type-specific response. Overall these experiments revealed a highly complex immune landscape promoted by multiple cell types within the mature root.

By contrast, single cell RNA-seq analysis of the *Arabidopsis* root meristem revealed that cell types were less strongly defined at the transcriptional level during early development, despite cell identity (based on cell files) being set after the first division (Dolan et al., 1993). The lack of clustering by cell type in the Drop-seq dataset suggests that meristem cell identity remains flexible, consistent with root regeneration studies (Efroni et al., 2016). DEG analysis of flg22 treatment at different developmental zones in the root tip was revealed a limited immune response, indicating that the meristem could be less responsive to flg22 compared to mature cells. However, given that many effects of flg22 treatment have been observed experimentally in the root meristem (Schäfer lab, unpublished), it is more likely that this limited response is due to the inherent low resolution of Drop-seq, and low capture rate of unbroken cells in this experiment.

This application of new single cell technology to the root meristem revealed the potential of single cell transcriptomics to examine root gene expression patterns across development. With further study, changes in gene expression gradients in response to a variety of treatments could be examined within the context of this developmental axis. Furthermore, these data could be expanded to incorporate additional developmental zones. In particular, by harvesting the elongation zone and mature zones specifically, the data could be aligned to the root meristem data enabling the examination of gene expression gradients across multiple developmental stages.

6.2 Root immunity signalling within wider plant physiology

The use of root cell type-specific transcriptomics to investigate the immune response to two elicitors, bacteria-derived flg22 and endogenous Pep1 revealed differential responses between cells and between treatments. This highlights a complex immune landscape within the root where each cell contributes differently to elicitor treatment. Pep1 induced a very strong DEG response, encompassing the broad up-regulation of immunity genes, and the down-regulation of growth associated genes. In contrast, flg22 induced fewer genes than Pep1, and suppressed very few genes. The almost complete overlap of the flg22 response with the Pep1 response implies that the flg22 responsive gene networks are a subset of the Pep1 networks.

As an endogenous signal, Pep1 signalling via PEPR is activated in response to damage. This could be the result of a number of different pathogen attacks, such as bacterial or fungal infiltration or herbivory by root resident organisms. In order to successfully mount a defence against such a wide variety of attacks, logically Pep1 signalling must encompass some aspects of other elicitor response networks. Furthermore, as an endogenous signal, Pep1 could represent a strong “alarm signal” (Poncini et al., 2017) that the root has undergone damage, whereas flg22 perception is the earliest response to bacterial attack. This is consistent with Yamaguchi and Huffaker (2011)'s hypothesis that “Pep peptides are secreted to amplify defence responses initiated by pathogen-associated molecular patterns (PAMPs)”. The amplification of PAMP-triggered defence responses by Pep1 signalling would require strong overlap of signalling, and therefore DEGs. Pep1 has been shown to help protect the root from multiple threats including increasing host resistance to bacterial or fungal pathogens (necrotrophic and biotrophic) and offering some protection against herbivores (Huffaker et al., 2011, 2013; Tintor et al., 2013). The extent to which this is organised by overlapping gene networks (as was observed for flg22) needs to be investigated through similar gene expression experiments examining the reponse to a wider variety of PAMPs such as chitin, lipopolysaccharides and elongation factor Tu (EF-Tu).

GO term analyses also revealed disparate functions in cell type-specifically expressed genes. Analysis of DEGs uniquely expressed in each cell type following elicitor treatment revealed that each cell type might be specialising in a different function, in addition to a large shared gene response with other cells. The epidermis-specifically enriched genes were characterised by up-regulated canonical immune responses, and down-regulated growth terms. The specifically expressed genes in the cortex and pericycle were more associated with metabolism and peptide transport.

The use of GO terms to assign function to gene sets is inherently limited by previous experiments (Lewis, 2017). Genes are assign to GO terms based on the observed expression in previous gene expression studies. In plant science, there is a considerable bias towards leaf studies over root studies, and therefore key components of the root immune machinery may not be annotated in GO terms. Furthermore, much of the root literature is based on whole root studies. If a key immunity gene is cell type-specifically expressed, then the signal of that gene may not be detected in whole root studies, and therefore would not have been assigned to a GO term. This would be particularly likely for TFs, as these are typically expressed at low levels. One of the main contributions of this study will be to enhance the annotation of root-specifically expressed genes to immune associated GO terms. However, the continual expansion of GO terms creates additional problems. As more genes are associated to the wider array of GO terms, the number of terms enriched in each gene set increases such that it is harder to decipher which terms are relevant, and which ones are likely false-positives (Lewis, 2017). These difficulties highlight the importance of utilising multiple analysis methods to aid the interpretation of GO terms, including the specific analysis of potential regulatory mechanisms, using tools such as PMET, or additional experiments such as ATAC-seq or ChIP-seq to identify open chromatin and potential TF binding (Gligorijević and Pržulj, 2015).

Furthermore, the context of the differential gene expression must be considered, in terms of whether the enriched GO terms make sense spatially and developmentally. As more single cell studies are performed, the Gene Ontology could be enriched or combined with spatial and developmental information in order to increase the specificity in GO enrichment analyses.

The strong and specific up-regulation of immunity terms in epidermis specifically expressed genes implies that the epidermis responded more strongly to the immune trigger than the cortex and pericycle. There are several potential explanations for this observation. Firstly, it could simply be a dose effect, in that the epidermis will have had slightly longer or more substantial exposure to the elicitors. Under this hypothesis, the inner cells would only contact elicitor peptides that had diffused through the outer layer, or had penetrated via natural "wounds" such as epidermal cracks at the site of emerging lateral roots or at the elongation zone. Secondly, the stronger immune response in the epidermis could be due to bias in GO terms (as described above), or thirdly it could be a reflection of innate differences in sensitivity to PAMPs between cell types. Wyrsch et al. (2015) observed that all root cell types were able to respond swiftly (within 10 minutes) in isolation to flg22 elicitation. Furthermore, Wyrsch et al. (2015) observed that "the intensity of the immune responses did not always correlate with the expression level of the FLS2 receptor, but depended on the expressing tissue". This observation supports the idea that PAMP perception and sensitivity varies between different cell types, and that our observed differences are less likely to be the result of a dose response.

Prior to Wyrsch et al.'s paper, it was speculated that the epidermis would be less responsive than inner tissues. Intuitively, a highly active immune response in the epidermis would lead to constant activation of PTI following exposure to microbes in the soil, leading to negative effects such as growth inhibition (Heil, 2002; Heil and Baldwin, 2002). It would also inhibit potential beneficial interactions with mutualists (Faulkner and Robatzek, 2012). However this hypothesis is inconsistent with the highly active epidermis immune response reported in this thesis and observed by Wyrsch et al. One potential explanation for this discrepancy could be the sterility of the experimental growth system. In all the experiments in this thesis, and within the cited literature, plants were grown in a sterile environment. The first exposure in the lab-grown plants life cycle to a biotic stimulus is upon the treatment with an elicitor. By contrast, under normal growth conditions roots are constantly exposed to low, non-threatening levels of microbes throughout development, which could result in the attenuation of epidermis immune responses. A compelling hypothesis is that under normal growth conditions, the epidermis only reacts to a certain threshold of elicitation, or is activated primarily through cell-cell communication, rather than receptor activation.

This study showed significant cell type-specificity in three root cell types. Production of a comprehensive assessment of cell type-specificity in root immune responses requires further study. It would be particularly interesting to repeat the experiment with additional cell types, such as the endodermis, which represents the barrier between the outer tissues and the vasculature, as FLS2 is most strongly expressed in stele tissue under mock conditions (Beck et al., 2014; Poncini et al., 2017). It would also be ideal to repeat the pericycle experiment, in order to determine whether the limited response that was observed is a biological feature (contradicting Wyrsch et al.'s findings), or an artefact of the experiment resulting from high levels of noise between replicates.

Throughout the PMET analyses, a striking pattern occurred in the pairing of promoter motifs for known stress-regulatory TF families with motifs for developmental TFs in each cell type-specific gene network. Moreover, certain TF combinations prevailed in specific cell types in a treatment-specific manner. For instance, WRKY TFs might have a more prominent function in regulating epidermis-specific networks together with specific developmental TFs, with WRKY/MYC and WRKY/ATHB paired motifs regulating cell identity networks and KAN/WRKY and ANAC/WRKY paired motifs regulating cell immunity networks. In turn, cortex function relies on MYC, ATHB, and PIF TFs as central regulators that combine with WRKY TFs to regulate cortex-specific immunity and with AHLs to regulate cortex identity-specific gene networks. The observation of development TFs involvement in cell type-specific immune networks is consistent with similar cell type-specific responses to abiotic stresses. In a meta-analysis of four abiotic stresses, Iyer-Pascuzzi et al. (2011) observed that whilst there was no universal abiotic stress response, there were core stress response (CSR) genes within each cell type. These CSRs included known developmental TFs such as *LONESOME HIGHWAY*, *SCHIZORIA* and interactors with developmental TFs such as SCR. This suggests that all plant root defence responses are strongly linked to identity and development, and that the interlinking of stress and identity or development transcription factors is the mechanism by which robust cell type-specific networks are managed.

The PMET sensitivity study in Chapter 3 revealed unexpected differences between up- and down-regulated immune genes which prompted the thought that the PMET results might be revealing an underlying structure that defines the difference between up- and down-regulation in these gene sets. It is known that different plant TFs, (and likely in other systems too) have positional preferences relative to the TSS (Yu et al., 2016), and the pattern observed in our dataset is consistent with their observed results. The up-regulated TFs dominated by WRKYs and AT-hook containing TFs such as AHLs bound motifs preferentially positioned closer to the TSS, whereas the down-regulated motifs, dominated by MYB and ATHB TFs were shown to be preferentially expressed further away (Yu et al., 2016). Positional preference TF binding positions suggests that the switch between up and down could be regulated by structural changes in the chromatin rather than a model of direct competition between TFs.

6.3 Relevance of root immune signalling in the context of the plant

Three of the characteristic responses to immunity are MITOGEN-ACTIVATED PROTEIN KINASE (MAPK) phosphorylation, a ROS burst and the induction of defence genes such as *PLANT DEFENSIN 1.2 (PDF1.2).* These responses are observed both above ground and below ground. Both the roots and leaves have been shown to respond to bacterial flg22, fungal chitin and endogenous Pep1, but the roots are insensitive to bacterial elf18 (Ranf et al., 2011). This suggests that PRRs are organised on an organ specific spatio-temporal basis, likely to reflect the likelihood of encountering said PAMP in each organ. Furthermore both leaves and roots display organ-specific immune responses. Roots increase camalexin production and exudation (Millet et al., 2010), and leaves activate Ca^{2+} and ROS signalling through chloroplasts. Finally, the distinct natures of leaves and root immunity is highlighted by the fact that roots can activate an immune response in isolation (Wyrsch et al., 2015). These observations show that the roots immune response is at least partially distinct from that of the leaves.

Within leaves, an emerging pattern is the global, massive over-representation of nuclear encoded chloroplast genes (Nomura et al., 2012; de Torres Zabala et al., 2015). The chloroplast is a primary site for the production of immune signals such as ROS and calcium, and innate immune signalling pathways branch via chloroplasts. They also play a vital role in providing photosynthesis-derived carbon sources and energy needed for defence including the biosynthesis of several defence hormone precursors (reviewed in Stael et al. (2015)). Approximately 50% of ROS are produced in chloroplasts, the other half are produced by apoplastic peroxidases (O'Brien et al., 2012). Given that the chloroplast is a key part of the immune response in leaves, it is particularly interesting to investigate root immune signalling to discover how the defence response is enacted in the absence of chloroplast-mediated signalling.

Whilst a large portion of immune signalling occurs via the chloroplasts, there are immune signalling pathways in the leaves that bypass the chloroplast including MAPK signalling cascades, and CDPK signalling (regulated by apoplastic Ca^{2+}).

The Ca^{2+} burst following PAMP perception in the root has been described as resembling a low concentration flg22 response in the leaves (Ranf et al., 2011). This was suggested to be a reflection of lower expression of *FLAGELLIN SENSING 2 (FLS2)*, but could also be explained by the lack of chloroplasts, and therefore a smaller source of Ca^{2+} in the roots. The up-regulation of multiple *MAPKs*, MAP kinase kinase kinases and leucine-rich repeat receptor kinases (LRR-RKs) in roots (Chapter 3) implies that immune signalling utilises these non-chloroplastic signalling mechanisms (rather than having an analogous signalling method via a different organelle).

Whilst the root can elicit an immune response in isolation, it does not generally act in isolation and the long range transport of key defence hormones has been shown between leaves and roots (reviewed in Park et al. (2017)), and calcium has been shown to signal from root to shoot (Choi et al., 2014). As such the most complete model of immune signalling considers that both leaves and roots are supported by the other, through the transport and storage of energy and metabolites. Multi-level regulation from system to cell type-specific or even single cell level is advantageous as it results in resilient networks. Whilst these networks may require a higher degree of coordination through numerous regulatory and signalling proteins), maintaining these networks adds to the robustness and flexibility required for a plant to adapt to constantly changing environmental stimuli and conditions.

6.4 The application of single cell technologies to plant studies

The experiment in Chapter 5 demonstrates the feasibility of using Drop-seq to perform high-throughput single cell sequencing on plant roots. The experiment showed that some cells types and developmental stages can be identified even from small relatively small numbers of captured cells. However, this chapter also shows that there are significant challenges, some of which are plant-specific, that have to be overcome in order to achieve sufficiently high resolution to tackle clear biological questions. In addition to general improvements to single cell studies, the specific expertise of single cell studies in plants requires further refinement. In particular, improved methods to process plant cells without damage and thus reducing the number of broken cells will render vast improvements in the cell capture rate in

Drop-seq.

With further enhancement of single-cell technologies, and more expertise in processing plant tissue, more studies will be able to compare gene expression between cell types without relying on fluorescent markers (as were used in Chapter 4). Unlike in FACS-based studies, in a single cell approach all cells come from the same batch of plants. One of the shortcomings of FACS-based studies is that each cell type is produced by a different marker line, therefore the cells are captured from different sets of plants. In this situation, it is easy to introduce batch effects between cell types that may not be apparent from the analysis. By extracting all cells from the same batch of plants, the chance of batch effects is reduced, which increases the cohesion and integrity of the cell type gene expression patterns. However, these studies are reliant on harvesting sufficient tissue such that all cell types are sufficiently represented in the final dataset (accounting for low cell capture, and broken cells). In particular, performing differential gene expression on single cells require a vast number of cells in order for the significance testing to be sufficiently robust.

Drop-seq was the first single cell technology to use microfluidics combined with a barcoding system to enable parallel, high-throughput sequencing of single cell transcriptomes (Macosko et al., 2015). This ground-breaking new approach transformed single cell analysis. However, as the first iteration of a new technology, Drop-seq has a large number of drawbacks, primarily the low capture rate. This is a particular problem when it comes to processing small organs such as root meristems, as it is labour intensive to harvest enough tissue even to capture a small number of viable cells. The other major limitation of Drop-seq (and single cell transcriptomic technologies in general) is the low resolution of transcriptomes that are produced. For Drop-seq, the current detection level is limited to, on average, 10,000 reads per cell (Zhang et al., 2018). This means that only the most highly expressed genes are detected, which particularly limits the resolution of differential gene expression analysis.

There are two, recently developed, alternative systems for single-cell transcriptomes using microfluidics; inDrop (Klein et al., 2015; Zilionis et al., 2017) and 10X Chromium (Zheng et al., 2017). These recently developed alternatives have yielded significant improvements in cell capture, inDrop which captures 80% of cells (at a significant cost to mRNA capture) and the commercial 10X Chromium platform yields up to 65% capture efficiency (Zheng et al., 2017), with an improved sensitivity to gene detection (20000 reads from 4,000 genes on average per cell, Zhang et al. (2018)). This higher resolution is achieved at significantly higher cost compared to both Drop-seq and inDrop. However, the increased capture efficiency and gene detection sensitivity makes 10X the ideal platform for future studies where tissue is limited such as plant roots.

