Regulatory Networks in Plant Stem Cells:

An Integrated Bioinformatic and

Developmental Biology Analysis

Thesis submitted for the award Doctor of Philosophy (Biosciences)

Alexander James Murison

2013

I declare that, except where indicated by specific reference, the work submitted is the result of my own investigation and the views expressed are my own

Alexander James Murison

I declare that no portion of the work presented has been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award

Alexander James Murison

SHOOT MERISTEMLESS (*STM*) encodes a transcription factor in *Arabidopsis* essential for ensuring correct stem cell fate. *STM* is known to impinge on a number of key regulatory processes such as cytokinin synthesis and the cell cycle, and interacts with other core regulatory genes such as *CUP-SHAPED COTYLEDON1* (*CUC1*).

In this study inducible *STM* over-expression and RNAi-mediated downregulation over a time course experiment have been used to identify the genes which form *STM's* gene regulatory network (GRN). These results reveal for the first time how *STM* over-expression and knockout phenotypes are mediated and identified the temporal order of transcriptomic changes following *STM* over-expression. A Bayesian network approach further refined the GRN - identifying conditional dependencies among regulated TFs and core signalling components from an independent dataset (>2,000 experiments). Predictions of direct targets from the network have been tested, demonstrating a high degree of accuracy.

Interplay between *STM* and *CUC1* is a biologically interesting sub-module of the *STM* GRN, with unusual dynamics. Via gene expression and microscopy experiments it has been shown that *STM* positively regulates the *CUC1*-targetting microRNA *miR164c*. Mathematical modelling approaches show that this is consistent with a model in which the boundary is the site of highest *STM* mRNA production via CUC1, and STM movement with *miR164c* upregulation produces the observed spatial distributions of both proteins. These relationships recast the boundary zone as a particularly dynamic region of the shoot apical meristem (SAM) and significantly develop our understanding of *STM* developmental context.

Chapter Index

Figure Index

Abbreviations

ANOVA	Analysis of Variance
AUX	Auxin
cDNA	Complementary deoxyribonucleic acid
ChIP	Chromatin Immunoprecipitation
CHX	Cycloheximide
CK	Cytokinin
Col-0	Columbia (*Arabidopsis* ecotype)
CZ	Central Zone
DAG	Days After Germination
DAG	Directed Acyclic Graph
DAS	Days After Sowing
DEX	Dexamethasone
DMSO	Dimethylsulfoxide
DNA	Deoxyribonucleic Acid
DNAase	Deoxyribonuclease
EBI	European Bioinformatics Institute
EMSA	Electromotive shift assay
EV	Empty Vector
GA	Giberellic Acid
GM	Growth Media
GO	Gene Ontology
GR	Glucocorticoid Receptor
GRN	Gene Regulatory Network

GUI	Graphical User Interface
IQR	Interquartile Range
KAN	Kanamycin
LIMMA	Linear Models for Microarray Data
Ler	Landsberg Erecta (*Arabidopsis* ecotype)
Log	Logarithm
mRNA	messenger Ribonucleic Acid
NGS	Next Generation Sequencing
OC	Organizing Centre
ODE	Ordinary Differential Equation
OE	Over-Expression
PCA	Principal Components Analysis
PM	Perfect Match
PPT	Phosphinotrycin
PZ	Peripheral Zone
qRT-PCR	Quantitative (Reverse Transcriptase) Real Time Polymerase Chain Reaction
RMA	Robust Multichip Average
RNAi	Ribonucleic acid interference
SAM	Shoot Apical Meristem
SOM	Self Organizing Map
TAIR	The Arabidopsis Information Resource
TF	Transcription Factor
Wt	Wild Type

1.1.1 The Arabidopsis shoot apical meristem (SAM)

Plant growth is largely indetermate and characterised by the continuous production of new organs. This depends on pools of undifferentiated stem cells maintained within regular, organized structures known as meristems. Meristems are regions of undifferentiated, mitotically active cells, located primarily at the tips of the shoot and the root and are responsible for post-embryonic plant growth. As plants continue to initiate new organs in a modular manner throughout their lifecycle, it is essential for their correct growth and development that stem cells numbers and cell fate within meristems are both tightly controlled so that there is a supply of stem cells constantly available for continued organogenesis.

During embryogenesis, two meristems are formed at opposite poles of the developing embryo. Although subsequent meristematic tissue is formed, ultimately all post-embryonic plant tissues are derived from one of these two meristems. The root apical meristem (RAM) will subsequently initiate almost all root tissues, while the shoot apical meristem (SAM) forms the ultimate source of all postembryonic aerial tissue such as leaves and flowers. Despite parallels in their functions these two meristems have different structures and regulatory mechanisms, being distinct at the transcriptomic and organizational levels (Review: Barton, 2010).

In *Arabidopsis thaliana* (Arabidopsis) the presumptive SAM is first visible during embryogenesis as a cluster of cells which lie between the nascent cotyledons. As the cotyledons form as the embryo moves through the heart and torpedo stages, the SAM begins to take on a regular dome shape, with the pool of undifferentiated stem cells maintained near the top of the SAM dome. From this point, the SAM is characterized as a shallow dome shaped structure, with a pool of slowly dividing stem cells at the apex.

(Bowman and Eshed, 2000)

The post-germination Arabidopsis SAM has a highly regular morphological structure (Steeves and Sussex, 1989). It classically consists of three cell layers, the uppermost two (L1 and L2) are collectively referred to as the tunica, while the lower layers of cells, L3, covered by the tunica, is referred to as the corpus of the meristem. It can also be divided into a series of zones by cell characteristics. The undifferentiated stem cells are maintained within the central zone (CZ) at the apex of the SAM dome (Fletcher and Meyerowitz, 2000). Slow cell division gradually results in cells entering the peripheral zone (PZ) or organogenic zone. Within the PZ, localized accumulation of the phytohormone auxin leads to an outgrowth (primordium) with more rapid cell division which develops into a new organ such as a leaf (Heisler et al, 2005). A boundary zone between the CZ and PZ can be identified by having a distinct transcriptomic state from those cells on either side, expressing genes such as the *CUP-SHAPED COTYLEDON* family of transcription factors (Gordon et al, 2007).

Below the CZ lies the organizing centre (OC) of the meristem and a rib zone (RZ) from which the pith of the stem is formed. The OC is the source of signalling components which are necessary to maintain numbers of undifferentiated stem cells (Laux et al, 1996, Mayer et al, 1998) in the overlaying CZ. A schematic representation and overlay on an image of the Arabidopsis SAM is shown in Figure 1.1 reproduced from (Bowman and Eshed 2000).

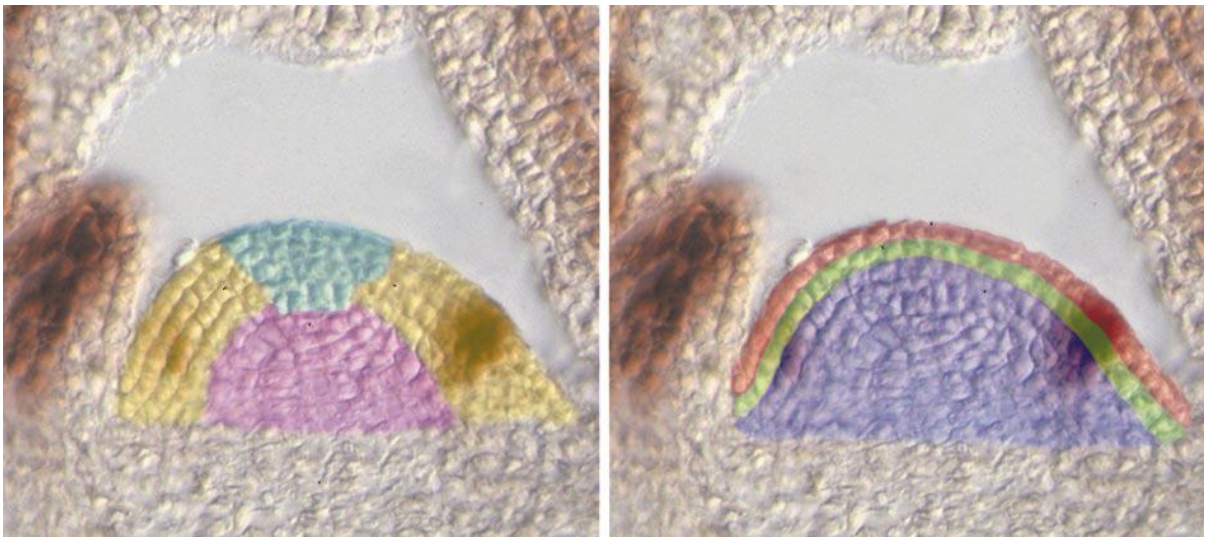


Figure 1.1 – Regionalization in the SAM.

A) Schematic of SAM regions. The rib zone containing the organizing centre is shown in purple, the central zone in cyan, and the peripheral zone is marked in yellow, reproduced from Bowman and Eshed (2000).

B) Schematic of SAM layers. The L1 layer is shaded red, the L2 in green and the remainder of the SAM is shaded in blue. Reproduced from Bowman and Eshed (2000).

1.1.2 Hormone balance in the SAM

Cytokinin (CK) is an adenine-derived phytohormone which is strongly associated with both promotion of cell division and growth, though it impinges on numerous other cellular processes such as senescence and branching (Müller & Sheen, 2007). It has been known since the 1950s that application of CK to cell culture is sufficient to stimulate growth and proliferation (Skoog & Miller, 1957), though the exact effect depends upon the ratio of cytokinin to the phytohormone auxin. In tissue culture, at higher CK to auxin ratios, shoot development is stimulated, in the converse situation, root development is stimulated, and at intermediate concentrations undifferentiated callus proliferates.

In this way the ratio of CK and auxin determines the mode of growth promoted by CK (Skoog & Miller, 1957).

CK is perceived by histidine kinase receptors such as *AHK4/CRE1/WOL* which upon recognition of CK trigger signal induction across the plasma membrane via histidine phosphotransfer proteins (Inoue et al, 2001; Müller & Sheen, 2007). Phosphorylation of B-type ARABIDOPSIS RESPONSE REGULATOR (ARR) proteins subsequently trigger a transcriptional response to CK detection. There are two classes of ARRs, B-Type *ARRs* are transcriptional activators, while A-Type ARRs are induced by the B-types and lack a DNA binding domain, and produce a negative feedback response upon CK production, though the exact mechanism remains unclear (Beuchel et al, 2009). Thus negative feedback ensures that CK levels are kept in balance within the SAM, and ensures tight control of the first cell cycle checkpoint.

As a requirement for continuous stem cell maintenance, stem cells must be kept in a mitotically active, undifferentiated state, and CK appears to play an important role in this. Laufs et al (1998) showed by laser scanning confocal microscopy that there are clear differences in the mitotic indices across the SAM, with lowest mitotic activity in the central zone, while mitotic activity was increased at the site of presumptive primordia, thus these differences imply that progression through the cell cycle is a tightly regulated process with spatial distinctions.

Arabidopsis has a relatively complex set of cell-cycle related genes. Despite the relatively large number of components impinging upon the decision by a cell to undergo mitotic cycling (Menges et al, 2005), understanding the process can be simplified by focussing on the two major checkpoints which exist during the cell cycle (G_1-S and G_2-M). The G_1-S checkpoint (demarcating the beginning of new DNA synthesis after a first gap phase) is dependent upon *D-TYPE CYCLINS (CYCDs),* which form a complex with

CYCLIN DEPENDENT KINASES (CDK) such as CDKA and hyperphosphorylate the *RETINOBLASTOMA-RELATED* (*RBR)* protein to trigger the activation of *E2F* family transcription factors. *E2F* transcription factors subsequently upregulate genes responsible for DNA synthesis (reviewed in DeJager & Murray, 1999). Over-expression of *CYCD3;1* leads to increased mitotic cycling and hence must be tightly regulated in order to prevent over-proliferation (Dewitte et al, 2003).

There is a close linkage between CK and regulation of the cell cycle through the three *CYCD3* family genes *CYCD3;1*, *CYCD3;2* and *CYCD3;3*. Increased CK signalling triggers the expression of the *CYCD3* genes (Riou-Khamlichi et al,1999; Dewitte et al, 2007), ectopic *CYCD3;1* results in callus able to grow and green without exogenous CK (Riou-Khamlichi et al, 1999), and CK responses are reduced in the absence of *CYCD3* (Dewitte et al, 2007). The CYCD3 genes are all expressed at various levels in the SAM. This, combined with their response to CK suggests they may have an important role in mediating stem cell division rates in response to CK, as recently shown by Scofield et al (2013).

1.1.3 Organogenesis and auxin within the SAM

As the purpose of the SAM is to house a pool of stem cells from which new organs can be derived continuously during the life cycle of the plant, proper control of organogenesis is central to SAM function. Organogenesis in plants is a highly regular process, although the exact pattern of organ emergence (phyllotaxis) varies between plant species. In Arabidopsis, the first two primordia emerge in opposite phyllotaxis, but subsequent organ primordia are formed in a regular spiral pattern around the SAM with a divergence angle between new primordia of 137.5^{o} – the golden ratio (Steeves & Sussex, 1989).

The key hormone regulating phyllotaxis in plants is auxin, indole-3-acetic acid or IAA.

The first observable event in organogenesis is the formation of an auxin maximum at the site of the next primordium (Review: Carraro et al, 2006). Auxin is produced in more basal tissues, and its presence within the SAM is a result of auxin transport (Reinhardt et al, 2003; Smith et al, 2006). Within the SAM auxin becomes concentrated at the sites of organ primordia formation as a consequence of polar auxin transport mediated by the PIN proteins. These are cell membrane bound proteins, which direct unidirectional auxin efflux through their presence on specific membranes of cells (Geldner et al, 2001). Hence the direction of auxin flow can be predicted from the polarised location of PIN proteins (primarily PIN1 in the shoot). In auxin transport mutants such as *pinoid,* and *pin1 (pin formed1)*, the failure of polar auxin flows produces a barren pin-shaped inflorescence (Bennet et al, 1995; Okada et al, 2001) which can be rescued (induced to form primordia) by exogenous auxin application. Thus, proper phyllotaxis is dependent upon the proper formation of auxin maxima (Reinhardt et al, 2003).

Remarkably, the highly regular pattern of auxin maxima and subsequent organogenesis can be produced by relatively simple assumptions. Mathematical modelling has shown that it is sufficient for auxin transport to occur against a concentration gradient (from lower to higher regions of auxin concentration) to produce the observed patterning of auxin and correct positioning of new organ primordia (Jonsson et al, 2006), but as of yet there is no clear proof of the exact proteins involved, or whether the true biological mechanism is as simple – as a consequence simulation of auxin dynamics has proceeded faster than our understanding of the underlying biology (Vernoux et al, 2010), and mathematical modelling is still playing a key role in determining possible modes of action (e.g. Fozard et al, 2013).

Incipient primordia initially recruit cells from the L1 layer of the SAM, with around 60

cells eventually expressing primordium markers (Grandjean et al, 2004). The first detectable physical changes observed in cells involved in primordium formation are L2 cells which expand and then divide periclinally. Consistent with this change in state the site of incipient primordia a number of transcriptomic changes take place, notably the repression of *KNOTTED1-like HOMEOBOX (KNOX)* family gene expression (see below; Long et al, 1996). As well as, and probably as a consequence of, changes in the transcriptome of incipient primordia, cell expansion and proliferation rates are both faster at the meristem flank than at the meristem summit (Grandjean et al, 2004; Reddy et al, 2004).

1.1.4 Gene expression patterns reflect distinct morphological zones in the SAM

Gene expression differences support the existence of distinct function zones within the SAM. The small *CLAVATA3* (*CLV3*) ligand is often used as a marker for the stem cells and hence the CZ, whereas the *WUSCHEL* (*WUS)* gene is restricted to the OC (Mayer et al, 1998). The *CUP-SHAPED COTYLEDON* family of genes (*CUC1, CUC2, CUC3)* are restricted to the boundary between the meristem and organ primordia (Aida et al, 1997). Finally, genes such as *FILAMENTOUS FLOWER* demarcate the emerging primordia in the PZ (Sawa et al, 1999). These markers, which also represent genes with key functions in their specific zones, have been used to express fluorescent reporters that enabled the flow sorting of the different cell populations and their analysis using microarrays. This revealed that the different regions have clear and distinct global transcriptomic states (Yadav et al, 2009), which reflect the distinctions between the various morphological regions. Thus it can be clearly seen that there are a number of genes whose expression reflects the morphological structure of the SAM and many of these genes which define morphological patterning have vital roles in proper SAM formation.

1.1.5 The *WUS-CLV* loop maintains stem cell numbers in the SAM

The *WUSCHEL* (*WUS*) gene is essential for SAM formation. In *wus* mutants, the SAM repeatedly initiates and terminates as an aberrant flat structure (Laux et al, 1996). *WUS* is expressed in the OC and appears to co-ordinate expression of a number of genes which are necessary for specifying stem cell identity, as without its expression, the overlying pool of stem cells is not maintained. In particular studies investigating downstream genes of *WUS* (Liebfried et al, 2005; Busch et al, 2010) demonstrated that *WUS* has a pronounced effect upon the response to the phytohormone cytokinin, by directly repressing the expression of multiple *A-TYPE ARRs* which are induced by CK and dampen CK responses (Beuchel et al, 2009). However, since *WUS* does not move out of its domain of expression in the organizing centre, this implies that the stem cell identify signal produced must affect cells in the layer above it – i.e. if functions in a non-cell autonomous manner (Groβhardt & Laux, 2003).

Although the exact means by which this is accomplished remain unclear (Rieu & Laux, 2009), *WUS* induces expression of the small signalling peptide *CLAVATA3* (Laufs et al, 1998) in the overlying stem cells. Indeed *CLV3* is regarded as a CZ marker due to its restricted localization to CZ stem cells. *CLV3* acts through a pair of kinase receptors (*CLAVATA1, CLAVATA2)* (Clark et al, 1993; Clark et al, 1997; Jeong et al, 1999) and feeds back to repress *WUS* expression in the OC. The *clv1* and *clv3* mutants produce a similar phenotype (Clark et al, 1993; Clark et al, 1995) with an enlarged shoot apical meristem with larger numbers of undifferentiated stem cells than the wild type SAM. *CLAVATA2* forms a receptor complex with *CLAVATA1* (Jeong et al, 1999) though the mutant phenotype for *clv2* is similar it is less severe than that of *clv1* or *clv3* (Kayes and Clark, 1998). The *CLAVATA* receptor complex (consisting of *CLV1* and *CLV2)* recognises *CLV3* following its export to the apoplastic space between cell walls -

transport to the apoplast permits CLV3 to transport between cells while still being recognized by the extracellular domain of the *CLV* complex (Rojo et al, 2002). Upon recognition of *CLV3*, the *CLV* complex triggers the downregulation of *WUS* (Brand et al, 2002), explaining the *clv* mutant phenotype, which mirrors the effect of overexpressing *WUS*.

Thus, *CLV* and *WUS* together form a regulatory loop, the disruption of which has dramatic impacts upon SAM formation. *WUS* promotes stem cell identify in overlaying cells, a process moderated by its up-regulation of *CLV3. CLV3* which traffics through the apoplast to the CZ and binds to the CLV complex triggering repression of *WUS*. This restricts *WUS* to the region beneath the stem cell population, however, as *WUS* is necessary to upregulate *CLV3*, homeostasis is maintained (Mayer et al, 1998, Schoof et al, 2000).

1.1.6 The *CUP-SHAPED COTYLEDON* genes are boundary genes also essential for initial SAM formation

The three NAC family *CUP-SHAPED COTYLEDON* genes also play a vital role in SAM specification. *CUC1, CUC2* and *CUC3* encode transcription factors with a high degree of functional redundancy evidenced by the requirement of double knockouts to display a strong phenotype (Aida et al, 1997; Hibara et al, 2006). Their name derives from the characteristic cup-shaped fused cotyledons which form in the double mutants of *cuc1* and *cuc2*, these lack a SAM and failing to develop beyond seedling developmental stage. (Aida et al, 1997). Single *cuc* mutants lack a strong phenotype only displaying occasional cotyledonary fusion along a single edge (Aida et al, 1997).

The *CUP-SHAPED COTYLEDON* genes are originally expressed throughout the SAM during embryogenesis, but are later restricted to the boundary region between

emerging primordia and the SAM itself (Aida et al, 1999.) While other factors may be involved, a key player in restricting *CUC1* expression is the miR164 family of microRNAs (*miR164a, b* and *c*) which degrade *CUC1* and *CUC2* mRNA (Laufs et al, 2004), miR164 over-expression phenocopies the *cuc1 cuc2* double mutant, indicating that this is sufficient to ablate *CUC1* and *CUC2* expression (Laufs et al, 2004). Conversely, the expression of microRNA-resistant *CUC1* is sufficient to permit its expansion into the expression domain of *miR164c* in the centre of the SAM (Sieber et al, 2007) demonstrating that *miR164* activity is required to maintain the normal expression pattern of *CUC* genes.

1.2 *SHOOT MERISTEMLESS* is a core regulator of SAM function

The *SHOOT MERISTEMLESS (STM)* gene was identified in mutants that lack a SAM, and so germinate to produce only a pair of cotyledons with fused bases (Barton & Poethig, 1993). Cloning of Arabidopsis *STM* showed that it encodes a *KNOTTED1-LIKE HOMEOBOX (KNOX)* transcription factor sharing close homology with its maize orthologue *KNOTTED1*, from which the name of its gene family originates (Long et al, 1996). The first mutant identified was the strong loss-of-function *STM* mutant *stm-1* which does not develop a SAM (Barton and Poethig, 1993). Subsequently an allelic series of *stm* mutants have been identified, many of which are less severe, such as the point mutant *stm-2* in which the SAM forms, but is repeatedly aborted as stem cells are consumed into emerging primordia. Subsequently, new meristematic tissue is initiated from leaf axils, and growth continues in an abort-retry pattern (Clark et al, 1996). There are further weak point mutants such as *stm-6* (Endrizzi et al, 1996) which have very mild phenotypic defects – in *stm-6* there is occasional cotyledonary fusion and often the SAM terminates prematurely during the inflorescence stage. These genetic studies showed that *STM* was essential for SAM formation as its complete absence precluded

the formation of a SAM. They also demonstrated that it plays a continuing role in SAM maintenance, preventing the inappropriate formation of organs consuming the pool of stem cells housed in the SAM (Clark et al, 1996).

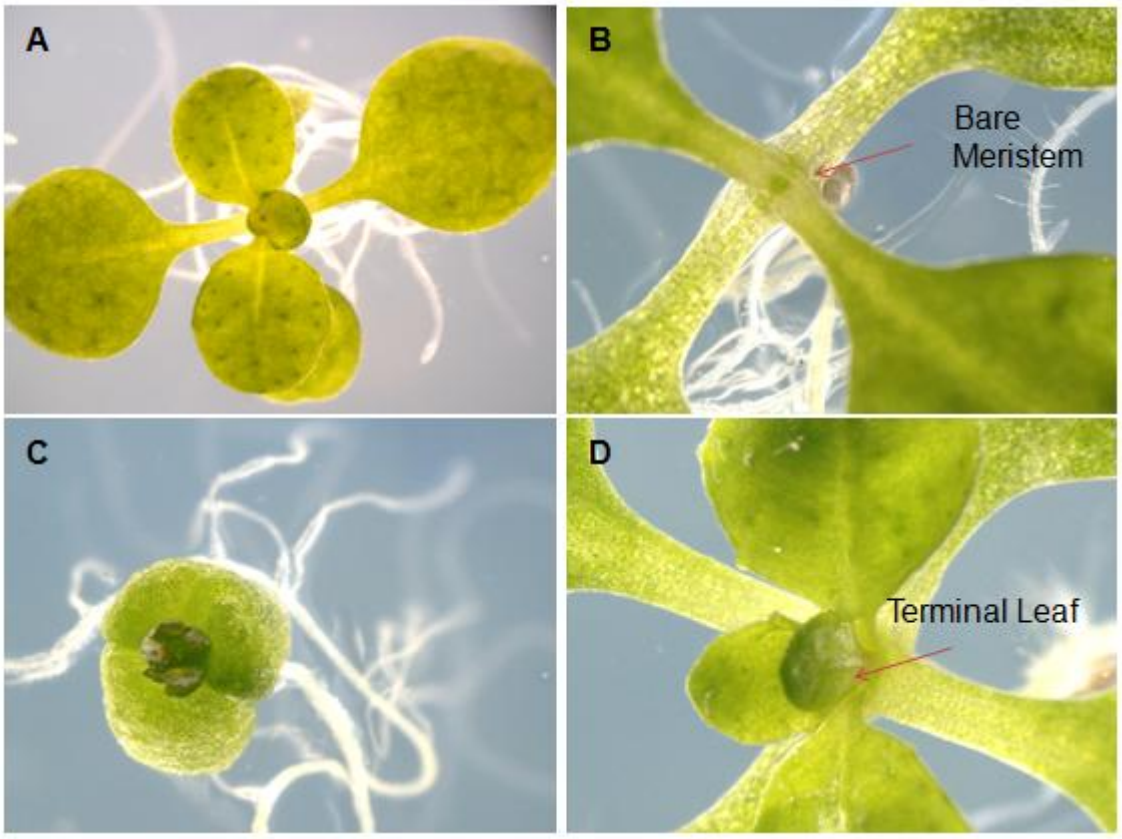


Figure 1.2 Over-Expression, knockout and mutant phenotypes of *STM.*
Images of 15 DAS Arabidopsis seedlings from A) wild type *Landsberg erecta*, B) constitutive *STM*-RNAi expressing line $SP^{1\text{-}8}$ – note the bare shoot apex phenotype. C) inducible ectopic *STM* over-expressor, S^{38} which has been grown on media containing 60μM DEX, D) *stm-2* mutant – note terminal leaf phenotype where stem cells have been consumed into emerging leaf. Images similar to those reported in Scofield et al (2007; 2013).

STM over-expression produces an equally striking phenotype (Figure 1.2 C; Scofield et al, 2007; 2013). In plants with constitutive ectopic *STM* over-expression, extremely small plants are observed, with extensive leaf lobbing, inhibited cell expansion and differentiation – resulting in cells in organs that more closely resemble meristematic cells than differentiated cells (Scofield et al, 2013). Dramatically, in *STM* over-expressing plants ectopic formation of new SAMs is observed upon the adaxial leaf

surface (Lenhard et al, 2002; Brand et al, 2002; Gallois et al, 2002). *STM* over-expresson alone has been shown to be sufficient to promote ectopic SAM formation (Brand et al, 2002; Scofield et al, 2013), though there was originally some disagreement over whether it was sufficient as Gallois et al (2002) suggested *WUS* over-expression was also required. Together these data suggest *STM* has a role in both sustaining a meristematic cell fate as well as promoting and maintaining meristem function.

While both *WUS* and *STM* are necessary to maintain correct meristem organization, *STM* does not depend upon *WUS* to trigger the formation of ectopic SAMs in over-expression lines (Brand et al, 2002), and ectopic expression of *WUS* also did not depend upon *STM* to produce observed phenotypes (Lenhard et al, 2002), and *STM* cannot compensate for the complete absence of *WUS* (Lenhard et al, 2002). However, multiple studies have shown that *STM* and *WUS* have synergistic effects, as over-expression of *STM* in a *wus* mutant background reduces the rate of ectopic meristem formation (Brand et al, 2002). Thus, *STM* is classed as a key regulator in the Arabidopsis SAM with interplay between *STM* and *WUS* whose nature is not fully understood.

1.2.1 *STM* expression within the SAM

Several studies have examined in detail the expression domain of *STM*. In situ analyses in seedlings reported that *STM* mRNA was found throughout the meristem apart from incipient primordia (Long et al, 1996; Long and Barton 1998; Aida et al, 1999). *STM* expression in the SAM is preceded by the *CUC* genes, which are initially expressed throughout the presumptive SAM until the early globular stage of embryogenesis, but subsequently become excluded from the CZ of the meristem and restricted to the boundary following *STM* activation (Aida et al, 1999). Post germination, *STM* is strongly downregulated at sites of incipient organ primordia formation, following

the formation of an auxin maximum, as shown in both fluorescent reporter studies (Gordon et al, 2007) and using in situ mRNA hybridisation analyses (Long et al, 1996), complete absence of *STM* was observed in incipient primordia. It is interesting to note that no intermediate stage was observed, which suggests that the downregulation is also fairly rapid.

There is some confusion however in the literature regarding the detailed expression pattern of *STM* using reporters driven by the *STM.* Laufs et al (2004) suggested using a two component system in which pSTM drove the alcohol-inducible *AlcR* transcription factor that *STM* promoter activity appeared focused in the boundary. Subsequent work by Gordon et al (2007) suggested a broader expression domain for a *pSTM:::STM-VENUS* line, however, it also appeared to show higher expression towards the boundary of the SAM. Part of the discrepancy may be between promoter and protein expression and could be explained by *STM* trafficking. Kim et al (2003) demonstrated that when STM was expressed exclusively in the L1 layer of the SAM, an *STM* protein reporter fusion was able to traffic into the L2 layer. Thus it is known that *STM* can move throughout the SAM, although the range of movement observed was less than for its maize ortholog *KN1* (Kim et al, 2002). Thus, as *STM* can move through the SAM, it is not necessarily the case that its protein would be detected in the domains of maximum promoter expression. However, there remains a discrepancy between mRNA in situ hybridisation experiments and fluorescent reporters, with the former suggesting a broader distribution of *STM* mRNA.