6.5 Outlook for cell type-specific transcriptomics in plants

The advent of next generation sequencing technology revolutionised cell and molecular biology. In a similar manner single cell technologies are in the process of changing the way that scientists investigate biological systems. In human biology, single cell approaches are being used to develop a human cell atlas (Regev et al., 2017) which aims to create an "ID card for each cell type" and a "3D map of how cell types form tissues". This vast dataset will be used to find out how changes in this map underlie health and disease. By demonstrating that single cell RNA-seq can be applied to the plant system, this thesis paves the way for a similar scale project in plants. In addition to the study of meristematic tissue in this thesis, Drop-seq has also been used to identify distinct populations of mature root cell types (Shulse et al., 2018). The development of a plant cell atlas could be applied to understand how the interactions between cells influence function, how cells act within a complex system to create an organ and how cells react as a network to disease or changes in environmental conditions. A plant single cell atlas would supplement the existing plant atlases that integrate microarray data of whole plant (and some cell type-specific) studies (Petryszak et al., 2015; Waese et al., 2017). These methods enable the discovery of novel markers that can define smaller and smaller subgroups of cells. These novels markers can then be used in the molecular biology lab to isolate cell groups in a highly specific manner. Having identified specific markers, FACS and single cell could even be combined to investigate heterogeneity within cell types.

In addition to investigating gene expression at a single cell level, the future of molecular research and in particular the prediction and validation of high-resolution gene regulatory networks will be through the integration of multi-omics datasets. Brady et al. (2011) developed a stele-enriched gene regulatory network on a tissue level based on the integration of miRNA, protein-protein interactions and protein-DNA interactions with gene expression data. The continued development of single cell ATAC-seq (Buenrostro et al., 2015) and ChIP-seq (Rotem et al., 2015) will enable the prediction of single cell gene regulatory networks. However, the current high expense of these high-throughput methods maintains the need for *in silico* predictions and modelling using methods such as PMET, in order to direct experimentation more efficiently.

Along with other motif studies, PMET could be used to build a cell type-specific ontology of transcription factor combinatorics, which could be used to decipher gene regulation in a similar manner to how gene ontology studies are used to determine function. By recursively combining regulatory information with functional data about individual transcription factors and spatio-temporal gene expression patterns, the fundamental principles defining plant stress responses can be defined. Increased understanding of immunity and other stresses in model organisms such as *Arabidopsis* informs crop research and ultimately genome engineering to develop crops that are resistant to pathogens, or other environmental stresses.

Appendix A

List of ribosomal proteins

Table A.1: Table of ribosomal proteins removed from cell type-specific RNA-seq analysis

Gene ID	Gene Acronym	Description
AT1G01100	RPP1A	60S acidic ribosomal protein P1-1
AT1G01860	PFC1	rRNA adenine N(6)-methyltransferase
AT1G02780	RPL19A	60S ribosomal protein L19-1
AT1G02830	RPL22A	Putative 60S ribosomal protein L22-1
AT1G04270	RPS15A	40S ribosomal protein S15-1
AT1G04480	RPL23A	At2g33370
AT1G05190	RPL6	50S ribosomal protein L6, chloroplastic
AT1G06380		Ribosomal protein L1p/L10e family
AT1G07070	RPL35AA	60S ribosomal protein L35a-1
AT1G07320	RPL4	RPL4
AT1G07770	RPS15AA	AT1G07770 protein
AT1G08360	RPL10AA	60S ribosomal protein L10a-1
AT1G08845		Ribosomal L18p/L5e family protein
AT1G09590	RPL21A	60S ribosomal protein L21-1
AT1G09690	RPL21A	60S ribosomal protein L21-1
AT1G12220	RPS5	Disease resistance protein
AT1G12960	RPL27AA	Putative 60S ribosomal protein L27a-1
AT1G14205		Ribosomal L18p/L5e family protein
AT1G14320	RPL10A	60S ribosomal protein L10-1
AT1G15250	RPL37A	60S ribosomal protein L37-1
AT1G15930	RPS12A	40S ribosomal protein S12-1
AT1G16740		50S ribosomal protein L20
AT1G16870		Mitochondrial 28S ribosomal protein S29-like protein
AT1G17560	HLL	50S ribosomal protein HLL, mitochondrial
AT1G18540	RPL6A	60S ribosomal protein L6-1
AT1G22780	RPS18C	40S ribosomal protein S18
AT1G23290	RPL27AB	60S ribosomal protein L27a-2
AT1G23410	RPS27AA	Ubiquitin-40S ribosomal protein S27a-1
AT1G25260		Ribosome assembly factor mrt4
AT1G26880	RPL34A	60S ribosomal protein L34-1
AT1G26910	RPL10B	60S ribosomal protein L10-2

Continued on next page

Table A.1 – *Continued from previous page*

Gene ID	Gene Acronym	Description
AT1G27400	RPL17A	60S ribosomal protein L17-1
AT1G29040		50S ribosomal protein L34
AT1G29070	RPL34	50S ribosomal protein L34, chloroplastic
AT1G29965	RPL18AA	60S ribosomal protein L18a-1
AT1G29970	RPL18AA	60S ribosomal protein L18A-1
AT1G30230		Translation elongation factor EF1B/ribosomal protein S6 family protein
AT1G31817	NFD3	Probable ribosomal protein S11, mitochondrial
AT1G32990	RPL11	50S ribosomal protein L11, chloroplastic
AT1G33120	RPL9B	60S ribosomal protein L9-1
AT1G33140	RPL9B	60S ribosomal protein L9-1
AT1G33850		40S ribosomal protein S15
AT1G34030	RPS18C	40S ribosomal protein S18
AT1G35680	RPL21	50S ribosomal protein L21, chloroplastic
AT1G36240	RPL30A	Putative 60S ribosomal protein L30-1
AT1G41880	RPL35AB	60S ribosomal protein L35a-2
AT1G43170	ARP1	60S ribosomal protein L3-1
AT1G48350	RPL18	EMB3105
AT1G48830	RPS7A	40S ribosomal protein S7-1
AT1G52300	RPL37B	60S ribosomal protein L37-2
AT1G52930	BRIX1-2	Ribosome biogenesis protein BRX1 homolog 2
AT1G54217		Ribosomal protein L18ae family
AT1G54770		Fcf2 pre-rRNA processing protein
AT1G56045	RPL41C	60S ribosomal protein L41
AT1G57540		40S ribosomal protein
AT1G57660	RPL21E	60S ribosomal protein L21-2
AT1G57670		Disease resistance protein RPS4, putative
AT1G57860	RPL21E	60S ribosomal protein L21-2
AT1G58380	RPS2A	At1g59359
AT1G58684	RPS2B	40S ribosomal protein S2-2
AT1G58983	RPS2B	40S ribosomal protein S2-2
AT1G59359	RPS2B	40S ribosomal protein S2-2
AT1G61580	ARP2	60S ribosomal protein L3-2
AT1G64510		Translation elongation factor EF1B/ribosomal protein S6 family protein
AT1G66580	RPL10C	60S ribosomal protein L10-3
AT1G66890		50S ribosomal-like protein
AT1G67430	RPL17B	60S ribosomal protein L17-2
AT1G68590		30S ribosomal protein 3-1, chloroplastic
AT1G69620	RPL34B	60S ribosomal protein L34-2
AT1G70600	RPL27AC	60S ribosomal protein L27a-3
AT1G72270		CONTAINS InterPro DOMAIN/s: Ribosome 60S biogenesis N-terminal (InterPro:IPR021714);
AT1G72370	RPSAA	40S ribosomal protein SA
AT1G74050	RPL6C	60S ribosomal protein L6-3
AT1G74060	RPL6B	60S ribosomal protein L6-2
AT1G74270	RPL35AC	60S ribosomal protein L35a-3
AT1G74970	RPS9	30S ribosomal protein S9, chloroplastic
AT1G75350	RPL31	50S ribosomal protein L31, chloroplastic
AT1G77750	RPS13	Small ribosomal subunit protein S13, mitochondrial

Continued on next page

Table A.1 – *Continued from previous page*

Gene ID	**Gene Acronym**	**Description**
AT1G77940	RPL30B	60S ribosomal protein L30-2
AT1G78630	RPL13	50S ribosomal protein L13, chloroplastic
AT1G79850	RPS17	30S ribosomal protein S17, chloroplastic
AT1G80750	RPL7A	60S ribosomal protein L7-1
AT2G01250	RPL7B	60S ribosomal protein L7-2
AT2G03130		50S ribosomal protein L7/L12
AT2G03810		18S pre-ribosomal assembly protein gar2-like protein
AT2G04390	RPS17A	40S ribosomal protein S17-1
AT2G05220	RPS17B	40S ribosomal protein S17-2
AT2G07675		Ribosomal protein S12/S23 family protein
AT2G07696	RPS7	At2g07696
AT2G07734		Alpha-L RNA-binding motif/Ribosomal protein S4 family protein
AT2G09990	RPS16A	40S ribosomal protein S16-1
AT2G16360		Ribosomal protein S25 family protein
AT2G16930		50S ribosomal protein L27
AT2G17360	RPS4A	40S ribosomal protein S4-1
AT2G18020	RPL8A	EMB2296
AT2G18400		Putative ribosomal protein L6
AT2G19720	RPS15AB	40S ribosomal protein S15a-2
AT2G19730	RPL28A	60S ribosomal protein L28-1
AT2G19740	RPL31A	60S ribosomal protein L31-1
AT2G19750	RPS30A	40S ribosomal protein S30
AT2G20060		50S ribosomal protein L4
AT2G20450	RPL14A	60S ribosomal protein L14-1
AT2G21290		30S ribosomal protein S31, mitochondrial
AT2G21580	RPS25B	40S ribosomal protein S25-2
AT2G24090	RPL35	50S ribosomal protein L35, chloroplastic
AT2G25210		Ribosomal protein L39 family protein
AT2G27530	RPL10AB	60S ribosomal protein L10a-2
AT2G27720		60S acidic ribosomal protein family
AT2G28815		60S ribosomal protein L16-like, mitochondrial
AT2G31610	RPS3A	40S ribosomal protein S3-1
AT2G32060	RPS12C	40S ribosomal protein S12
AT2G32220	RPL27A	60S ribosomal protein L27-1
AT2G33370	RPL23A	At2g33370
AT2G33450	RPL28	PRPL28
AT2G33800	RPS5	30S ribosomal protein S5, chloroplastic
AT2G34480		Ribosomal protein L18ae/LX family protein
AT2G34520	RPS14	At2g34520
AT2G36160	RPS14A	40S ribosomal protein S14-1
AT2G36170	RPL40B	Ubiquitin-60S ribosomal protein L40-1
AT2G36620	RPL24A	60S ribosomal protein L24-1
AT2G37190	RPL12A	60S ribosomal protein L12-1
AT2G37270	RPS5A	40S ribosomal protein S5-1
AT2G37600	RPL36A	60S ribosomal protein L36-1
AT2G37990		Ribosome biogenesis regulatory protein homolog
AT2G38140	RPS31	30S ribosomal protein S31, chloroplastic
AT2G39140	SVR1	Putative ribosomal large subunit pseudouridine synthase SVR1, chloroplastic

Continued on next page

Table A.1 – *Continued from previous page*

Gene ID	**Gene Acronym**	**Description**
AT2G39390	RPL35B	60S ribosomal protein L35-2
AT2G39460	RPL23AA	60S ribosomal protein L23a-1
AT2G39590	RPS15AC	40S ribosomal protein S15a-3
AT2G40010	RPP0A	60S acidic ribosomal protein P0-1
AT2G40205	RPL41C	60S ribosomal protein L41
AT2G40360	BOP1	Ribosome biogenesis protein BOP1 homolog
AT2G40510	RPS26B	40S ribosomal protein S26
AT2G40590	RPS26A	40S ribosomal protein S26
AT2G41840	RPS2C	40S ribosomal protein S2-3
AT2G42740	RPL11A	60S ribosomal protein L11-1
AT2G43030	RPL3A	50S ribosomal protein L3-1, chloroplastic
AT2G43460	RPL38B	60S ribosomal protein L38
AT2G44120		Ribosomal protein L30/L7 family protein
AT2G44860		Probable ribosome biogenesis protein RLP24
AT2G45710	RPS27A	40S ribosomal protein S27-1
AT2G47110	RPS27AB	UBQ6
AT2G47420	DIM1A	rRNA adenine N(6)-methyltransferase
AT2G47570		Ribosomal protein L18e/L15 superfamily protein
AT2G47610	RPL7AA	60S ribosomal protein L7a-1
AT3G01160		Pre-rRNA-processing ESF1-like protein
AT3G02080	RPS19A	40S ribosomal protein S19-1
AT3G02190	RPL39B	60S ribosomal protein L39-2
AT3G02560	RPS7B	40S ribosomal protein S7-2
AT3G03600	RPS2	RPS2
AT3G04230	RPS16B	40S ribosomal protein S16-2
AT3G04400	RPL23A	At2g33370
AT3G04770	RPSAb	40S ribosomal protein SA
AT3G04840	RPS3AA	40S ribosomal protein S3a-1
AT3G04920	RPS24A	40S ribosomal protein S24-1
AT3G05560	RPL22B	60S ribosomal protein L22-2
AT3G05590	RPL18B	RPL18
AT3G06680		60S ribosomal protein L29
AT3G06700	RPL29A	60S ribosomal protein L29-1
AT3G07110		Ribosomal protein L13 family protein
AT3G08520	RPL41C	60S ribosomal protein L41
AT3G09200	RPP0B	60S acidic ribosomal protein P0-2
AT3G09500	RPL35A	60S ribosomal protein L35-1
AT3G09630	RPL4A	60S ribosomal protein L4-1
AT3G09680	RPS23A	40S ribosomal protein S23-1
AT3G10090	RPS28A	40S ribosomal protein S28-1
AT3G10610	RPS17C	40S ribosomal protein S17-3
AT3G10950	RPL37AB	Putative 60S ribosomal protein L37a-1
AT3G11120	RPL41C	60S ribosomal protein L41
AT3G11250	RPP0C	60S acidic ribosomal protein P0-3
AT3G11510	RPS14B	At3g11510
AT3G11940	RPS5B	40S ribosomal protein S5-2
AT3G11964	RRP5	rRNA biogenesis protein RRP5
AT3G12370		50S ribosomal protein L10
AT3G12915		Ribosomal protein S5/Elongation factor G/III/V family protein

Continued on next page

Table A.1 – *Continued from previous page*

Gene ID	**Gene Acronym**	**Description**
AT3G13120	RPS10	30S ribosomal protein S10, chloroplastic
AT3G13580	RPL7D	Ribosomal protein L30/L7 family protein
AT3G13882		Ribosomal protein L34
AT3G14600	RPL18AC	60S ribosomal protein L18a-3
AT3G15190	RPS20	PRPS20
AT3G15460	BRIX1-1	Ribosome biogenesis protein BRX1 homolog 1
AT3G16080	RPL37C	60S ribosomal protein L37-3
AT3G16780	RPL19B	60S ribosomal protein L19-2
AT3G17465	RPL3B	50S ribosomal protein L3-2, chloroplastic
AT3G17626		Structural constituent of ribosome
AT3G18740	RPL30C	60S ribosomal protein L30-3
AT3G18880		40S ribosomal protein S17-like
AT3G19800		Large ribosomal RNA subunit accumulation protein YCED homolog 2, chloroplastic
AT3G19810		Large ribosomal RNA subunit accumulation protein YCED homolog 1, chloroplastic
AT3G22230	RPL27B	60S ribosomal protein L27-2
AT3G22300	RPS10	40S ribosomal protein S10, mitochondrial
AT3G22660	EBP2	Probable rRNA-processing protein EBP2 homolog
AT3G22980		Ribosomal protein S5/Elongation factor G/III/V family protein
AT3G23390	RPL36AB	60S ribosomal protein L36a
AT3G23620		Ribosome production factor 2 homolog
AT3G24830	RPL13AB	60S ribosomal protein L13a-2
AT3G25520	ATL5	RPL5A
AT3G25920	RPL15	50S ribosomal protein L15, chloroplastic
AT3G26360		Ribosomal protein S21 family protein
AT3G27160	GHS1	Ribosomal protein S21 family protein
AT3G27830	RPL12A	RPL12-A
AT3G27840	RPL12B	50S ribosomal protein L12-2, chloroplastic
AT3G27850	RPL12C	50S ribosomal protein L12-3, chloroplastic
AT3G28500	RPP2C	60S acidic ribosomal protein P2-3
AT3G28900	RPL34C	60S ribosomal protein L34-3
AT3G43980	RPS29C	40S ribosomal protein S29
AT3G44010	RPS29C	Ribosomal protein S14p/S29e family protein
AT3G44590	RPP2D	Acidic ribosomal protein P2-like
AT3G44890	RPL9	RPL9
AT3G45030	RPS20A	40S ribosomal protein S20-1
AT3G46040	RPS15AD	RPS15AD
AT3G47370	RPS20B	40S ribosomal protein S20-2
AT3G48930	RPS11A	40S ribosomal protein S11-1
AT3G48960	RPL13C	60S ribosomal protein L13
AT3G49010	RPL13B	60S ribosomal protein L13
AT3G49460		60S acidic ribosomal protein-like protein
AT3G49910	RPL26A	60S ribosomal protein L26-1
AT3G51190	RPL8B	60S ribosomal protein L8-2
AT3G52580	RPS14C	40S ribosomal protein S14-3
AT3G52590	RPL40B	Ubiquitin-60S ribosomal protein L40-1
AT3G53020	RPL24B	60S ribosomal protein L24-2
AT3G53430	RPL12B	60S ribosomal Protein L12-like