STM has a number of homologues with a high degree of sequence similarity in Arabidopsis, including *KNOTTED IN ARABIDOPIS THALIANA1/ BREVIPEDICELLUS* (*KNAT1/BP*), *KNAT2* and *KNAT6* which together with *STM* comprise a phylogenetic clade of 'Class-I *KNOX* genes' (Long et al, 1996; Bharathan et al, 1999). These

homologues have distinct but partially overlapping expression patterns with *STM*. Sequence conservation of the DNA binding homeodomain region of the protein is over 80% for *KNAT1/BP* and *KNAT2* – suggesting that these genes might have common targets (Long et al, 1996). Indeed, ectopic expression of *KNAT1/BP* can substitute for *STM* in meristem formation (Byrne et al, 2002; Scofield et al, 2013), though this functional redundancy is not observed in *stm* mutants due to competitive regulation of *KNAT1/BP* by *STM* and *ASYMMETRIC LEAVES1* (*AS1)* which promote and repress them respectively (Byrne et al, 2002; Scofield and Murray 2006).

Since *STM* is preceded by *CUC* expression in the embryonic SAM, this suggests it lies genetically downstream of *CUC* expression. Supporting this, the meristemless phenotype of the *cuc1 cuc2* double mutant is accounted for by lack of *STM* expression (Aida et al, 1997; Aida et al, 1999). The milder cotyledonary fusion phenotype of individual *cuc* mutants is also enhanced by weak *stm* mutations, indicating a close relationship between the two (Aida et al, 1999). During the work on this project, Spinelli et al (2011) also showed via electrophoretic mobility shift assays (EMSA) that *STM* directly binds the sequence in the promoter of *CUC1*.

1.2.2 *STM* recognition sequence and dimerization partners

As *STM* is a transcription factor, it functions through modulating the expression of other genes. Various studies have analyzed the cis-regulatory sequences recognized by *STM* and its orthologues (Hake et al, 2004; Spinelli et al, 2011) and all seem to feature a TGAC core as the minimal consensus sequence recognized by KNOX1 proteins. Needless to say, this is an extremely short sequence which is likely to occur frequently throughout the genome, and while this appears to be the basis for the consensus sequence, it is not the sole determinant of whether a potential site is an actual binding site. For example Spinelli et al (2011) demonstrated that in a promoter at which *STM*

binding triggered transcription there are multiple TGAC motifs within the binding site, though in this case, the presence of conserved flanking sequences important for binding was also suggested.

An additional complication in determining likely binding sites of *STM* is posed by the fact that STM forms heterodimers with *BEL1-LIKE HOMEODOMAIN* proteins, such as *BEL1* (Bellaoui et al, 2001), potentially through the conserved MEINOX domain (Figure 1.2). A large yeast-2-hybrid screen confirmed that *STM* can bind to seven out of 11 *BEL1-LIKE HOMEODOMAIN* proteins tested (Hackbusch et al, 2005). As the consequences of heterodimerization of these proteins with *STM* on its recognition sequence are uncertain, it is unknown how this affects *STM* downstream targets, or how the differing expression domains of *STM*'s binding partners may modulate its effect through different regions of its native expression domain.

Figure 1.3 Schematic of *STM* Protein.
The structure of the *STM* protein showing the MEINOX homeodomain in blue and the ELK domain necessary for nuclear import (Cole et al, 2006).

1.2.3 *SHOOT MERISTEMLESS* and its relationship to cytokinin biosynthesis

STM has been shown that to up-regulate a number of genes related to CK biosynthesis. The *ISOPENTENYL TRANSFERASE (IPT)* gene family, encode proteins responsible for the rate limiting, initial stage in CK biosynthesis. *IPT5* and *IPT7* have been both shown to be upregulated following *STM* induction (Jasinski et al, 2005; Yanai et al, 2005), although in the experiment performed by Jasinski et al (2005), the induction of *IPT5* was transient. The two experiments showed different dynamics of *STM* induction of *IPT7*. Yanai et al (2005) showed substantial induction between 2 and 24 hours, while

Jasinski et al (2005) showed only high level induction in plants induced for 10 days. Thus, although there is agreement that *IPT* gene expression is promoted by *STM*, there is disagreement about the rate and dynamics.

Consistent with this induction of *IPT* genes, *STM* promotes the expression of *A-TYPE ARABIDOPSIS RESPONSE REGULATOR* genes through induction of CK biosynthesis (Yanai et al, 2005; Jasinski et al, 2005). As these genes provide negative feedback on CK responses and are themselves promoted by CK signalling, their up-regulation is a strong indicator that *STM* is genuinely increasing the levels of CK present in the SAM (Beuchel et al, 2009). Scofield et al (2013) further showed that this regulation of ARRs depends on CK.

Part of *STM*'s function is therefore mediated by the promotion of CK biosynthesis. Yanai et al (2005) demonstrated that endogenous application of the cytokinin zeatin was sufficient to induce recovery via the formation of new meristems in the strong *stm-1* and *stm-11* mutants over half of the time. Additionally, when expressing *IPT7* under the *STM* promoter in an *stm-1* mutant, almost all plants recovered and formed new meristems (Yanai et al, 2005). As CK has been shown to promote *CYCD* genes and thus passage through the G_1-S checkpoint of the cell cycle, we would expect that this may be a means through which *STM* is affecting the cell cycle. However, it should be noted that the CK-rescued meristems in these plants lacked any proper organisation and effectively acted by increasing leaf production, albeit in a disorganised manner.

Finally, *STM*'s role in CK biosynthesis is known to be a conserved function among its close orthologues. Bolduc et al (2012) demonstrated that promoters of a number of CK-related genes are directly bound by *STM*'s maize homolog *KN1*, including components of the CK signalling pathway such as the orthologue of *WOL* which had not previously been identified as downstream of *KNOX1* genes.

1.2.4 *STM* represses gibberellic acid (GA) biosynthesis genes and is repressed by auxin

The phytohormone gibberellic acid (GA) is strongly associated with stem growth and elongation, as plants defective in GA signalling, such as the *GIBBERELLIC ACID INSENSITIVE (GAI)* mutant, demonstrate a dwarf phenotype (Silverstone et al, 2007), and constitutive GA response mutants such as *SPINDLY* show elongated stems. Correct *STM* function is known to depend upon GA as the *spindly-5* mutation enhances the intermediate *stm-2* mutant – leading to a failure of organogenesis resumption and failure of cotyledon separation (Hay et al, 2002). GA fulfils other roles aside from growth and elongation, in processes ranging from germination to fruit development (Review: Schwechheimer, 2008).

In tobacco and rice, *STM*'s orthologs have been shown to bind directly and repress the expression of a *GA20OXIDASE* gene which catalyses the rate limiting first stage in GA synthesis (Kusaba et al, 1998; Sakamoto et al, 2001). Similarly in maize, Bolduc et al (2012) showed that *KN1* directly binds to and represses *GA20OXIDASE*. *KNAT1/BP* and *KN1* have been shown to repress *GA20OXIDASE1* in-vivo when ectopically expressed in Arabidopsis leaves (Hay et al, 2002). Thus, it has been shown in multiple species that *GA20OXIDASES* are directly repressed by *KNOX1* proteins.

In Arabidopsis it has been shown that over-expression *STM* induces the expression of the *GA2OXIDASE* genes *GA2OX2* and *GA2OX4* - which catalyze the deactivation of GA in the SAM (Jasinski et al, 2005).It is also known that reducing GA activity by over-expression of GA catabolic enzyme *CYTOKININ OXIDASE 3* enhances *KNOX1* mutant phenotypes, while if GA is added exogenously *KNOX1* OE phenotypes are suppressed, and SAM function is disrupted (Yanai et al, 2005, Jasinski et al, 2005). However, direct repression of *GA20OXIDASE* genes has not been demonstrated in Arabidopsis, as it

has been in other species.

In contrast to its relatively direct effects on CK or GA levels and responses, *STM* has not been shown to modulate auxin concentrations, however, it is at least indirectly affected by the formation of auxin maxima, since *STM* is sharply downregulated upon the formation of a new auxin maximum (Heisler et al, 2005). *STM* would therefore appear to be in some way regulated by auxin, potentially through promotion of CK. Increased CK levels trigger the repression of PIN proteins, disrupting the auxin efflux pathway (Pernisová et al, 2008), meanwhile CK is repressed at sites of auxin maxima, resulting in an antagonistic relationship between AUX and CK (Su et al, 2011), where auxin is required for organ formation, but high CK is required for maintaining a pool of undifferentiated stem cells

1.3 Bioinformatics approaches

1.3.1 Systems wide approaches are increasingly necessary to interpret and understand complex phenotypes

Traditionally, attempts to understand molecular biology have taken a highly reductionist approach to understand the genetic and biochemical basis of individual phenotypes. While many of the major breakthroughs in biology over the past century have proven the strength of reductionist approaches, such as the dissection of core metabolic pathways, and identification of clear disease markers, in some fields of study we are approaching a point where systems approaches are necessary in order to garner a full understanding of biological phenotypes.

Even if we only focus on the molecular level, and attempt to exclude the interactions between organism and environment from our studies through experimental design, it is unclear that traditional genetic approaches can lead us to a full understanding of every

gene's function. Due to a mixture of genetic redundancy, feedback mechanisms and other effects, many genes have no observable phenotype when knocked out. For other genes, perturbation introduces a range of pleiotropic effects which can make the identification of core functionality difficult. In the case for *STM* – which has an extremely complex mutant phenotype which impinges on a wide range of developmental processes – it is clear that a large number of downstream processes will be affected by its perturbation. We thus have to conclude that the functionality of a single protein does not necessarily lead to a single phenotypic trait or defect. Rather, the interplay between different genes can lead to complex phenotypes, often with different functionality in different developmental stages and different tissues. In the *WUS-CLV* loop we have already seen an example where the functionality of a gene crucial to SAM function cannot be understood without fully considering not only its interactions with other genes and the interplay between different tissue types, but also how it is regulated at a spatial level.

The advent of high-throughput technologies such as microarrays and next-generation sequencing for transcriptomic data, has led to an explosion of data – indeed a single experiment performed using the ATH1 microarray platform in Arabidopsis yields data for over 22,000 genes. A large amount of this data has since become publically available. For example at the time of writing, the amount of transcriptomics data available in the Gene Expression Omnibus (Edgar and Lash, 2003) now exceeds 850,000 microarrays. As a consequence, particularly for model organisms, we now have access to sufficient data to enable the application of several systems-wide approaches to understanding genetic data which were not possible before.

Arabidopsis thaliana is a model species frequently used in plant genetics. It was initially selected for partially reductionist reasons as it has a very small genome compared to

many plant species (although its relatively quick growth cycle and ease of cultivation were also advantages). In spite of this, as a model species the amount of publically available Arabidopsis data is substantial, (for example data from over 21,000 ATH1 microarrays stored at the Gene Expression Omnibus at the time of writing). This makes Arabidopsis a particularly tractable organism for applying systems approaches because of the amount of freely available data which can be used to perform in-silico inference.

It has become commonplace for bioinformaticians to categorize data sources according to their 'omic' level – i.e. the genomic (DNA), transcriptomic (mRNA), proteomic (protein) etc. Different *in silico* and *in vivo* approaches exist to understand each omic level systematically, and to investigate the links between these omic levels. As we wish to investigate the means by which the transcription factor *STM* achieves its function, we expect to observe the clearest response to *STM* perturbation at the transcriptomic level. Thus, it is necessary to use methods for inferring the relationships between genes in order to properly understand and interpret mRNA expression data.

1.3.2 Biological tools for generating transcriptomic data - qRT-PCR, microarray and high throughput sequencing technology

A number of biological tools are available for investigating transcriptomic data both on a high throughput and gene by gene basis. Quantitative Real Time PCR (qRT-PCR) is a well established technique in which the exponential amplification by PCR of a sequence specifically recognized by selected primers is coupled to a fluorescent reporter system. The fluorescence measured can then be quantified absolutely, or more simply, relative to a control sample (Livak and Schmittingen, 2001). While qRT-PCR is accurate, it is not a high-throughput method. However, in the case of a small number of hypotheses, it is often more appropriate than the use of other techniques such as microarrays or next generation sequencing.

DNA microarrays are a high-throughput adaptation of northern blotting. They use the principle of specific base-pair hybridization, consisting of a large number of probesets (designed to specifically bond with a target sequence) spotted onto a solid surface. When washed with fluourescently labelled target sequences, the relative fluorescence provides a readout of the amount of target bound to the substrate and hence its relative abundance in the sample. In addition to custom made microarrays which contain custom-designed probesets, a number of commercially available species-specific microarrays have been designed which consist of probesets which capture as much of the transcriptome of the target species as possible. Thus, by using these microarrays, it is possible to obtain relative abundance values for most of the mRNA species present in a given sample. Microarrays are giving way to high-throughput sequencing methods (RNA-Seq) which depend upon the ability of modern sequencing techniques to sequence a large number of DNA or RNA molecules in which have the advantage of capturing all possible matches, such as structural variants, alternative splicing point mutants or transcripts which may not be represented by probesets on a microarray, though the computational requirements are far larger (normally requiring an HPC to be performed efficiently) and the analysis procedures are more complex. As such microarrays remain competitive in price, and specialised or custom microarrays can be used to perform experiments to investigate phenomenon not captured on a standard array, for instance copy number variation or alternative splicing. Additionally the large body of microarray data now available in the public domain in repositories such as Gene Expression Omnibus (Edgar and Lash, 2003) means that an enormous store of information is available for researchers with the tools to analyse microarray data, which for some time will outnumber the data held for NGS experiments.

1.3.3 Statistical Analysis of Microarray Data

While analysis of microarray data is a well-studied field, no dominant methodology for microarray analysis has emerged. This is true for all stages of microarray data analysis;

- **Image Analysis** – The first stage of microarray data analysis depends upon extracting relative expression data based around images capturing probe intensities across microarrays. A number of competing methodologies exist including
    - **MAS5 -** Affymetrix's MAS5 algorithm which takes advantage of the presence of mismatch probes on Affymetrix arrays to estimate directly the amount of background or cross-hybridization for each probeset which can be subtracted from the "perfect match" probes, in addition to subtracting a background noise level estimated for the neighbourhood of the microarray around the probeset (Hubbell et al, 2002). Commonly, expression levels are normalized across different arrays in experiments using a scaling factor.
    - **Robust Multichip Average (RMA)** – Background correction in RMA is based around the assumption that the observed signal is an additive combination of an exponentially distributed true signal component and a normally distributed error component. Depending upon the implementation of the algorithm, a number of methods can be used to derive the parameters of the combination of these components (rate of exponential growth, mean and variance of error) as well as the expected true signal given observed signal. The data is commonly then normalized across arrays using quantile normalization and individual probe

expressions summarised into a probeset expression level through fitting a linear additive model consisting of terms for probe affinity, expression level and error using median polish (Irrizzary et al, 2003).

- **Identification of significantly differentially expressed genes -** A large number of methods for identification of differentially expressed genes exist. While standard biostatistical methods such as t-tests and ANOVA have been applied with success to microarray data, there are a number of less common methods which have also found use in analysis of microarray data. In particular, LIMMA was designed to overcome power limitations in analysis of microarray data as more standard statistical tests suffer when analysing experiments with very low numbers of replicates, 2-3 being common numbers. A moderated t-statistic for differential expression of each gene is calculated using a linear model. In calculating the moderated t-statistic, sample variances are pooled into a variance estimate which results in better performance with small sample sizes (Smyth et al, 2004).

- **Multiple Testing** – Given the large number of statistical tests performed when analysing microarray data – even when using a stringent p-value cut-off – we expect a large number of false positives to be detected. This problem can in general either be accounted for by taking a more stringent p-value (for instance by calculating a new p-value cut-off using the Bonferroni correction or similar (Bonferroni C, 1935; Holm, 1979) or by correcting p-values according to the number of tests performed (for instance, by correcting p-values according to the expected false discovery rate which can be estimated from the distribution of observed p-values; Benjamini & Hochberg, 1995; Storey & Tibshirani, 2003).

In contrast to microarray data, analysis of qRT-PCR data is more clearly understood.

Absolute quantitation can be performed by comparing against a standard curve, however, this is usually not required in gene expression studies as we are mostly interested in the difference in expression between two samples. In this case relative quantitation can be performed based around the number of cycles it takes for the observed fluorescence from a given sample and primer set to reach a fixed threshold value. This cycle number can be normalised against a reference gene which does not show variability between the samples in question (*ACTIN2* is commonly used in Arabidopsis), and the difference in corrected cycle numbers compared between different samples. This method is referred to as the -ΔΔCt Method (Livak & Schmittingen, 2001).

1.3.4 Data mining of transcriptomic data

While statistical analyses of microarray data demonstrate whether a probeset is differentially expressed at a given p-value, this often provides a researcher with just a list of significant genes, when what is more often more pertinent is to find associations between those genes. A wide range of data mining algorithms have been developed to identify patterns within complex datasets, a development which has been particularly fuelled by the need for large commercial organisations to understand and make best use of ‘big data’. In particular, although some use of classification and regression algorithms have been made to interpret biological data, it is mostly clustering techniques that have been applied to mine transcriptomic data (Review: Slonim, 2002).

Hierarchical clustering techniques in the form of heatmaps are the most widely recognised form of microarray data mining. In hierarchical clustering a hierarchical tree in which closer branches imply closer association is built up, either by starting from individual data points and recursively agglomerating them with similar objects, or by dividing the entire group of objects recursively into separate branches. In either case

one needs to define a definition of similarity (distance measure, e.g. Euclidian distance) as well as a decision as to whether that distance will be taken from the centre, average distance or closest members of two clusters when deciding to agglomerate or divide that cluster.

Other non-hierarchical methods of clustering are widely used, for example k-means clustering has seen widespread use owing to its simplicity. From a number of (usually) randomly generated centroids in n-dimensional space (where n is the dimension of the data under consideration), data points are assigned to groups according to their nearest centroid. The centroids are updated to the geometric centre of their group, and the process repeated until the clustering algorithm stabilizes. However, it is somewhat biased in that the user must specify the number of clusters into which the data will be separated, which is often something that cannot be determined unambiguously, and as a consequence biologists are beginning to use more complex methods of clustering to identify patterns within data.

Self-Organizing maps in particular are a mapping of each point of data in a dataset to a node in a user-defined grid (Kohonen, 1990; Tamayo et al, 1999). Each node in the grid contains a randomly initialised codebook vector of the same dimensions as the data, and by some distance measure data points are recursively assigned to the node closest to them. At this point the codebook vector is adjusted to more closely match the data point being assigned to it. Once all data points have been assigned, they are removed and this process is repeated for a large number of steps – until the codebook vectors converge to a stable state. In this manner the data points are clustered into nodes, however the patterns underlying the nodes are generated spontaneously as the map is being learned. Additionally, the codebook vectors provide a representation of the data points contained within each node which can be directly related back to the initial data

or even clustered.

1.3.5 Biological Methods for inferring Gene Regulatory Networks

Since transcription factor activity is an inherently hierarchical process – one TF will be upstream of its downstream targets, one convenient way to represent TF activity is via a directed graph. In such a representation nodes represent transcription factors or other genes, and directed edges represent regulatory relationships between those genes.

Gene regulatory networks are difficult to infer using gene expression data alone, a process sometimes referred to as “reverse engineering”. The most accurate method available for determining the binding sites of a transcription factor biologically is using chromatin Immunoprecipication (ChIP) (Orlando, 2000). In ChIP, factors associated with DNA are bound to it by crosslinking using formaldehyde, and then the DNA is sheared by chemical or mechanical means. The fragments of chromatin bound to the associated factors can be obtained by immunoprecipitating the desired factor using an appropriate antibody and then reversing the DNA crosslinking. The DNA obtained can then be analyzed using qRT-PCR, microarrays or sequencing. If the entirety of the purified fraction is analyzed (genome wide ChIP), then this is also a high throughput method for inferring direct relationships between one gene and others.

Even in the case that genome wide ChIP is performed, it is difficult to extend the process to study entire systems. Either appropriate antibodies of sufficient quality must be generated for each gene of interest, or lines with appropriate epitopes attached to the genes of interest must be generated. Thus in order to perform a systematic search for associations between genes it is useful to apply in silico methods first to generate predictions which can subsequently be tested. Moreover, it is clear from a number of studies that the binding of a TF does not necessarily indicate that it regulates the

adjacent gene (for example Boruc et al, 2011).

1.3.6 In-Silico Methods for inferring Gene Regulatory Networks

There has recently been a large amount of interest in the use of in-silico methods for inferring gene regulatory networks, particularly as the number of datasets publically available has dramatically increased. While a number of methods exist, the majority of published methods are either based around using correlation based metrics or Bayesian approaches.

Correlation based metrics include simpler procedures such as co-expression analysis - ranking genes according to Pearsons correlation co-efficient, and assigning undirected edges between nodes based around the degree of their correlation. However, these methods have been extended to include more sophisticated elements such as the inclusion of machine learning elements to refine predictions. The principal weakness of correlation based methods is the dictum that “correlation does not equal causation.” Directionality in such graphs is usually impossible to define and indirect regulatory relationships, or simply co-presence in the same tissue types may lead genes to be classed as associated. Though they have been used to effect, to predict regulatory modules and relationships in gene expression studies (examples: Fukuskima et a, 2012, Menashe et al, 2013)

A number of methods for inferring networks using Bayesian approaches also exist. Bayesian networks are directed graphs in which edges encode conditional dependency relationships between nodes. Importantly they consist of a network structure and a joint conditional probability distribution over the nodes within the network. A number of scoring metrics exist for computing how well a dataset fits an observed network structure. Given an appropriately large training dataset, it is then possible to use

optimization procedures to iterate efficiently over the possible structures (the number of which is super-exponential to the number of nodes) to predict the network structure which best explains the training dataset (Heckerman, 1995). A number of variations exist on the standard static Bayesian network structural inference procedure such as dynamic Bayesian network structural inference – in which if sufficient temporal resolution is available, the conditional dependencies between different temporal states of a network can also be predicted. (Example: Godsey, 2013)

Several caveats must be borne in mind when using Bayesian networks to infer structure. One limitation is that the scoring metrics used can only distinguish between equivalency classes of graphs, the subset of graphs where the reversal of any number of edges not forming part of a ‘v-structure’ - two edges leading to the same node – has taken place. All graphs within the same equivalency class will score identically given the same dataset, thus directionality cannot usually be inferred. (Heckerman et al, 1995) Another is that the network will find associations between any dataset if they are present – as with most data mining and machine learning algorithms, relationships may even be detected in random noise. Thus, associations cannot be taken as anything other than predictions which require testing. Finally, overfitting of a network structure to a given dataset is a problem – exacerbated by the fact that it can be difficult to determine whether it has occurred.

However, Bayesian network structural inference provides one major advantage in that it is easy to incorporate prior knowledge into a network. This can be done either by forcing edges between nodes, by forbidding edges between nodes, by specifying a specific initial network structure or by accounting for perturbations in scoring network structures. This makes Bayesian networks particularly amenable to application of the iterative systems biology lifecycle to continuously derive and refine predictions. In this

manner the results from validation experiments can then be fed in to a new round of data mining or modelling in order to further refine and improve upon the predictions. This iterative cycle between prediction, validation and refinement forms the systems biology lifecycle which can be used to continuously drive improvements in the modelling of a given system.

1.4 Project aims and objectives

The aim of this project is to define and analyse the gene regulatory network of *STM* using microarray analysis, and to test the predicted regulated components.

While it is currently known that *STM* impinges on a number of developmental processes, such as CK and GA biosynthesis, there is not at present a clear understanding of how regulation of core genes in the SAM may be affected by *STM*. The core aim of this project is to apply systems biology approaches to develop and refine a broad understanding of the GRN regulated by *STM,* with the intention of building for the first time a picture of how *STM* expression affects global gene expression. This will produce a model of interactions between *STM* and downstream genes and processes which can be used to guide further experiments into *STM* function in SAM development and maintenance. This project thus consists of three objectives: The identification of components in the GRN regulated by *STM*, identification of associations between those components, and finally the detailed modelling of a selected submodule.

1.4.1 Identify the components of the GRN Regulated by *STM* using transcriptomic analysis of a timecourse of STM up- and down-regulation

Isolated interactions between *STM* and SAM-specific genes and processes have been subsequently identified by previous studies such as Spinelli et al (2011). However, a

broader understanding of the function of *STM* requires a broader picture of how processes are affected following *STM* perturbation. By using a mixture of ectopic over-expression of *STM* and RNAi-mediated downregulation of *STM* the effects of perturbing *STM* expression in both directions and both inside and outside of its native expression domain can be identified.

A novel feature of this approach is that the dynamic behaviour of genes following *STM* perturbation has not been studied previously. It is desirable to know not just the overall effects of *STM* perturbation upon the Arabidopsis transcriptome, but to be able to identify the order in which genes and processes are regulated. Without a proper understanding of the dynamics observed following *STM* perturbation it is impossible to place *STM* in its proper developmental context.

1.4.2 Predict associations between genes in the *STM* GRN

Understanding the temporal behaviour of genes and processes downstream of *STM* can be supplemented by understanding how those processes are associated with one another. To do this in-silico techniques can be applied to gene expression data in order to mine key associations between downstream genes and processes.

Predictions from these approaches can be tested experimentally and the results subsequently used to refine the existing models. Thus, by an iterative system-wide approach, it is possible to obtain and refine a model of overall associations and dynamics between genes and processes regulated by *STM*. By necessity parts of this will be putative, however, by continuous validation and refinement the model can be improved in an iterative fashion and used to guide future experiments.

1.4.3 Investigate the Behaviour and Dynamics of Core Submodules

Having derived a plausible network of interactions, the behaviour and dynamics of core

submodules will be investigated in greater detail. As can be seen with the *WUS-CLV* regulatory loop, or auxin patterning in the SAM discussed earlier, simply knowing the components involved in a feedback loop is often insufficient to understand how they properly function in their biological context. Modelling approaches can often provide a clearer and deeper understanding of how those components may combine to generate phenotypic effects.

In order to build up a clearer picture of *STM* function in the SAM – a selected core submodule can be examined in greater detail. With both molecular biology and mathematical modelling approaches it is possible to derive a better understanding of how the dynamic behaviour between components in the submodule affect the overall functionality of *STM*.

2.1.1 Plant Lines and Growth Conditions

For growth *in vitro*, seeds were surface-sterilised in a class-2 sterile tissue culture hood with a 5 minute wash with 70% ethanol and treatment for a further 5 minutes with 20% bleach (20% bleach and 5µl Silwet in 50mL). Seeds were washed 5 times with MilliQ (sterile) water and dried for 30 minutes prior to sowing. All plants used in these experiments were sown on GM Agar (For 1ml 4.6g Murashige & Skoog Medium (with vitamins, Melford, Ipswich), 15g Sucrose, 0.5g MES buffer (Sigma-Aldrich, St Louis), up to 1ml with H_20) and grown under constant light at 295K.

Lines grown on soil for crossing were sown to a mix of 2/3 soil and 1/3 sand and grown at constant temperature of 295K with a 16 hour photoperiod and watered twice per week.

Lines used which have previously been described include S^{38} - a dexamethasone-(DEX) inducible *STM* over-expression line constructed using the TGV system. 2^{37} - an empty vector control line for S38 constructed using the TGV system and SP1-1 - an inducible STM RNAi line constructed using the TGV system (Scofield et al, 2007.)

The pmiR164c::VENUS and pmiR164a::VENUS lines were described in Sieber et al, (2007) and obtained via The Nottingham Arabidopsis Stock Centre (NASC). p*CUC1::GUS* and *pCUC1::CUC1-GUS* were gifted from Mitsuhiro Aida (NAIST, Japan), and the *35S::STM-GR* line was gifted from Rüdiger Simon (Dusseldorf, Germany) and first described in Brand et al. 2002.

Female S^{38} plants were manually crossed to male pCUC1::GUS lines, by emasculating the female parents, and subsequently pollinating the naked gynaecium of multiple flowers on that plant with pollen from the male line. Individual siliques from the initial

cross containing F1 seed were harvested, and in subsequent generations (F2 and F3) individual plants were harvested separately. Progeny were selected using kanamycin (kan) and phosphinothrycin (ppt) resistance and/or for *STM* over-expression by growth on GM agar plus 60 µM DEX. F3 progeny were selected for p*CUC1::GUS* by staining seedlings from individual batches. 100% of offspring - from batch 15500 (n>12) were both *STM* and *GUS* positive. The same procedure was followed for female *pCUC1::CUC1-GUS* and male S^{38}. 100% of offspring - from batch 14201 were both *STM* and *GUS* positive. Lines 15500 and 14201 were used for all microscopy of these lines within this thesis.

Female S38 plants were crossed to male pmiR164c::VENUS-and pmiR164a::VENUS lines described in (Sieber et al, 2007) using the same method described above and grown on GM agar with 60µM DEX. Lines were screened for *STM* overexpression phenotype and VENUS expression checked by confocal microscopy. 100% of F3 S38 x pmiR164c::VENUS for lines 16256 and 16253 were VENUS positive and *STM* positive and were subsequently used for microscopy experiments. S38 x pmiR164a::VENUS F3 line 16290 was used for microscopy experiments. 12/12 screened plants from this line were venus positive, however the line was still segregating for S38.

2.1.2 Gene expression analysis by real-time qRT-PCR

Plants for qRT-PCR analysis were harvested and frozen immediately in liquid nitrogen. Harvested plants were ground to a fine powder in a mortar and pestle with liquid nitrogen and stored at -80^{o}C prior to RNA extraction. RNA for qRT-PCR experiments was extracted using Tripure reagent (Roche, West Sussex) according to the manufacturer's instructions and treated with DNAaseI (1µl of DNAaseI added to RNA for 30 minutes at 37^{o}C, DNAase was subsequently deactivated by heat-treatment at 95^{o}C for 2 minutes or using the manufacturer's DNAse1 removal solution) and checked

for quality on a 1% agarose gel. Concentrations were measured on a Nanodrop UV spectrophotometer and cDNA synthesised using a Retroscript kit (Life Technologies, New York) using 2µg total RNA . RNA was denatured at 95^{o}C and annealed to oligoDT primers, incubated at 42^{o}C for 1 hour with dNTPs, Reverse Transcriptase enzyme and RNAase inhibitor. Reverse Transcriptase denatured by heat treatment and final cDNA diluted 20x with MilliQ water.

The Abgene SYBR green master mix (Thermo Fisher) was used for qRT-PCR assays with the Qiagen/ Corbett Life Science Rotorgene 6000 real-time PCR machine - 5µl of SYBR Green, 2.5µl of equal Forward/Reverse primer mix (0.3uM each primer), and 2.5µl cDNA were applied to each well and qRT-PCR was then carried out on a Rotor Gene 6000 (Corbett Life Science, Cambridgeshire). Samples were held at 95^{o}C for 15m then cycled between 95^{o}C for 30s, 55^{o}C for 30s and 72^{o}C for 60s (after which fluorescence was acquired) for up to a maximum of 45 cycles. Finally a melt profile was obtained by ramping by 1^{o}C each 5s from 72^{o}C to 95^{o}C. For data analysis purposes, cycles before 10 were eliminated, dynamic slope and slope corrections were applied. Relative fold changes were obtained using the $\Delta\Delta C_t$ method (Livak & Schmittingen, 2001) with ACTIN2 (ACT2) as a normalization factor. All primers used in qRT-PCR experiments are listed in Appendix 1.

2.1.3 qRT-PCR for miRNA

Small RNA was extracted from plant tissue using the miRVANA kit (Life Technologies, New York) with 1ng of total RNA. Reverse Transcription was performed using miR164c, miR164a/b and snoR85 specific primers following kit protocol – (TaqMan MicroRNA Reverse Transcription Kit , Life Technologies, New York).

Taqman probes for miR164c and miR164a/b were obtained from (Life Technologies, New York), and a TaqMan probe for the small Non-coding RNA SnoR85 was used as a control as it has been used previously as a control for microRNA experiments (Brady et al, 2011). qRt-PCR experiments were performed on a Rotor Gene 6000 (Corbett Life Science, Cambridgeshire) using volumes specified by TaqMan small RNA assay (Life Technologies, New York). After 10 minutes at 95^{o}C for enzyme activation, 40 PCR cycles were performed with 15 seconds at 95^{o}C to denature the sample, followed by a 60^{o}C annealing/extension stage. Relative fold changes were obtained using the $\Delta\Delta C_t$ method (Livak & Schmititngen, 2001) with ACT2 as a normalization factor as for standard mRNA qRT-PCR analysis.

2.1.4 GUS Staining

Samples were harvested and immersed in acetone for 10 minutes. Plants were washed with 100mM K-Phosphate buffer (pH 7.0) and immersed in GUS staining buffer (50mM Potassium Phosphate, 1mg/ml 5-Bromo-4-chloro-3-indolyl-β-D-glucuronide in Dimethylformamide, 0.5mM K-Fe potassium ferricyanide, 0.5mM Fe-K potassium ferrocyanide, 0.05% Triton X-100). Lines being stained for *pCUC1::GUS* were stained for 2 hours at 37^{o}C, lines being stained for *pCUC1::CUC1-GUS* were stained overnight at 37^{o}C. Plants were cleared in 100% methanol overnight prior to visualization.

2.1.5 Microscopy

Lines visualised by ordinary or fluorescence microscopy were visualized on a Leica MZ16F dissecting microscope (Leica, Solms, Germany) at either 10x, 20x or 40x magnification as indicated, using a YFP filter. Lines visualized for confocal microscopy were visualized on a Zeiss 710 Meta Confocal Microscope (Zeiss, Jena, Germany).

2.1.6 Direct Target Experiment Induction

Induction of plants for direct target induction was performed as follows - 5ml of 60μM DEX in DMSO with 5μl Silwet was used to treat DEX treated plants. This was either poured gently onto the agar plate and gently spread to cover the plate or sprayed directly onto the plate inside a fume hood. The same procedure was followed for Cycloheximide (CHX) DEX treated plants, however in this case, a solution of 60μM CHX and of 60μM DEX in DMSO was used. For CHX treated plants, 60μM CHX was added without DEX, finally for Mock treated plants the equivalent volume of DMSO was added instead of DEX or CHX.

2.1.7 Sample preparation for Microarray Experiments

Total RNA was isolated according to the same procedure as qRT-PCR analysis and RNA integrity and quality was examined on a Bioanalyser (Agilent). RNA was then sent for hybridization to ATH1 microarrays (Affymetrix, Santa Clara) by the NASC microarray service (NASC, Nottingham).

2.2.1 Time Course Microarray Analysis

Expression data was extracted for all time course microarrays using the RMA algorithm (Irrizzary et al, 2003). Initially the MAS5 algorithm (Hubbell et al, 2002) was also investigated but for several genes, including *STM* - whose expression is restricted to a narrow domain within the meristem, MAS5 returned absent calls in control and RNAi lines due to overall low levels of expression. Thus RMA was used preferentially. Expression data was normalized using Quantile normalization, with median polish applied (Bolstad et al, 2003).

A number of quality control metrics were applied using the affyQCreport module (Parman et al, 2013). These are shown in Fig 2.1. Fig 2.1a shows a boxplot of all PM

intensities, as can be seen no arrays show dramatically different intensities. Figure 2.1b shows the 3'/5' ratios for the ACT and GADPH control probes. As can be seen, the RNAi 9d1, showed significant differences between its GADPH ratio. No arrays differ dramatically in terms of the percentage of present/absent calls or the average background intensity. Thus although 9dRNAi 1 failed on one QC metric, it passed the remaining measures and was not regarded as needing exclusion.

LIMMA (Smyth et al, 2007) was used to compute the likelihood of differential expression of genes between the following contrasts:-

1. 8h OE - 8h EV
2. 24h OE - 24h EV
3. 72h OE - 72h EV
4. 9d OE - 9d EV
5. 72h RNAi - 72h EV
6. 9d RNAi - 9d EV
7. stm-2 - wt

The test correction applied by (Benjamini & Hochberg, 1995) was used to compute adjusted p-values, and a threshold of 0.01 was selected for significance across all datasets.

R scripts used to perform these analyses are given in Appendix 2

A

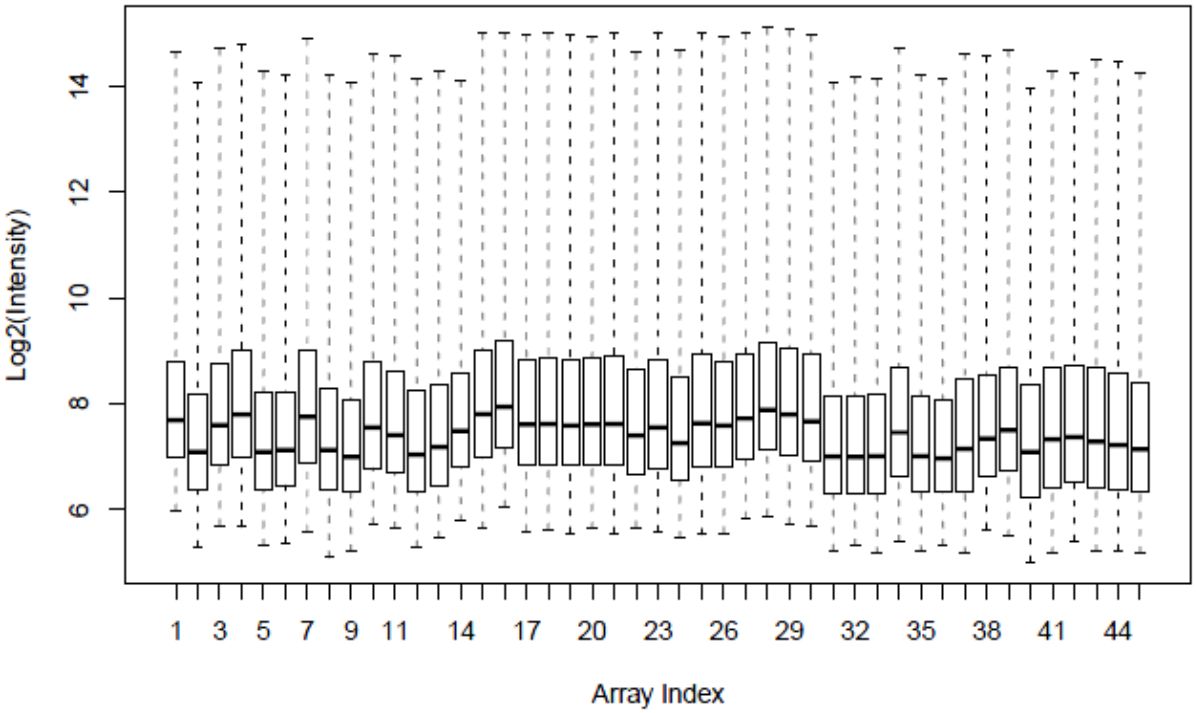


Figure 2.1 - Time Course Arrays QC.

A) Boxplot of all PM Probe $\log_2$ signal intensities across all microarrays for the Time course , wt and *stm*-2. Indexes correspond to arrays are as follows: 1-3 Empty Vector, 0h, 4-6 STM+, 0h, 7-9 RNAi, 0h, 10-12 Empty Vector, 8h, 13-15 STM+, 8h,16-18 wt, 19-21 Empty Vector, 24h, 22-24 STM+, 24h, 25-27 stm-2 mutant , 28-30 Empty Vector, 72h 31-33 STM+, 72h, 34-36 RNAi, 72h, 37-39 Empty Vector, 9d, 40-42 STM+, 9d, 43-45 RNAi, 9d – where STM+ is shorthand for *STM* overexpression lines, RNAi is shorthand for *STM* RNAi lines. Numbers are as show in B.

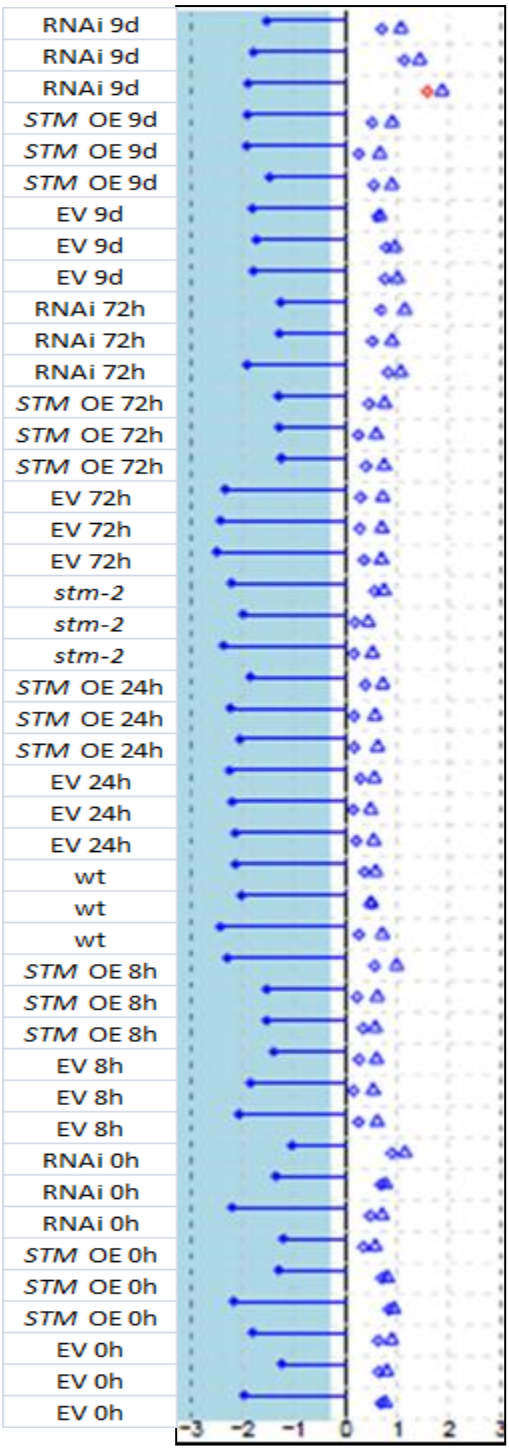
B) Percentage of present/absent calls, average background intensity.

C) ACT (circles)and GADPH (triangles) 3'/5' ratios for same arrays. red indicates an outlier.

B

% Present	Background	Line	Number
69.62%	44.08	RNAi 9d	45
70.91%	45.6	RNAi 9d	44
70.69%	45.84	RNAi 9d	43
71.21%	51.41	*STM* OE 9d	*42*
70.65%	46.34	*STM* OE 9d	*41*
69.33%	40.02	*STM* OE 9d	*40*
68.7%	57.78	EV 9d	39
69.07%	57.45	EV 9d	38
69.06%	44.27	EV 9d	37
67.28%	48.91	RNAi 72h	36
67.18%	46.91	RNAi 72h	35
68.78%	50.08	RNAi 72h	34
67.81%	44.97	*STM* OE 72h	*33*
67.11%	45.56	*STM* OE 72h	*32*
68.5%	44.97	*STM* OE 72h	*31*
69.07%	64.9	EV 72h	30
69.52%	69.37	EV 72h	29
68.87%	77.39	EV 72h	28
69.82%	67.61	*stm-2*	*27*
69%	59.18	*stm-2*	*26*
71.74%	60.88	*stm-2*	*25*
70.16%	55.22	*STM* OE 24h	*24*
70.6%	59.32	*STM* OE 24h	*23*
70.48%	58.38	*STM* OE 24h	*22*
70.32%	63.06	EV 24h	21
69.95%	63.04	EV 24h	20
69.14%	63.62	EV 24h	19
69.26%	63.17	wt	18
69.25%	64.66	wt	17
69.37%	78.7	wt	16
69.53%	66.57	*STM* OE 8h	*15*
67.41%	67.85	*STM* OE 8h	*14*
69.13%	50.44	*STM* OE 8h	*13*
68.11%	46.38	EV 8h	12
68.73%	59.8	EV 8h	11
69.28%	62.74	EV 8h	10
66.76%	47.57	RNAi 0h	9
68.22%	45.67	RNAi 0h	8
69.03%	60.63	RNAi 0h	7
66.87%	51.17	*STM* OE 0h	*6*
67.77%	47.78	*STM* OE 0h	*5*
69.55%	65.48	*STM* OE 0h	*4*
67.86%	61.36	EV 0h	3
66.92%	48.44	EV 0h	2
68.89%	77.29	EV 0h	1

C



2.2.2 Direct Target Microarray Data Analysis

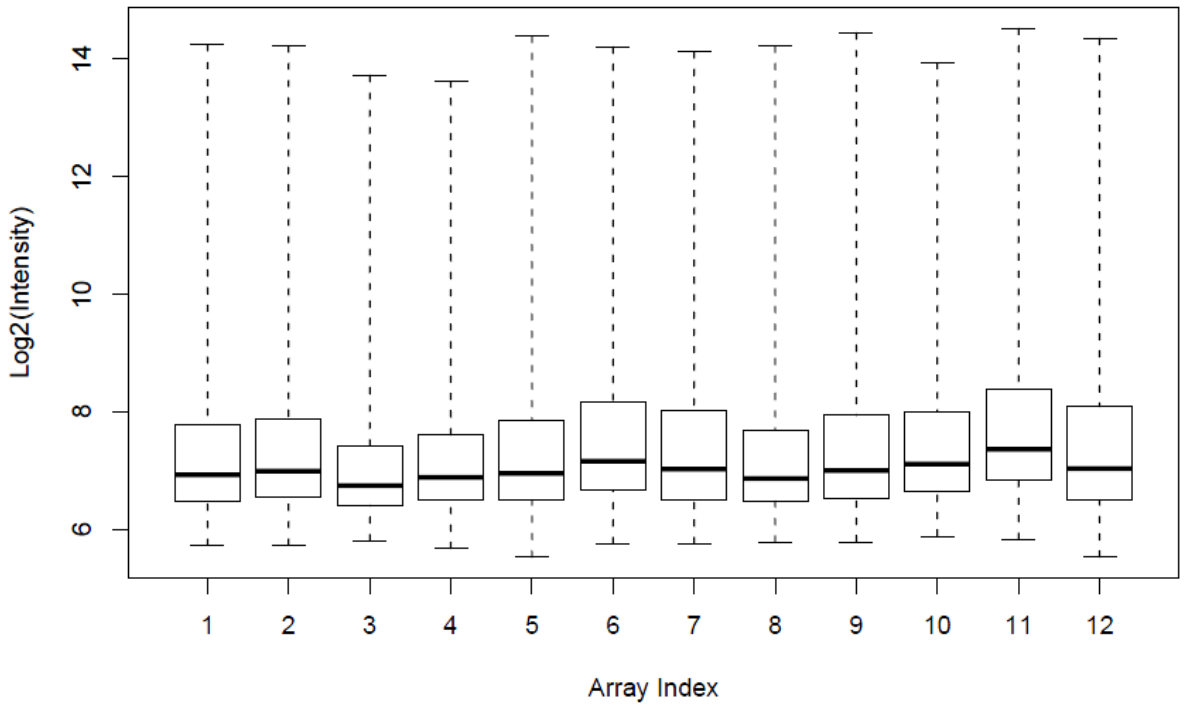
In order to ensure that data was as comparable as possible with previous time course data, the same procedure was applied to all the direct target microarray experiments. Quality control data are shown in Figure 2.2

LIMMA was used to compute the likelihood of differential expression of genes between the following contrasts:-

1. DEX - Mock
2. CHX + DEX - CHX
3. Mock – CHX

The test correction applied by (Benjamini & Hochberg, 1995) was used to compute adjusted p-values. R Scripts used to perform this analysis are given in Appendix 3

A



B

%	Backaround	Line	Number
7.28%	60.4	Mock	12
67.95%	78.27	CHX + DEX	11
67.03%	69.48	CHX	10
66.83%	64.14	DEX	9
65.27%	65.3	Mock	8
67.65%	62.31	CHX + DEX	7
68.58%	69.9	CHX	6
66.63%	63.77	DEX	5
64.13%	65.6	CHX + DEX	4
65.34%	63.44	CHX	3
65.70%	66.25	DEX	2
62.68%	61.18	Mock	1

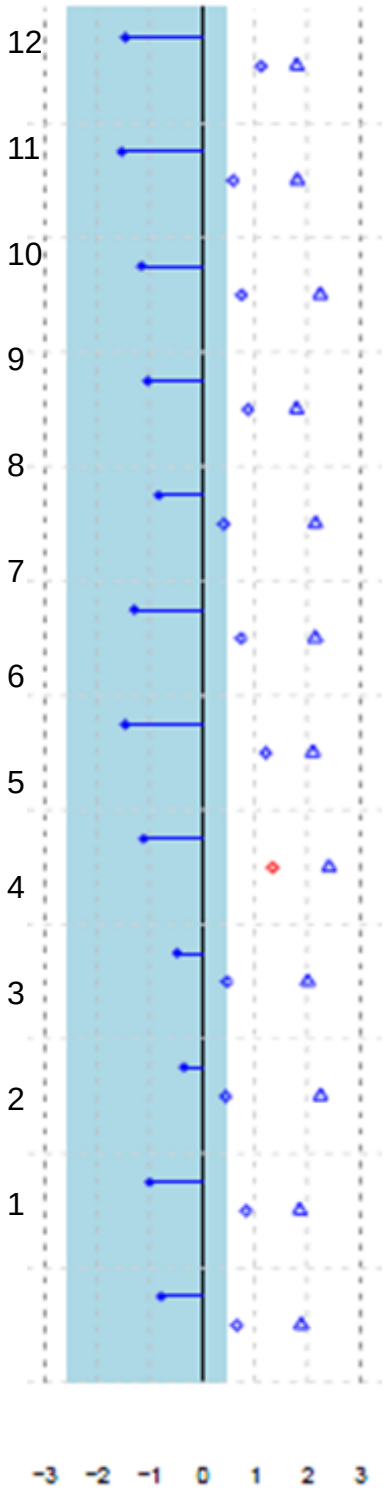
Figure 2.2 - Direct Target Arrays QC.

A) Boxplot of all PM Probe $\log_2$ signal intensities across all microarrays for the Time course , wt and *stm*-2. Indexes correspond to arrays are as follows: 1,8,12 – 3h Mock treated plants, 2,5,9 – 3h DEX treated plants, 3,6,10 – 3h CHX treated plants, 4,7,11 – 3h CHX+DEX treated plants. Numbers are as show in B.

B) Percentage of present/absent calls , average background intensity.

C) ACT (circles)and GADPH (triangles) 3'/5' ratios for same arrays. red indicates an outlier. Numbers are as shown in B.

C



2.2.3 Hormone Microarray Analysis and *CYCD3* OE Microarray Analysis

Affymetrix '.cel' files for various hormone treatments as performed by (Goda et al, 2008) were obtained from TAIR. Additionally, '.cel' files were obtained for microarrays describing *CYCD3* constitutive over-expression (DeWitte et al, 2003). These were subsequently analyzed in the same way as previous microarray experiments - R Scripts used to perform this analysis are given in Appendix 4.

2.2.4 Meta Analysis

Omnibus p-values were calculated using both Fisher's Inverse Chi Squared test and Stouffer's Z test. As it is possible for a gene to undergo negative feedback over the time course experiment, and we were interested in identifying genes falling into this pattern of regulation, p-values were not separated by the direction of fold change. Omnibus p-values were subsequently corrected by False Discovery Rate correction as described in (Storey, 2002) using the q-value package in R. R scripts for the meta analysis are provided in Appendix 5.

2.2.5 Microarray Data Preparation for Bayesian Network Analysis

Using R to connect to the Gene Expression Omnibus via the ArrayExpress Module (Kaufmann, 2012), all microarrays annotated for seedling or shoot apex tissue as of 17/3/2013 were obtained. The list was manually curated, and any arrays which had obviously contained root tissue or were not at the seedling developmental stage were removed. These arrays were thus treated as having been enriched for appropriate tissue and are listed in Appendix 6. Due to memory constraints, RMAExpress was used to extract expression values (RMA) and normalize data (quantile normalization with median polish) from the .cel files. (2003, http://stat-www.berkeley.edu/bolstad/RMAExpress/RMAExpress.html)

Expression values for all TFs identified in the robust response meta analysis were

extracted from the dataset and discretized as follows:-

2 $E_{ij} - E_{avi} > 1$

0 $E_{avi} - E_{ij} > 1$

1 Otherwise

where E_{ij} is the expression of probe i in condition j and E_{avi} is the expression of probe i over all conditions. I.e. 2 if it is more than 2-fold greater than the average expression, 0 if it is more than 2-fold less than the average expression, 1 otherwise.

2.2.6 Bayesian Network Structural Inference

Consensus networks were inferred from the datasets described in Section 2.1.5 using BANJO version 2 (Hartemink A, http://www.cs.duke.edu/~amink/software/banjo/). Networks were run for 1 hour, with maximum parent count of 5 (constrained for memory considerations), using simulated annealing to search through the solution space, over a maximum of 10,000 restarts, initial simulated annealing temperature of 10000, a cooling factor of 0.7, reannealing temperature of 800, a maximum of 2500 accepted networks before cooling, a maximum of 10000 proposed networks before cooling and a minimum of 500 accepted networks before re-annealing. 20 top-scoring networks were generated and a consensus network produced via influence scores.

In constrained networks a list of forbidden edges was supplied. In forced networks a list of forced edges was supplied. Networks were visualized using Cytoscape (Smoot et al, 2011).

2.2.7 Self Organizing Map

A self organizing map (SOM) was inferred for the time course data using the Kohonen module in R (Wehrens, 2007). Full tables of codebook vectors and assignments of genes to nodes are in Appendix along with R scripts used to generate the SOM 7.

2.2.8 Principal Components Analysis (PCA)

Principal Components Analysis was performed in R using the inbuilt dist and cmdscale commands. The resulting mapping was visualized in 2D using the biplotGUI (LeGrange et al, 2009) package for R. Full scripts used to generate the Principal Components Analysis are found in Appendix 8.

2.2.9 Stochastic Modelling

Stochastic Modelling was performed in R as described in Chapter 5. Events were assigned a relative likelihood based around multiplication of rate parameters and quantities of dependent species (proteins, mRNA molecules or Promoters - modelled as discrete units) and ordered. The time to an event was modelled by the Poisson distribution and at each iteration of the model a time to next event (change in quantity of a species) was generated, the event was selected by generating a random number between 0 and the sum of all likelihoods. The event which occurred was determined by addition of ordered event probabilities until further addition would exceed the random number. The event whose probability was to be added was subsequently performed and quantities of target species adjusted accordingly. Scripts used to perform the analysis are provided in Appendix 9.

2.2.10 ODE Modelling using COPASI

Copasi was used to produce a 2-compartment ODE model as described in Chapter 5 (Hoops et al, 2006). Simulations were run as a 1000 time unit time course and this was sufficient for the model to reach a steady state under all parameters considered. Model files are provided in Appendix 10.

2.2.11 GO Enrichment Analysis

GO Enrichment analysis was performed within DAVID (Dennis Jr et al, 2003) using the default Arabidopsis background set. Data was exported into Cytoscape using the BINGO (Maere et al, 2005) module to create network visualizations. Full tables are provided in Appendix 11.

2.2.12 Reanalysis of spatial gene expression data from Yadav et al., 2009

As original '.cel' files were not available, the MAS5 expression values obtained by Yadav et al (2009), were log-transformed to base 2 and subsequently analyzed in LIMMA as in previous microarray analyses using the following contrasts.

1. CLV-sorted cells vs FIL-sorted cells
2. CLV-sorted cells vs WUS-sorted cells
3. WUS-sorted cells vs FIL-sorted cells

P-Values were corrected using the qvalue module in R.

All p-values less than 0.05 were considered statistically significant. Genes considered statistically significant in any contrast were categorized as follows.

4. CLV Specific: Significant in at least one of CLV vs FIL or CLV vs WUS, expression in CLV higher than FIL and WUS. This category considered CZ enriched.

5. WUS Specific: Significant in at least one of WUS vs FIL or WUS vs CLV, expression in WUS higher than CLV and FIL. This category considered OC enriched.

6. FIL Specific: Significant in at least one of FIL vs WUS or FIL vs CLV. Expression higher in FIL than CLV and WUS. This category considered Primordium Enriched

7. CLV+FIL: Significant in CLV vs WUS and FIL vs WUS. Expression lowest in WUS. This category considered L1 enriched.

8. CLV+WUS: Significant in CLV vs FIL and WUS vs FIL. Expression lowest in FIL. This category considered considered enriched for genes which are primordium excluded.

9. FIL+WUS: Significant in FIL vs CLV and WUS vs CLV. Expression lowest in CLV. This category considered enriched for CZ-excluded genes.

10. No Data: Any gene which was reported as absent by MAS5 in all contrasts

11. Ubiquitous: Any gene reported as present by MAS5 in all contrasts but with no significantly differential expression. This category considered enriched for genes expressed throughout the SAM.

Scripts used to perform the reanalysis can be found in Appendix 12.

2.2.13 Co-Expression Analysis

The CressExpress tool (Srinivasasaingendra et al, 2008) was used on version 3 RMA arrays, on ATH1 probeset 260632_at (*STM)* using all arrays annotated as any of the following:-

1. hypocotyls
2. cotyledons
3. meristems
4. true leaves 1 and 2
5. true leaves primordium initials
6. shoot apex, vegetative + young leaves
7. shoots
8. shoot apices
9. shoot apex, inflorescence (after bolting)
10. whole shoot
11. shoot apex, transition (before bolting)
12. shoot apex, vegetative
13. Aerial parts of whole plants

GO Annotations for all probe sets with a r^2 greater than 0.75 were obtained from TAIR, and all probe sets annotated for transcription factor binding activity were selected for Bayesian network structural inference. Full results of co-expression analysis can be found in Appendix 14.

Chapter 3 - Transcriptome-wide responses to *STM* perturbation

3.1 Introduction

High-throughput omic technologies have dramatically changed the way that molecular biology research is conducted since the completion of the first genome projects. In theory, the ability to simultaneously analyze all responses at a given omic level to a particular perturbation should help minimize variation from performing multiple experiments, permit serendipity in discovering unexpectedly differentially expressed targets, and dramatically increase the amount of available data. In practice, the relatively high cost of such techniques, the statistical difficulties in analyzing large numbers of simultaneous experiments and their inherent noisiness remain barriers to their effective implementation, for which careful analysis is required to resolve.

DNA microarrays were one of the earliest omic technologies to find widespread use. Analogous in some ways to a highly parallel RNA blot, they consist of a large number of short, labelled DNA probes bound to a chip which hybridise with complementary fluorescently labelled sample DNA and fluoresce. In Arabidopsis, the Affymetrix ATH1 array has found extensive use, as a standardised array platform suitable for use in many different projects studying this model organism. Based on Affymetrix GeneChip technology, the ATH1 array is a single-channel array which allows for simultaneous measurement of RNA levels for over 22,500 probesets each consisting of 11 25-mer probe pairs, corresponding to around 24,000 gene products (ATH1 Gene Chip Datasheet, Affymetrix). The ATH1 array has been used extensively to study the Arabidopsis transcriptome, thousands of microarray datasets being available in public repositories using this platform.