Continued on next page

Table A.1 – *Continued from previous page*

Gene ID	Gene Acronym	Description
AT3G53740	RPL36B	60S ribosomal protein L36
AT3G53870	RPS3B	40S ribosomal protein S3-2
AT3G53890	RPS21B	40S ribosomal protein S21-1
AT3G54210	RPL17	50S ribosomal protein L17, chloroplastic
AT3G55170	RPL35C	60S ribosomal protein L35-3
AT3G55280	RPL23AB	60S ribosomal protein L23a-2
AT3G55750	RPL35AD	60S ribosomal protein L35a-4
AT3G56020	RPL41C	60S ribosomal protein L41
AT3G56340	RPS26C	40S ribosomal protein S26
AT3G56910	PSRP5	50S ribosomal protein 5, chloroplastic
AT3G57490	RPS2D	40S ribosomal protein S2-4
AT3G58700	RPL11C	60S ribosomal protein L11-2
AT3G59540	RPL38B	60S ribosomal protein L38
AT3G59650		Mitochondrial ribosomal protein L51/S25/CI-B8 family protein
AT3G60245	RPL37AC	60S ribosomal protein L37a-2
AT3G60770	RPS13A	40S ribosomal protein S13-1
AT3G61110	RPS27B	40S ribosomal protein S27
AT3G61111		40S ribosomal protein S27
AT3G62250	RPS27AC	Ubiquitin-40S ribosomal protein S27a-3
AT3G62870	RPL7AB	60S ribosomal protein L7a-2
AT3G63190	RRF	Ribosome-recycling factor, chloroplastic
AT3G63490	RPL1	50S ribosomal protein L1, chloroplastic
AT4G00100	RPS13B	40S ribosomal protein S13-2
AT4G01310	RPL5	50S ribosomal protein L5, chloroplastic
AT4G02230	RPL19C	60S ribosomal protein L19-3
AT4G09012		Mitochondrial ribosomal protein L27
AT4G09800	RPS18C	40S ribosomal protein S18
AT4G10450	RPL9D	60S ribosomal protein L9-2
AT4G12600		Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein
AT4G13170	RPL13AC	60S ribosomal protein L13a-3
AT4G14245		Structural constituent of ribosome protein
AT4G14250		structural constituent of ribosome
AT4G14320		Zinc-binding ribosomal protein family protein
AT4G15000	RPL27C	60S ribosomal protein L27
AT4G15770		60S ribosome subunit biogenesis protein NIP7 homolog
AT4G16030		Probable ribosomal protein
AT4G16720	RPL15A	Ribosomal protein L15
AT4G17390	RPL15B	60S ribosomal protein L15-2
AT4G17560		50S ribosomal protein L19-1, chloroplastic
AT4G17610		tRNA/rRNA methyltransferase (SpoU) family protein
AT4G18100	RPL32A	60S ribosomal protein L32-1
AT4G18730	RPL11C	60S ribosomal protein L11-2
AT4G21460		Ribosomal protein S24/S35, mitochondrial
AT4G23620		Ribosomal protein L25/Gln-tRNA synthetase, anti-codon-binding domain-containing protein
AT4G25730		Putative rRNA methyltransferase
AT4G25740	RPS10A	40S ribosomal protein S10-1
AT4G25890	RPP3A	60S acidic ribosomal protein P3-1
AT4G26090	RPS2	Disease resistance protein RPS2

Continued on next page

Table A.1 – *Continued from previous page*

Gene ID	Gene Acronym	Description
AT4G26230	RPL31B	Putative ribosomal protein
AT4G27010		CONTAINS InterPro DOMAIN/s: Ribosome 60S biogenesis N-terminal (InterPro:IPR021714);
AT4G27090	RPL14B	AT4G27090 protein
AT4G29390	RPS30A	40S ribosomal protein S30
AT4G29410	RPL28C	60S ribosomal protein L28-2
AT4G29430	RPS15AE	40S ribosomal protein S15a-5
AT4G30150		CONTAINS InterPro DOMAIN/s: Nucleolar 27S pre-rRNA processing, Urb2/Npa2
AT4G30800	RPS11B	40S ribosomal protein S11-2
AT4G30930	RPL21M	NFD1
AT4G31700	RPS6A	40S ribosomal protein S6-1
AT4G31985	RPL39C	60S ribosomal protein L39-1
AT4G33865	RPS29C	40S ribosomal protein S29
AT4G34555	RPS25D	40S ribosomal protein S25-3
AT4G34620	RPS16-1	SSR16
AT4G34670	RPS3AB	40S ribosomal protein S3a
AT4G34730		Probable ribosome-binding factor A, chloroplastic
AT4G35490	MRPL11	At4g35490
AT4G36130	RPL8C	60S ribosomal protein L8-3
AT4G39200	RPS25E	40S ribosomal protein S25-4
AT5G02440		60S ribosomal protein L36
AT5G02450	RPL36C	60S ribosomal protein L36-3
AT5G02610		Ribosomal L29 family protein
AT5G02870	RPL4D	60S ribosomal protein L4-2
AT5G02960	RPS23B	40S ribosomal protein S23-2
AT5G03850	RPS28A	40S ribosomal protein S28-1
AT5G04800	RPS17D	40S ribosomal protein S17-4
AT5G07090	RPS4B	40S ribosomal protein S4-2
AT5G09490	RPS15B	40S ribosomal protein S15-2
AT5G09500	RPS15C	40S ribosomal protein S15-3
AT5G09510	RPS15D	40S ribosomal protein S15-4
AT5G10070		Probable ribosome biogenesis protein At5g10070
AT5G10360	RPS6B	40S ribosomal protein S6
AT5G11750		Ribosomal protein L19 family protein
AT5G13510	RPL10	50S ribosomal protein L10, chloroplastic
AT5G14290		Mitochondrial ribosomal protein L37
AT5G14320	RPS13	30S ribosomal protein S13, chloroplastic
AT5G15200	RPS9B	40S ribosomal protein S9-1
AT5G15390		rRNA methylase-like protein
AT5G15520	RPS19B	40S ribosomal protein S19-2
AT5G15550	WDR12	Ribosome biogenesis protein WDR12 homolog
AT5G15750		Alpha-L RNA-binding motif/Ribosomal protein S4 family protein
AT5G16130	RPS7C	40S ribosomal protein S7-3
AT5G16200		50S ribosomal protein-like protein
AT5G17870	PSRP6	plastid-specific 50S ribosomal protein 6
AT5G18380	RPS16C	40S ribosomal protein S16-3
AT5G19025		Ribosomal protein L34e superfamily protein

Continued on next page

Table A.1 – *Continued from previous page*

Gene ID	Gene Acronym	Description
AT5G19720		Ribosomal protein L25/Gln-tRNA synthetase, anti-codon-binding domain-containing protein
AT5G20160		Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein
AT5G20180		Ribosomal protein
AT5G20290	RPS8A	40S ribosomal protein S8
AT5G22440	RPL10AC	60S ribosomal protein L10a-3
AT5G23535		50S ribosomal protein L24
AT5G23740	RPS11C	RPS11-BETA
AT5G23900	RPL13D	60S ribosomal protein L13
AT5G24490		30S ribosomal protein
AT5G24510		60S acidic ribosomal protein family
AT5G27700	RPS21C	40S ribosomal protein S21
AT5G27770	RPL22C	60S ribosomal protein L22-3
AT5G27850	RPL18C	60S ribosomal protein L18-3
AT5G28060	RPS24B	40S ribosomal protein S24-2
AT5G30495		Fcf2 pre-rRNA processing protein
AT5G30510	RPS1	30S ribosomal protein S1, chloroplastic
AT5G35530	RPS3C	40S ribosomal protein S3-3
AT5G39600		39S ribosomal protein
AT5G39740	RPL5B	RPL5B
AT5G39800		Mitochondrial ribosomal protein L27
AT5G39850	RPS9C	40S ribosomal protein S9-2
AT5G40040	RPP2E	60S acidic ribosomal protein P2-5
AT5G40950	RPL27	50S ribosomal protein L27, chloroplastic
AT5G41520	RPS10B	40S ribosomal protein S10-2
AT5G43640	RPS15E	40S ribosomal protein S15-5
AT5G44710		37S ribosomal protein S27
AT5G45250	RPS4	Disease resistance protein RPS4
AT5G45775	RPL11C	60S ribosomal protein L11-2
AT5G46160	HLP	50S ribosomal protein HLP, mitochondrial
AT5G46420		16S rRNA processing protein RimM family
AT5G46430	RPL32B	60S ribosomal protein L32-2
AT5G46470	RPS6	Disease resistance protein RPS6
AT5G47190		50S ribosomal protein L19-2, chloroplastic
AT5G47320	RPS19	40S ribosomal protein S19, mitochondrial
AT5G47700	RPP1C	60S acidic ribosomal protein P1-3
AT5G47930	RPS27D	40S ribosomal protein S27-3
AT5G47940		40S ribosomal protein S27
AT5G48760	RPL13AD	60S ribosomal protein L13a-4
AT5G51610		50S ribosomal protein L11-like
AT5G52370		28S ribosomal S34 protein
AT5G52490	FIB3	Putative rRNA 2'-O-methyltransferase fibrillarin 3
AT5G52650	RPS10C	40S ribosomal protein S10-3
AT5G53920		Ribosomal protein L11 methyltransferase-like protein
AT5G54600	RPL24	SVR8
AT5G56670	RPS30A	40S ribosomal protein S30
AT5G56710	RPL31C	60S ribosomal protein L31-3
AT5G56940	RPS16-2	30S ribosomal protein S16-2, chloroplastic/mitochondrial
AT5G57060		60S ribosomal L18a-like protein
AT5G57290	RPP3B	60S acidic ribosomal protein P3-2

Continued on next page

Table A.1 – *Continued from previous page*

Gene ID	**Gene Acronym**	**Description**
AT5G58420	RPS4D	40S ribosomal protein S4-3
AT5G58990		28S ribosomal S34 protein
AT5G59240	RPS8B	40S ribosomal protein S8
AT5G59850	RPS15AA	AT1G07770 protein
AT5G60670	RPL12C	60S ribosomal protein L12-3
AT5G61170	RPS19C	40S ribosomal protein S19-3
AT5G61330		rRNA processing protein-related
AT5G62300	RPS20A	40S ribosomal protein S20-1
AT5G63070	RPS15F	40S ribosomal protein S15-6
AT5G64140	RPS28C	RPS28
AT5G65220	RPL29	AT5G65220 protein
AT5G66360	DIM1B	Ribosomal RNA small subunit methyltransferase, mitochondrial
AT5G67510	RPL26B	60S ribosomal protein L26-2
ATMG00080	RPL16	Rpl16
ATMG00210	RPL5	At2g07725
ATMG00290	RPS4	mitochondrial ribosomal protein S4
ATMG00560	RPL2	Rpl2
ATMG00980	RPSL2	Ribosomal protein S12/S23 family protein
ATMG01270	RPS7	At2g07696
ATCG00050	RPS16	30S ribosomal protein S16, chloroplastic
ATCG00160	RPS2	30S ribosomal protein S2, chloroplastic
ATCG00330	RPS14	30S ribosomal protein S14, chloroplastic
ATCG00380	RPS4	30S ribosomal protein S4, chloroplastic
ATCG00640	RPL33	50S ribosomal protein L33, chloroplastic
ATCG00650	RPS18	30S ribosomal protein S18, chloroplastic
ATCG00660	RPL20	50S ribosomal protein L20, chloroplastic
ATCG00065	RPS12A	ribosomal protein S12A
ATCG00750	RPS11	30S ribosomal protein S11, chloroplastic
ATCG00760	RPL36	50S ribosomal protein L36, chloroplastic
ATCG00770	RPS8	30S ribosomal protein S8, chloroplastic
ATCG00780	RPL14	50S ribosomal protein L14, chloroplastic
ATCG00790	RPL16	50S ribosomal protein L16, chloroplastic
ATCG00800	RPS3	30S ribosomal protein S3, chloroplastic
ATCG00810	RPL22	50S ribosomal protein L22, chloroplastic
ATCG00820	RPS19	ribosomal protein S19
ATCG00830	rpl2-A	50S ribosomal protein L2, chloroplastic
ATCG00840	RPL23-A	50S ribosomal protein L23, chloroplastic
ATCG00900	RPS7-A	30S ribosomal protein S7, chloroplastic
ATCG00905	RPS12C	ribosomal protein S12C
ATCG01020	RPL32	50S ribosomal protein L32, chloroplastic
ATCG01120	RPS15	30S ribosomal protein S15, chloroplastic
ATCG01230	RPS12B	ribosomal protein S12B
ATCG01240	RPS7-A	30S ribosomal protein S7, chloroplastic
ATCG01300	RPL23-A	50S ribosomal protein L23, chloroplastic
ATCG01310	rpl2-A	50S ribosomal protein L2, chloroplastic

Appendix B

Full PMET results for cell-type specific identity and immunity genes

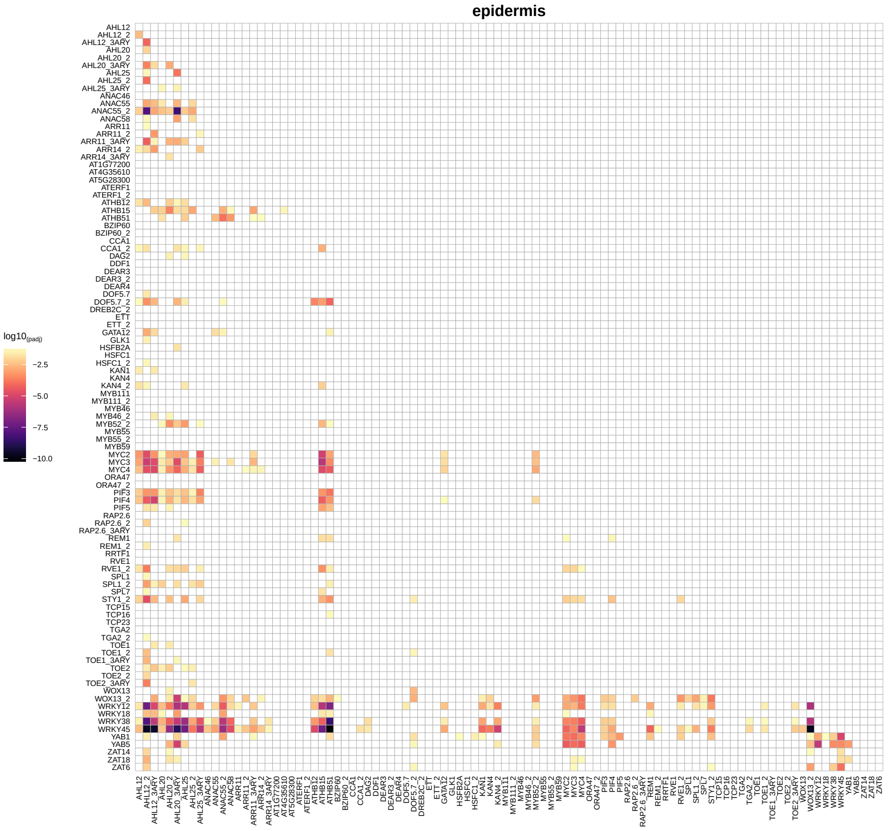


Figure B.1: Heat map to show the paired motif enriched in the 1000bp promoter upstream of the transcription start site (TSS) plus the 5' untranslated region (UTR) region in the top 521 epidermis identity genes in the epidermis. Colour indicates significance of association for values of $p \leq 0.05$, $p > 0.05$ are coloured as white.

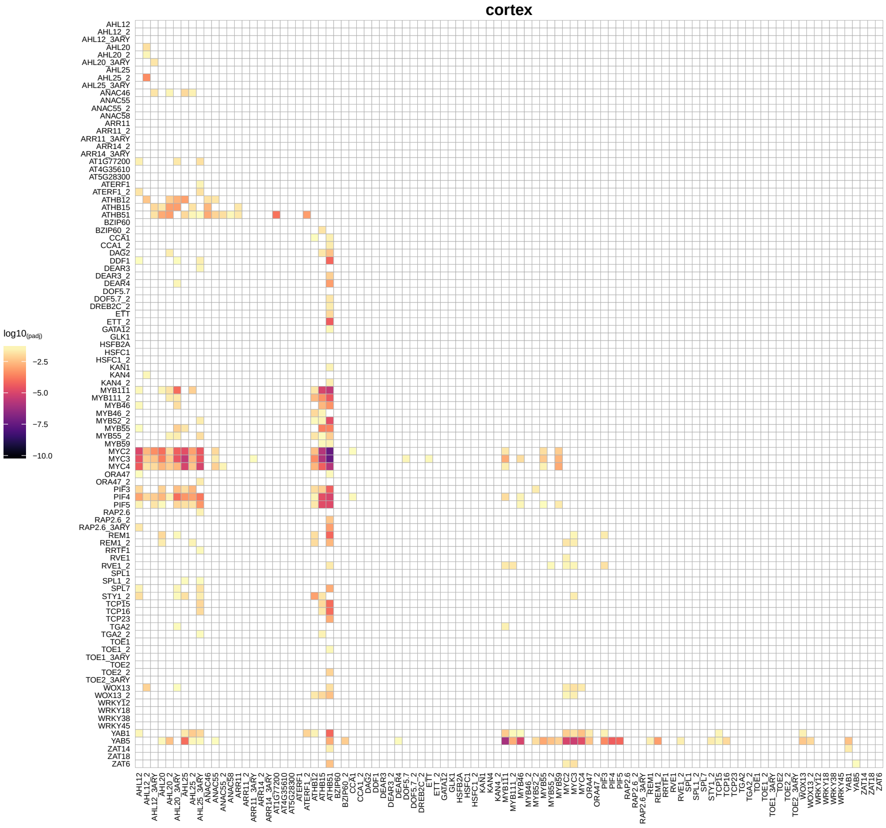


Figure B.2: Heat map to show the paired motif enriched in the 1000bp promoter upstream of the TSS plus the 5’ UTR region in the top 521 cortex identity genes in the cortex. Colour indicates significance of association for values of $p \leq 0.05$, $p > 0.05$ are coloured as white.

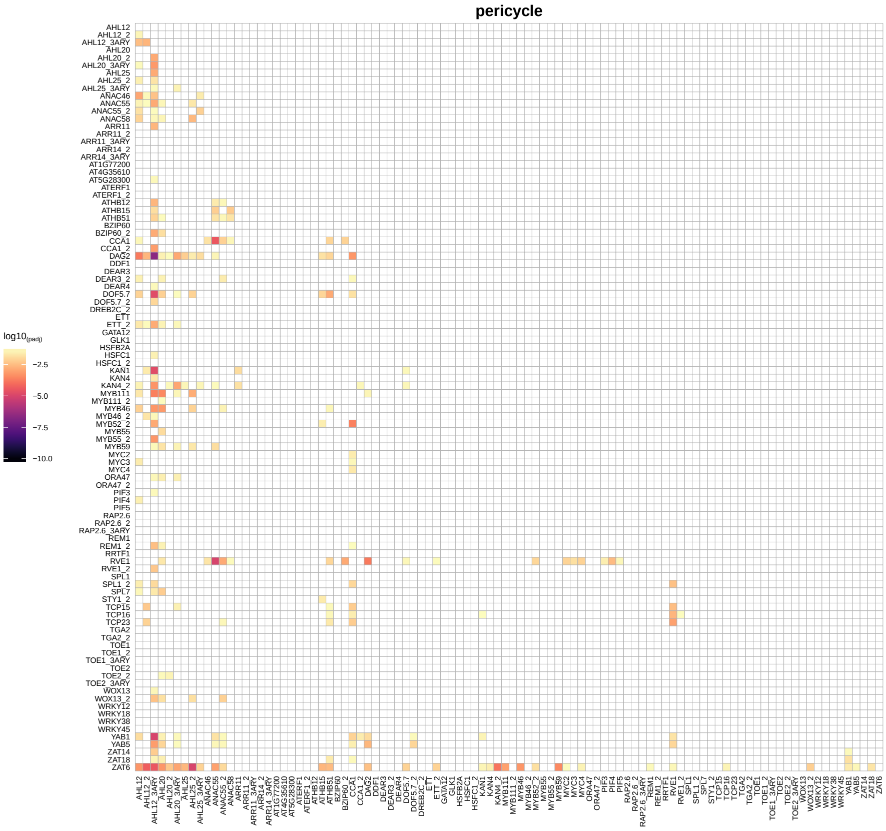


Figure B.3: Heat map to show the paired motif enriched in the 1000bp promoter upstream of the TSS plus the 5' UTR region in the top 521 pericycle identity genes in the pericycle. Colour indicates significance of association for values of $p \leq 0.05$, $p > 0.05$ are coloured as white.

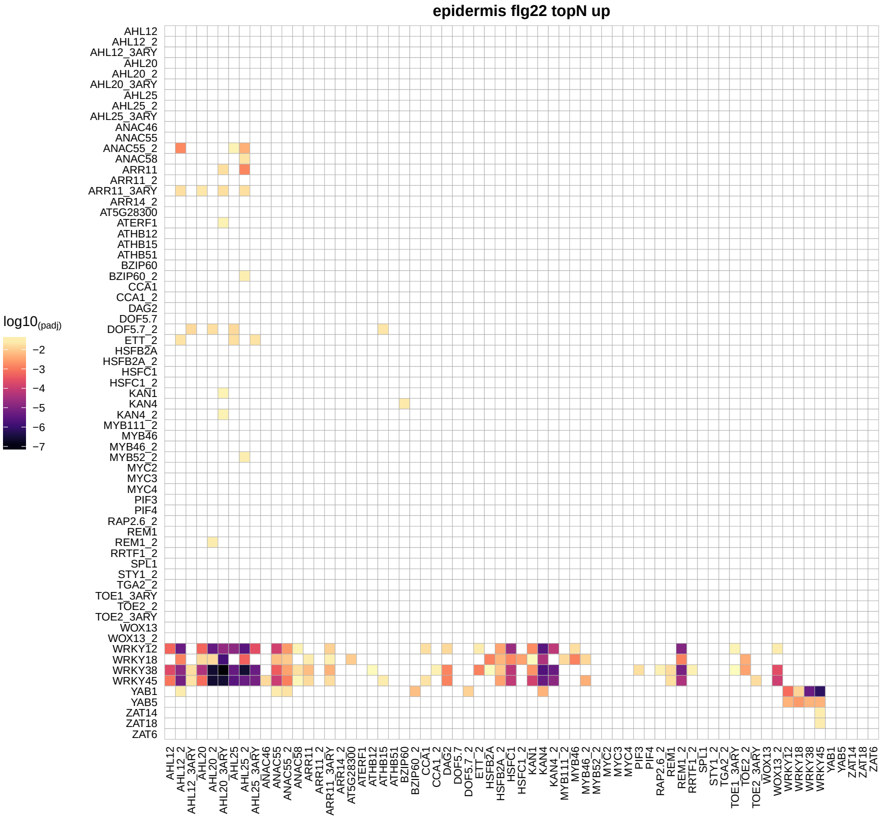


Figure B.4: Heat map to show the paired motif enriched in the 1000bp promoter upstream of the TSS plus the 5' UTR region in the top 128 flg22 responsive genes in the epidermis . Colour indicates significance of association for values of $p \leq 0.01$, $p > 0.01$ are coloured as white.

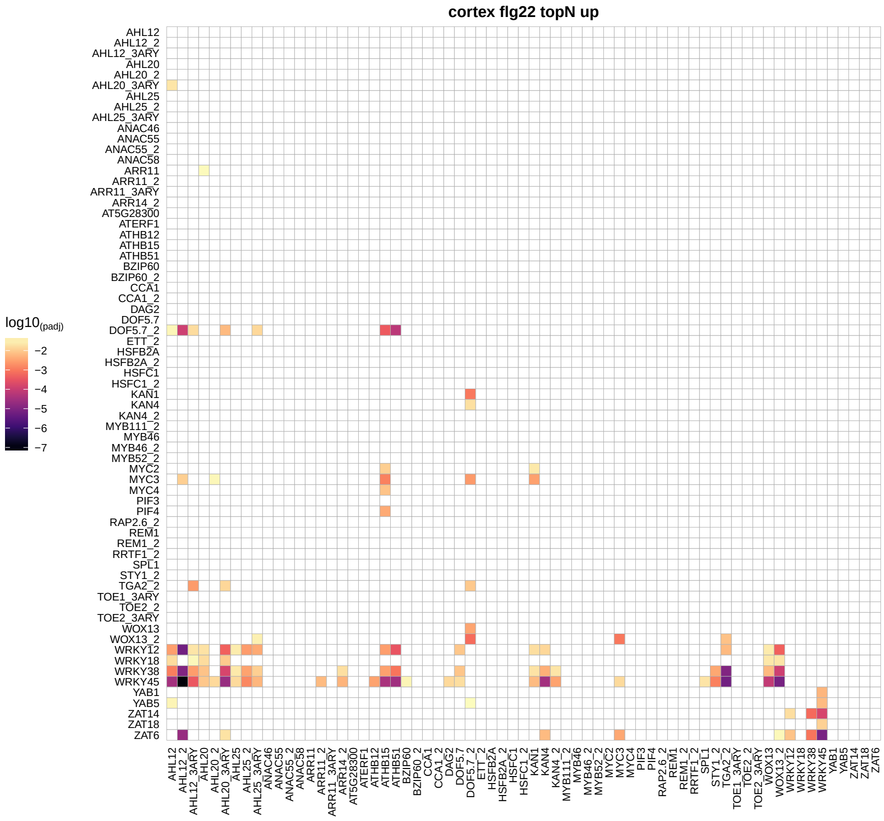


Figure B.5: Heat map to show the paired motif enriched in the 1000bp promoter upstream of the TSS plus the 5' UTR region in the top 128 flg22 responsive genes in the epidermis . Colour indicates significance of association for values of $p \leq 0.01$, $p > 0.01$ are coloured as white.

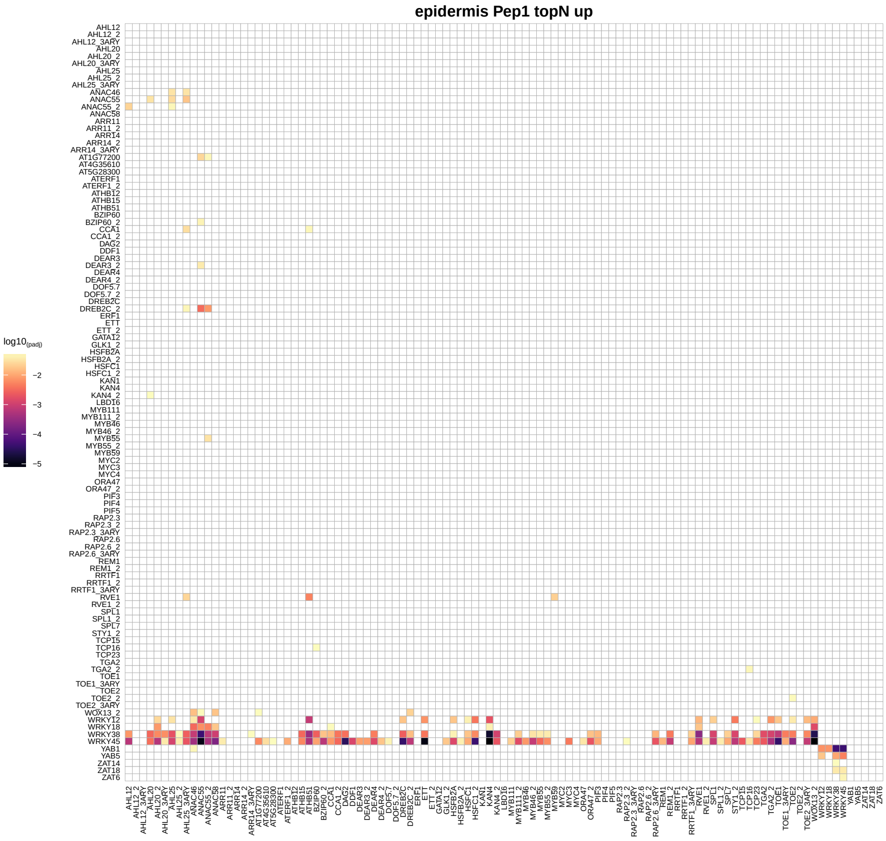


Figure B.6: Heat map to show the paired motif enriched in the 1000bp promoter upstream of the TSS plus the 5' UTR region in the top 365 Pep1 induced genes in the epidermis . Colour indicates significance of association for values of $p \leq 0.01$, $p > 0.01$ are coloured as white.

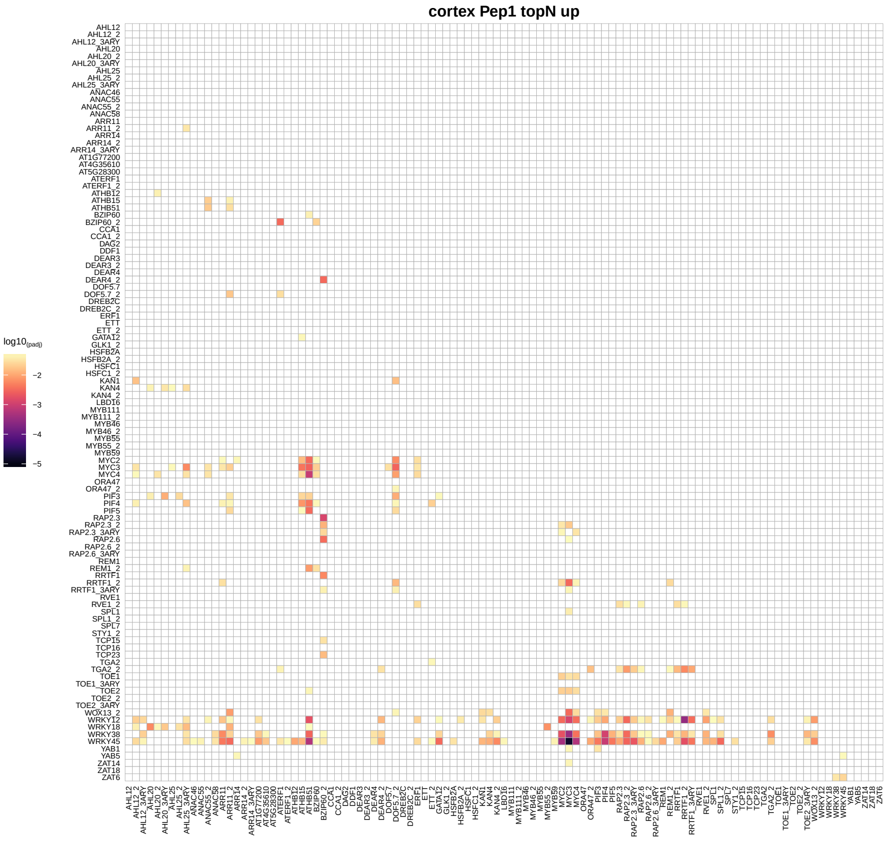


Figure B.7: Heat map to show the paired motif enriched in the 1000bp promoter upstream of the TSS plus the 5' UTR region in the top 365 Pep1 induced genes in the epidermis . Colour indicates significance of association for values of $p \leq 0.01$, $p > 0.01$ are coloured as white.