Due to the ability of the ATH1 array to capture a large number of gene expression changes, it is particularly useful when analyzing complex phenotypes such as that observed following *STM* over-expression or down-regulation (reviewed in Scofield and Murray, 2006). Additionally the fact that the measurements obtained correspond to a consistent subset of gene products, allows for comparison to other array experiments performed on this platform. In this chapter, I will detail the results of analysis of a time-course experiment of *STM* over-expression and down-regulation using the ATH1 array platform.

While previously published experiments have used microarrays to examine transcriptome-wide responses to *STM* perturbation (Spinelli et al, 2011), and there are publically available datasets relating to *STM* perturbation which were not discussed in their source paper (Liebfried et al, 2005), these have suffered from a combination of low power due to small numbers of replicates, and looked only at a single time point. To overcome the limitations of previous studies, we have examined how the transcriptome is perturbed using inducible *STM* up-regulation and RNAi mediated down-regulation lines (lines described in Scofield et al, 2007). These experiments have been performed at multiple time points, so that consistency of expression across multiple lines and lengths of induction can be analyzed.

Microarray experiments have been conducted for some time, and the bioinformatics community has placed strong emphasis on the creation of repositories such as the Gene Expression Omnibus (GEO) (Edgar et al, 2003). These store large amounts of microarray data which, thanks to the widespread use of the ATH1 platform, have made available a large number of experimental data interrogating molecular phenotypes related to known phenotypes of *STM* perturbation, such as cytokinin responses. Where

possible, publicly available data has been used to further interrogate how *STM* over-expression and knock-down relates to these known molecular phenotypes.

3.2 Analysis of *STM* Over Expression and RNAi Mediated Downregulation Time Course

As previously mentioned, other groups have previously attempted to identify genes in the GRN regulated by *STM*. This study has improved on these previous experiments by performing a large time course experiment, monitoring both transcriptomic responses to *STM* induction from 8 hours to 9 days and RNAi mediated *STM* knock-down from 72 hours to 9 days. *STM* was induced ectopically, while *STM* was down-regulated via inducible RNAi and thus only takes place within the native domain of *STM.* This has enabled both placement of the phenotypic changes observed following *STM* induction in their developmental context - by tying them down to temporal dynamics - and to circumvent the power issues experienced by previous experiments - by permitting the combination of results from multiple time points to look for consistent responses. To provide an additional end point control, a dataset contrasting the *stm-2* mutant was compared against a wild type control.

3.2.1 Lines used and Experimental Design

The 2-component TGV expression system had been used to create dexamethasone (DEX)-inducible *STM* Over-Expression (OE) and *STM* RNAi lines as described (Scofield et al, 2007). Previously in the Murray lab, these lines had been used to generate a microarray time course to study how genes respond to inducible *STM* up and down-regulation. As an additional end-point control, microarrays for the *stm-2* mutant and wild type control were also generated. The experimental design is shown in Figure 3.1**.** 8 or 24 hour time points were not generated for the RNAi line as previous

experiments had suggested that 72 hours were required to be able to reliably detect *STM* down-regulation (Scofield S, personal communication).

In order to demonstrate that the experiment had functioned as expected, the change in levels of *STM* compared to the appropriate empty vector control was evaluated using qRT-PCR. The results of this analysis are shown in Figure 3.1b - as can be seen in the *STM* OE lines, RNAi lines and *stm-2,* there was a detectable difference in expression observed between each line and its appropriate control, thus confirming that at all time points - *STM* was clearly differentially expressed and that the experimental approach had worked.

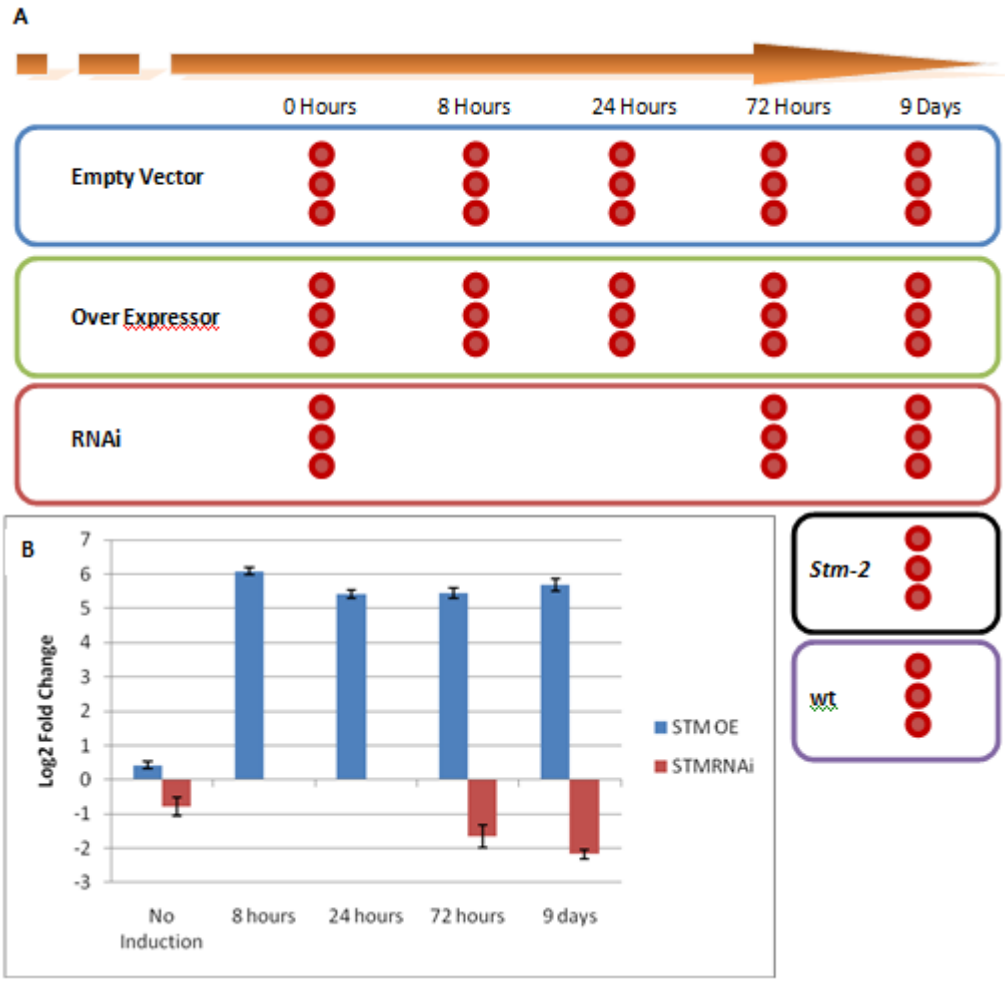


Figure 3.1 - Experimental Design and Validation of *STM* induction.
A) Experimental design of time course microarray experiment. Replicates are indicated by red circles, with horizontal axis indicating the length of induction, all plants were harvested at 9 DAG. For uninduced lines the horizontal axis only represents age of plants at harvest.
B) qRT-PCR validation of *STM* levels at 0, 8, 24, 72 hours and 9 days of induction in the OE lines and 0h, 72h and 9d of induction in the RNAi lines relative to their respective empty vector controls. Fold changes computed via -ΔΔCt Method normalized against an *ACTIN2* control on the same samples used for the microarray time course. Each contrast is composed of 3 technical replicates with error bars indicating standard error of the mean.

3.2.2 Genome Wide Response to *STM* over-expression.

The *STM* over-expression time course consisted of 4 time points - 8 hours, 24 hours, 72 hours and 9 days of induction of *STM* by DEX. All plants were 9 days old at harvesting, and a DEX-treated empty-vector line was used as a control at each time point. At the early time points, 8 and 24 hours, 92 and 321 probesets respectively passed a

significance threshold of $p < 0.01$. However in the later time points, 4133 probesets were significantly differentially expressed at 72 hours ($p < 0.01$), and 8,070 probesets passed the threshold by 9 days. As there are 22746 probesets represented on the ATH1 array this implied that at 9 days, approximately 35% of the Arabidopsis transcriptome had responded to *STM* induction. Given the severity of the *STM* over-expression phenotype at 9 days, this is not an implausible result, however it complicated isolating transcriptional responses to *STM* from responses to broader phenotypic changes at this time point. Full data has been provided in Appendix 13.

Despite the large numbers of probesets significantly differentially expressed at 9 days, it could be seen that the majority which were differentially expressed following *STM* perturbation at any time point remained significantly differentially expressed at subsequent time points. Figure 3.2 shows that the overlap between adjacent time points was particularly strong, with 80% of probesets significantly differentially expressed at 8 hours, still differentially expressed at 24 hours, 81% of those at 24 hours still significant at 72 hours, and 65% of 72 hour probesets still significant at 9 days. 73% of probesets significant at 8 hours are significant at all other time points, suggesting that the early response to *STM* induction was sustained. Interestingly as can be seen in Figure 3.2, it was not only the number of genes themselves whose differential expression was sustained, the direction of change was highly consistent between adjacent time points, which provided corroboration that the time course was capturing a consistent phenotype and that the expression of only a relatively small number of significantly differentially expressed probesets oscillated in their direction of response following *STM* induction which could indicate complex expression dynamics.

To further investigate the biological significance of these changes, volcano plots were used to plot the range of fold changes observed against p-values for significant

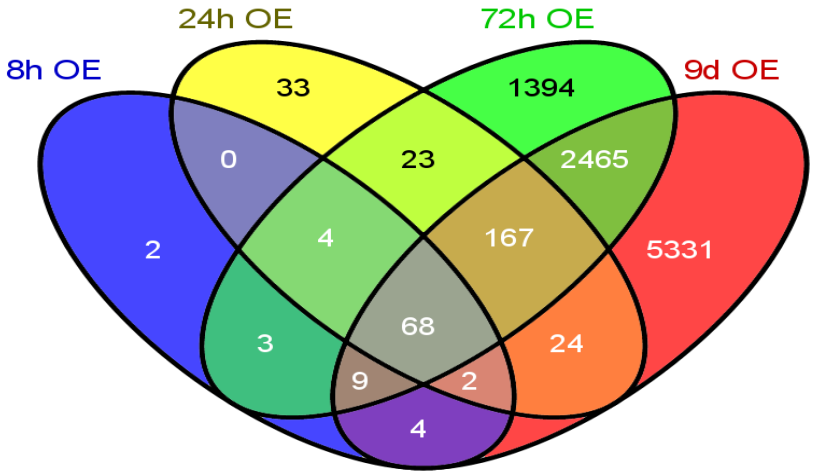
probesets in each biological contrast at each time point. As can be seen in Figure 3.3A, at all time points the majority of significantly differentially expressed probesets showed less than a 2-fold change in either direction, thus fold changes observed were small.

There is little variation in the observed fold change for *STM* after 24 hours. At 8 hours, it is upregulated 4.7 fold on the $\log_2$ scale, at 24 and 72 hours this remains roughly constant at 4.2-fold and 4.4-fold at 9 days on the $\log_2$ scale which is lower than observed by qRT-PCR. The induction of the probeset corresponding to *STM* is dramatically higher than the induction of the next nearest probeset at the 8 hour time point, *BETA-GLUCOSIDASE45* (Chapelle et al, 2012), which is only upregulated by 1.5 and *STM* remains more strongly induced than any other probeset until 9 days, when six are more strongly induced. Among these six are *BGLU45* and *CXE17,* both of which already show greater than a $\log_2$ fold change at 8 hours, suggesting that this response is sustained following *STM* induction.

Thus we can see from Figure 3.3b that while *STM* perturbs the expression of a large number of genes, the number for which a strong inductive or repressive effect is observed is quite a small proportion of the total, and it required most of the time course to elapse before any inductive effect on other genes of the same magnitude as *STM* was observed, probably due to the inducible promoter showing strong induction.

There is very little reciprocal expression between points over the earlier stages of the time course. All genes between the 8 and 24 hour time points are differentially expressed in the same direction (i.e. upregulated or downregulated.) Only 2 genes are reciprocally regulated between the 72 hour time points and the earlier time points. While there is some reciprocal regulation of expression after 9 days of induction with DEX, it is still comparatively low, with 3, 10 and 222 genes reciprocally expressed between 9 days and the 8, 24 and 72 hour time points respectively.

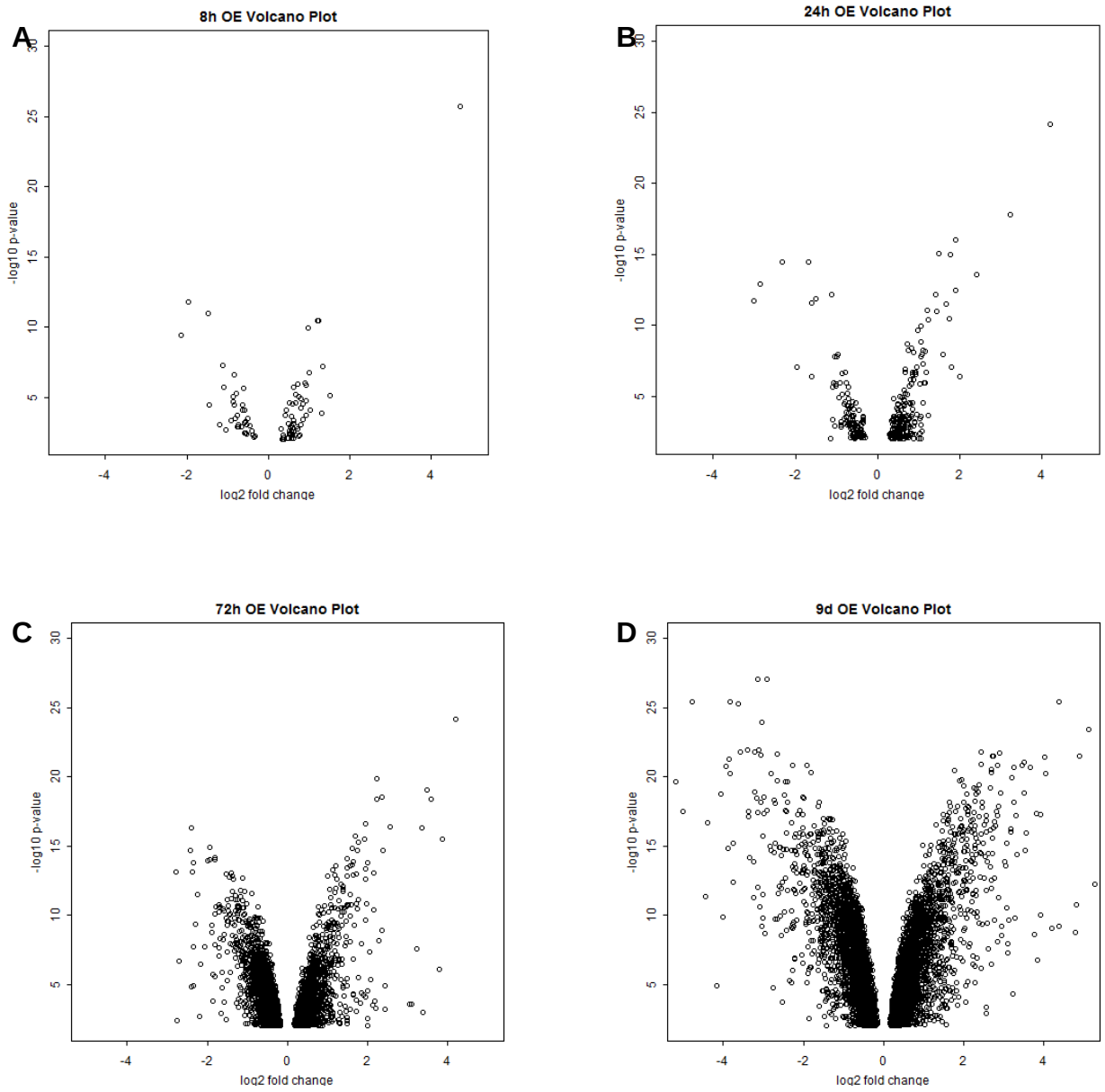
A



B

	8 & 24h	24 & 72h	72h & 9d
Number of overlapping genes	74	260	2,487
Proportion of overlapping genes showing consistent response	100%	99%	92%

Figure 3. 2 - Consistency of response of genes to *STM* induction.
A) A Venn Diagram displaying the size of the overlaps between the various time points on the over-expression time course.
B) A Table showing the number and proportion of genes either significantly up-regulated, or significantly down-regulated in pairs of adjacent time points on the over-expression time course along with how many overlapping genes are consistently either up-regulated or down-regulated between each pair of time points.



E	8 hours		24 hours		72 hours		9 days	
Log_2 Fold Change >1	**8**	**7**	**17**	**35**	**168**	**203**	**676**	**710**
Log_2 Fold Change >0.8	**6**	**9**	**12**	**27**	**142**	**121**	**372**	**273**

Figure 3.3 – Volcano Plots showing magnitude of Induction vs significance across the OE time course.

Volcano plots of point estimates of log_2 fold changes calculated using LIMMA/RMA vs -log_{10} times adjusted p-values for lines over-expressing *STM* for A) 8, B)24, C)72 hours and D) 9 days. Only shown are those probesets with a q-value <0.01 at each over expression contrast. The numbers of probesets which are upregulated (green) or down-regulated (red) with a given fold change at each time point are shown in E.

3.2.3 Gene Ontology Enrichment Analysis of the *STM* Over-Expression Time course

Figure 3.4 shows the results of gene ontology (GO) (Ashburner et al, 2000) enrichment analysis for biological processes across the over-expression time course. Enriched categories have been grouped and annotated manually according to their most significantly enriched leaf (terminal) nodes. Full classifications can be found in Appendix 11.

As can be seen, while the specific categories enriched varies across the time point, several groups of categories are enriched at multiple time points. This suggests that a number of processes are perturbed by *STM* over-expression which remain perturbed. In particular, categories related to anatomical structure morphogenesis (such as shoot, cotyledon and reproductive structure development) are significantly enriched at all time points. At 8 hours, the categories enriched relating to anatomical structure morphogenesis are non-specific, however by 24 hours the terminal leaves of the enriched GO graph show that a large number of shoot development terms are enriched along with terms for regionalization. Shoot development terms remain enriched at 72 hours and 9 days, along with a terms relating to development of reproductive tissues and leaf development. This suggests that the immediate developmental effect of *STM* induction is on shoot development, and that it is only after longer-term induction of *STM* that consequences on reproductive development are significantly enriched, which fits with the known role for *STM* in reproductive development (Scofield et al, 2007).

Other categories enable us to identify the order in which *STM* over-expression perturbs normal development or developmental process and mechanisms. Hormone responses become enriched from 24 hours onwards, while at 24 hours and 72 hours regulation of transcription is enriched. This suggests that from these time points onwards, a

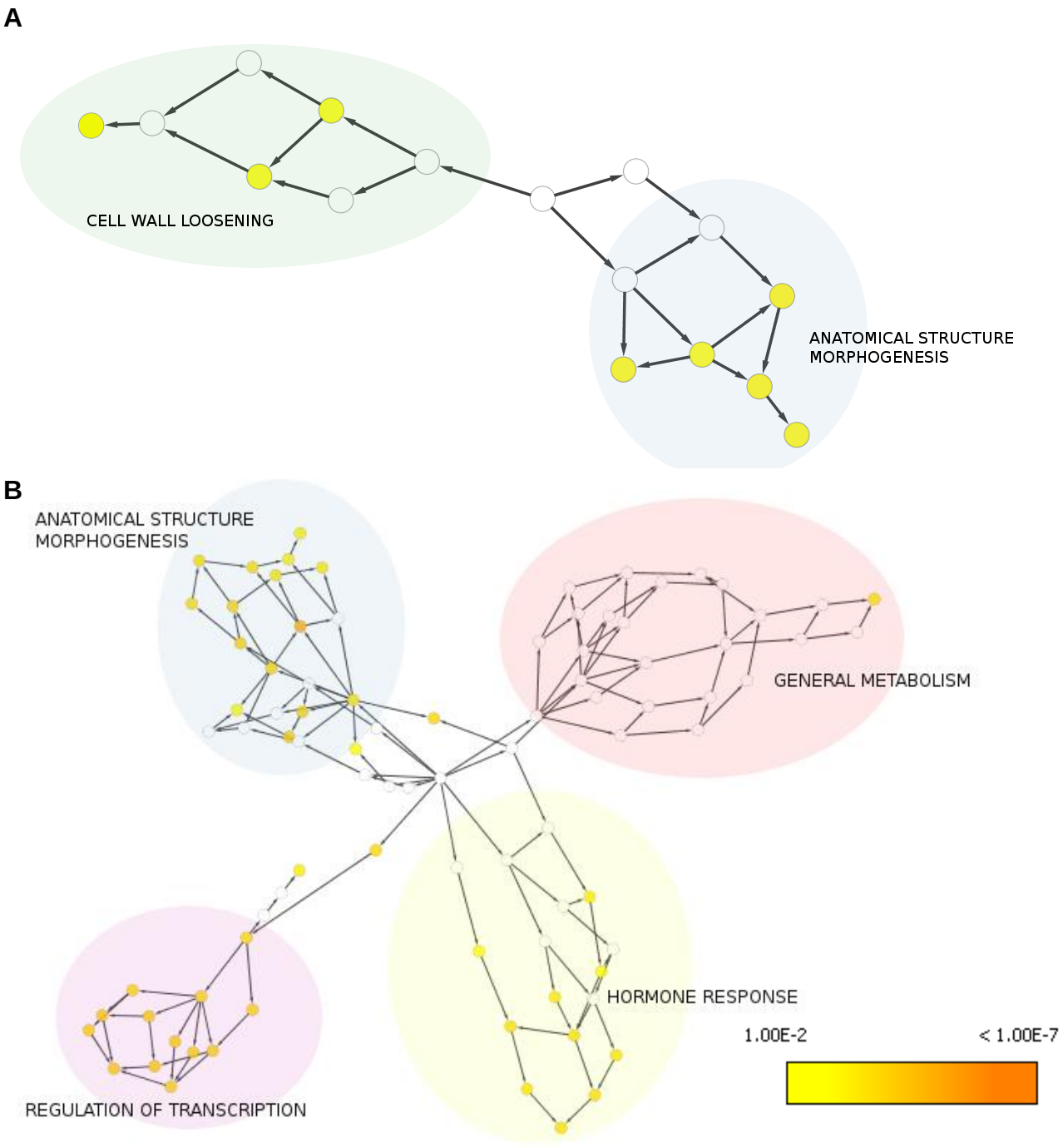
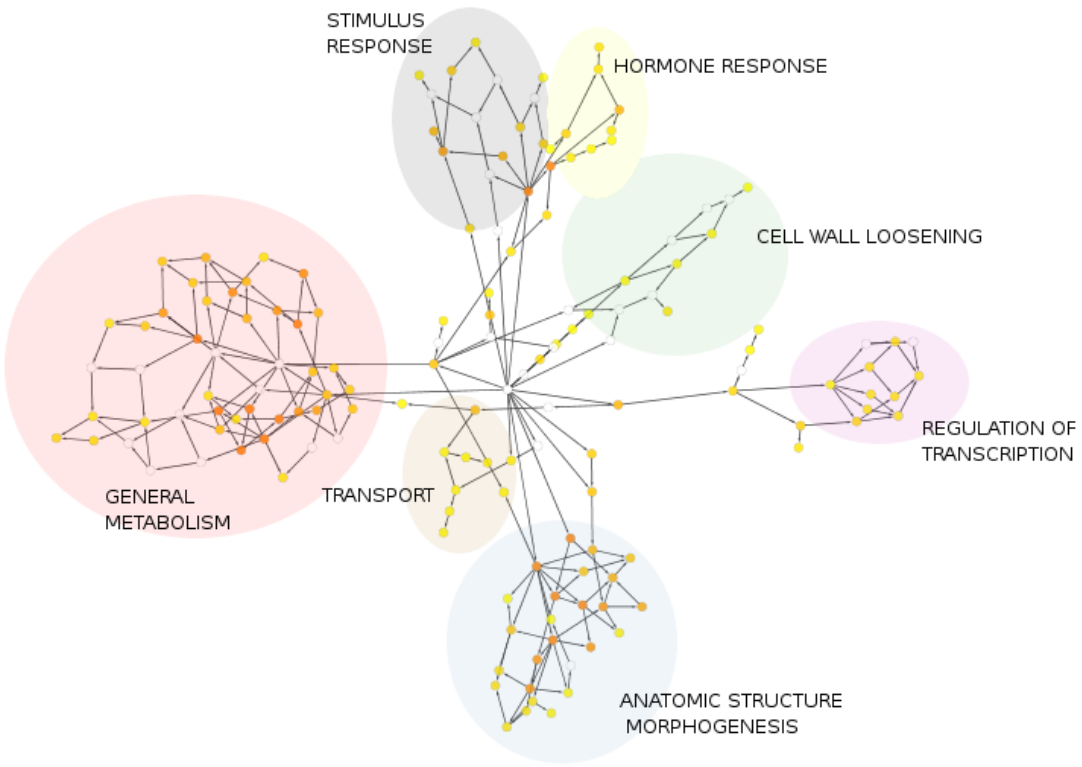
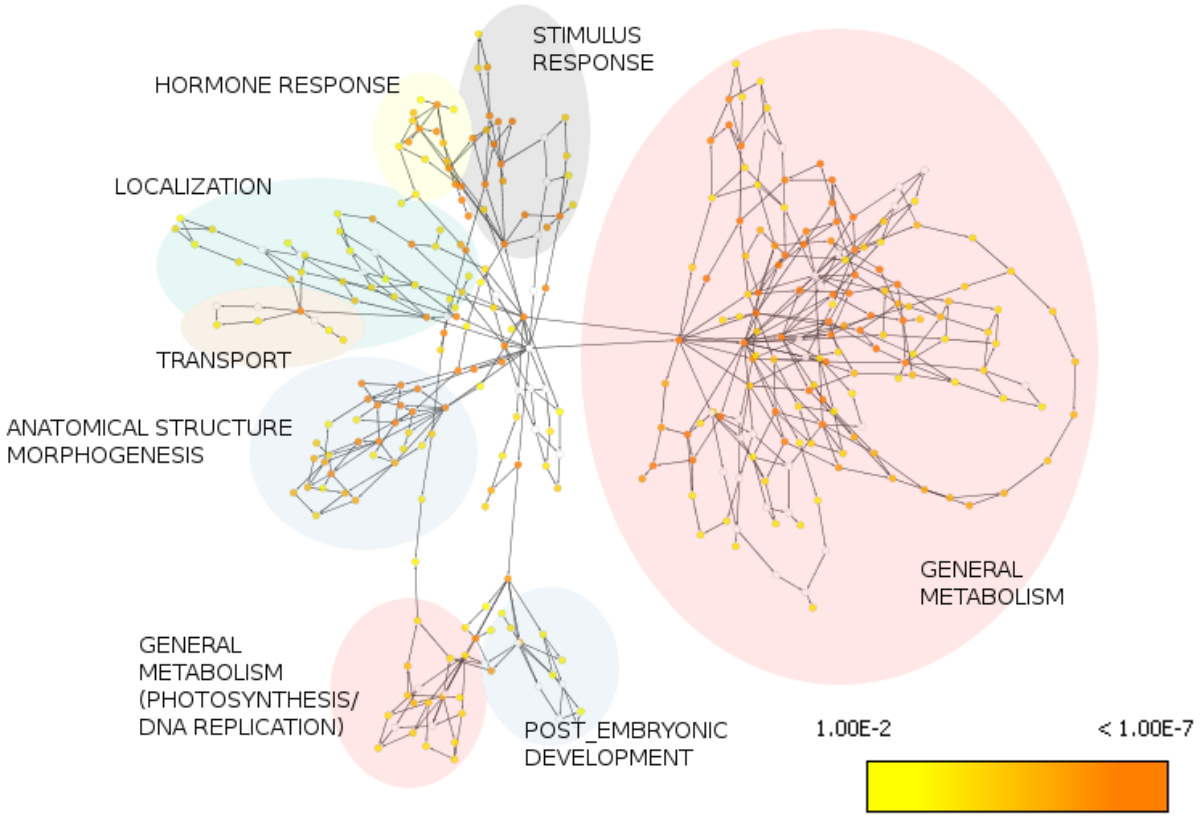


Figure 3.4 Biological Process GO Enrichment for OE time course.
DAVID was used to identify all biological process gene ontology categories curated by TAIR among all genes statistically significant ($p<0.01$) at each time point on the *STM* over-expression time course (A – 8 hours, B – 24 hours, C- 72 hours, D – 9 days). All categories with $p<0.01$ or lower (hypergeometric distribution) were considered significantly enriched and BINGO was used to output graphs showing the relationships between statistically significant categories. Coloured nodes are statistically significantly enriched, uncoloured nodes are not and are included to show the connections between categories. The more orange the enriched node, the lower the p-value as shown in the key. In each graph, categories have been manually grouped into similar processes, a full table of enriched categories is given in Appendix 11.

C


STIMULUS
RESPONSE
HORMONE RESPONSE
CELL WALL LOOSENING
REGULATION OF
TRANSCRIPTION
GENERAL
METABOLISM
TRANSPORT
ANATOMIC STRUCTURE
MORPHOGENESIS

D


STIMULUS
RESPONSE
HORMONE RESPONSE
LOCALIZATION
TRANSPORT
ANATOMICAL STRUCTURE
MORPHOGENESIS
GENERAL
METABOLISM
GENERAL
METABOLISM
(PHOTOSYNTHESIS/
DNA REPLICATION)
POST_EMBRYONIC
DEVELOPMENT
1.00E-2
< 1.00E-7

significant proportion of the *STM* over-expression phenotype may be mediated through modulating other transcription factors or through modulation of hormone responses at the transcriptional level. At 24 and 72 hours, GO terms related to GA response are the only hormone response terms specifically referring to a given phytohormone. However, probesets corresponding to some cytokinin (CK) and auxin responsive genes are also significantly differentially expressed at these time points (such as *TYPE-A ARABIDOPDIS RESPONSE REGULATOR* and *AUXIN RESPONSE FACTOR* genes; Tiwari et al, 2003; Beuchel et al, 2009).