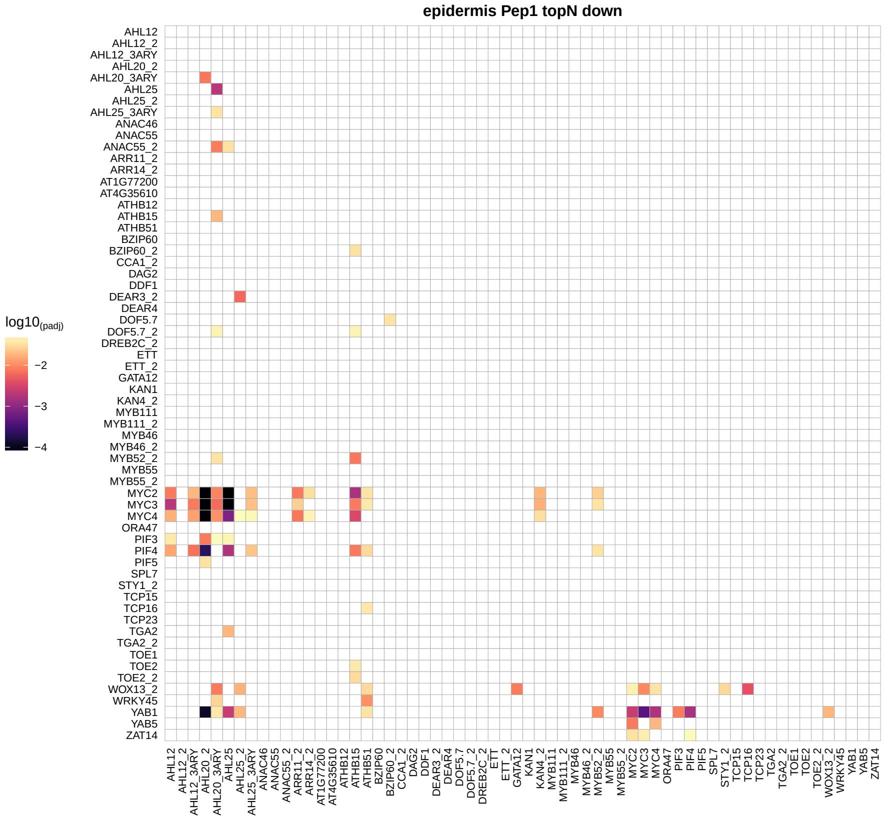


Figure B.8: Heat map to show the paired motif enriched in the 1000bp promoter upstream of the TSS plus the 5' UTR region in the top 365 Pep1 induced genes in the epidermis . Colour indicates significance of association for values of $p \leq 0.01$, $p > 0.01$ are coloured as white.

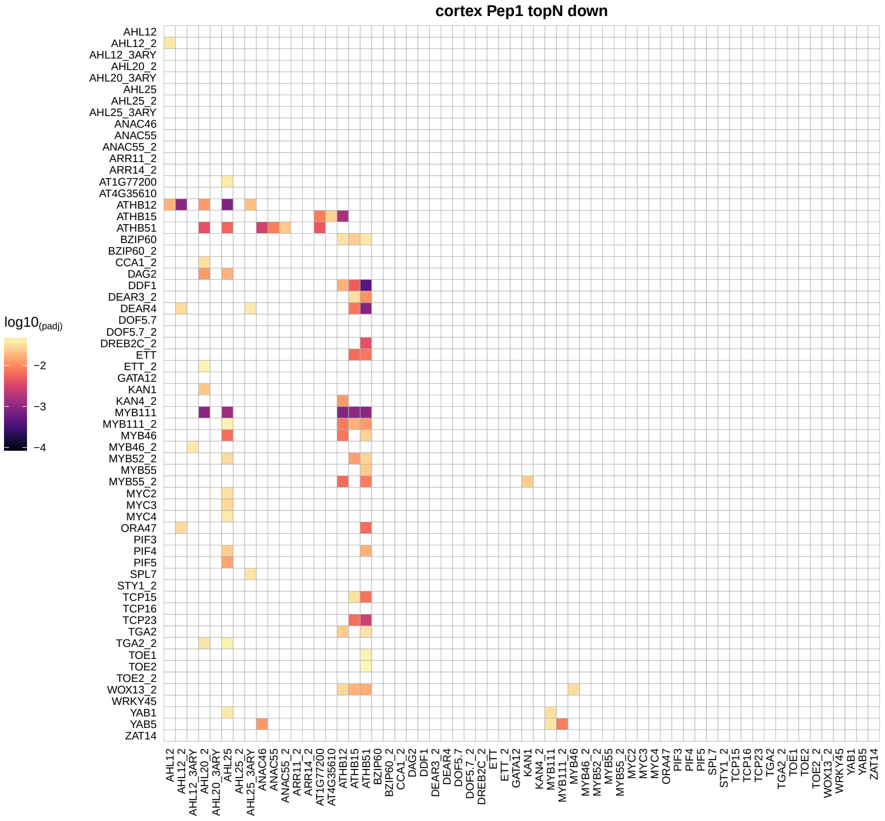


Figure B.9: Heat map to show the paired motif enriched in the 1000bp promoter upstream of the TSS plus the 5' UTR region in the top 337 Pep1 induced genes in the epidermis . Colour indicates significance of association for values of $p \leq 0.01$, $p > 0.01$ are coloured as white.

Bibliography

Patrick Achard, Andi Gusti, Soizic Cheminant, Malek Alioua, Stijn Dhondt, Frederik Coppens, Gerrit T.S. Beemster, and Pascal Genschik. Gibberellin signaling controls cell proliferation rate in *Arabidopsis*. *Current Biology*, 19(14):1188–1193, 2009.

Stein Aerts, Peter Van Loo, Gert Thijs, Yves Moreau, and Bart De Moor. Computational detection of cis-regulatory modules. *Bioinformatics*, 19(Supplement 2):ii5–ii14, 2003.

Mitsuhiro Aida, Dimitris Beis, Renze Heidstra, Viola Willemsen, Ikram Blilou, Carla Galinha, Laurent Nussaume, Yoo Sun Noh, Richard Amasino, and Ben Scheres. The *PLETHORA* genes mediate patterning of the *Arabidopsis* root stem cell niche. *Cell*, 119(1):119–120, 2004.

James Aist and William Bushnell. Invasion of plants by powdery mildew fungi, and cellular mechanisms of resistance. In *The Fungal Spore and Disease Initiation in Plants and Animals*, pages 321–345. Springer, 1991.

Ayelet Alpert, Lindsay Moore, Tania Dubovik, and Shai Shen-Orr. Alignment of single-cell trajectories to compare cellular expression dynamics. *Nature Methods*, 15(4):267, 2018.

Simon Andrews. FastQC [software], 2010. Retrieved from `http://www.bioinformatics.babraham.ac.uk/projects/fastqc`.

Tallulah Andrews and Martin Hemberg. Identifying cell populations with scRNASeq. *Molecular Aspects of Medicine*, 59(2018):114–122, 2017.

Tsuneaki Asai, Guillaume Tena, Joulia Plotnikova, Matthew Willmann, Wan Ling Chiu, Lourdes Gomez-Gomez, Thomas Boller, Frederick Ausubel, and Jen Sheen. MAP kinase signalling cascade in *Arabidopsis* innate immunity. *Nature*, 415(6875):977–983, 2002.

Gwenael Badis, Michael Berger, Anthony Philippakis, Andrew Gehrke, Savina Jaeger, Esther Chan, Anastasia Vedenko, Xiaoyu Chen, Hanna Kuznetsov, Chi-fong Wang, Daniel Newburger, Quaid Morris, Timothy Hughes, and Martha Bulyk. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935):1720–1723, 2009.

Bastiaan Bargmann, Steffen Vanneste, Gabriel Krouk, Tal Nawy, Idan Efroni, Eilon Shani, Goh Choe, Jiří Friml, Dominique Bergmann, Mark Estelle, and Kenneth Birnbaum. A map of cell type-specific auxin responses. *Molecular Systems Biology*, 9(1):688–688, 2013.

Sebastian Bartels and Thomas Boller. Quo vadis, Pep? Plant elicitor peptides at the crossroads of immunity, stress, and development. *Journal of Experimental Botany*, 66(17):5183–5193, 2015.

Sebastian Bartels, Martina Lori, Malick Mbengue, Marcel Van Verk, Dominik Klauser, Tim Hander, Rainer Böni, Silke Robatzek, and Thomas Boller. The family of Peps and their precursors in *Arabidopsis*: differential expression and localization but similar induction of pattern-triggered immune responses. *Journal of Experimental Botany*, 64(17):5309–5321, 2013.

Juan I Fuxman Bass, John S Reece-Hoyes, and Albertha JM Walhout. Performing yeast one-hybrid library screens. *Cold Spring Harbor Protocols*, 2016(12):pdb–prot088955, 2016.

Laura Baxter, Aleksey Jironkin, Richard Hickman, Jay Moore, Christopher Barrington, Peter Krusche, Nigel P Dyer, Vicky Buchanan-Wollaston, Alexander Tiskin, Jim Beynon, et al. Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. *The Plant Cell*, pages tpc–112, 2012.

Martina Beck, Ines Wyrsch, James Strutt, Rinukshi Wimalasekera, Alex Webb, Thomas Boller, and Silke Robatzek. Expression patterns of *FLAGELLIN SENSING 2* map to bacterial entry sites in plant shoots and roots. *Journal of Experimental Botany*, 65(22):6487–6498, 2014.

Paweł Bednarek, Mariola Piślewska-Bednarek, Aleš Svatoš, Bernd Schneider, Jan Doubský, Madina Mansurova, Matt Humphry, Chiara Consonni, Ralph Panstruga, Andrea Sanchez-Vallet, Antonio Molina, and Paul Schulze-Lefert. A glucosinolate metabolism pathway in living plant cells mediates broad-spectrum antifungal defense. *Science*, 323(5910):101–106, 2009.

Philip Benfey and Ben Scheres. Root development. *Current Biology*, 10(22):R813–R815, 2000.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 5(1):289–300, 1995.

Rahul Bhosale, Veronique Boudolf, Fabiola Cuevas, Ran Lu, Thomas Eekhout, Zhubing Hu, Gert Van Isterdael, Georgina M Lambert, Fan Xu, Moritz K Nowack, et al. A spatiotemporal dna endoploidy map of the arabidopsis root reveals roles for the endocycle in root development and stress adaptation. *The Plant Cell*, 30(10):2330–2351, 2018.

Kenneth Birnbaum, Dennis E Shasha, Jean Wang, Jee Jung, Georgina Lambert, David Galbraith, and Philip Benfey. A gene expression map of the *Arabidopsis* root. *Science*, 302(5652):1956–1960, 2003.

Kenneth Birnbaum, Jee Jung, Jean Wang, Georgina Lambert, John Hirst, David Galbraith, and Philip Benfey. Cell type-specific expression profiling in plants via cell sorting of protoplasts from fluorescent reporter lines. *Nature Methods*, 2(8):615–619, 2005.

Anthony Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

Thomas Boller and Georg Felix. A renaissance of elicitors: Perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. *Annual Review of Plant Biology*, 60(1): 379–406, 2009.

Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

Martin Bonke, Siripong Thitamadee, Ari Pekka Mähönen, Marie-Theres Hauser, and Ykä Helariutta. APL regulates vascular tissue identity in *Arabidopsis*. *Nature*, 426(6963):181, 2003.

Frèdèric Bouchè. *Arabidopsis* - root cell types, 2017. URL `https://figshare.com/articles/Arabidopsis_-_Root_cell_types/4688752/1`.

Matthieu Bourdon, Julien Pirrello, Catherine Cheniclet, Olivier Coriton, Mickaël Bourge, Spencer Brown, Adeline Moïse, Martine Peypelut, Valérie Rouyère, Jean-Pierre Renaudin, et al. Evidence for karyoplasmic homeostasis during endoreduplication and a ploidy-dependent increase in gene transcription during tomato fruit growth. *Development*, 139(20):3817–3826, 2012.

John Bowman. The YABBY gene family and abaxial cell fate. *Current Opinion in Plant biology*, 3(1): 17–22, 2000.

Siobhan Brady, David Orlando, Ji-Young Lee, Jean Wang, Jeremy Koch, José Dinneny, Daniel Mace, Uwe Ohler, and Philip Benfey. A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science*, 318(5851):801–806, 2007.

Siobhan M Brady, Lifang Zhang, Molly Megraw, Natalia J Martinez, Eric Jiang, S Yi Charles, Weilin Liu, Anna Zeng, Mallorie Taylor-Teeples, Dahae Kim, et al. A stele-enriched gene regulatory network in the arabidopsis root. *Molecular Systems Biology*, 7(1):459, 2011.

Emily Breeze, Elizabeth Harrison, Stuart McHattie, Linda Hughes, Richard Hickman, Claire Hill, Steven Kiddle, Youn-sung Kim, Christopher Penfold, Dafyd Jenkins, Cunjin Zhang, Karl Morris, Carol Jenner, Stephen Jackson, Brian Thomas, Alexandra Tabrett, Roxane Legaie, Jonathan Moore, David Wild, Sascha Ott, David Rand, Jim Beynon, Katherine Denby, Andrew Mead, and Vicky Buchanan-Wollaston. High-resolution temporal profiling of transcripts during *Arabidopsis* leaf senescence reveals a distinct chronology of processes and regulation. *The Plant Cell*, 23(3):873–894, 2011.

Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah Teichmann, John Marioni, and Marcus Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093, 2013.

Christian Breuer, Luke Braidwood, and Keiko Sugimoto. Endocycling in the path of plant development. *Current Opinion in Plant Biology*, 17:78–85, 2014.

The Broad Institute. Picard tools v2.0.1, 2018. URL `http://broadinstitute.github.io/picard/`.

Lester Brown. *World on the Edge: How to Prevent Environmental and Economic Collapse*. Routledge, 2012.

Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486, 2015.

Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5): 411–420, 2018.

Jun Cao, Korbinian Schneeberger, Stephan Ossowski, Torsten Günther, Sebastian Bender, Joffrey Fitz, Daniel Koenig, Christa Lanz, Oliver Stegle, Christoph Lippert, et al. Whole-genome sequencing of multiple arabidopsis thaliana populations. *Nature Genetics*, 43(10):956, 2011.

Hanbo Chen and Paul C. Boutros. Venndiagram: A package for the generation of highly-customizable venn and euler diagrams in r. *BMC Bioinformatics*, 12(1):35, 2011.

Chia-Yi Cheng, Vivek Krishnakumar, Agnes Chan, Françoise Thibaud-Nissen, Seth Schobel, and Christopher Town. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *The Plant Journal*, 89(4):789–804, 2017.

Delphine Chinchilla, Cyril Zipfel, Silke Robatzek, Birgit Kemmerling, Thorsten Nürnberger, Jonathan Jones, Georg Felix, and Thomas Boller. A flagellin-induced complex of the receptor FLS2 and BAK1 initiates plant defence. *Nature*, 448(7152):497–500, 2007.

Won-Gyu Choi, Masatsugu Toyota, Su-Hwa Kim, Richard Hilleary, and Simon Gilroy. Salt stress-induced ca2+ waves are associated with rapid, long-distance root-to-shoot signaling in plants. *Proceedings of the National Academy of Sciences*, 111(17):6497–6502, 2014.