Cell wall loosening terms are enriched early in the timecourse at 8 hours and then again at 72 hours. At 8 hours, *STM* OE represses a number of expansins (*EXLB1, EXPA5, EXPA11*) (Review: Cosgrove et al, 2002), which promote cell growth through control of cell wall plasticity, and induces one extensin (*EXT4)* which has been shown to have a role in promoting stem thickness. As *STM* OE plants exhibit reduced cell size, this suggests a possible, rapid mechanism for this process. *EXPA11* and *EXT4* remain significantly differentially expressed at all time points showing the same direction of change, suggesting that this effect is maintained at subsequent time points even though the gene category is not significantly enriched. Cell cycle progression is also significantly enriched – importantly this category contained *SIAMESE* repression (Churchman et al, 2006), a gene which contains a motif shared with the *KRP* genes which encode inhibitors of CDK activity. In addition, three *KRP* genes were differentially expressed by *STM* at 72 hours – *KRP1, 2* and *5*. *SIAMESE* is known to promote endoreduplication along with *KRP1*, which has been shown to rescue the *SIAMESE* phenotype. Thus the joint downregulation of *SIAMESE* (0.3 fold on the $\log_2$ scale) and *KRP1* (0.32 fold on the $\log_2$ scale) suggests that this could explain the reduced endoreduplication phenotype reported by Scofield et al (2013) in *STM* OE, though the

low levels of fold change make this a weak argument without further experimental analysis.

Over the time course a number of categories relating to more general metabolic processes, transport, localization and non-hormonal stimulus responses become enriched. However, with a few exceptions it is difficult to relate these to known phenotypes of *STM* over-expression. For example, while polar auxin transport terms are enriched at 72 hours, the remaining transport terms are generic terms such as amino acid transport, though this may include auxin transport as auxin transporters are amino acid permeases (Tegeder & Ward, 2012), a number of which, such as *AAP7* are differentially expressed following *STM* induction. It is interesting that this is observed at the same point in the time series as auxin responsive genes are induced. Similarly the non-hormone stimulus categories include terms such as response to nematodes, which are likely to be due to common stress response pathways or possibly artefacts due to the experimental conditions in which the plants were grown.

3.2.4 Statistical Analysis of RNAi time course

The *STM* RNAi time course consisted of 72 hour and 9 day time points contrasted against DEX-treated empty vector controls. 8 and 24 hour time points were not possible as experiments performed by Simon Scofield had suggested that 72 hours was required to detect at least a 2-fold downregulation of *STM*. At a significance threshold of $p < 0.01$, 157 genes were significantly differentially expressed at 72h, and 389 at 9 days. However, neither of the RNAi datasets were particularly similar to one another, as can be seen in Figure 3.5. The likelihood of the overlap between the 72 hour and 9 day RNAi datasets having occurred by chance was only just less than 0.05, and only 5 genes were differentially expressed in common between them (*SHOOT MERISTEMLESS, YABBY5, SAMC2*, At5G37890 and At1G06740 (*YABBY5:* Meister et

al, 2005; *SAMC2:* Palmieti et al, 2006)) . Additionally, while the datasets have some commonality with the 8 hour over-expression dataset, they are most similar to the 72 hour and 9 day datasets. Of the five genes overlapping between the RNAi time points, 3 are reciprocally regulated, suggesting that either negative feedback plays a bigger role, or that the samples from the two time points are quite different in transcriptomic terms. Given that we expect there to be reciprocal expression between the RNAi lines and the OE time course, it was unexpected that at 9 days of overexpression, we only observe 50/190 genes overlapping reciprocally expressed between the OE and RNAi time courses. The contrasts which appear to be most reciprocal are the 9 day RNAi against the 72 hour OE contrast and the 72 hour RNAi contrast against the 9 day E (figure 3.5). This suggests that the 72h RNAi may be more comparable at all time points to the OE experiment than the 9d RNAi, however, the overlap is still less than 50% of the dataset and thus we cannot regard the RNAi line as a straightforward opposite to the OE lines.

One conclusion from this is that the RNAi time points appear to be closer to the later *STM* over-expression time points. This suggests that while downregulation was only clearly detected at a 2-fold level by 72 hours, we are detecting changes among other genes by that stage. This suggests that less than 2-fold perturbation of *STM* is sufficient to induce a response at the transcriptional level.

Another interpretation of this discrepancy between the over-expression and RNAi datasets may be that while in the over-expression lines *STM* is acting ectopically, in the RNAi datasets, *STM* will only have been repressed in tissues it was already expressed. For genes with expression domain both within and outside the meristem (i.e. not meristem specific), the disruption of *STM* expression by RNAi was only able to affect their expression in the part of their domain co-incident with *STM* expression. Thus changes in expression may have been masked as the power of the experiment may be

insufficient to detect a small change in the SAM as most expression signal comes from outside this domain. Both these issues would affect the ability of the microarray experiment to detect significant differences in expression. This also brings into question the sensitivity of the microarrays used, as shown in Figure 3.1, by qRT-PCR *STM* is downregulated by around 20% from wild type levels. This discrepancy suggests that the arrays may be underestimating the true level of downregulation.

Observed fold changes in the RNAi time course are smaller on average than for the overexpression time course as can be seen in Figure 3.6, and the range of p-values obtained are closer to those seen in the 8 and 24 hour datasets. This is unsurprising since as previously stated the over-expression lines resulted in ectopic expression of *STM* whereas the RNAi line is affecting a narrower expression domain.

A

	8H OE	24H OE	72H OE	9D OE	72H RNAi	9D RNAi	stm-2	Total
8H OE		74	84	83	2	4	16	92
24H OE	74		262	261	2	6	34	321
72H OE	84	262		2709	93	95	307	4133
9D OE	83	261	2709		52	190	419	8070
72H RNAi	2	2	93	52		5	1	157
9D RNAi	4	6	95	190	5		10	385
stm-2	16	34	307	419	1	10		599
Total	92	321	4133	8070	157	385	599	

B

	8h OE		24h OE		72h OE		9d OE	
72 h RNAi	2	0	1	1	1	92	34	19
9 day RNAi	2	2	3	3	71	24	50	140

C

	8H OE	24H OE	72H OE	9D OE	72H RNAi	9D RNAi	*stm-2*
8H OE		0	0	0	0.0251	0.0193	0.0168
24H OE	0		0	0	0.374	0.289	2.47×10^{-13}
72H OE	0	0		0	0	6.21×10^{-10}	0
9D OE	0	0	0		0.001	0	0
72H RNAi	0.0251	0.374	0	0.001		0.049	0.921
9D RNAi	0.0193	0.289	6.21×10^{-10}	0	0.049		0.413
stm-2	0.0168	2.47×10^{-13}	0	0	0.921	1.1.	

Figure 3.5 – Size and significance of overlap between time points analyzed.

A) The number of probesets common to each contrast in the experiment described in Figure 3.1. Green indicates the overlap is significant ($p<0.01$, hypergeometric distribution), Orange indicates the overlap is almost significant ($p<0.05$, hypergeometric distribution) and red indicates that the overlap is not significant ($p>0.05$, hypergeometric distribution). Grey bars indicate total number of genes significant ($p<0.01$) in each contrast.

B) The number of genes reciprocally expressed (purple columns) and non-reciprocally expressed (orange columns) between each pair of RNAi and OE time points.

C) The probabilities that the overlap observed occurred by chance, calculated using the hypergeometric distribution. Colouring is given as per Figure 3.5A.

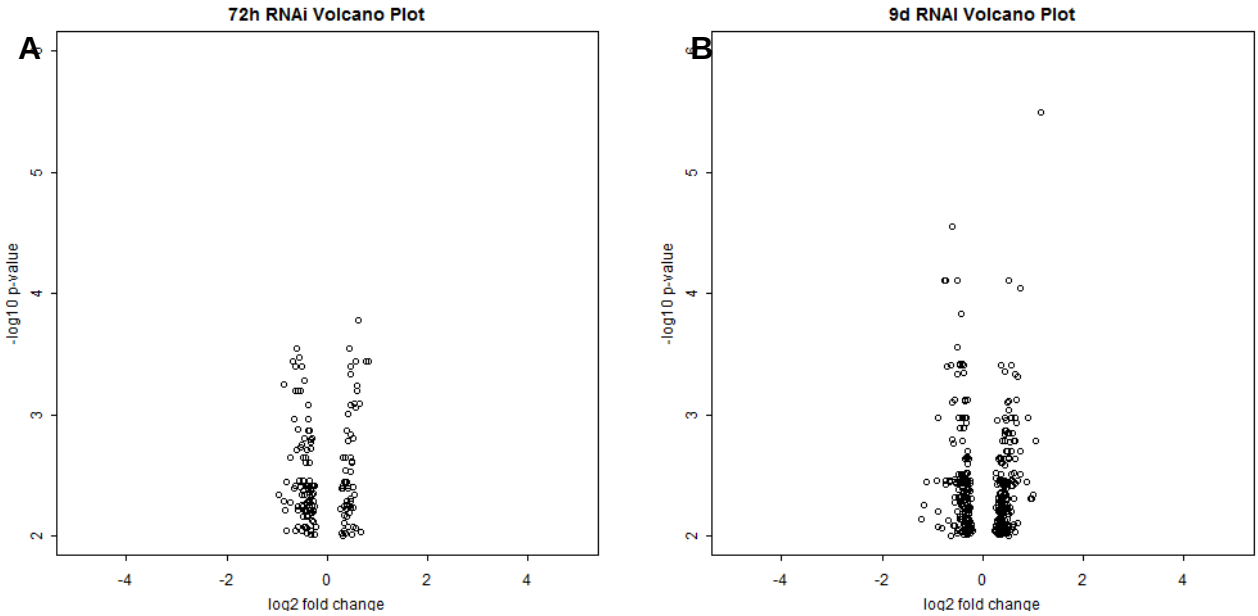


Figure 3.6 – Volcano Plots showing magnitude of Induction vs significance across the RNAi time course.
Volcano plots of point estimates of $\log_2$ fold changes calculated using LIMMA/RMA vs -$\log_{10}$ times adjusted p-values for *STM* RNAi lines induced for A) 72 hours and B) 9 days. Only shown are those probesets with a q-value <0.01 at each over expression contrast.

3.2.5 Statistical analysis of *stm-2* dataset

While the RNAi experiment was designed as a reciprocal time course experiment to the *STM* over-expression time course, its lack of power and small number of time points made it more difficult to use for this purpose. Given this, the *stm-2* mutant vs wild type contrast was investigated for its potential use as a reciprocal experiment. As can be seen in Figure 3.5, the *stm-2* dataset has a large and statistically significant overlap ($p<0.01$) with all of the over-expression datasets. In particular, 70% of the 599 probesets identified as significantly differentially expressed in the *stm-2* dataset are also significantly differentially expressed in the 9 day *STM* OE dataset and in the equivalent contrast with the 72 OE hour dataset, 50% of probesets are shared.

Interestingly, only 44 of the 419 genes show reciprocal expression between the 9 day OE and the *stm-2* datasets and only 19 of the 307 shared with the 72 hour OE dataset. This discrepancy may be partially due to the fact that in the mild *stm-2* mutant there is periodic abort-retry of meristems, which may result in the presence of multiple meristems in one sample. This could lead to enrichment rather than a depletion of meristematic genes. However it could also be due to feedback responses to a depletion of *STM*. A similar effect was observed in the RNAi 9 day dataset, where only 50 out of 190 genes were reciprocally expressed compared to the 9 day OE line.

As we can see from Figure 3.7, the range of observed fold changes and p-values is similar to the RNAi lines and earlier *STM* over-expression datasets. Higher maximum p-values and fold changes were observed for over-expressed than down-regulated genes.

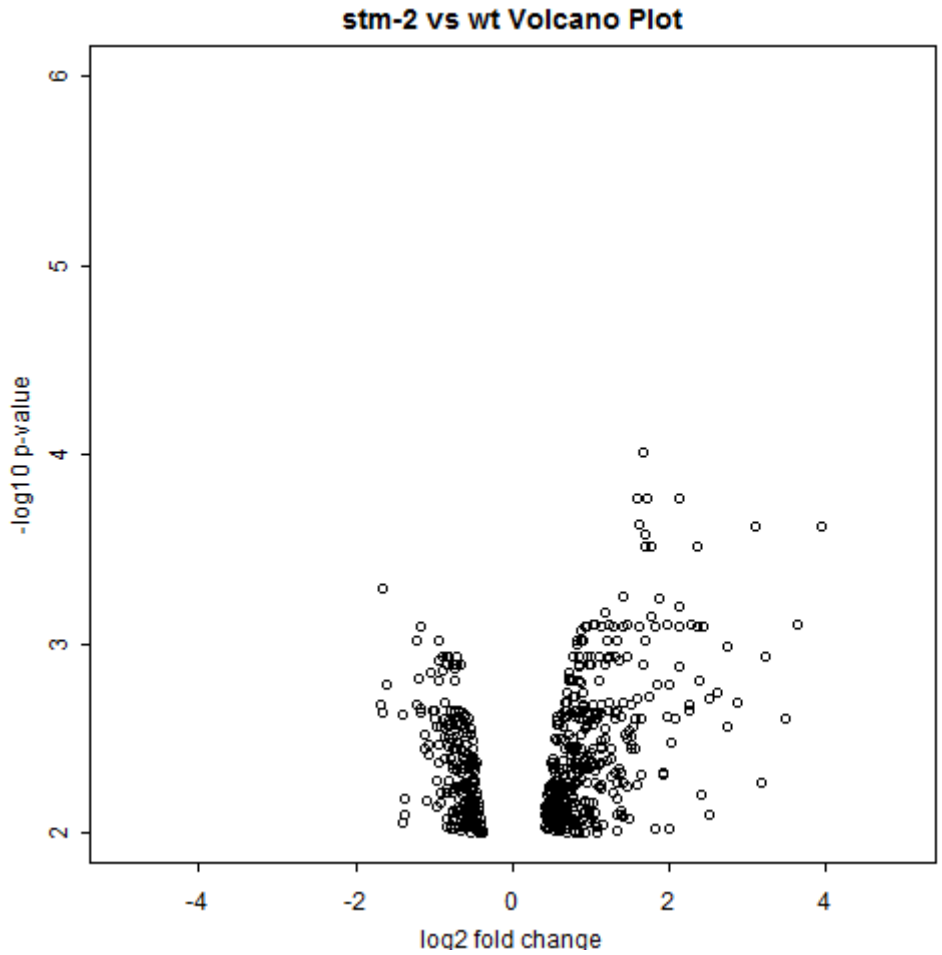


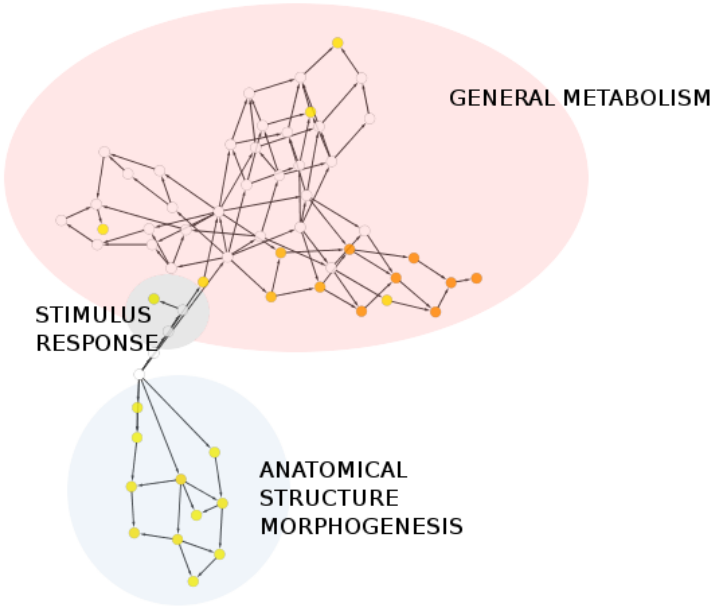
Figure 3.7 – Volcano Plots showing magnitude of Induction vs significance in the *stm-2* vs *wt* contrast .
Volcano plots of point estimates of $\log_2$ fold changes calculated using LIMMA/RMA vs -$\log_{10}$ times adjusted p-values for *stm-2* vs *wt* contrast. Only shown are those probesets with a q-value <0.01 at each over expression contrast.

3.2.6 GO Enrichment of RNAi Time course and *stm-2*

Compared to the over-expression time course, the number of significantly enriched GO terms is far lower in the RNAi and stm-2 datasets. No biological process terms are significantly enriched for the 72 hour RNAi line, and for the 9 day RNAi time point, the main interesting responses are related to anatomical structure morphogenesis. As can be seen in Figure 3.8, the most specific GO terms within this group relate to reproductive tissue development, supporting the argument that the RNAi time course has more in common with the later *STM* over-expression time points.

For the *stm-2* dataset, the results for the GO enrichment analysis superficially show more in common with the *STM* 9 day and 72 hour time points in terms of the classes of GO terms enriched, particularly as terms relating to anatomical structure morphogenesis and hormone response are enriched. However, the specific categories are somewhat different. Stomatal development is the most specific leaf node on the GO graph for anatomical structure morphogenesis and the hormone stimulus terms do not contain specific enrichment for GA, CK or auxin.

A



B

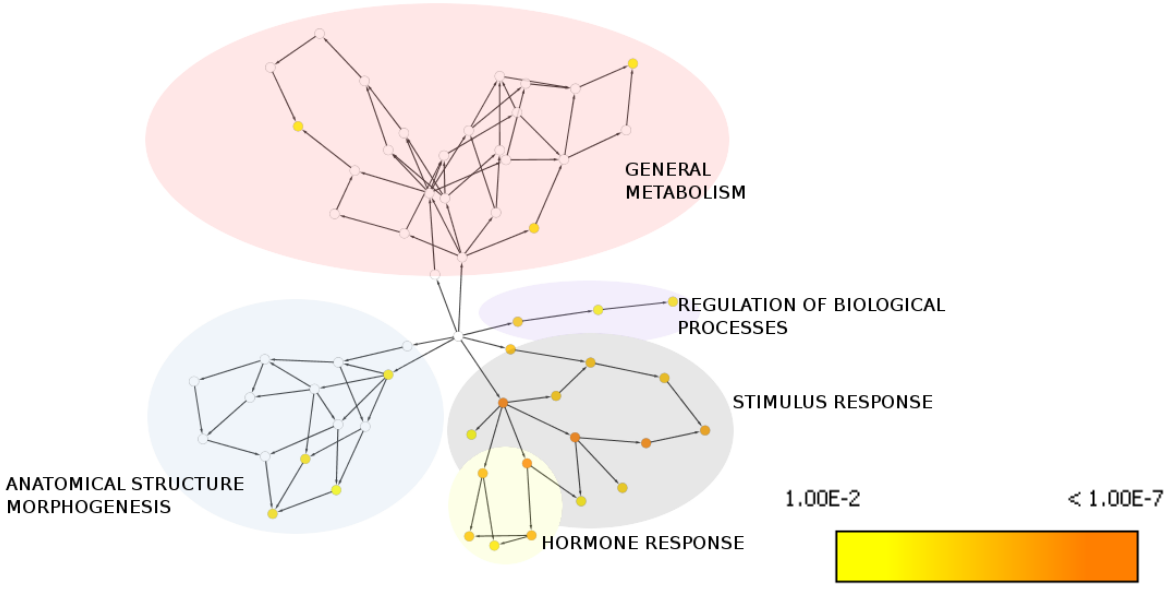


Figure 3.8 Biological Process GO Enrichment for 9 day *STM* RNAi and *stm-2* vs *wt* experiment.
DAVID was used to identify all biological process gene ontology categories curated by TAIR among all genes statistically significant ($p<0.01$) at A – 9 day *STM* RNAi, B – *stm-2* vs *wt*. All categories with $p<0.01$ or lower (hypergeometric distribution) were considered significantly enriched and BINGO was used to output graphs showing the relationships between statistically significant categories. Coloured nodes are statistically significantly enriched, uncoloured nodes are not and are included to show the connections between categories. The more orange the enriched node, the lower the p-value as shown in the key. In each graph, categories have been manually grouped into similar processes, a full table of enriched categories is given in Appendix 11.

3.2.7 Statistical Analysis of a separate *STM* induction system

As it was unclear how similar the OE and RNAi lines were to one another, an additional experiment was performed using an *35S*:*STM-GR* line obtained from Rüdiger Simon (Brand et al, 2002). In this line, *STM* fused to the rat glucocorticoid receptor is ectopically and constitutively expressed throughout the plant. It is excluded from the cell nucleus through interactions with heat shock proteins 90/70 until dexamethasone (DEX) is applied, whereupon the association with HSPs is disrupted and STM-GR translocates to the nucleus. Thus, induction is far more rapid than in the 2-component system. As it is expressed constitutively, we do not expect to observe transcriptional induction of *STM* expression itself using this system unless *STM* autoregulates.

207 genes were significantly differentially expressed at a p-value of 0.01 or lower. While the size of the overlaps were less than expected, as seen in Figure 3.9, they are still highly statistically significant at all time points apart from with the RNAi lines. This further corroborated the hypothesis that the RNAi lines were more similar to the later time points than the earlier ones. While it appears there is a common subset of genes which have been identified between the two over-expression experiments, roughly 75% of the genes identified in the 8 hour OE dataset were not identified as significant in this experiment.

	8h OE	24h OE	72h OE	9d OE	72h RNAi	9d RNAi	*stm-2*
Overlap	23	53	125	148	0	3	27
p-value	0	0	0	0	0.76	0.46	4.38×10^{-13}

Table 3.9 – Overlap Between *STM* time course experiment and *STM-GR* experiment.
The top row indicates the number of genes present in the overlap between the *STM*-GR microarray experiment and each point on the Time course microarray experiment described in Figure 3.1. The bottom row indicates the probability that the overlap observed occurred by chance, calculated using the hypergeometric distribution. Significant values ($p<0.01$) are shaded in green, non-significant values are shaded in red.

To investigate the similarity between the different datasets, multidimensional scaling was applied to the Euclidian distance between raw expression values from each pair of datasets. The mapping of these to 2-dimensional space can be seen in Figure 3.10. The distance between each point represents the typical $\log_2$ fold change observed between pairwise samples between each dataset. As can be seen, this suggests that the variability between the datasets can be viewed in terms of level (or duration) of *STM* over-expression (y-axis) and also in terms of the *STM* induction or activation method (i.e whether the lines used were *STM-GR* or otherwise (x-axis)). The difference between the two over-expression systems is likely not only due to the dynamics of each system, but also the fact that each is in a separate genetic background (TGV based lines in *Landsberg erecta,* STM-GR in Col-0), suggesting that minor differences in experimental conditions and wild type phenotypes could have a role in determining the differences in expression between the two datasets. However, since the overlap between the over-expression datasets using the TGV system and the GR system is almost impossible to have occurred by chance, this suggests that two experiments are measuring a large proportion of "true" effects. It should also be borne in mind that the

two expression systems have different dynamics, as the rapid translocation of already-translated *STM-GR* into the nucleus would permit less time for negative feedback responses to take place. However, since the mock datasets also cluster with the DEX-treated datasets, it is unlikely that this explains the differences between the various datasets. An additional confounding factor may be that the experiments were performed by different individuals in different locations (Cardiff and Cambridge).

The positioning of the 3h DEX samples along dimension two suggests that these samples are most closely related to the 8 and 24 hour OE samples generated via the TGV line in terms of observed fold changes. This is corroborated by the volcano plot of fold changes against p-values for this contrast (Figure 3.11) which shows a similar distribution of fold changes to the 8 and 24h OE TGV datasets. The positioning of the RNAi lines among the empty vector lines corroborates the evidence that these lines had lower power to detect significant effects due to the lower fold changes observed. While the average fold changes for the *stm-2* line appear to be closer to the early TGV OE time points, this does not invalidate the observation that the overlap between groups of significant genes is more similar to the later TGV OE time points.

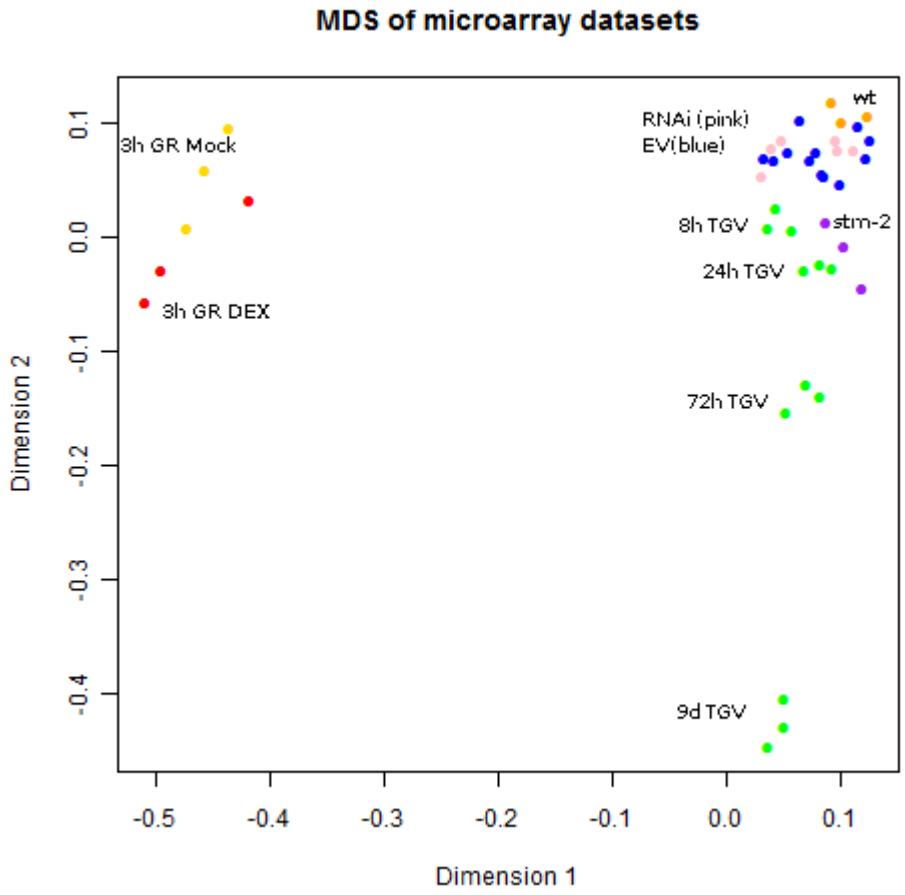


Figure 3.10 - MDS plot of *STM* over-expression, mutant, RNAi and control microarray datasets.
MDS plot of log_2 expression values of all probesets computed using RMA in each STM over-expression, mutant, RNAi and control line described in this chapter. *STM* over-expression lines (green), RNAi lines (pink), Empty Vector Controls (blue), Mock treated *STM-GR* (yellow), DEX treated *STM-GR* (red), *stm-2 (*purple) and wild type (orange).

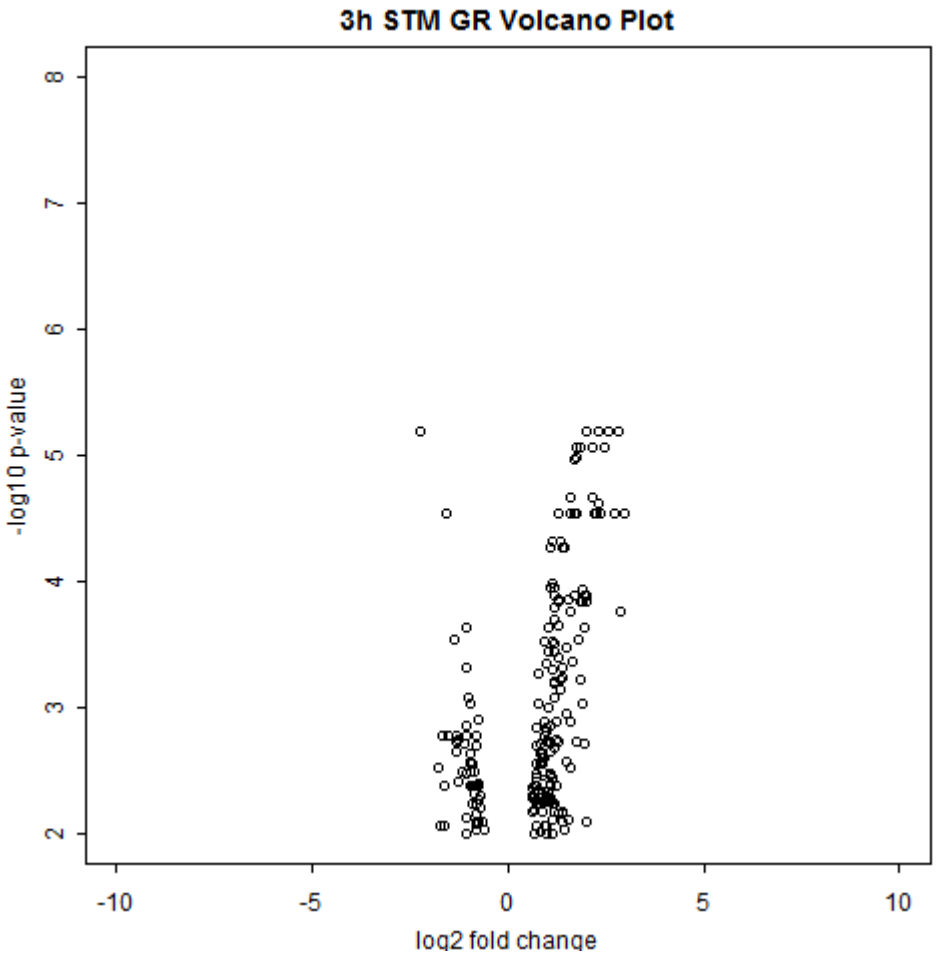


Figure 3.11 – Volcano Plot showing magnitude of Induction vs significance in the 3h *STM-GR* Over-expression experiment.
Volcano plots of point estimates of $\log_2$ fold changes calculated using LIMMA/RMA vs -$\log_{10}$ times adjusted p-values for 35S::*STM*-GR lines treated with 60μM DEX vs Mock treated plants. Only shown are those probesets with a q-value <0.01 at each over expression contrast.