Jake Conway, Alexander Lex, and Nils Gehlenborg. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940, 2017.

Daniel Couto and Cyril Zipfel. Regulation of pattern recognition receptor signalling in plants. *Nature Reviews Immunology*, 16(9):537, 2016.

Marta de Torres Zabala, George Littlejohn, Siddharth Jayaraman, David Studholme, Trevor Bailey, Tracy Lawson, Michael Tillich, Dirk Licht, Bettina Bölter, Laura Delfino, William Truman, John Mansfield, Nicholas Smirnoff, and Murray Grant. Chloroplasts play a central role in plant defence and are targeted by pathogen effectors. *Nature Plants*, 1(6):15074, 2015.

Lieven De Veylder, John C Larkin, and Arp Schnittger. Molecular control and function of endoreplication in development and physiology. *Trends in Plant Science*, 16(11):624–634, 2011.

Etienne Delannoy, Monique Le Ret, Emmanuelle Faivre-Nitschke, Gonzalo Estavillo, Marc Bergdoll, Nicolas Taylor, Barry Pogson, Ian Small, Patrice Imbault, and José Gualberto. *Arabidopsis* tRNA adenosine deaminase arginine edits the wobble nucleotide of chloroplast tRNA Arg(ACG) and is essential for efficient chloroplast translation. *The Plant Cell*, 21(7):2058–2071, 2009.

Raffaele Dello Ioio, Kinu Nakamura, Laila Moubayidin, Serena Perilli, Masatoshi Taniguchi, Miyo Morita, Takashi Aoyama, Paolo Costantino, and Sabrina Sabatini. A genetic framework for the control of cell division and differentiation in the root meristem. *Science*, 322(5906):1380–1384, 2008.

José Dinneny, Terri Long, Jean Wang, Jee Jung, Daniel Mace, Solomon Pointer, Christa Barron, Siobhan Brady, John Schiefelbein, and Philip Benfey. Cell identity mediates the response of *Arabidopsis* roots to abiotic stress. *Science*, 320(5878):942–945, 2008.

Alexander Dobin, Carrie Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29 (1):15–21, 2013.

Liam Dolan, Kees Janmaat, Viola Willemsen, Paul Linstead, Scott Poethig, Keith Roberts, and Ben Scheres. Cellular organisation of the *Arabidopsis thaliana* root. *Development*, 119(1):71–84, 1993.

Bruno Dombrecht, Gang Peng Xue, Susan Sprague, John Kirkegaard, John Ross, James Reid, Gary Fitt, Nasser Sewelam, Peer Schenk, John Manners, and Kemal Kazan. MYC2 differentially modulates diverse jasmonate-dependent functions in *Arabidopsis*. *The Plant Cell*, 19(7):2225–2245, 2007.

Lina Duan, Jose Sebastian, and Jose Dinneny. Salt-stress regulation of root system growth and architecture in *Arabidopsis* seedlings. In *Plant Cell Expansion*, pages 105–122. Springer, 2015.

Christophe Dunand, Michèle Crèvecoeur, and Claude Penel. Distribution of superoxide and hydrogen peroxide in *Arabidopsis* root and their influence on root development: possible interaction with peroxidases. *New Phytologist*, 174(2):332–341, 2007.

Nidel Dyer, Vahid Shahrezaei, and Daniel Hebenstreit. LiBiNorm: an HTSeq analogue with improved normalisation of SMART-seq2 data and library preparation diagnostics. available online at: https://warwick.ac.uk/fac/sci/lifesci/research/libinorm, 2018.

Idan Efroni and Kenneth Birnbaum. The potential of single-cell profiling in plants. *Genome Biology*, 17(1): 65, 2016.

Idan Efroni, Pui Leng Ip, Tal Nawy, Alison Mello, and Kenneth Birnbaum. Quantification of cell identity from single-cell gene expression profiles. *Genome Biology*, 16(1):9, 2015.

Idan Efroni, Alison Mello, Tal Nawy, Pui-Leng Ip, Ramin Rahni, Nicholas DelRose, Ashley Powers, Rahul Satija, and Kenneth Birnbaum. Root regeneration triggers an embryo-like sequence guided by hormonal interactions. *Cell*, 165(7):1721–1733, 2016.

Ruth Eichmann and Patrick Schäfer. Growth versus immunity—a redirection of the cell cycle? *Current Opinion in Plant Biology*, 26:106–112, 2015.

Daphne Ezer, Nicolae Radu Zabet, and Boris Adryan. Homotypic clusters of transcription factor binding sites: A model system for understanding the physical mechanics of gene expression. *Computational and Structural Biotechnology Journal*, 10(17):63–69, 2014.

Seth Falcon and Robert Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 2(2):257–258, 2007.

Christine Faulkner and Silke Robatzek. Plants and pathogens: putting infection strategies and defence mechanisms on the map. *Current Opinion in Plant Biology*, 15(6):699–707, 2012.

Georg Felix, Juliana D. Duran, Sigrid Volko, and Thomas Boller. Plants have a sensitive perception system for the most conserved domain of bacterial flagellin. *The Plant Journal*, 18(3):265–276, 1999.

Nuria Fernández-Bautista, Lourdes Fernández-Calvino, Alfonso Munoz, and Mar Castellano. HOP3 a new regulator of the ER stress response in *Arabidopsis* with possible implications in plant development and response to biotic and abiotic stresses. *Plant Signaling and Behavior*, 12(5):e1317421, 2017.

Patricia Fernández-Calvo, Andrea Chini, Gemma Fernández-Barbero, José-Manuel Chico, Selena Gimenez-Ibanez, Jan Geerinck, Dominique Eeckhout, Fabian Schweizer, Marta Godoy, José Manuel Franco-Zorrilla, Laurens Pauwels, Erwin Witters, María Isabel Puga, Javier Paz-Ares, Alain Goossens, Philippe Reymond, Geert De Jaeger, and Roberto Solano. The *Arabidopsis* bHLH transcription factors MYC3 and MYC4 are targets of JAZ repressors and act additively with MYC2 in the activation of jasmonate responses. *The Plant Cell*, 23(2):701–715, 2011.

María Lorena Falcone Ferreyra, Alejandro Pezza, Jordane Biarc, Alma L Burlingame, and Paula Casati. Plant l10 ribosomal proteins have different roles during development and translation under uv-b stress. *Plant Physiology*, pages pp–110, 2010.

Daniela Fichtenbauer, Xianfeng Morgan Xu, Dave Jackson, and Friedrich Kragler. The chaperonin CCT8 facilitates spread of tobamovirus infection. *Plant Signal Behaviour*, 7(3):318–321, 2012.

Pascale Flury, Dominik Klauser, Birgit Schulze, Thomas Boller, and Sebastian Bartels. The anticipation of danger: Microbe-Associated Molecular Pattern perception enhances AtPep-triggered oxidative burst. *Plant Physiology*, 161(4):2023–2035, 2013.

José Franco-Zorrilla, Irene López-Vidriero, José Carrasco, Marta Godoy, Pablo Vera, and Roberto Solano. DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proceedings of the National Academy of Sciences*, 111(6):2367–2372, 2014.

Edith Francoz, Philippe Ranocha, Huan Nguyen-Kim, Elisabeth Jamet, Vincent Burlat, and Christophe Dunand. Roles of cell wall peroxidases in plant development. *Phytochemistry*, 112:15–21, 2015.

Martin C Frith, Ulla Hansen, and Zhiping Weng. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, 17(10):878–889, 2001.

Martin C Frith, Michael C Li, and Zhiping Weng. Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Research*, 31(13):3666–3668, 2003.

Lillian Fritz-Laylin, Nandini Krishnamurthy, Mahmut Tör, Kimmen Sjölander, and Jonathan Jones. Phylogenomic analysis of the receptor-like proteins of rice and *Arabidopsis*. *Plant Physiology*, 138(2):611–623, 2005.

Ushio Fujikura, Gorou Horiguchi, María Rosa Ponce, José Luis Micol, and Hirokazu Tsukaya. Coordination of cell proliferation and cell expansion mediated by ribosome-related processes in the leaves of arabidopsis thaliana. *The Plant Journal*, 59(3):499–508, 2009.

Yu Geng, Rui Wu, Choon Wei Wee, Fei Xie, Xueliang Wei, Penny Mei Yeen Chan, Cliff Tham, Lina Duan, and José Dinneny. A spatio-temporal understanding of growth regulation during the salt stress response in *Arabidopsis*. *The Plant Cell*, 25(6):2132–2154, 2013.

Sandip Ghuge, Andrea Carucci, Renato Rodrigues-Pousada, Alessandra Tisi, Stefano Franchi, Paraskevi Tavladoraki, Riccardo Angelini, and Alessandra Cona. The apoplastic copper AMINE OXIDASE1 mediates jasmonic acid-induced protoxylem differentiation in *Arabidopsis* roots. *Plant Physiology*, 168(2): 690–707, 2015.

Miriam Gifford, Alexis Dean, Rodrigo Gutierrez, Gloria Coruzzi, and Kenneth Birnbaum. Cell-specific nitrogen responses mediate developmental plasticity. *Proceedings of the National Academy of Sciences*, 105(2):803–808, 2008.

Miriam Gifford, Joshua Banta, Manpreet Katari, Jo Hulsmans, Lisa Chen, Daniela Ristova, Daniel Tranchina, Michael Purugganan, Gloria Coruzzi, and Kenneth Birnbaum. Plasticity regulators modulate specific root traits in discrete nitrogen environments. *PLoS Genetics*, 9(9):e1003760, 2013.

Vladimir Gligorijević and Nataša Pržulj. Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*, 12(112):20150571, 2015.

Lourdes Gómez-Gómez and Thomas Boller. FLS2: An LRR receptor–like kinase involved in the perception of the bacterial elicitor flagellin in *Arabidopsis*. *Molecular Cell*, 5(6):1003–1011, 2000.

Lourdes Gómez-Gómez, Georg Felix, and Thomas Boller. A single locus determines sensitivity to bacterial flagellin in *Arabidopsis thaliana*. *The Plant Journal*, 18(3):277–284, 1999.

Jennifer Gottwald, Patrick Krysan, Jeffery Young, Ray Evert, and Michael Sussman. Genetic evidence for the in planta role of phloem-specific plasma membrane sucrose transporters. *Proceedings of the National Academy of Sciences*, 97(25):13979–13984, 2000.

Charles Grant, Timothy Bailey, and William Stafford Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

Jesper Grønlund, Alison Eyres, Sanjeev Kumar, Vicky Buchanan-Wollaston, and Miriam Gifford. Cell specific analysis of *Arabidopsis* leaves using Fluorescence Activated Cell Sorting. *Journal of Visualized Experiments*, (68):e4214, 2012.

Kay Gully, Tim Hander, Thomas Boller, and Sebastian Bartels. Perception of *Arabidopsis* AtPep peptides, but not bacterial elicitors, accelerates starvation-induced senescence. *Frontiers in Plant Science*, 6:14, 2015.

Marc Halfon, Ana Carmena, Stephen Gisselbrecht, Charles Sackerson, Fernando Jiménez, Mary Baylies, and Alan Michelson. Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors. *Cell*, 103(1):63–74, 2000.

Nathaniel Hawker and John Bowman. Roles for Class III HD-Zip and *KANADI* genes in *Arabidopsis* root development. *Plant Physiology*, 135(4):2261–2270, 2004.

Antje Heese, Dagmar R Hann, Selena Gimenez-Ibanez, Alexandra ME Jones, Kai He, Jia Li, Julian I Schroeder, Scott C Peck, and John P Rathjen. The receptor-like kinase serk3/bak1 is a central regulator of innate immunity in plants. *Proceedings of the National Academy of Sciences*, 104(29):12217–12222, 2007.

Martin Heil. Ecological costs of induced resistance. *Current Opinion in Plant Biology*, 5(4):345–350, 2002.

Martin Heil and Ian T Baldwin. Fitness costs of induced resistance: emerging experimental support for a slippery concept. *Trends in Plant Science*, 7(2):61–67, 2002.

Yrjo Helariutta, Hidehiro Fukaki, Joanna Wysocka-Diller, Keiji Nakajima, Jee Jung, Giovanni Sena, Marie-Theres Hauser, and Philip Benfey. The *SHORTROOT* gene controls radial patterning of the *Arabidopsis* root through radial signaling. *Cell*, 101(5):555–567, 2000.

Richard Hickman, Claire Hill, Christopher Penfold, Emily Breeze, Laura Bowden, Jonathan Moore, Peijun Zhang, Alison Jackson, Emma Cooke, Findlay Bewicke-Copley, et al. A local regulatory network around three nac transcription factors in stress responses and senescence in *Arabidopsis* leaves. *The Plant Journal*, 75(1):26–39, 2013.

Yosef Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4): 800–802, 1988.

Alisa Huffaker, Nicole Dafoe, and Eric Schmelz. ZmPep1, an ortholog of *Arabidopsis* elicitor peptide 1, regulates maize innate immunity and enhances disease resistance. *Plant Physiology*, pages pp–110, 2011.

Alisa Huffaker, Gregory Pearce, Nathalie Veyrat, Matthias Erb, Ted Turlings, Ryan Sartor, Zhouxin Shen, Steven Briggs, Martha Vaughan, Hans Alborn, Peter Teal, and Eric Schmelz. Plant elicitor peptides are conserved signals regulating direct and indirect antiherbivore defense. *Proceedings of the National Academy of Sciences*, page 201214668, 2013.

Michaela Hundertmark and Dirk Hincha. LEA (late embryogenesis abundant) proteins and their encoding genes in *Arabidopsis thaliana*. *BMC Genomics*, 9(1):118, 2008.

Yoon Sun Hur, Ji Hyun Um, Sunghan Kim, Kyunga Kim, Hee Jung Park, Jong Seok Lim, Woo Young Kim, Sang Eun Jun, Eun Kyung Yoon, Jun Lim, Masaru Ohme-Takagi, Donggiun Kim, Jongbum Park, Gyung Tae Kim, and Choong Ill Cheon. *ARABIDOPSIS THALIANA* HOMEOBOX 12 (ATHB12), a homeodomain-leucine zipper protein, regulates leaf growth by promoting cell expansion and endoreduplication. *New Phytologist*, 205(1):316–328, 2015.

Michael Ilegems, Véronique Douet, Marlyse Meylan-Bettex, Magalie Uyttewaal, Lukas Brand, John Bowman, and Pia Stieger. Interplay of auxin, KANADI and Class III HD-ZIP transcription factors in vascular tissue formation. *Development*, 137(6):975–984, 2010.

Tomislav Ilicic, Jong Kyoung Kim, Aleksandra Kolodziejczyk, Frederik Otzen Bagger, Davis James McCarthy, John Marioni, and Sarah Teichmann. Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, 17(1):29, 2016.

Anjali Iyer-Pascuzzi, Terry Jackson, Hongchang Cui, Jalean Petricka, Wolfgang Busch, Hironaka Tsukagoshi, and Philip Benfey. Cell identity regulators link development and stress responses in the *Arabidopsis* root. *Developmental Cell*, 21(4):770–782, 2011.

Sophie Jacobs, Bernd Zechmann, Alexandra Molitor, Marco Trujillo, Elena Petutschnig, Volker Lipka, Karl-Heinz Kogel, and Patrick Schäfer. Broad-spectrum suppression of innate immunity is required for colonization of *Arabidopsis* roots by the fungus *Piriformospora indica*. *Plant Physiology*, 156(2):726–740, 2011.

Elena Jeworutzki, Rob Roelfsema, Uta Anschütz, Elzbieta Krol, Theo Elzenga, Georg Felix, Thomas Boller, Rainer Hedrich, and Dirk Becker. Early signaling through the *Arabidopsis* pattern recognition receptors FLS2 and EFR involves Ca^{2+}-associated opening of plasma membrane anion channels. *The Plant Journal*, 62(3):367–378, 2010.

Peng Jiang and Mona Singh. CCAT: combinatorial code analysis tool for transcriptional regulation. *Nucleic Acids Research*, 42(5):2833–2847, 2013.