3.2.8 GO Enrichment of *STM-GR* experiment

Gene Ontology enrichment for the 3h DEX-Mock contrast provides a much clearer link to *STM*'s over-expression and mutant phenotype than the 8 hour or 24 hour TGV datasets, which appears to be most consistent with fewer secondary effects within this time frame. In particular, as can be seen in Figure 3.12, we see that in the anatomical structure morphogenesis categories, the terminally enriched leaf node corresponds to

meristem development. Additionally, cell fate specification is enriched in a separate-developmentally related cluster.

In addition to this a large number of patterning and regionalization terms contain genes such as *CUP-SHAPED COTYLEDON1 (CUC1*) (Aida et al, 1997) and *ASYMMETRIC LEAVES1 (AS1;* Byrne et al, 2002) - though the effect upon *AS1* is small - both of which were upregulated and which have been shown to have genetic interactions with *STM* and other *KNOX* family proteins. We also see a strong transcription factor response, 39 transcription factors are significantly differentially expressed in this dataset, which comprises a statistically significant enrichment of this category.

In common with the 8 hour TGV OE contrast, we again observe cell-wall loosening terms enriched, and in common with all subsequent TGV OE contrasts we see an enriched hormone response is found. In this case, the terminal enriched leaf nodes correspond to gibberellin stimulus, as was the also seen at 24 and 72 hours in the TGV *STM* OE time course.

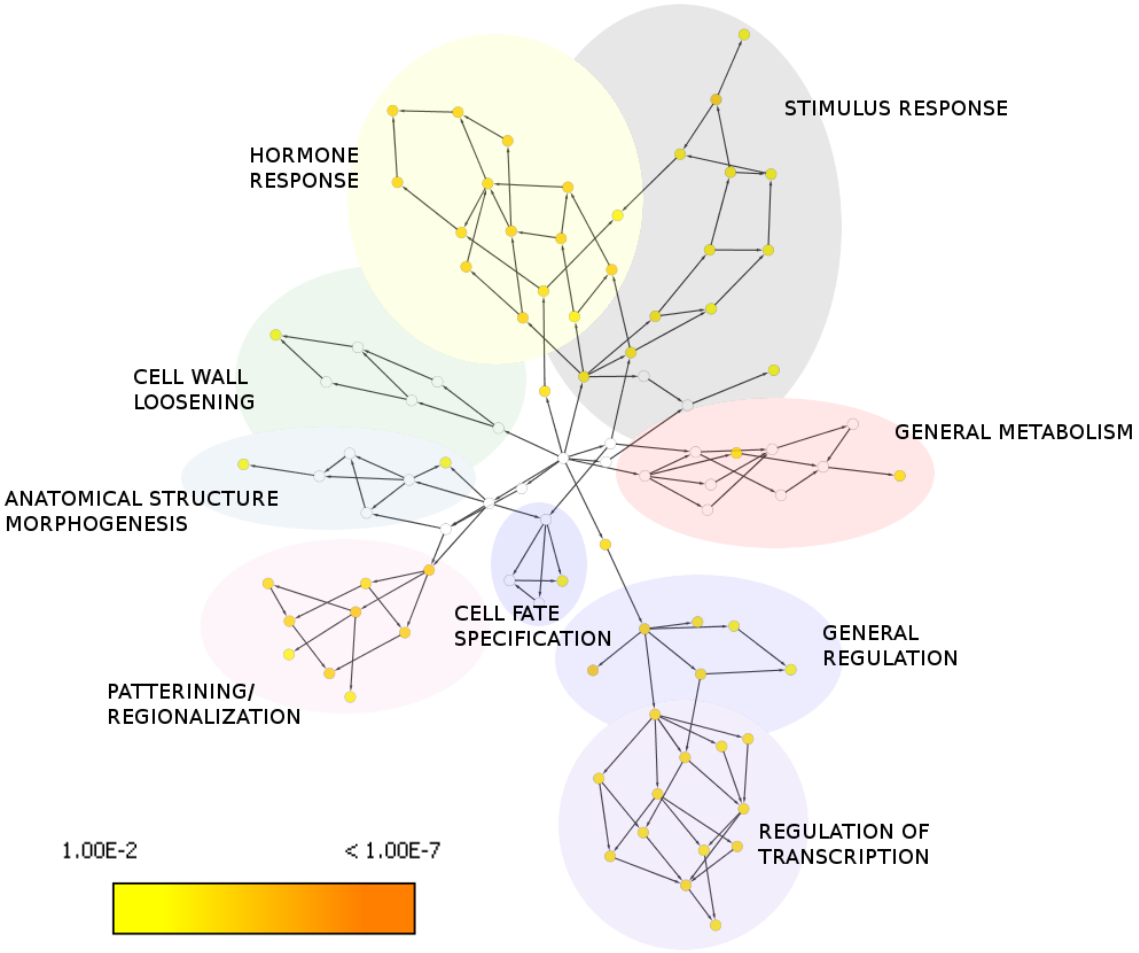


Figure 3.12 - Biological Process GO Enrichment for *STM-GR* experiment.
DAVID was used to identify all biological process gene ontology categories curated by TAIR among all genes statistically significant ($p<0.01$) following 3h of DEX induction in the *35S::STM-GR* line compared to mock treated plants. All categories with $p<0.01$ or lower (hypergeometric distribution) were considered significantly enriched and BINGO was used to output graphs showing the relationships between statistically significant categories. Coloured nodes are statistically significantly enriched, uncoloured nodes are not and are included to show the connections between categories. The more orange the enriched node, the lower the p-value as shown in the key. In each graph, categories have been manually grouped into similar processes, a full table of enriched categories is given in Appendix 11.

3.2.9 Overlap between the *STM* time course and previous studies of *STM* induction

STM induction has been previously examined by Spinelli et al (2011) using ethanol-inducible lines. In contrast to the time course experiment described here, they generated microarrays at one 12 hour time point, from one line inducing *STM* alone, and one inducing the *STM* coding sequence fused in frame to the herpes simplex virus transactivation domain VP16 (Spinelli et al, 2011). The plants used were two-week old seedlings, and they used two biological replicates. As the seedlings were two weeks old there may also have been some reproductive tissue present in their dataset.

While it would be expected that the datasets should be similar some key differences would be anticipated, partially due to the difference in age of plants between the two experiments (14 days compared to 9 days), but mainly due to the decreased number of replicates. The lower number of replicates will have negatively affected their power to detect significant effects, a problem exacerbated by the fact that these authors subsequently took the intersection of the two sets of significant genes in generating their list of significantly differentially expressed genes. Thus if either of their experiments had been less effective, by taking the intersection they could never detect genes significant outside of the lower-power dataset.

129 genes were defined as significantly differentially expressed in their experiment, and as can be seen from Figure 3.13, the overlap between all time course data points is significant at $p<0.01$. Importantly the overlap between their dataset and the RNAi lines, although small is also statistically significant, this provides additional confidence that the RNAi lines are providing meaningful data. As data was only available for significantly differentially expressed genes within their dataset and no raw expression data was

available, MDS was not performed. Overall, both experiments have produced similar results, which provides additional validation of the time course methodology, however, as they chose statistical methodologies which limit their power, and chose a different developmental stage, it is expected that the time course experiment should have captured a more robust set of differentially expressed genes, relevant to early growth and development.

	8h OE	24h OE	72h OE	9d OE	72h RNAi	9d RNAi
Overlap Size	17	35	78	92	5	8
p-value	0	0	0	0	0.000233	0.000306

Table 3.13 – The Overlap between *STM* OE and RNAi Time course and Spinelli et al (2012) Experiments.
The top row indicates number of probesets present in the overlap between those probesets called statistically significant in the Spinelli et al (2012) experiment and each point on the *STM* over-expression and RNAi time course experiment ($q<0.01$ as computed using Limma and RMA with FDR correction). The p-value given on the bottom Row indicates the likelihood that the overlap observed occurred by chance, calculated using the hypergeometric distribution.

3.3 Meta-Analysis of *STM* Time Course Experiment

3.3.1 Rationale for Meta Analysis

Meta-analysis is the integration of results from multiple comparable experiments to compute a combined p-value for the likelihood that the joint null hypothesis being tested is rejected. To be comparable, such experiments must be testing the same null hypothesis, and be independent of one another. Meta-analysis has been used to remove false positives which may be detected in a single dataset but not others, reduce

false discovery rate by demanding reproducibility between experiments and to boost the power to detect smaller changes which occur consistently throughout the datasets at a level which would not otherwise be considered significant.

Across the microarray time course, although the overlap between datasets which should capture similar information was unlikely to have occurred by chance, it was clear from the differences between the large number of significantly differentially expressed genes (particularly at later over-expression time points) that the results were noisy and there was heterogeneity between the response to *STM* upregulation and the response to *STM* downregulation. By applying meta-analysis procedures to appropriate combinations of contrasts, it was thought to be possible to identify those genes which robustly changed over specific independent subsets of experiments selected as testing equivalent null hypotheses. Three independent combinations of experiments were considered as plausibly testing comparable null hypotheses, these were:-

- **Null Hypothesis 1: Gene X does not show a rapid response to *STM* differential expression. i.e. Rapid Response**
    - Datasets: 8h OE - 8h EV, 24h OE - 24h EV, 3h *STM-GR* vs Mock
    - Rationale: The earliest OE time points should capture the most rapidly differentially expressed *STM* targets. The 8h dataset was almost completely encapsulated by the 24h dataset, however it was unclear whether all rapidly responding targets were expressed at a sufficiently differential level in the 8h dataset to be called significant. Thus we argue that the 24 hour OE dataset may also be considered as enriched for rapidly responding targets, and may have captured those missed in the 8h dataset due to marginal significance. Multi-dimensional scaling had previously suggested that compared to the

other *STM* perturbed lines, while the *STM-GR* experiment was similar to the earlier time points along one principal component, the difference in expression system meant that it was not as directly comparable as hoped. By using a more stringent meta-analysis procedure, it was expected that only the most significant results would be maintained.

- **Null Hypothesis 2: Gene X does not show a robust response to *STM* upregulation. i.e. Robust Response**
    - Datasets: 8h OE - 8h EV, 24h OE - 24h EV, 72h OE - 72h EV
    - Rationale: As most genes show consistent changes in direction of expression between these datasets, a meta-analysis of genes differentially expressed across all three should reveal those genes robustly perturbed by *STM*. The 9 day OE contrast is excluded since the severity of its phenotype compared to the empty vector precludes certainty that differentially expressed genes are due to *STM* levels alone. Its inclusion was thus likely to skew the dataset further towards identification of far downstream effects. While this is also true at 72 hours, the effect appears to be less severe and in order to detect changes further on in the time course it was necessary to include at least one later time point.
- **Null Hypothesis 3: Gene X is not differentially expressed in phenotypic plants with perturbed *STM* expression**
    - Datasets: 9d OE - 9d EV, 9d RNAi - 9d EV, stm-2 - wt
    - Rationale: All these plants show phenotypic consequences of long-term STM perturbation The number of differentially expressed genes between all three is dramatically different. By performing a meta-analysis we would hope to

identify only the most consistently differentially expressed genes in *STM*-perturbed plants.

Fisher's inverse χ^2 squared test is a simple omnibus multiple test procedure, it combines p-values according to the formula:

$$P_{combined} = -2\Sigma_i \ln(p_i)$$

Where p_i is the p-value of interest in experiment i. This is then compared against the χ^2 distribution with 2k degrees of freedom, where k is the number of experiments. It has been commonly used in gene expression studies as although many older papers do not provide full experimental data, p-values for differential expression are usually available. The principal drawback to using Fisher's inverse χ^2 test is that information on fold changes is lost when calculating the omnibus p-value. When applying it to a 2-tailed distribution, a rejection of the null hypothesis allows no assumption as to homogeneity in direction of the effect size, however as some genes may have a fluctuating response to *STM* induction though the time course, we do not consider it to be a problem if a gene is called significant which is both significantly up and down regulated by *STM* at different time points as it may reflect complex temporal expression dynamics incorporating feedback responses.

Stouffer's Z test is a closely related omnibus procedure in which p-values are converted into the quantiles of a standard normal distribution. The average of these quantiles is then used as a test statistic against the normal distribution. It is somewhat less sensitive to single highly significant p-values skewing the final result than Fisher's inverse χ^2 test.

For each of the meta-analyses both methods were applied and compared. Corrected p-values from the previous analysis were used as input for the analyses, and corrected q-

values were subsequently calculated for the omnibus p-values using the method devised by Storey, 2002.

3.3.2 Overview of Rapid Response meta-analysis and choice of statistical analysis methods

Fisher's method in the rapid response meta analysis identified 119 significantly differentially expressed genes at a q-value of <0.01, while Stouffer's proved more selective and identified 98 genes as statistically significant. As the numbers in both datasets is low, and the intention was to derive a list of putative early responding targets, the increased sensitivity of Fisher's test was preferred to the additional specificity of Stouffer's procedure.

As can be seen in Figure 3.14 a number of key developmental processes are enriched following GO enrichment analysis of the rapid response meta analysis. These include categories related to organogenesis and regionalization, consistent with *STM* playing a critical role in meristem organization. The enrichment of carbohydrate synthesis genes is also consistent with regulation of growth as this would by necessity require regulation of the core components of the cell wall. That a number of genes related to transport are upregulated was unexpected, however, this is not inconsistent with the known phenotype of *STM*.

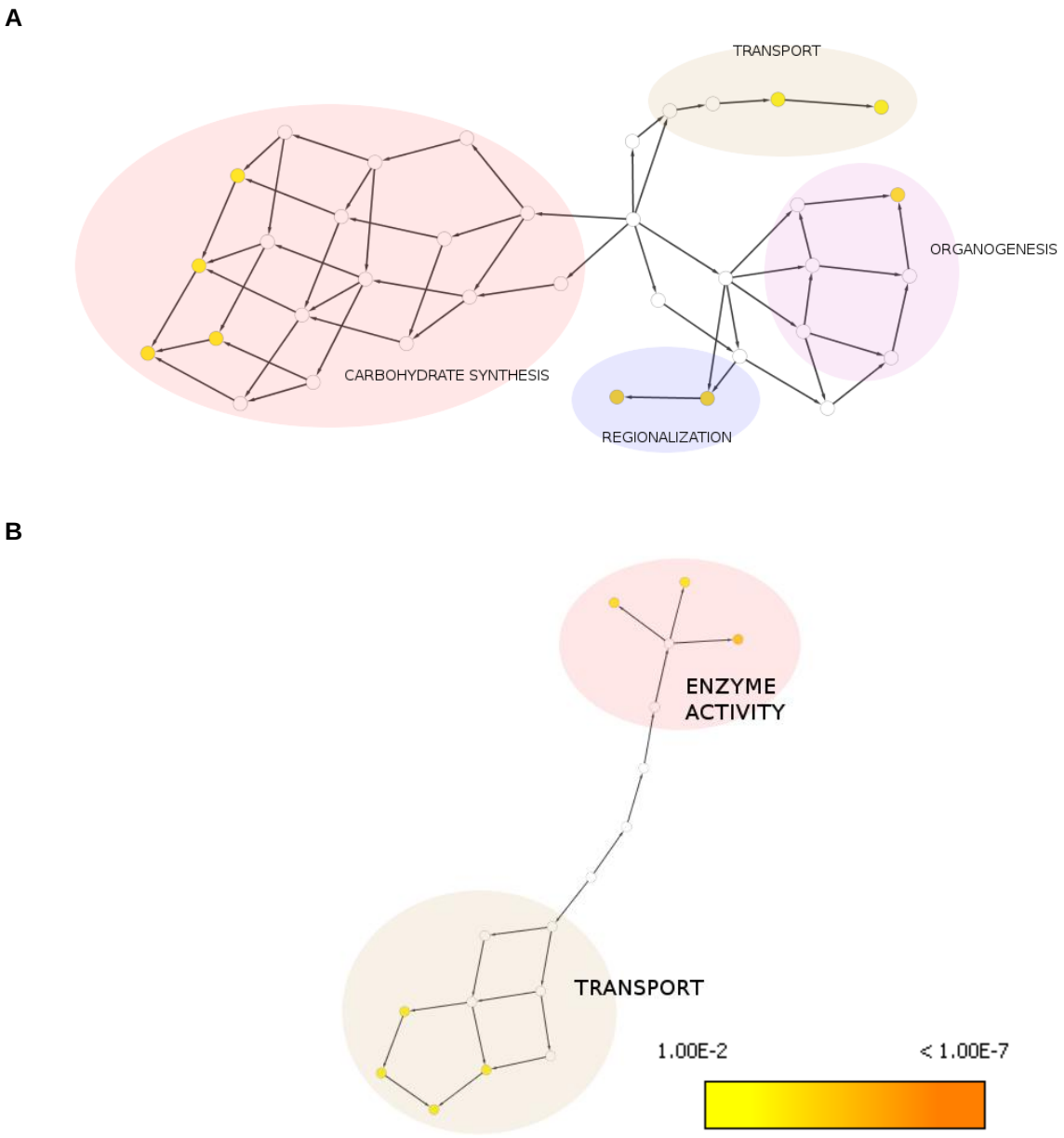


Figure 3.14 – GO Enrichment for Rapid Response Meta Analysis.

DAVID was used to identify all A) biological process and B) Molecular Function gene ontology categories curated by TAIR among all genes statistically significant ($q<0.01$) in the rapid response meta analysis. All categories with $p<0.01$ or lower (hypergeometric distribution) were considered significantly enriched and BINGO was used to output graphs showing the relationships between statistically significant categories. Coloured nodes are statistically significantly enriched, uncoloured nodes are not and are included to show the connections between categories. The more orange the enriched node, the lower the p-value as shown in the key. In each graph, categories have been manually grouped into similar processes, a full table of enriched categories is given in Appendix 11.

3.3.3 Overview of Robust Response Meta-Analysis and choice of statistical analysis methods

Stouffer's Z was used to compute omnibus p-values for the Robust Response Meta-Analysis. The intention here was to identify the subset of genes which are robustly changed across the three non-phenotypic time points on the *STM* OE time course. As such, Stouffer's increased robustness to individual outliers was considered essential, as larger fold changes might be expected in the 72 hour OE for strongly responsive genes. This choice was supported by the fact that 407 genes were considered differentially expressed ($q<0.01$) using Stouffer's Z , whereas 1,112 genes were considered differentially expressed ($q<0.01$) using Fisher's Inverse χ^2 test. While genes identified in the Fisher's Inverse χ^2 test meta analysis may be true positives, the intention of the robust response meta analysis was to identify those genes which showed the most consistent and robust changes across the OE time course. As such Stouffer's Z was used to compute omnibus p-values, as it appeared to be the more stringent method.

As can be seen in Figure 3.15, when looking at the robust response meta analysis a strong developmental response can be observed, with many more developmental categories enriched than in the rapid response meta analysis. Importantly, the enrichment of transcription factor activity suggests that the robust phenotypic responses to *STM* over expression are mediated to a large effect through regulation of transcriptional regulators. Hormone responses are also enriched, suggesting that a more thorough analysis of key phytohormone responses to *STM* induction may elucidate further details of pathways modulated by *STM* over-expression.

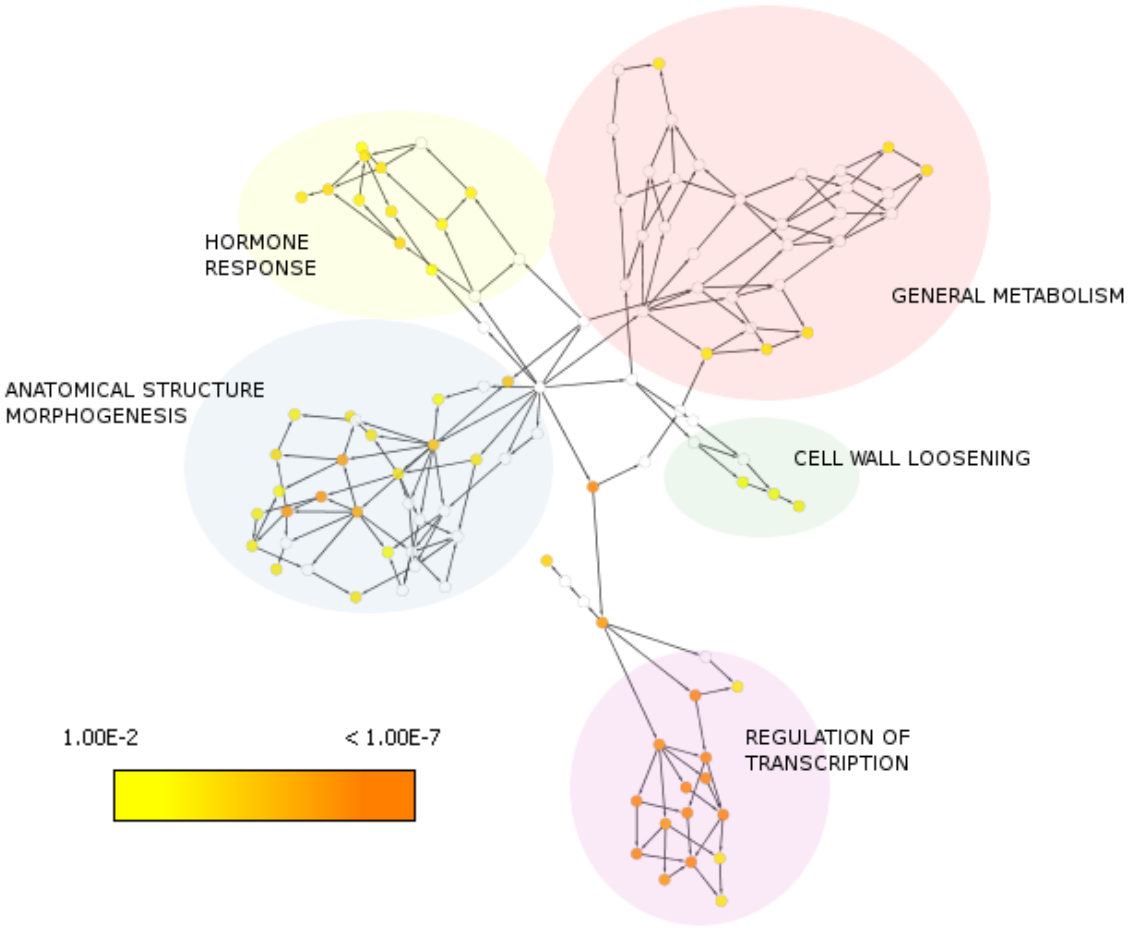


Figure 3.15 – Biological Process GO Enrichment for Robust Response Meta Analysis.
DAVID was used to identify all biological process and gene ontology categories curated by TAIR among all genes statistically significant ($q<0.01$) in the robust response meta analysis. All categories with $p<0.01$ or lower (hypergeometric distribution) were considered significantly enriched and BINGO was used to output graphs showing the relationships between statistically significant categories. Coloured nodes are statistically significantly enriched, uncoloured nodes are not and are included to show the connections between categories. The more orange the enriched node, the lower the p-value as shown in the key. In each graph, categories have been manually grouped into similar processes, a full table of enriched categories is given in Appendix 11.

3.3.4 Overview of Phenotypic Meta-Analysis and choice of statistical analysis methods

Due to the extent of the 9 day over-expression phenotype, it was expected that a meta-analysis procedure applying Fisher's Inverse χ^2 test to this dataset would result in a high number of significantly differentially expressed genes. 6,399 genes were classed as differentially expressed by this meta analysis procedure as opposed to 4,193 using Stouffer's Z test. While this is still a large number of genes, the more conservative Stouffer's method is more likely to represent a consistent phenotypic response to long-term *STM* induction than the Fisher's Inverse χ^2 test procedure and was hence used to analyze the effects of *STM* perturbation on phenotypic plants.

Due to the large number of enriched categories for biological processes, to ensure that only the most confident enrichments were examined, a p-value cutoff of 0.01 was selected for the GO enrichment analysis of the phenotypic meta analysis. The results for which, colour coded by category, are shown in Figure 3.16. As can be seen, a large number of the enriched categories related to general metabolism or stimulus responses. Similar to the robust meta-analysis, a transcriptional response can be observed, as well as a hormone response, though these are connected to nodes relating to more general regulatory and response processes respectively. This is plausibly a consequence of downstream phenotypic effects not directly mediated by *STM* over-expression or knock-out. Interestingly, transport related categories are once again significantly enriched which was also the case in the rapid response but not the robust meta analysis.

3.3.5 Overlap of Meta-Analysis Datasets and time course microarray data

Figure 3.17 shows the size of the overlaps and the likelihood that those overlaps occurred by chance for the three meta-analyses and the other microarray experiments we have previously described. As can be seen, all meta analyses overlap to a statistically significant degree with the over-expression and mutant experiments. This is unsurprising for the rapid and robust meta-analyses since we have already noted that genes which are differentially expressed at one time point in our time course commonly remain statistically significant throughout, thus there is a large amount of overlap between all datasets which we would expect to be present in a meta analysis which combined all of them.

Interestingly, the rapid response meta analysis has some statistically significant overlap with the 72h RNAi dataset, whereas the 9d RNAi dataset has statistically significant overlap with the robust and phenotypic meta analyses. Though the size of the overlaps (with the exception of the phenotypic meta analysis) is small, this provides some reassurance that the RNAi datasets are capturing sequential responses, despite the low amount of concordance which was observed between each RNAi dataset individually.

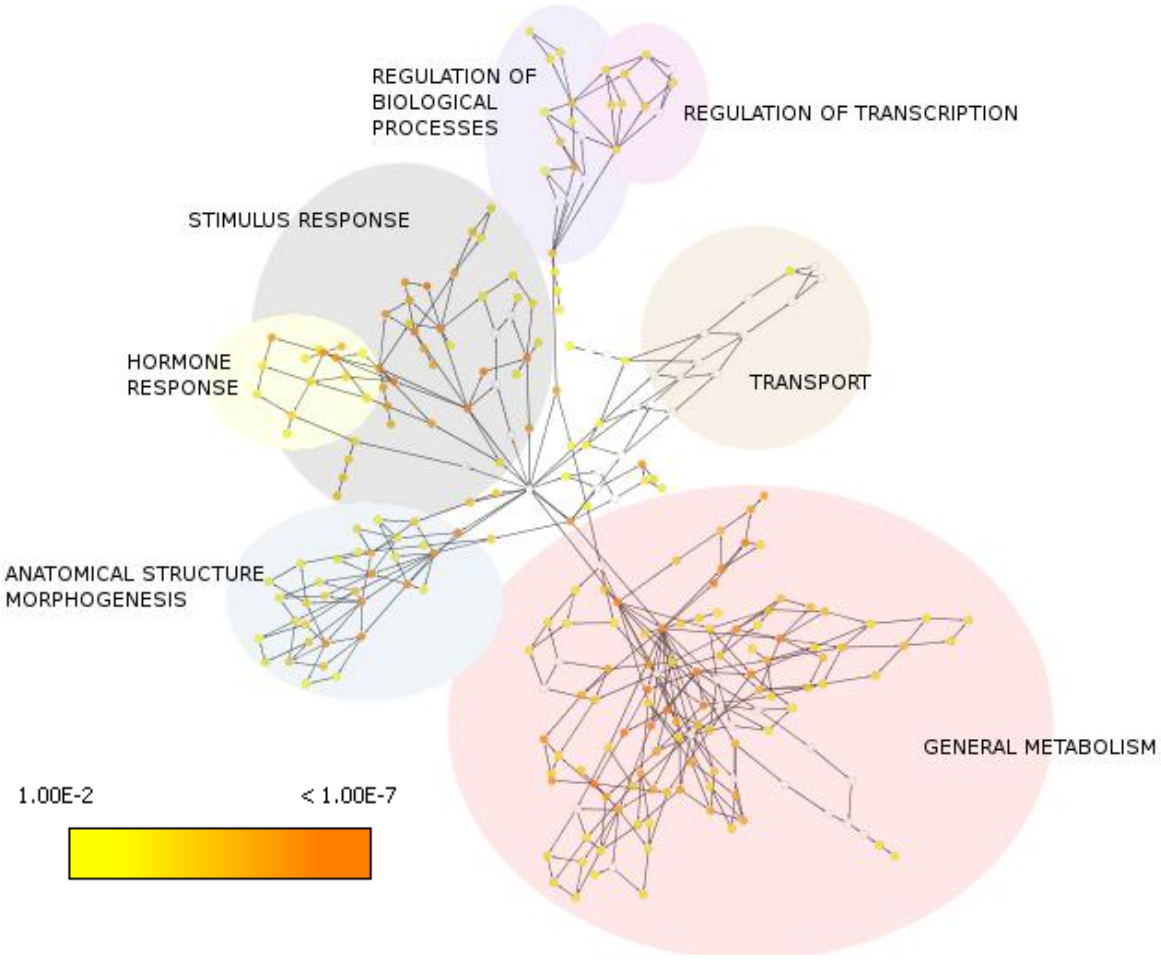


Figure 3.16 Biological Process GO Enrichment for Phenotypic Meta Analysis. DAVID was used to identify all biological process and gene ontology categories curated by TAIR among all genes statistically significant (q<0.01) in the phenotypic meta analysis. All categories with p<0.01 or lower (hypergeometric distribution) were considered significantly enriched and BINGO was used to output graphs showing the relationships between statistically significant categories. Coloured nodes are statistically significantly enriched, uncoloured nodes are not and are included to show the connections between categories. The more orange the enriched node, the lower the p-value as shown in the key. In each graph, categories have been manually grouped into similar processes, a full table of enriched categories is given in Appendix 11.