Öjvind Johansson, Wynand Alkema, Wyeth W Wasserman, and Jens Lagergren. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, 19:i169–i176, 2003.

Jonathan Jones and Jeffery Dangl. The plant immune system. *Nature*, 444(7117):323–329, 2006.

Guillaume Junion, Mikhail Spivakov, Charles Girardot, Martina Braun, Hilary Gustafson, Ewan Birney, and Eileen Furlong. A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell*, 148(3):473–486, 2012.

Kaori Kashi, Lindsey Henderson, Alessandro Bonetti, and Piero Carninci. Discovery and functional analysis of lncRNAs: methodologies to investigate an uncharacterized transcriptome. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1859(1):3–15, 2016.

Tomonori Kawano. Roles of the reactive oxygen species-generating peroxidase reactions in plant defense and growth induction. *Plant Cell Reports*, 21(9):829–837, 2003.

A Kel, Tatiana Konovalova, T Waleev, Evgeny Cheremushkin, O Kel-Margoulis, and Edgar Wingender. Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations. *Bioinformatics*, 22(10):1190–1197, 2006.

Jeong Hoe Kim and Hans Kende. A transcriptional coactivator, AtGIF1, is involved in regulating leaf growth and morphology in *Arabidopsis*. *Proceedings of the National Academy of Sciences*, 101(36):13374–13379, 2004.

Dominik Klauser, Gaylord Desurmont, Gaétan Glauser, Armelle Vallat, Pascale Flury, Thomas Boller, Ted Turlings, and Sebastian Bartels. The *Arabidopsis* Pep-PEPR system is induced by herbivore feeding and contributes to JA-mediated plant defence against herbivory. *Journal of Experimental Botany*, 66(17): 5327–5336, 2015.

Sabrina Kleessen, Roosa Laitinen, Corina M Fusari, Carla Antonio, Ronan Sulpice, Alisdair R Fernie, Mark Stitt, and Zoran Nikoloski. Metabolic efficiency underpins performance trade-offs in growth of arabidopsis thaliana. *Nature Communications*, 5:3537, 2014.

Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

Jae Heung Ko, Won Chan Kim, and Kyung Hwan Han. Ectopic expression of MYB46 identifies transcriptional regulatory genes involved in secondary wall biosynthesis in *Arabidopsis*. *The Plant Journal*, 60(4): 649–665, 2009.

Paula Korkuć, Jozefus Schippers, and Dirk Walther. Characterization and identification of cis-regulatory elements in *Arabidopsis thaliana* based on SNP information. *Plant Physiology*, 164(1):181–200, 2013.

Kevin C Kregel. Heat shock proteins: modifying factors in physiological stress responses and acquired thermotolerance. *Journal of Applied Physiology*, 92(5):2177–2186, 2002.

Elzbieta Krol, Tobias Mentzel, Delphine Chinchilla, Thomas Boller, Georg Felix, Birgit Kemmerling, Sandra Postel, Michael Arents, Elena Jeworutzki, Khaled Al-Rasheid, Dirk Becker, and Rainer Hedrich. Perception of the *Arabidopsis* danger signal peptide 1 involves the pattern recognition receptor AtPEPR1 and its close homologue AtPEPR2. *Journal of Biological Chemistry*, 285(18):13471–13479, 2010.

Chian Kwon, Christina Neu, Simone Pajonk, Hye Sup Yun, Ulrike Lipka, Matt Humphry, Stefan Bau, Marco Straus, Mark Kwaaitaal, Heike Rampelt, Farid El Kasmi, Gerd Jürgens, Jane Parker, Ralph Panstruga, Volker Lipka, and Paul Schulze-Lefert. Co-option of a default secretory pathway for plant immune responses. *Nature*, 451(7180):835–840, 2008.

Harri Lähdesmäki, Alistair G Rust, and Ilya Shmulevich. Probabilistic inference of transcription factor binding from multiple data sources. *PLoS ONE*, 3(3):e1820, 2008.

Ji-Young Lee, Juliette Colinas, Jean Y Wang, Daniel Mace, Uwe Ohler, and Philip Benfey. Transcriptional and posttranscriptional regulation of transcription factor expression in *Arabidopsis* roots. *Proceedings of the National Academy of Sciences*, 103(15):6055–6060, 2006.

Myeong Min Lee and John Schiefelbein. WEREWOLF, a MYB-related protein in arabidopsis, is a position-dependent regulator of epidermal cell patterning. *Cell*, 99(5):473–483, 1999.

Pablo Leivar and Peter Quail. PIFs: pivotal components in a cellular signaling hub. *Trends in Plant Science*, 16(1):19–28, 2011.

Laura Lewis, Krzysztof Polanski, Marta de Torres-Zabala, Siddharth Jayaraman, Laura Bowden, Jonathan Moore, Christopher Penfold, Dafyd Jenkins, Claire Hill, Laura Baxter, Satish Kulasekaran, William Truman, George Littlejohn, Justyna Prusinska, Andrew Mead, Jens Steinbrenner, Richard Hickman, David Rand, David Wild, Sascha Ott, Vicky Buchanan-Wollaston, Nick Smirnoff, Jim Beynon, Katherine Denby, and Murray Grant. Transcriptional dynamics driving MAMP-triggered immunity and pathogen effector-mediated immunosuppression in *Arabidopsis* leaves following infection with *Pseudomonas syringae* pv tomato DC3000. *The Plant Cell*, 27(11):3038–3064, 2015.

Suzanna E Lewis. The vision and challenges of the gene ontology. In *The Gene Ontology Handbook*, pages 291–302. Springer, 2017.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25 (16):2078–2079, 2009.

Catherine Lilley, Dong Wang, Howard Atkinson, and Peter Urwin. Effective delivery of a nematode-repellent peptide using a root-cap-specific promoter. *Plant Biotechnology Journal*, 9(2):151–161, 2011.

Qing Lin, Yohei Ohashi, Mariko Kato, Tomohiko Tsuge, Hongya Gu, Li-Jia Qu, and Takashi Aoyama. GLABRA2 directly suppresses basic helix-loop-helix transcription factor genes with diverse functions in root hair development. *The Plant Cell*, 27:2894–2906, 2015.

Cynthia Lincoln, James Britton, and Mark Estelle. Growth and development of the *axr1* mutants of *Arabidopsis*. *The Plant Cell*, 2(11):1071–1080, 1990.

Zixu Liu, Ying Wu, Fan Yang, Yiyue Zhang, She Chen, Qi Xie, Xingjun Tian, and Jian-Min Zhou. BIK1 interacts with PEPRs to mediate ethylene-induced immunity. *Proceedings of the National Academy of Sciences*, 110(15):6205–6210, 2013.

Michael Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

Chunli Ma, Jie Guo, Yan Kang, Kohei Doman, Anthony Bryan, Frans Tax, Yube Yamaguchi, and Zhi Qi. AtPEPTIDE RECEPTOR2 mediates the AtPEPTIDE1-induced cytosolic Ca^{2+} rise, which is required for the suppression of *Glutamine Dumper* gene expression in *Arabidopsis* roots. *Journal of Integrative Plant Biology*, 56:684–694, 2014.

Shisong Ma and Hans J Bohnert. Integration of *Arabidopsis thaliana* stress-related transcript profiles, promoter structures, and cell-specific expression. *Genome Biology*, 8(4):R49, 2007.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

Evan Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison Bialas, Nolan Kamitaki, Emily Martersteck, John Trombetta, David Weitz, Joshua Sanes, Alex Shalek, Aviv Regev, and Steven McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Jocelyn Malamy and Philip Benfey. Analysis of *SCARECROW* expression using a rapid system for assessing transgene expression in *Arabidopsis* roots. *The Plant Journal*, 12(4):957–963, 1997.

Alan Marchant, Joanna Kargul, Sean May, Philippe Muller, Alain Delbarre, Catherine Perrot-Rechenmann, and Malcolm Bennett. AUX1 regulates root gravitropism in *Arabidopsis* by facilitating auxin uptake within root apical tissues. *The EMBO Journal*, 18(8):2066–2073, 1999.

Alan Marchant, Rishikesh Bhalerao, Ilda Casimiro, Jan Eklöf, Pedro Casero, Malcolm Bennett, and Goran Sandberg. AUX1 promotes lateral root formation by facilitating indole-3-acetic acid distribution between sink and source tissues in the seedling. *The Plant Cell*, 14(3):589–597, 2002.

Akane Matsushita, Tsuyoshi Furumoto, Sarahmi Ishida, and Yohsuke Takahashi. AGF1, an AT-hook protein, is necessary for the negative feedback of AtGA3ox1 encoding GA 3-oxidase. *Plant Physiology*, 143 (3):1152–1162, 2007.

Jessica Messmer McAbee, Theresa Hill, Debra Skinner, Anat Izhaki, Bernard Hauser, Robert Meister, Venugopala Reddy, Elliot Meyerowitz, John Bowman, and Charles Gasser. *ABERRANT TESTA SHAPE* encodes a KANADI family member, linking polarity determination to separation and growth of *Arabidopsis* ovule integuments. *The Plant Journal*, 46(3):522–531, 2006.

Cornelia Meckbach, Rebecca Tacke, Xu Hua, Stephan Waack, Edgar Wingender, and Mehmet Gültas. PC-TraFF: identification of potentially collaborating transcription factors using pointwise mutual information. *BMC Bioinformatics*, 16(1):400, 2015.

Jia Meng, Shou Jiang Gao, and Yufei Huang. Enrichment constrained time-dependent clustering analysis for finding meaningful temporal transcription modules. *Bioinformatics*, 25(12):1521–1527, 2009.

Luana Micallef and Peter Rodgers. eulerAPE: Drawing area-proportional 3-Venn diagrams using ellipses. *PLoS ONE*, 9(7):e101717, 2014.

Todd Michael, Todd Mockler, Ghislain Breton, Connor McEntee, Amanda Byer, Jonathan Trout, Samuel Hazen, Rongkun Shen, Henry Priest, Christopher Sullivan, Scott Givan, Marcelo Yanovsky, Fangxin Hong, Steve Kay, and Joanne Chory. Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. *PLoS Genetics*, 4(2):e14, 2008.

Yves Millet, Cristian Danna, Nicole Clay, Wisuwat Songnuan, Matthew Simon, Daniéle Werck-Reichhart, and Frederick Ausubel. Innate immune responses activated in *Arabidopsis* roots by Microbe-Associated Molecular Patterns. *The Plant Cell*, 22(3):973–990, 2010.

Tatiana Mishina and Jürgen Zeier. Pathogen-associated molecular pattern recognition rather than development of tissue necrosis contributes to bacterial induction of systemic acquired resistance in *Arabidopsis*. *The Plant Journal*, 50(3):500–513, 2007.

M. A. Moreno-Risueno, R. Sozzani, G. G. Yard mc, J. J. Petricka, T. Vernoux, I. Blilou, J. Alonso, C. M. Winter, U. Ohler, B. Scheres, and P. N. Benfey. Transcriptional control of tissue formation throughout root development. *Science*, 350(6259):426–430, 2015.

Arnab Mukhopadhyay, Bart Deplancke, Albertha JM Walhout, and Heidi A Tissenbaum. Chromatin immunoprecipitation (chip) coupled to detection by quantitative real-time pcr to study transcription factor binding to dna in caenorhabditis elegans. *Nature Protocols*, 3(4):698, 2008.

Andreas Müller, Changhui Guan, Leo Gälweiler, Petra Tänzler, Peter Huijser, Alan Marchant, Geraint Parry, Malcolm Bennett, Ellen Wisman, and Klaus Palme. AtPIN2 defines a locus of *Arabidopsis* for root gravitropism control. *The EMBO Journal*, 17(23):6903–6911, 1998.

Rasika Mundade, Hatice Gulcin Ozer, Han Wei, Lakshmi Prabhu, and Tao Lu. Role of chip-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle*, 13(18):2847–2852, 2014.

Keiji Nakajima, Giovanni Sena, Tal Nawy, and Philip Benfey. Intercellular movement of the putative transcription factor SHR in root patterning. *Nature*, 413(6853):307–311, 2001.

Tal Nawy, Ji-Young Lee, Juliette Colinas, Jean Wang, Sumena Thongrod, Jocelyn Malamy, Kenneth Birnbaum, and Philip Benfey. Transcriptional profile of the *Arabidopsis* root quiescent center. *The Plant Cell*, 17(7):1908–1925, 2005.

Vladimir Nekrasov, Jing Li, Martine Batoux, Milena Roux, Zhao Hui Chu, Severine Lacombe, Alejandra Rougon, Pascal Bittel, Marta Kiss-Papp, Delphine Chinchilla, Peter Van Esse, Lucia Jorda, Benjamin Schwessinger, Valerie Nicaise, Bart Thomma, Antonio Molina, Jonathan Jones, and Cyril Zipfel. Control of the pattern-recognition receptor EFR by an ER protein complex in plant immunity. *The EMBO Journal*, 28(21):3428–3438, 2009.

James Nemesh and Steve McCarroll. Drop-seq alignment cookbook, 2016. URL `http://mccarrolllab.com/dropseq/`.

Hironari Nomura, Teiko Komori, Shuhei Uemura, Yui Kanda, Koji Shimotani, Kana Nakai, Takuya Furuichi, Kohsuke Takebayashi, Takanori Sugimoto, Satoshi Sano, et al. Chloroplast-mediated activation of plant immune signalling in arabidopsis. *Nature Communications*, 3:926, 2012.

Jose Antonio O'Brien, Arsalan Daudi, Paul Finch, Vernon S Butt, Julian P Whitelegge, Puneet Souda, Frederick M Ausubel, and G Paul Bolwell. A peroxidase-dependent apoplastic oxidative burst in cultured arabidopsis cells functions in mamp-elicited defence. *Plant Physiology*, pages pp–111, 2012.

Kyoko Ohashi-Ito and Dominique Bergmann. Regulation of the *Arabidopsis* root vascular initial population by *LONESOME HIGHWAY*. *Development*, 134(16):2959–2968, 2007.

Ronan C. O'Malley, Shao-shan Carol Huang, Liang Song, Mathew Lewsey, Anna Bartlett, Joseph Nery, Mary Galli, Andrea Gallavotti, and Joseph Ecker. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, 165(5):1280–1292, 2016.

Shree Pandey and Imre Somssich. The role of WRKY transcription factors in plant immunity. *Plant Physiology*, 150(4):1648–1655, 2009.

Swati Parekh, Christoph Ziegenhain, Beate Vieth, Wolfgang Enard, and Ines Hellmann. The impact of amplification on differential expression analyses by RNA-seq. *Scientific Reports*, 6:25533, 2016.

Jiyoung Park, Youngsook Lee, Enrico Martinoia, and Markus Geisler. Plant hormone transporters: what we know and what we would like to know. *BMC Biology*, 15(1):93, 2017.

Sandhya Payankaulam, Li Li, and David Arnosti. Transcriptional repression: conserved and evolved features. *Current Biology*, 20(17):R764–R771, 2010.

Benjamin Péret, Kamal Swarup, Alison Ferguson, Malvika Seth, Yaodong Yang, Stijn Dhondt, Nicholas James, Ilda Casimiro, Paula Perry, Adnan Syed, Haibing Yang, Jesica Reemmer, Edward Venison, Caroline Howells, Miguel Perez-Amador, Jeonga Yun, Jose Alonso, Gerrit Beemster, Laurent Laplaze, Angus Murphy, Malcolm Bennett, Erik Neilsen, and Ranjan Swarup. *AUX/LAX* genes encode a family of auxin influx transporters that perform distinct functions during *Arabidopsis* development. *The Plant Cell*, pages 2874–2885, 2012.

Robert Petryszak, Maria Keays, Y Amy Tang, Nuno A Fonseca, Elisabet Barrera, Tony Burdett, Anja Füllgrabe, Alfonso Muñoz-Pomer Fuentes, Simon Jupp, Satu Koskinen, et al. Expression atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Research*, 44(D1):D746–D752, 2015.

Carloalberto Petti, Meera Nair, and Seth DeBolt. The involvement of J-protein AtDjC17 in root development in Arabidopsis. *Frontiers in Plant Science*, 5:532, 2014.

Elena K Petutschnig, Alexandra ME Jones, Liliya Serazetdinova, Ulrike Lipka, and Volker Lipka. The lysin motif receptor-like kinase (lysm-rlk) cerk1 is a major chitin-binding protein in arabidopsis thaliana and subject to chitin-induced phosphorylation. *Journal of Biological Chemistry*, 285(37):28902–28911, 2010.

Yitzhak Pilpel, Priya Sudarsanam, and George Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics*, 29(2):153–159, 2001.

Roger Pique-Regi, Jacob F Degner, Athma A Pai, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. *Genome Research*, 21(3):447–455, 2011.

Julien Pirrello, Cynthia Deluche, Nathalie Frangne, Frédéric Gévaudant, Elie Maza, Anis Djari, Mickaël Bourge, Jean-Pierre Renaudin, Spencer Brown, Chris Bowler, et al. Transcriptome profiling of sorted endoreduplicated nuclei from tomato fruits: how the global shift in expression ascribed to dna ploidy influences rna-seq data normalization and interpretation. *The Plant Journal*, 93(2):387–398, 2018.

Lorenzo Poncini, Ines Wyrsch, Valérie Dénervaud Tendon, Thomas Vorley, Thomas Boller, Niko Geldner, Jean Pierre Métraux, and Silke Lehmann. In roots of *Arabidopsis thaliana*, the damage-associated molecular pattern AtPep1 is a stronger elicitor of immune signalling than flg22 or the chitin heptamer. *PLoS ONE*, 12(10):e0185808, 2017.

Aloisie Poulíèková, Petra Mazalová, Radim J Vašut, Petra Šarhanová, Jiøí Neustupa, and Pavel Škaloud. Dna content variation and its significance in the evolution of the genus micrasterias (desmidiales, streptophyta). *PLoS ONE*, 9(1):e86247, 2014.

Michael Prigge, Denichiro Otsuga, José Alonso, Joseph Ecker, Gary Drews, and Steven Clark. Class III homeodomain-leucine zipper gene family members have overlapping, antagonistic, and distinct roles in *Arabidopsis* development. *The Plant Cell*, 17(1):61–76, 2005.

Zhi Qi, Rajeev Verma, Chris Gehring, Yube Yamaguchi, Yichen Zhao, Clarence Ryan, and Gerald Berkowitz. Ca^{2+} signaling by plant *Arabidopsis thaliana* Pep peptides depends on AtPepR1, a receptor with guanylyl cyclase activity, and cGMP-activated Ca^{2+} channels. *Proceedings of the National Academy of Sciences*, 107(49):21193–21198, 2010.

Stefanie Ranf, Lennart Eschen-Lippold, Pascal Pecher, Justin Lee, and Dierk Scheel. Interplay between calcium signalling and early signalling elements during defence responses to microbe-or damage-associated molecular patterns. *The Plant Journal*, 68(1):100–113, 2011.

Stefanie Ranf, Nicolas Gisch, Milena Schäffer, Tina Illig, Lore Westphal, Yuriy A Knirel, Patricia M Sánchez-Carballo, Ulrich Zähringer, Ralph Hückelhoven, Justin Lee, et al. A lectin s-domain receptor kinase mediates lipopolysaccharide sensing in arabidopsis thaliana. *Nature Immunology*, 16(4):426, 2015.

Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. Science forum: the human cell atlas. *Elife*, 6:e27041, 2017.

Silke Robatzek, Delphine Chinchilla, and Thomas Boller. Ligand-induced endocytosis of the pattern recognition receptor fls2 in *Arabidopsis*. *Genes and Development*, 20(5):537–542, 2006.

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, 2011.

Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edger: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2009.

José Rodríguez-Martínez, Aaron Reinke, Devesh Bhimsaria, Amy Keating, and Aseem Ansari. Combinatorial bZIP dimers display complex DNA-binding specificity landscapes. *eLife*, 6, 2017.

Assaf Rotem, Oren Ram, Noam Shoresh, Ralph A Sperling, Alon Goren, David A Weitz, and Bradley E Bernstein. Single-cell chip-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, 33(11):1165, 2015.

Paul Rushton, Imre Somssich, Patricia Ringler, and Qingxi Shen. WRKY transcription factors. *Trends in Plant Science*, 15(5):247–258, 2010.

Sabrina Sabatini, Renze Heidstra, Marjolein Wildwater, and Ben Scheres. SCARECROW is involved in positioning the stem cell niche in the *Arabidopsis* root meristem. *Genes and Development*, 17(3):354–358, 2003.

Robert Sablowski and Marcelo Carnier Dornelas. Interplay between cell growth and cell cycle in plants. *Journal of Experimental Botany*, 65(10):2703–2714, 2013.

Yusuke Saijo, Nico Tintor, Xunli Lu, Philipp Rauf, Karolina Pajerowska-Mukhtar, Heidrun Häweker, Xinnian Dong, Silke Robatzek, and Paul Schulze-Lefert. Receptor quality control in the endoplasmic reticulum for plant innate immunity. *The EMBO Journal*, 28(21):3439–3449, 2009.

Silvia Santopolo, Alessandra Boccaccini, Riccardo Lorrai, Veronica Ruta, Davide Capauto, Emanuele Minutello, Giovanna Serino, Paolo Costantino, and Paola Vittorioso. DOF AFFECTING GERMINATION 2 is a positive regulator of light-mediated seed germination and is repressed by DOF AFFECTING GERMINATION 1. *BMC Plant Biology*, 15(1):72, 2015.

Ananda Sarkar, Marijn Luijten, Shunsuke Miyashima, Michael Lenhard, Takashi Hashimoto, Keiji Nakajima, Ben Scheres, Renze Heidstra, and Thomas Laux. Conserved factors regulate signalling in *Arabidopsis thaliana* shoot and root stem cell organizers. *Nature*, 446(7137):811–814, 2007.

Serge Savary, Andrea Ficke, Jean-Noël Aubertot, and Clayton Hollier. Crop losses due to diseases and their implications for global food production losses and food security. *Food Security*, 4(4):519–537, 2012.

Fabian Schweizer, Patricia Fernández-Calvo, Mark Zander, Monica Diez-Diaz, Sandra Fonseca, Gaétan Glauser, Mathew G Lewsey, Joseph R Ecker, Roberto Solano, and Philippe Reymond. *Arabidopsis* basic helix-loop-helix transcription factors MYC2, MYC3, and MYC4 regulate glucosinolate biosynthesis, insect performance, and feeding behavior. *The Plant Cell*, 25(8):3117–3132, 2013.

Roded Sharan, Ivan Ovcharenko, Asa Ben-Hur, and Richard M Karp. CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics*, 19:i283–i291, 2003.

Christine Shulse, Benjamin Cole, Gina Turco, Yiwen Zhu, Siobhan Brady, and Diane Dickel. High-throughput single-cell transcriptome profiling of plant cell types. *bioRxiv*, 2018.

Matthew Slattery, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordân, and Remo Rohs. Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences*, 39(9):381–399, 2014.

Viggo Smedegaard-Petersen and Karl Tolstrup. The limiting effect of disease resistance on yield. *Annual Review of Phytopathology*, 23(1):475–490, 1985.

Vaclav Smil. *Feeding the world: A challenge for the twenty-first century*. MIT press, 2001.

Erin Sparks. *Arabidopsis* seedling cartoon. figshare. figure., 2017.

Simon Stael, Przemyslaw Kmiecik, Patrick Willems, Katrien Van Der Kelen, Nuria S Coll, Markus Teige, and Frank Van Breusegem. Plant innate immunity–sunny side up? *Trends in Plant Science*, 20(1):3–11, 2015.

Yvonne Stahl, René H. Wink, Gwyneth C. Ingram, and Rüdiger Simon. A signaling module controlling the stem cell niche in *Arabidopsis* root meristems. *Current Biology*, 19(11):909–914, 2009.

Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl_1):D535–D539, 2006.

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

Ralf Stracke, Oliver Jahns, Matthias Keck, Takayuki Tohge, Karsten Niehaus, Alisdair Fernie, and Bernd Weisshaar. Analysis of production of flavonol glycosides-dependent flavonol glycoside accumulation in *Arabidopsis thaliana* plants reveals MY11-, MYB12- and MYB111-independent flavonol glycoside accumulation. *New Phytologist*, 188(4):985–1000, 2010.

Richard N Strange and Peter R Scott. Plant disease: a threat to global food security. *Annual Review of Phytopathology*, 43:83–116, 2005.

Kushal Suryamohan and Marc S Halfon. Identifying transcriptional cis-regulatory modules in animal genomes. *Wiley Interdisciplinary Reviews: Developmental Biology*, 4(2):59–84, 2015.

William Swindell, Marianne Huebner, and Andreas Weber. Transcriptional profiling of *Arabidopsis* heat shock proteins and transcription factors reveals extensive overlap between heat and non-heat stress response pathways. *BMC Genomics*, 8(1):125, 2007.

Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, 2014.

Nico Tintor, Annegret Ross, Kazue Kanehara, Kohji Yamada, Li Fan, Birgit Kemmerling, Thorsten Nürnberger, Kenichi Tsuda, and Yusuke Saijo. Layered pattern receptor signaling via ethylene and endogenous elicitor peptides during *Arabidopsis* immunity to bacterial infection. *Proceedings of the National Academy of Sciences*, 110(15):6211–6216, 2013.

Jan Van de Velde, Ken Heyndrickx, and Klaas Vandepoele. Inference of transcriptional networks in *Arabidopsis* through conserved noncoding sequence analysis. *The Plant Cell*, 26(7):2729–2745, 2014.

Claudia Van den Berg, Viola Willemsen, Willem Hage, Peter Weisbeek, and Ben Scheres. Cell fate in the *Arabidopsis* root meristem determined by directional signalling. *Nature*, 378(6552):62–65, 1995.

Jimmy Vandel, Oceane Cassan, Sophie Lebre, Charles-Henri Lecellier, and Laurent Brehelin. Probing transcription factor combinatorics in different promoter classes and in enhancers. *bioRxiv*, page 197418, 2017.

Klaas Vandepoele, Tineke Casneuf, and Yves Van de Peer. Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics. *Genome Biology*, 7(11):R103, 2006.

Klass Vandepoele, Mauricio Quimbaya, Tine Casneuf, Lieven De Veylder, and Yves Van de Peer. Unraveling transcriptional control in *Arabidopsis* using cis-regulatory elements and coexpression networks. *Plant Physiology*, 150(2):535–546, 2009.

Vladimir Vovk. Combining p-values via averaging. *arXiv*, 2012.

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

Jamie Waese, Jim Fan, Asher Pasha, Hans Yu, Geoffrey Fucile, Ruian Shi, Matthew Cumming, Lawrence Kelley, Michael Sternberg, Vivek Krishnakumar, Erik Ferlanti, Jason Miller, Chris Town, Wolfgang Stuerzlinger, and Nicholas Provart. ePlant: Visualizing and exploring multiple levels of data for hypothesis generation in plant biology. *The Plant Cell*, 29(8):1806–1821, 2017.

Liam Walker, Clare Boddington, Dafyd Jenkins, Ying Wang, Jesper T Grønlund, Jo Hulsmans, Sanjeev Kumar, Dhaval Patel, Jonathan D Moore, Anthony Carter, Siva Samavedam, Giovanni Bomono, David S Hersh, Gloria M Coruzzi, Nigel J Burroughs, and Miriam L Gifford. Root architecture shaping by the environment is orchestrated by dynamic gene expression in space and time. *The Plant Cell*, tpc.00961.2016, 2017.

Ludo Waltman and Nees Jan Van Eck. A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B*, 86(11):471, 2013.

J. Wan, X.-C. Zhang, D. Neece, K. M. Ramonell, S. Clough, S.-y. Kim, M. G. Stacey, and G. Stacey. A lysm receptor-like kinase plays a critical role in chitin signaling and fungal resistance in *Arabidopsis*. *The Plant Cell*, 20(2):471–481, 2008.

Todd Wasson and Alexander Hartemink. An ensemble model of competitive multi-factor binding of the genome. *Genome Research*, 19(11):2101–2112, 2009.

Jos Wendrich, Barbara Möller, Song Li, Shunsuke Saiga, Rosangela Sozzani, Philip Benfey, Bert De Rybel, and Dolf Weijers. Framework for gradual progression of cell ontogeny in the *Arabidopsis* root meristem. *Proceedings of the National Academy of Sciences*, 114(42):E8922–E8929, 2017.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `http://ggplot2.org`.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Kyoung-Jae Won, Bing Ren, and Wei Wang. Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biology*, 11(1):R7, 2010.

Ines Wyrsch, Ana Domínguez-Ferreras, Niko Geldner, and Thomas Boller. Tissue-specific *FLAGELLIN-SENSING 2* (*FLS2*) expression in roots restores immune responses in *Arabidopsis fls2* mutants. *New Phytologist*, 206(2):774–784, 2015.

Shuping Xing, Niklas Wallmeroth, Kenneth W Berendzen, and Christopher Grefen. Techniques for the analysis of protein-protein interactions in vivo. *Plant Physiology*, 171(2):727–758, 2016.

Yube Yamaguchi and Alisa Huffaker. Endogenous peptide elicitors in higher plants. *Current Opinion in Plant Biology*, 14(4):351–357, 2011.

Yube Yamaguchi, Gregory Pearce, and Clarence Ryan. The cell surface leucine-rich repeat receptor for AtPep1, an endogenous peptide elicitor in *Arabidopsis*, is functional in transgenic tobacco cells. *Proceedings of the National Academy of Sciences*, 103(26):10104–10109, 2006.

Yube Yamaguchi, Alisa Huffaker, Anthony Bryan, Frans Tax, and Clarence Ryan. PEPR2 is a second receptor for the Pep1 and Pep2 peptides and contributes to defense responses in *Arabidopsis*. *The Plant Cell*, 22(2):508–522, 2010.

Mingzhu Yin, Yanping Wang, Lihua Zhang, Jinzhu Li, Wenli Quan, Li Yang, Qingfeng Wang, and Zhulong Chan. The *Arabidopsis* Cys2/His2 zinc finger transcription factor ZAT18 is a positive regulator of plant tolerance to drought stress. *Journal Of Experimental Botany*, 68(11):2991–3005, 2017.

Chun-Ping Yu, Jinn-Jy Lin, and Wen-Hsiung Li. Positional distribution of transcription factor binding sites in arabidopsis thaliana. *Scientific Reports*, 6:25164, 2016.

Wenli Zhang, Tao Zhang, Yufeng Wu, and Jiming Jiang. Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in *Arabidopsis*. *The Plant Cell*, pages tpc–112, 2012.

Xiannian Zhang, Tianqi Li, Feng Liu, Yaqi Chen, Jiacheng Yao, Zeyao Li, Yanyi Huang, and Jianbin Wang. Comparative analysis of droplet-based ultra-high-throughput single-cell rna-seq systems. *Molecular Cell*, 2018.

Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017.

Ruiqin Zhong, Elizabeth Richardson, and Zheng-Hua Ye. The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in *Arabidopsis*. *The Plant Cell*, 19(9):2776–2792, 2007.

Qing Zhou and Wing H Wong. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proceedings of the National Academy of Sciences*, 101(33):12114–12119, 2004.

Rapolas Zilionis, Juozas Nainys, Adrian Veres, Virginia Savova, David Zemmour, Allon M Klein, and Linas Mazutis. Single-cell barcoding and sequencing using droplet microfluidics. *Nature Protocols*, 12(1):44, 2017.

Cyril Zipfel, Silke Robatzek, Lionel Navarro, Edward Oakeley, Jonathan Jones, Georg Felix, and Thomas Boller. Bacterial disease resistance in *Arabidopsis* through flagellin perception. *Nature*, 428(6984):764–767, 2004.

Cyril Zipfel, Gernot Kunze, Delphine Chinchilla, Anne Caniard, Jonathan DG Jones, Thomas Boller, and Georg Felix. Perception of the bacterial pamp ef-tu by the receptor efr restricts agrobacterium-mediated transformation. *Cell*, 125(4):749–760, 2006.