3.3.6 Robust Transcription Factor Responses to *STM* induction revealed by meta analysis

There are a number of enriched GO categories relating to TF activity in both the robust and phenotypic meta-analysis. Since we have argued that the robust meta analysis represents the most stable subset of genes responsive to *STM* it makes sense to

investigate in greater detail, the subset of TFs differentially expressed in this meta-analysis.

A

Numbers	8h OE	24h	72h	9d	72h RNAi	9d	stm-2	3h GR	Total
Rapid	68	114	114	110	2	3	17	50	119
Robust	84	237	401	375	3	11	60	64	407
Phenotypic	72	225	1910	4034	23	174	429	112	4193
Total	92	321	4133	8070	157	385	599	207	

B

	8h OE	24h	72h	9d	72h RNAi	9d	stm-2	3h GR
Rapid	0	0	0	0	0.049	0.14	$2.2x10^{-9}$	0
Robust	0	0	0	0	0.31	0.044	0	0
Phenotypic	0	0	0	0	0.87	0	0	0

Figure 3.17 – Overlap between Meta Analysis and Time Course/*STM-GR* experiments.
A) Size of overlaps between the three meta analysis categories and the other contrasts performed. OE lines are coloured in green, RNAi lines are coloured in red, the mutant line is coloured in blue.
B) Likelihood of those overlaps having occurred by chance according to the hypergeometric distribution. Green cells indicate significant results ($p<0.05$), red cells indicate non-significant results ($p>0.05$)

Interestingly, there are several transcription factor families in which several closely related genes are represented, notably, the *CUP-SHAPED COTYLEDON* family, in which the closely related *CUC1* and *CUC3* genes are significantly differentially upregulated following *STM* induction. The TCP DOMAIN PROTEIN gene family is represented by three *TCP* genes, *TCP3, 4* and *10,* all of which fall within the same phylogenetic clade, as well as *TCP24* (Cubas et al, 1999, Koyama et al, 1997, Koyama et al, 2010). All of these *TCP* genes are significantly downregulated by *STM* over-expression. Several of *STM*'s *KNOTTED* family homologues, *KNAT1/BP* and *KNAT2* (Long et al, 1996, Bharathan et al, 1999) are significantly differentially expressed in the

robust meta-analysis, along with *AS1* which is a known repressor of *KNAT* gene expression. Two competing models of *AS1* interaction with *STM* and *KNAT1* have been proposed. In the first, *STM* represses *AS1* which in turn represses *KNAT1*, in the other *STM* and *AS1* competitively regulate *KNAT1* (*STM* upregulating and *AS1* repressing it) (Byrne et al, 2002; Scofield & Murray, 2006). In this experiment, we see a mild upregulation of AS1 which suggests that the first model is not correct. Interestingly, all three are upregulated by *STM* induction. The less closely related homeobox genes *HB-7* and *HB25 (*Henriksson et al, 2005) are also significant in the robust meta analysis. Finally, both *BLADE-ON-PETIOLE* family genes, *BOP1* and *BOP2* (Ha et al, 2007), are present and upregulated by *STM* induction.

This consistent response on multiple genes in the same family suggests that *STM* induction broadly alters expression of multiple genes with similar function in boundary specification (*CUC1/CUC3)*, leaf polarity *(BOP1/BOP2)*, and leaf formation (*TCP3,4,10,24)*. Since in these cases, the genes are closely related and in some cases represent almost all of their particular phylogenetic clade, this suggests that the transcriptional response to *STM* induction has been established to broadly promote or repress developmental programs by regulating subsets of gene families simultaneously. *STM* also appears to regulate the expression of its KNOXI homologs *KNAT1/BP* and *KNAT2*, though in all cases it is unclear whether these effects are all true within *STM*'s native domain.

3.4 Evaluating the effects of *STM* on phytohormones and the cell cycle

3.4.1 The effects of *STM* on cytokinin

While *STM* is known to upregulate CK biosynthesis (Jasinski et al, 2005; Yanai et al, 2005; Scofield et al, 2013), this effect was not possible to deduce solely from the gene

ontology data, so the response to *STM* induction of several important gene families in CK biosynthesis and signalling was examined, with the intention of identifying broad trends in *STM*'s regulation of CK metabolism and response. The gene families considered were the *IPT* genes involved in CK biosynthesis, the CK oxidase (*CKX*) genes involved in CK degradation (Bartrina et al, 2011), the A-type *ARRs*, which are induced by CK and are negative regulators of CK signalling, and the *AHK* receptor gene family involved in CK signalling.

As can be seen in Figure 3.18, we see an early response from the CK-related histidine kinase gene WOODEN LEG (*WOL)/AHK4,* a histidine kinase responsible for transmitting CK responses across the plasma membrane (Muller & Sheen, 2007). Jasinski et al (2005) have shown that the *wol* mutant enhances the weak *bum1* allele of *STM*, leading to 60% of seedlings lacking a SAM. It was significantly differentially upregulated at all points on the time course except 24 hours where it only just missed the threshold for significance. The related gene *AHK1* was also significantly differentially upregulated at 72 hours and 9 days (Tran et al, 2007). This upregulation of *WOL* was observed prior to any statistically significant increase in CK biosynthesis genes, suggesting that the first CK-related transcriptional response to *STM* induction is to increase sensitivity to CK through upregulating the principal CK receptor.

IPT5 and *IPT7* have been shown to respond positively to *STM* induction (Jasinski et al, 2005; Yanai et al, 2005), however the experiments reported by these groups detected *IPT5* and *IPT7* induction far more rapidly than the time course used here, which only detected significant upregulation of these *IPT* genes after 9 days, though *IPT7 wa*s close to being significantly induced ($p<0.05$) at 72 hours. In our experiment, *IPT3* was the most rapidly upregulated *IPT* gene, showing a significant increase from 72 hours onwards, however, *IPT2* and *IPT9* were downregulated in the time course from 72

hours onwards. Thus the behaviour of the *IPT* genes was not completely consistent, though the *IPT* genes known to be upregulated by *STM we*re observed to be upregulated in this experiment, albeit later than previously reported. The complex dynamics observed may be because IPTs are repressed by CK (Miyawaki et al, 2004) hence at later timepoints, where there is more CK, IPT levels should fall, thus with complex feedback mechanisms at work we would not expect to see IPT levels change in a completely mechanistic manner.

In contrast to the *IPT* family the response for the CK catabolic genes *CKX1* and *CKX6* was consistent displaying significant downregulation at 72 hours and 9 days, suggesting that *STM* is not only affecting CK perception and synthesis, but also downregulating CK degradation.

Finally, the A-type ARRs, a family of transcription factors responsive to CK and which repress CK responses were examined. These genes are frequently used to detect changes in CK as their responses correlate closely to CK levels. *ARR6* and *ARR15* showed induction by *STM* at 24 hours, and *ARR6* remained significantly upregulated for the remainder of the time course, while *ARR15* returns to wild type levels. *ARR4, ARR5* and *ARR9* all showed significant upregulation from 72 hours onwards, and *ARR5* and *ARR6* were also significantly upregulated in the *stm-2* mutant, possibly owing to a complex phenotype. *ARR16* however, showed the opposite response to the other A-type *ARR* genes and was significantly downregulated by *STM* from 72 hours onwards. This agrees with the detection of *ARR6* upregulation in *STM* over-expressing leaves observed by Scofield et al (2013).

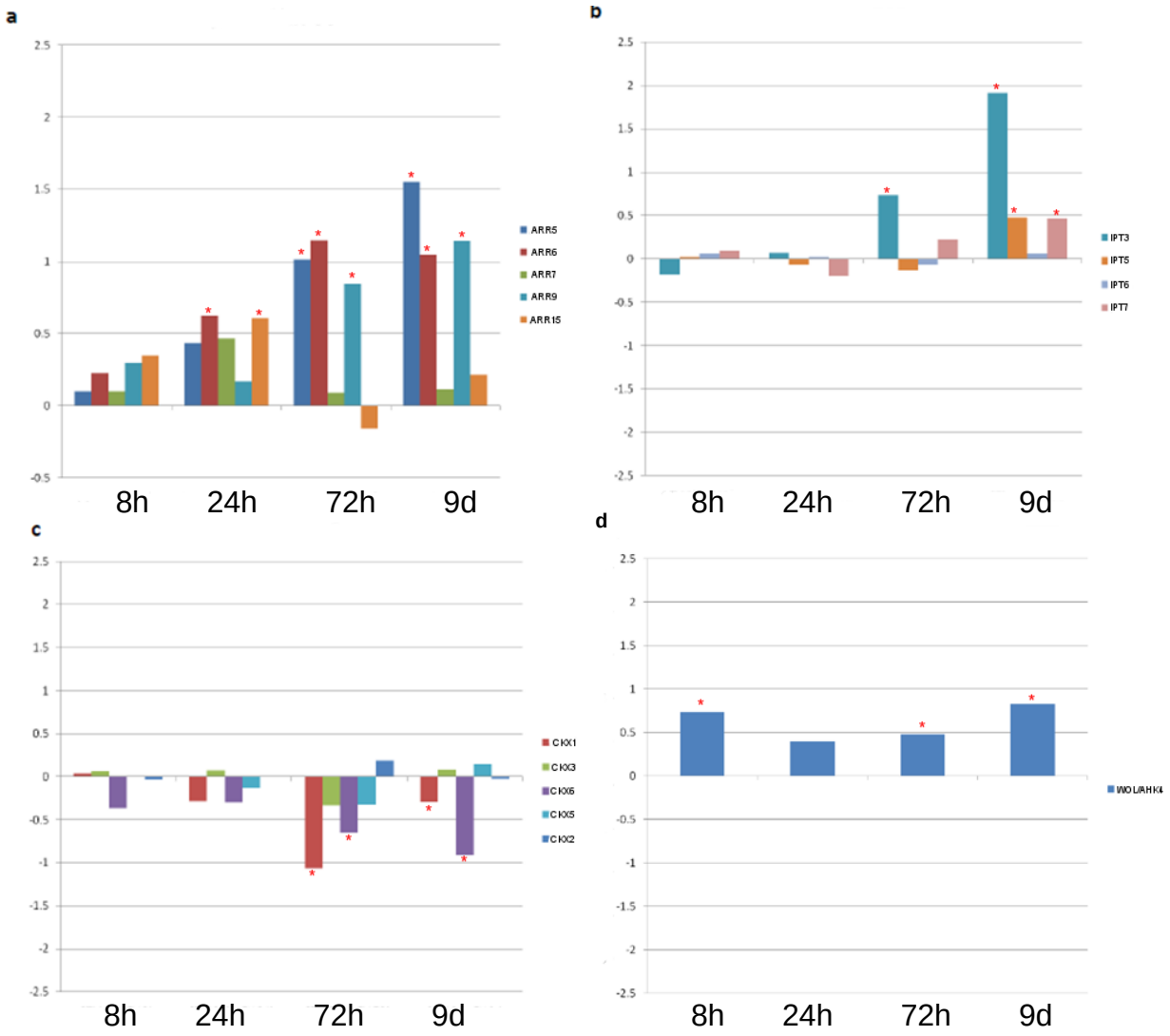


Figure 3.18 – CK responses following *STM* induction.
Point estimates of Log_2 Fold-Changes (y axis) calculated using LIMMA from microarrays of *STM* over-expressing plants vs mock treated plants of 4 classes of CK-responsive genes. X axis shows length of induction, either 8 hours, 24 hours, 72 hours or 9 days, A) A-Type ARRs, B) IPTs, C) CK Oxidases, and D) AHK genes are shown. Time points where genes are statistically significantly differentially expressed have been marked with a star.

3.4.2 Overlap of *STM* time course with transcriptomic data for cytokinin induction

While there are a few CK-related genes which show contradictory expression, for the most part, *STM* induction produced a consistent upregulatory effect upon CK biosynthesis, signalling and response genes, while downregulating CK catabolism. As such it was interesting to consider the overlap between *STM* and CK-treated plants to

identify which parts of the *STM* phenotype may be due to CK responses rather than transcriptomic perturbation by *STM.*

Goda et al (2008) examined genome-wide responses to hormone treatments as part of the AtGenExpress project. The original data from their research was re-analyzed using the same procedures as applied to the *STM* time course, and the response to 3 hours of zeatin (a form of CK) treatment in their experiment was compared to the *STM* time course (Figure 3.19)**.** 239 genes were significantly differentially expressed in the zeatin dataset at a p-value of 0.05. This less stringent p-value was used due to the lower power of their experiment with only 2 replicates. As can be seen, at all points on the OE time course, and in the *stm-2* mutant, the overlaps were highly unlikely to have occurred by chance and suggests that a significant proportion of the *STM* transcriptomic response may be mediated through CK.

	8h OE	24h OE	72h OE	9 day OE	72h RNAi	9 day RNAi	stm-2
Overlap Size	4	12	75	134	1	3	12
P value	0.0029	3.82×10^{-5}	2.34×10^{-7}	2.33×10^{-11}	0.49	0.59	0.011

Figure 3.19 – Overlap Between Time Course Experiment and CK Response Datasets.
The number of genes overlapping between Goda et al, 2008 CK dataset and the *STM* time course and the significance of the overlap as calculated by the hypergeometric distribution significant results (<0.05) are highlighted in green, non-significant results are highlighted in red.

We can potentially extract more useful information regarding the earlier CK-related responses on the OE time course, by using the overlap to investigate early responses. The most interesting result is that at 8 hours, *WOL* (Inoue et al, 2001) is the only named significant overlapping gene, but it is not part of the overlap at 24 hours, which instead

contains *ARR6* and *ARR15* –suggesting that positive feedback on CK may be replaced by negative feedback responses by 24 hours.

3.4.3 The Effects of *STM* on auxin

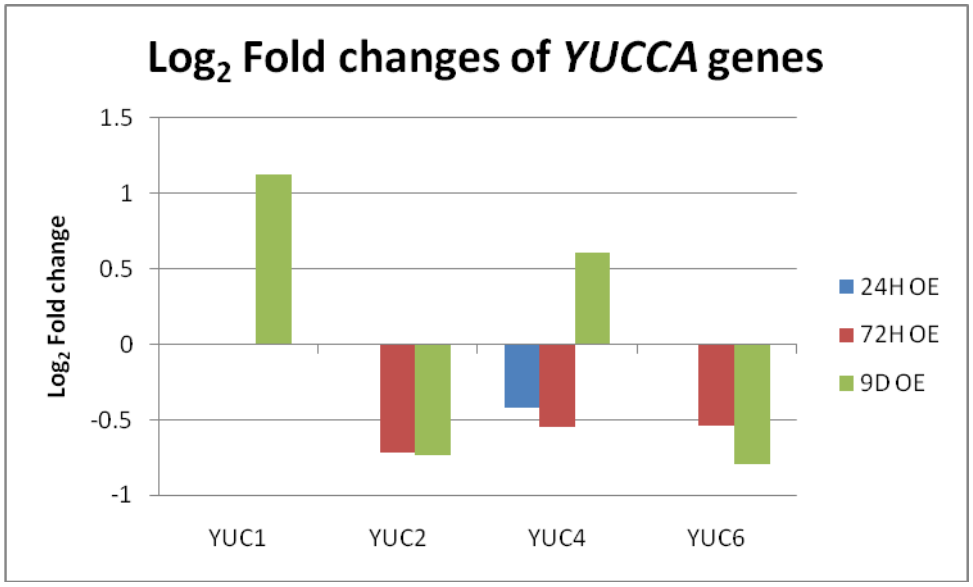
Since auxin and CK often act antagonistically, and the expression of *STM* is mutually exclusive to the auxin response domain in incipient and developing primordia, the possibility that *STM* would exert a repressive effect on auxin responses was examined. First the effect of *STM* on the *YUCCA* family of flavin monoozygenases genes was investigated, in particular *YUC1, YUC2, YUC4* and *YUC6* demonstrated by Cheng et al (2006) to play a key role in auxin biosynthesis. This family was selected as *YUC4* had already been identified by GO enrichment as an *STM*-responsive gene. A *YUCCA* gene has also been identified as a direct target of *KN1* in maize (Bolduc et al, 2012).

All 4 *YUCCA* genes known to play a key role in auxin biosynthesis are significantly differentially expressed during the *STM* OE time course (Figure 3.20A). At 24 and 72 hours, all significantly differentially expressed *YUC* genes are downregulated, consistent with *STM* inhibiting auxin synthesis. Although *YUC1* and *YUC* 6 are consistently downregulated at all significant time points, at 9 days *YUC2* and *YUC4* are significantly upregulated.

Secondly, as Bolduc et al, (2012) had identified a large number of AUX/IAA genes (Review: Lau et al, 2008) as direct targets of *KN1* in maize, the p-values and fold changes of the genes within this family across the time course were analyzed. As can be seen in Figure 3.20b, no AUX/IAA genes are detected as significantly differentially expressed until 24 hours. At this point only one gene (*PAP1/ IAA18*) is detected as significantly differentially expressed. It is also the only *AUX/IAA* gene identified in the robust meta analysis as significantly differentially expressed. It remains significantly

upregulated for the remainder of the time course. At 72 hours, *PAP1*/*IAA18* and *IAA28* are also upregulated. However *IAA8, 11* and 13 are significantly downregulated. Finally, as can be seen in Figure 3.20C, by 9 days a large number of *AUX/IAA* genes are significantly differentially expressed, amongst which only *PAP1/IAA18* is upregulated. No *AUX/IAA* genes were significantly differentially expressed in the RNAi or *stm-2* lines. This suggests that the timing of auxin regulation is much slower in Arabidopsis than in maize, and also that *AUX/IAA* genes are not consistently regulated in Arabidopsis as they appear to be in maize.

A



B

ID	8h OE	24h OE	72h OE	9d OE	72h RNAi	9d RNAi	stm-2
IAA1	0.92	0.80	0.22	$1.42x10^{-11}$	0.77	0.70	0.11
IAA2	0.52	0.51	0.01	$1.18x10^{-7}$	0.53	0.99	0.15
IAA4	0.50	0.87	0.03	0.00	0.92	0.99	0.01
IAA5	0.98	0.82	0.10	$1.30x10^{-7}$	0.84	0.78	0.43
IAA6	0.77	0.99	0.19	0.02	0.66	0.66	0.63
IAA7	0.46	0.96	0.20	$6.12x10^{-5}$	0.71	0.43	0.24
IAA8	0.94	1.00	0.02	0.00	0.28	0.48	0.04
IAA9	0.97	0.92	0.13	0.09	0.90	0.19	0.33
IAA11	0.61	0.85	$2.86x10^{-6}$	$2.09x10^{-6}$	0.19	0.79	0.81
IAA12	0.95	0.58	0.73	0.25	0.95	0.84	0.24
IAA13	0.91	0.47	$1.71x10^{-10}$	$3.71x10^{-10}$	0.67	0.53	0.02
IAA14	0.95	0.87	0.59	$7.38x10^{-9}$	0.85	0.80	0.59
IAA16	0.30	0.64	0.93	0.01	0.90	0.12	0.96
IAA17	0.79	0.75	0.94	$6.30x10^{-13}$	0.56	0.44	0.69
IAA17	0.99	0.87	0.45	$3.38x10^{-7}$	0.61	0.06	0.37
IAA18	0.96	0.13	$6.23x10^{-6}$	0.57	0.79	0.84	0.27
IAA19	0.41	0.18	0.01	$3.95x10^{-6}$	0.70	0.71	0.33
IAA20	0.42	0.70	0.62	0.16	0.88	0.29	0.86
IAA28	1.00	0.47	0.00	0.00	0.76	0.94	0.93
IAA29	0.95	0.69	0.35	0.00	0.85	0.82	0.16
IAA30	0.94	0.50	0.02	1.00	0.61	0.20	0.37
IAA31	0.95	0.94	0.59	0.81	1.00	0.74	0.53
IAA33	0.79	0.99	0.12	0.88	0.13	0.90	0.57
IAA34	0.79	0.78	0.93	0.04	0.22	0.62	0.57
PAP1/IAA18	0.20	0.00	$8.15x10^{-11}$	$8.18x10^{-16}$	0.54	0.96	0.88
PAP2/IAA8	0.10	0.08	0.00	0.02	0.78	0.70	0.10
SHY2/IAA3	0.62	0.96	0.04	$3.18x10^{-12}$	0.24	0.82	0.06

C

ID	8h OE	24h OE	72h OE	9d OE
IAA1	-0.06	-0.11	-0.19	-1.34
IAA2	-0.24	-0.22	-0.42	-0.94
IAA4	-0.22	-0.08	-0.32	-0.50
IAA5	0.01	-0.09	-0.22	-0.78
IAA6	-0.12	-0.02	-0.19	-0.31
IAA7	-0.36	-0.06	-0.31	-0.89
IAA8	0.03	0.00	-0.18	-0.24
IAA9	0.02	-0.03	-0.13	-0.13
IAA11	-0.15	-0.07	-0.62	-0.58
IAA12	-0.03	0.13	-0.04	-0.11
IAA13	-0.04	-0.12	-0.75	-0.67
IAA14	-0.06	0.13	-0.14	-1.55
IAA16	-0.26	-0.15	0.01	-0.30
IAA17	-0.11	0.12	0.01	-1.36
IAA17	0.01	0.04	-0.06	-0.41
IAA18	0.02	0.22	0.48	0.06
IAA19	-0.22	-0.27	-0.33	-0.62
IAA20	-0.21	-0.12	-0.07	-0.17
IAA28	0.00	0.21	0.54	-0.41
IAA29	-0.05	-0.18	-0.18	-0.65
IAA30	-0.04	-0.15	-0.25	0.00
IAA31	0.04	-0.04	0.07	-0.03
IAA33	-0.11	-0.01	0.20	-0.02
IAA34	0.10	0.09	0.01	-0.22
PAP1/IAA18	0.20	0.39	0.78	1.17
PAP2/IAA8	-0.34	-0.32	-0.41	-0.27
SHY2/IAA3	-0.16	0.03	-0.28	-1.28

Figure 3.20 – Auxin Responses to *STM* induction.

A) Point estimates of Log_2 Fold-Changes calculated using LIMMA from microarrays of *STM* over-expressing plants vs mock treated plants for *YUCCA* genes at those time points when they are significant.

B) p-values for *AUX/IAA* genes across the time course experiment. Significant changes are coloured green.

C) Observed Point estimates of Log_2-Fold changes calculated using LIMMA from microarrays of *STM* over-expressing plants vs mock treated plants for *AUX/IAA* genes across the time course experiment. Significant changes are coloured green if upregulated, red if downregulated.

3.4.4 Overlap of *STM* time course with transcriptomic data for auxin induction

Using the Goda et al (2008) dataset, genome-wide responses to auxin (IAA) treatment was examined. Following the same procedure as analyzing the CK dataset significant overlap was found at all time points except the 72 hour RNAi line in the over-expression time course and in the *STM* mutant (Figure 3.21) In the case of the RNAi line, the p value of 0.08 still implies a high confidence that the overlap is unlikely to have occurred by chance.

WOL, and multiple expansins (*EXPA5* and *EXPA11*) are among the overlaps with the 8h time point. However, *WOL* and *EXPA11* are both reciprocally regulated by *STM* (8 hours of induction) and auxin (*STM* induction upregulated *WOL* and downregulated *EXPA11*, while IAA treatment triggered their down and up regulation respectively). *EXPA5* was downregulated by both, but less so following *STM* induction. This suggests that the earliest component of *STM* induction is antagonistic to IAA treatment. The reciprocal expression of *WOL* suggests that this antagonism may occur through CK responses.

	8h OE	24h OE	72h OE	9 day OE	72h RNAi	9 day RNAi	stm-2
IAA	11	34	228	434	3	9	58
p-value	2.37×10^{-10}	0	0	0	0.08	0.008	0

Figure 3.21 - Overlap Between Time Course Experiment and IAA Response Datasets.
The number of genes overlapping between Goda et al, 2008 IAA dataset and the *STM* time course and the significance of the overlap as calculated by the hypergeometric distribution. Significant results ($p<0.05$) have been highlighted in Green, non-significant results have been highlighted in red.

3.4.5 The effects of *STM* on gibberellic acid

While STM's effect upon GA observed through GO data was more consistent than on auxin or CK, *STM*'s tobacco orthologue has been demonstrated to directly repress the expression of *GA20OXIDASE* genes (Sakamoto et al, 2001). As such, it was interesting to note that no *GA20OXIDASE*s were identified as significantly differentially expressed at 8 or 24 hours, suggesting that either the repression is a slow process, or *STM* does not directly repress *GA20OXIDASES*. To further investigate this the effects of *STM* up and downregulation across the time course on the *GA20OXIDASE* gene family were examined.

GA20OX1 and *GA20OX3* are significantly differentially expressed by 9 days and 72 hours respectively. However, while *GA20OX1* is downregulated as expected, *GA20OX3* is upregulated (Figure 3.22). There is also no sign of any *GA20OX* gene being significantly differentially expressed early in the time course, which would be consistent with the known behaviour of *STM*'s ortholog in tobacco. However, interestingly Wang et al, (2007) have shown that *OFP1*, a TF which is rapidly induced by *STM* directly represses the expression of *GA20OX1.* This is the only *GA20OX* gene which was observed as being downregulated following *STM* induction. This suggests the possibility that *STM* mediated downregulation of *GA20OX* may be occurring indirectly, via *OFP1* in contrast to other plant species.

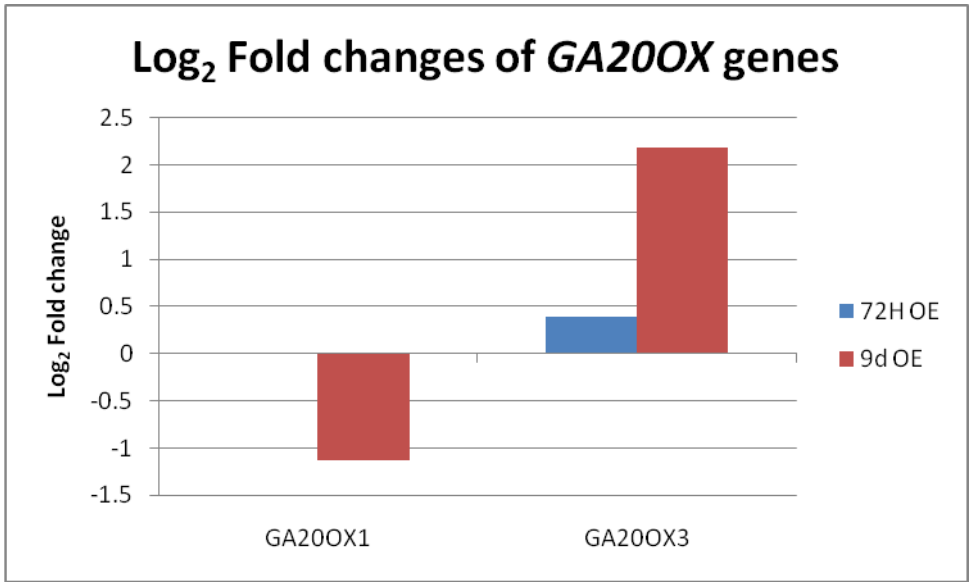


Figure 3.22 – *GA20OX* responses to *STM* induction.
Point estimates of Log_2 Fold-Changes calculated using LIMMA from microarrays of *STM* over-expressing plants vs mock treated plants for GA20OX genes at the time points they become significant on the OE time course.

3.4.6 Overlap between *STM* and *WUS-CLV* regulatory networks: mediation via hormone pathways

STM is not the only core stem cell regulator in the SAM. The *WUS-CLV* regulatory loop is commonly regarded as regulating stem cell identity (Brand et al, 2002; Lenhard et al, 2002), while *STM* is regarded as regulating stem cell fate. *WUS* encodes a homeodomain transcription factor and is the archetypal member of the *WOX* (*WUSCHEL-LIKE HOMEOBOX*) clade Amongst other genes, *WUS* represses expression of the A-Type ARRs which provide negative feedback on CK responses (Busch et al, 2010; Liebfried et al, 2005).

WUS over-expression has previously been studied using microarrays, as such there are two over-expression analyses available which can be compared to the *STM* OE time course (Liebfried at al, 2005; Busch et al, 2010). Liebfried et al (2005) examined the expression of genes using ethanol-inducible *WUS* over-expression lines following 12 hours of induction. The size of the overlaps, proportion of genes showing reciprocal

expression between *STM* and *WUS,* and probability of the overlaps occurring by chance are shown in Figure 3.23a. This suggests that the *WUS* over-expression datasets are most similar to the *STM* phenotypic datasets, although some similarity is observed when compared to the robust meta analysis, in the over-expression and *stm-2* lines slightly more genes were expressed reciprocally than otherwise.

Busch et al (2010) used a meta-analytical method to examine perturbations to the *WUS* GRN combining data from over-expression and loss-of function mutants. This identified 675 differentially expressed genes with a p-value of <0.01. Compared against the over-expression time course, this identified no time points which showed a statistically significant overlap (Figure 3.23b). As similar numbers of genes were identified in the overlaps to the previous analysis, this difference is presumably due mainly to the larger size of the *WUS* dataset, making it more likely that the overlaps observed would have occurred by chance.

A	8H	24H	72H	9D	72H	9D	stm-2
Overlap Size	1	5	35	75	0	1	25
P-value	0.111	0.015	0.018	5.36×10^{-6}	0.627	0.695	2.11×10^{-15}
Reciprocality	100%	60%	57%	61%	N/A	100%	44%

B	8H	24H	72H	9D	72H	9D	stm-2
Overlap Size	0	4	36	65	0	1	15
P-value	0.93	0.96	1	1	0.99	1	0.66
Reciprocality	N/A	25%	33%	42%	N/A	100%	66%

Figure 3.23 – Overlaps between Time Course and *WUS* responses.
The size of the overlaps, likelihood of those overlaps occurring by chance, and proportion which are reciprocally expressed between A)*STM* datasets and Liebfried et al, 2005 *WUS* datasets B) *STM* datasets and Busch et al, 2010 *WUS* datasets.

Thus it appears that any overlap between *STM* and *WUS* target gene sets is small, and that the early targets of each are dissimilar. One possible mechanism through which they interface was considered to be CK. Liebfried et al (2005) showed that *WUS* directly represses the expression of *ARR5, ARR6, ARR7* and *ARR15*. We have already shown that with the exception of *ARR7*, *STM* perturbs the expression of these genes, and as can be seen in Figure 3.18, at all time points where the observed differential expression was significant, these genes were upregulated by *STM*. Thus, *WUS* inhibits the expression of CK negative feedback genes, whereas these were upregulated following *STM* induction, probably as a consequence of increased CK biosynthesis. Thus, A-type ARRs represent a class of genes which are significantly, and reciprocally differentially expressed by *STM* and *WUS*. The targets of *STM* and *WUS* otherwise appear to have very little overlap. It is important to stress that this relationship is not necessarily

antagonistic, despite the reciprocal effects of both *STM* and *WUS* on the *A-TYPE ARR* genes. In the *STM* over-expressor the observed upregulation is likely to be a response to increased CK biosynthesis, and is consistent with higher CK levels. In *WUS* over-expressors the repression is direct, and thus consistent with promotion of CK signalling.

Aside from these *A-TYPE ARR* genes, an interesting difference in the response of both *STM* and *WUS* is the regulation of the *LONGIFOLIA* family of genes (Lee et al, 2006). *LNG1* and *LNG2* are both significantly differentially expressed at 72 hours following *STM* induction, and classed as significantly differentially expressed following *WUS* induction (Liebfried et al, 2005). However, following *WUS* induction they are upregulated, whereas at 72 hours following *STM* induction, both genes show over 2-fold downregulation. *LNG* proteins promote cell expansion towards the leaf, thus it is interesting to note that within the overlap at 72 hours (Lee et al, 2006), another of the shared genes between the two datasets which is downregulated by *STM* and upregulated by *WUS* is the expansin *EXPA8*.

3.4.7 The Effects of *STM* on the cell cycle

A list of 82 cell-cycle related genes in Arabidopsis was obtained from Menges et al (2005). Where expression data is available on the ATH1 array, the effect of *STM* over-expression and knock down was observed for these 82 genes and is shown in Figure 3.24. In the over-expression time course at 24 hours only *CYCD1;1* was significantly differentially upregulated. However by 72 hours, 29 cell-cycle genes were significantly differentially expressed and at 9 days 38 genes were significantly differentially expressed. Curiously, the behaviour of cell cycle genes at 72 hours and 9 days were quite different, in the 72 hour dataset all but 4 were downregulated, whereas in the 9 day dataset all but 6 were upregulated.

The CYCLIN B and D families were most widely perturbed at 72 hours, and of the B-Type cyclins only *CYCB2;1* was not significantly differentially downregulated by *STM*, although its expression was also found to be non-significantly downregulated. Among the D-Type CYCLINs, only *CYCD5;1* and *CYCD2;1* did not show perturbed expression. While *CYCD1;1* was upregulated following *STM* induction, all the remaining perturbed *CYCD*s were downregulated at 72 hours albeit with a low fold change.

CYCB genes regulate progression past the G2/M checkpoint, and thus regulate mitosis. *CYCD* genes regulate progression past the G1/S checkpoint, and thus

regulate entry into S-phase. Almost all of the *CYCD* and *CYCB* genes present were downregulated at 72 hours (with only *CYCD1;*1 upregulated), consistent with a broad inhibition of cell cycle progression, though the fold changes observed were small. Cell divisions in the meristems push stem cells towards the periphery where they divide, in the CZ the process is slow, and the speed of the cell cycle must be attenuated in order to prevent phyllotactic disruption. Thus, this could reflect a broad indirect function of STM in regulating the speed of entry into the cell cycle to maintain correct phyllotaxis.

However, by 9 days the situation was reversed, as all the differentially expressed B and D type cyclins were upregulated, with the exception of *CYCD4;2*. Given the large transcriptomic changes which were observed in the 9 day over-expression dataset and the large number of defence and metabolism GO categories enriched, it is plausible that the change in behaviour of the cyclins was due to downstream phenotypic effects, including the increased number of less differentiated cells, thus the effect on most cell cycle genes by *STM* appears different in the 72 hours and 9 day datasets.

Gene	24hOE	72h OE	9dOE	72h RNAi	Gene	24hOE	72h OE	9dOE	72h RNAi
CDC25		0.36	0.56		CYCD1;1	0.67	0.51	1.17	
CDKB1;1			0.37		CYCD2;1			0.80	
CDKB1;2			0.37		CYCD3;1		-0.39		
CDKB2;1			0.50		CYCD3;2		-0.47		
CDKB2;2		-0.31			CYCD3;3		-0.30	0.48	
CKL1			-0.33		CYCD4;1		-0.54		
CKL5			0.29		CYCD4;2		-0.57	-0.47	
CKS2			0.33		CYCD5;1			0.68	
CYCA1;1		-0.57	0.36		CYCD6;1		-0.89		
CYCA2;2			0.78		CYCL1			0.32	
CYCA2;3		-0.64	-0.27		CYL1		-0.38	-0.96	
CYCA2;4		-0.41			DEL2		-1.04		
CYCA3;1			0.36		DEL3		-0.46		
CYCA3;2			0.87	-0.47	DPa			0.25	
CYCB1;1		-0.49	0.48		E2Fb			0.27	
CYCB1;2		-0.61	0.51		E2Fc			0.46	
CYCB1;3		-0.41	0.41		KRP1		-0.32	-0.43	
CYCB1;5		-0.61	0.51		KRP2		0.71	1.45	
CYCB2;1		-0.46			KRP3			0.60	
CYCB2;2		-0.73	0.46	-0.37	KRP4			0.31	
CYCB2;3		-0.43	0.41		KRP5		1.29	1.45	
CYCB2;4		-0.69	0.41		KRP7			-0.33	
CYCB2;5		-0.43	0.41		Rb			0.44	
CYCB3;1		-0.49			WEE1			0.54	
CYCC1;1		-0.38							

Figure 3.24 – Effects of *STM* upon cell cycle genes as defined by Menges et al (2005).

The observed point $\log_2$ fold change estimates computed via LIMMA of cell cycle genes at the points where they are significantly differentially expressed (Red - downregulated, green - upregulated) for all cell cycle genes differentially expressed at any point in the OE or RNAi time course.

The same is not observed with the *KRP* genes, which are cell cycle progression inhibitors. Several were differentially expressed following *STM* induction, and though there is no consistent pattern of up-or-down regulation within the family as a whole, each *KRP* gene which is significantly differentially expressed at 72 hours and 9 days is perturbed in the same direction subsequently. 2/3 of the *KRP* genes are upregulated by *STM* with *KRP2* and *KRP5* showing greater than 2 fold upregulation by 9 days, an effect also confirmed using qRT-PCR (Scofield, personal communication). As the *KRPs* are negative cell cycle regulators, this suggests that at least this portion of the cell cycle apparatus behaves consistently between 72 hours and 9 days. However, for the dataset as a whole, between the 72 hour and 9 day time points, 10 out of the 18 genes significant at both time points were expressed reciprocally. This difference suggests that *STM*'s effects on the cell cycle are different in plants grown on DEX to those under short-term induction. One possibility is that this is a negative feedback response to slow-down of division post-72 hours owing to high *KRP* expression; this would amount to a feedback mechanism to trigger compensatory up regulation of positive regulators of the cell cycle.

In the Murray lab, microarrays for *CYCD* over-expressing leaves have previously been generated (Dewitte et al, 2007). *CYCD3;1* was constitutively over-expressed and compared against wild type old and young leaf tissues. In *CYCD* over-expressing tissue, endocycles are inhibited and mitotic cycling is favoured. If there was a reasonable degree of overlap between the *CYCD3* OE data and the *STM* OE time course, then this would suggest that we would expect to see similar phenotypic effects. The *CYCD3* OE data was analyzed using the same procedure as the *STM* microarray time course and the overlaps compared.

The overlap between the various *STM* OE datapoints and the agreement in the direction of change between differentially expressed genes is shown in Figure 3.25**.** There were a total of 487 significantly differentially expressed genes at p<0.01 in the young (undifferentiated) leaves D3 OE sample, and 1,710 in the old leaves D3 OE sample. It was clear from the size of the overlaps that there was a great deal of commonality between the two datasets, however it was particularly important to note that in all but the 72 hour OE dataset contrasted against the young D3 OE sample - the majority of genes show the same direction of response in both datasets. This is consistent with *STM* recapitulating a low endoreduplication phenotype, and we could also that *STM* ectopic expression is more similar to the over-expression of *CYCD3;1* in old tissue than young tissue.

A	8 Hours	24 Hours	72 Hours	9 Days
Number	17	52	579	1,102
Proportion	76.5%	69.2%	73.7%	90.2%
Significance	9.74×10^{-8}	4.44×10^{-16}	0	0
B	8 Hours	24 Hours	72 Hours	9 Days
Number	10	21	185	303
Proportion	70%	66.7%	34.5%	79.2%
Significance	1.34×10^{-6}	$5.84 \text{ z } 10^{-7}$	0	0

Figure 3.25- Overlap Between STM Time Course and (Dewitte et al, 2007) CYCD3 OE datasets.
The number of genes commonly differentially expressed between the *STM* OE time course and a) the Old tissue D3 Over-expressing experiment or B) the Young tissue D3 overexpressing experiment, and the proportion of those changed in the same direction. Finally the statistical significance of the overlap, as calculated by the hypergeometric distribution is shown

3.4.8 Evaluating similarity in hormone and transcriptional response between *STM* and its Maize ortholog *KN1*

Finally, we consider the analysis by Bolduc et al (2012), who performed a genome-wide ChIP study on *STM*'s maize ortholog *KN1*. Given the importance of KNOX genes in plant development and as severe mutations in *STM* produce plants which fail to develop a SAM, we anticipate a high amount of selection pressure and thus there should be a degree of evolutionary conservation with respect to the role and function of *KNOTTED1-LIKE HOMEOBOX* genes throughout the plant kingdom.

Given the caveats that the Bolduc et al (2012) paper is in a different species (and a monocot) and only designed to identify direct targets for *KN1* - the gene families containing members they identified as bound by *KN1* were examined rather than individual genes. As in our case, they found a large number of homeobox genes were being regulated by *KN1*. They also identified a number of NAC family transcription factors as being bound by *KN1*, a member of this family has already been shown to be directly induced by *STM* expression (Spinelli et al, 2011), however Bolduc et al (2012) could not conclude that any were direct targets as they were not differentially expressed in *kn1* mutants. Particularly interestingly, they identified an *AHK*-related gene highly similar to *WOL* – which as previously shown is one of the first CK-related genes to be differentially expressed over our time course.

Aside from this, with hormone regulation, there are large differences between the direct targets identified for *KN1* and the genes which would be present in our putative Arabidopsis GRN. This may represent a divergence in the role of *STM* between monocot and dicot species, particularly as a number of genes downstream of *STM* relate to cotyledon separation and organ polarity which are quite distinct processes between these two classes of plant species. They identified a direct effect of *KN1* upon

auxin which is far larger than the auxin-related response we have observed at early time points in Arabidopsis, although we do observe statistically significant overlap with an auxin induction dataset at all time points except the 72 hour RNAi. Similarly, no *GA-20OXIDASE* genes were significantly differentially expressed early in our time course, whereas both their *KN1* ChIP experiment, and previous work by Bolduc & Hake (2009) have shown it to be a directly repressed target of *KN1*.This difference may be one of timing, as there is clearly a statistically significant effect upon auxin, and it is stronger at later time points, and we do see *GA20OXIDASES* being downregulated later in the time course. However it appears that in Arabidopsis, this is unlikely to be a direct an effect of *KNOTTED1-LIKE HOMEOBOX* gene expression as in maize and tobacco (Sakamoto et al, 2001).

Importantly, these authors found that the number of targets bound by *KN1* was far in excess of the number of genes which were significantly differentially expressed in mutant gene expression experiments. This may be due to the extremely short TGAC core which forms the basis of the *KNOX* consensus sequence likely to occur randomly throughout the genome at relatively high frequency, although studies such as Spinelli et al, 2011 have shown that additional features improve target sequence recognition. This suggests that for direct target experiments in Arabidopsis we need to ensure that we are observing not just binding but induction or repression of a gene in order to conclude that it is a direct target.

3.5 Discussion

I have shown that long-term induction of *STM* induced a large phenotypic effect, which was reflected in the large number of transcriptomic changes identified by microarray analysis. However, it was also clear that much of this change was due to perturbation of defence responses and general metabolism, which may not reflect specific function of

STM in meristem development, but arise instead from experimental conditions. When compared to short-term induction of *STM,* the response to 9 days of *STM* induction also produced contradictory responses from several genes in key categories of *STM*-responsive processes, such as cell-cycle regulation and hormone responses. The RNAi lines appear to have had low power to detect significantly differentially expressed genes on an individual basis, but many of the biological processes identified as enriched were similar to those in the over-expression lines. Thus studying the over-expression lines at 8, 24 and 72 hours was considered the most useful approach to understanding *STM*'s function.

The initial response to *STM* induction at 8 and 24 hours was small, but by 72 hours a more general response was observable. Chronologically the enrichment of GO categories suggested that organogenesis and regionalization were the first core developmental processes perturbed by *STM* and that a large transcriptional response was observable from 24 hours onwards. Several families of genes, such as the *TCPs*, *CUC*s and *ARRs* show multiple closely related members responding similarly to *STM* induction, these were considered more likely candidates for regulating important processes downstream of *STM* as it was clear that a concerted response was being induced upon these gene families.

Meta-analytical techniques were used to identify smaller subsets of rapidly differentially expressed genes, robustly differentially expressed genes and genes differentially expressed in phenotypic plants. The GO enrichment of these meta-analyses matched well with the observed GO enrichment in the time course. The rapid response meta analysis and robust response meta analysis were both considered particularly interesting. The first was hypothesised to be more likely to contain direct targets of *STM*, while the second the most consistently responding genes to *STM* induction.

A clear effect upon the expression of cell cycle genes was observed, with *STM* appearing to regulate genes in a manner consistent with inhibition of endocycling, until the 9 day time point. This is consistent with the reported effect on endocycles (Scofield et al, 2013) however at 9 days, many differentially expressed genes were being expressed in a different manner to the rest of the time course. *CYCLIND1;1* was the earliest significantly differentially expressed cell cycle gene, and the effect of *STM* upon *CYCLIND1;1* expression was consistent and strong, making it an interesting candidate for further study.

Similarly, in depth analysis of key CK and auxin related genes confirmed that *STM* appears to be promoting CK biosynthesis and downregulating both CK catabolism and auxin biosynthesis. This is consistent with the published data on *STM* function regarding hormones. However, *STM* did not appear to directly repress *GA20OXIDASE* genes as is the case for its close homologue in tobacco. It was hypothesised that this may be mediated through *OFP1* - as this is an early responding gene to *STM* induction, known to directly repress *GA20OX1* (Wang et al, 2007).

As *STM* is not the only key regulator of SAM function and its function was contrasted with that of *WUS*, which is also required for proper meristem organization. Here it was found that the overlap between *STM* and *WUS* was small and mostly restricted to phenotypic plants, which it was concluded were possibly showing a response to grosser morphological changes occurring. Though the overlap was small, *WUS* is known to directly repress four A-type ARRs, which dampen CK signalling. *STM* upregulates all but one of these genes, suggesting that an interesting interface between *STM* and *WUS* may be through their modulation of CK levels or signalling respectively. As *STM* induces CK responses through a variety of means, and *WUS* dampens CK negative feedback, this suggests that the two genes exert a jointly positive effect upon CK

responses, which explains the synergistic effects observed by studies such as Gallois et al (2002).

However, while this study had suggested several gene families which are perturbed by *STM* and given a clearer idea of the chronological order in which biological processes are affected by *STM* over-expression, it was considered that data mining and machine learning techniques might be applicable to identify interesting associations between individual genes in order to connect temporal changes in the transcriptome to biological processes perturbed by *STM*. In particular the rapid and robust meta-analyses were considered useful starting points for identifying the genes between which associations might be mined.

Chapter 4 - The use and validation of data mining techniques to identify core genetic modules associated with *STM*.

4.1 Introduction

In Chapter 3, genes which comprise *STM*'s GRN were identified allowing insight into how *STM* affects Arabidopsis development, and the order in which events occur after its expression is induced. While this has expanded upon our knowledge of genes affected by STM mis-expression which may explain its phenotypic effects, this does not give clear information on how these genes interact downstream of *STM*.

As a transcription factor, STM functions by regulating the expression of other genes. The work of Chapter 3 showed that *STM* induces a response in which transcription factors are highly enriched. This suggests that if the links between *STM*, downstream transcription factors and other genes can be identified then we can begin to assemble the molecular basis of *STM*'s function.

There are, however, significant challenges in identifying these links. First, the microarray data available has limited temporal resolution (particularly as the gap between later points on the time course is large). It is also noisy due to the nature of high-throughput data analysis and there were differences in the magnitude of observed response in the RNAi and OE lines, since the RNAi line only perturbed *STM* expression within its native expression domain. Given these limitations, the robust response meta-analysis, which identified those genes most consistently responsive to *STM* is particularly interesting as it provides a candidate subset for further investigation. In this dataset, enrichment of the core GO categories expected from *STM*'s phenotype are observed - a strong transcriptional response, overlap with other core processes in the SAM, perturbation of hormone responses and metabolism and a potential link to the cell cycle, and thus we can be relatively confident that this dataset has captured a range of

the most robustly differentially expressed components of *STM*'s GRN. Here this meta-analysis is used as a starting point to infer and then validate relationships between this subset of genes.

To achieve this, a number of data mining and machine learning techniques such as Self organizing maps (Kohonen, 2001) and Bayesian Network Structural Inference (Heckerman et al, 1995) have been used to attempt to infer relationships between robust response genes. These techniques are not by themselves able to identify statistically significant relationships. Self-organizing maps have found large-scale use as an unsupervised method for clustering and classifying biological and other data (Tamayo et al, 1999). Bayesian networks have been the subject of extensive study within computer science particularly as the basis for decision making and classification processes. The mathematical tractability of deriving a scoring function for Bayesian networks enables the inference of network structures to be an effective way of identifying relationships within datasets and has begun to see use in systems biology as the amount of data available is increasingly making this technique more attractive. While they identify underlying patterns in the datasets provided, these algorithms can also identify patterns in noise, and are vulnerable to problems such as over-fitting the data to match input data better than real data.

To help validate and confirm predictions, a variety of molecular biology approaches have been used to validate predicted relationships from the data, including transient induction in the presence of cycloheximide (CHX) to identify genes whose response does not require *de novo* protein synthesis and which therefore represent putative direct targets. These CHX experiments use the *35S::STM-GR* line described in Chapter 3. CHX is a potent inhibitor of translation (Obrig et al, 1971). Co-application of CHX with DEX should prevent translation of STM-regulated RNAs, thereby precluding the

identification of indirect (or second order) targets when performing microarray or qRT-PCR analysis.

This Chapter thus had two core goals. First, data mining techniques are used to better understand how *STM* is functioning through downstream targets. Second molecular biology techniques are used to validate and evaluate the effectiveness of the predictive approaches employed.

4.2 Data mining of microarray data

4.2.1 Self organizing maps identify groups of co-regulated genes

Self Organizing Maps (SOMs) are widely used data mining tools for classifying data points (Kohonen, 2001) according to intrinsic properties and visualizing them in a spatially connected manner. Over a large number of iterations, data points are assigned to the nearest node on a grid of nodes, as defined by Euclidian distance from a codebook vector (initially created at random). After each addition, the codebook vector of the target node and those adjacent to it are adjusted to more closely match the data point most recently assigned. Over time, the SOM develops a structure, as adjacent nodes' codebook vectors become closer, and distant ones further. Thus in addition to classifying the data points, the spatial ordering of the nodes provides a form of clustering of the data points within them.

An 8x8 Self Organizing Map was created for all genes present in the Robust Response Meta Analysis at all time points in the OE and RNAi time courses with their Empty Vector controls. Two ways of visualizing the similarity between these nodes are shown in Figure 4.1A and B. In the Fig. 4.1A, nodes are coloured by the average Euclidian distance between each node and its neighbour's codebook vectors. As can be seen, *STM*'s node (node 1) is not particularly close even to its neighbouring nodes in

Euclidian space. It is the only gene assigned to its node, which is unsurprising given that its expression has been perturbed as part of the experiment.

Genes in nodes nearest to *STM* (spatially) are anticipated to be more likely to be direct targets of *STM* as their expression is similar; however, as in many of these samples *STM* expression has been dramatically perturbed it is unsurprising that few nodes show close codebook vectors in Euclidian space. A further feature anticipated for putative direct target would be that such genes would respond more rapidly to STM. Thus, in Figure 4.1B**,** nodes with > 0.9 Pearson correlation co-efficient with *STM*'s node's codebook vector have been highlighted with blue dots, and those with <-0.9 correlation have been highlighted with red dots. *STM*'s node has been highlighted with a pink dot. Thus we can identify those nodes containing genes where the direction of changes in expression more closely follows *STM*. Once again, this is an imperfect measure as in the over-expression lines *STM* is driven to saturation, whereafter its expression does not change much. Thus, for genes with subsequent negative feedback, we may not expect them to correlate perfectly with *STM*.

As can be seen in Figure 4.1B, the structure of the SOM has divided those nodes representing high values in OE lines (bottom left of the map) from those nodes which represent downregulation in the OE lines (top right or the map). However, we can also distinguish nodes on the basis of more specific characteristics such as those which show a slow but consistent increase across the time course (e.g. node 6) from nodes which show rapid and sustained increase (e.g. node 1).

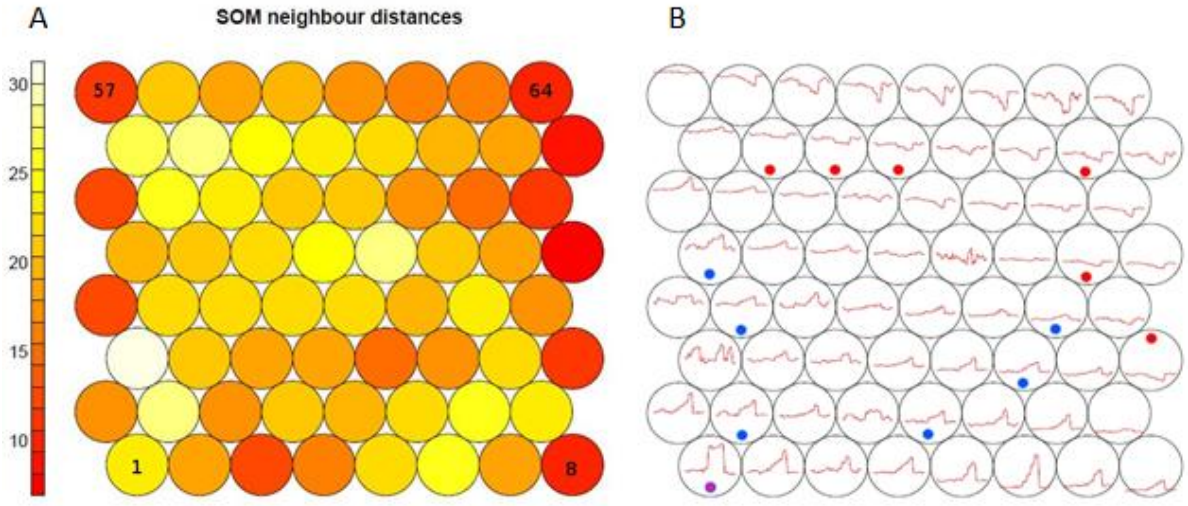


Figure 4.1 – Self organizing map of Robust Meta Analysis Genes.
A) An 8x8 self-organizing map created using the R package Kohonen with default settings. Nodes are numbered from 1 to 64 left to right, bottom to top. Colouring of each node indicates the average euclidian distance from the node to its nearest neighbour.
B) The same self-organizing map with codebook vectors displayed within each node. Codebook vectors represent expression from left to right in the empty vector lines, over-expression lines then RNAi lines. Nodes with dots are positively (blue) or negatively (red) correlated (> 0.9 pearson's correlation coefficient) with *STM*'s node (Purple). Assignments of genes to nodes are given in the Appendix.
C) A list of named genes within the positively (red) and negatively correlated (blue) nodes indicated by the equivalent coloured dots in B.

C

Positively Correlated		Negatively Correlated			
AGP7	ATMES1	AT1G19960	AT3G11930	AT5G64700	LURP1
ASL9	BGLU7	AT1G27210	AT3G15680	ATEXP11	NRPB9A
AT1G01070	BOP2	AT1G33440	AT3G54000	ATEXP3	OBP2
AT1G24530	CYP707A1	AT1G49370	AT3G60160	ATGSTU27	pde191
AT1G44760	DPL1	AT1G49470	AT3G60970	AtHB28	PRA1.F1
AT1G77660	ELF5A-1	AT1G49740	AT4G24810	ATPDHK	PROPEP6
AT2G28510	EXL4	AT1G56710	AT4G29030	ATTOC64-I	PYRR
AT2G38160	EXLA2	AT1G73620	AT4G31805	ATWHY1	sks4
AT3G12830	LSH4	AT1G80280	AT4G38520	ATXYL1	SOL1
AT3G24780	LSH5	AT2G15680	AT5G22390	BAM2	
AAT3G60450	MYBR1	AT2G28810	AT5G24580	CRF1	
AT4G16670	NLM8	AT2G36470	AT5G25490	E13L3	
AT5G42500	OFP1	AT3G01350	AT5G47800	FAD3	
AT5G42510	PARLL1	AT3G01750	AT5G49170	GATA18	
ATBAG3	SCPL32	AT3G01960	AT5G50740	GLN1;5	
AtIDD2	TUB1	AT3G10840	AT5G62680	LAX2	

We would expect genes which are in nodes most closely positively or negatively correlated by Pearson's R with *STM* to be more likely direct targets of *STM*. Figure 4.1C provides a list of genes which are present in nodes highly correlated with *STM* – as defined by a Pearson's correlation coefficient of >0.9 or <-0.9. These include a number of transcription factors such as *KNAT1/BP, OFP1* and *BOP2,* all upregulated and in nodes correlated with *STM*. A larger number of genes are present in nodes negatively correlelated with *STM* to a high degree – interestingly these include a pair of *EXPANSIN* genes – which coupled with their early upregulation over the time course suggests this process may be tightly regulated by *STM.*

We would expect those genes which are in nodes adjacent to *STM*'s node to be those which are most likely to have a sustained and consistent response to *STM* induction – by examining those nodes within a radius of 2 nodes from *STM*'s we can look at those genes which are likely to be closest to it across the time course. Appendix 7 contains a full list of all genes in all nodes; however it is interesting to note that *CYCD1;1* and a number of homeobox genes (*KNAT1, HB-7, HB13)* are within this subset of genes, even though these genes are not in nodes closely correlated with *STM*. These genes may therefore be unlikely to be direct targets, but nevertheless over time mirror changes in *STM* expression level. This approach is good for identifying genes with broadly similar expression dynamics but is limited by the magnitude of change in *STM*'s node since this is unlikely to be matched by other target genes. Hence while very useful for coregulation across the timecourse, it is less useful for capturing target genes with potentially complex expression dynamics.

4.2.2 Principal Component Analysis of robust response meta analysis data

To investigate whether the SOM may have overfitted the data – i.e. produced a pattern

too specific to the algorithm and input data, rendering it less true to the underlying biology, the time course data was analysed using Principal Component Analysis (PCA). If the same patterns were repeated through a different data mining methodology, we could be more confident that the results represent genuine patterns in the data and are not a result of overfitting. The expression values for each gene in the analysis were averaged across all replicates for the same condition and time point to produce averaged expression values for the EV, OE and RNAi lines at each time point. This was to make the PCA more easily interpreted in terms of the effects of each treatment at each time point.

The first two principal components jointly account for 97.5% of the variability in the input data, indicating that they explain almost all the differences in the data. As can be seen in the biplot shown in Figure 4.2, points further right along the x-axis generally have higher expression in the earlier time points of the OE, the EV or RNAi time points. Points further up on the y-axis generally have higher expression in the 72h or 9d OE lines than in the other lines. Thus, PCA suggests that the main divide between genes in the robust response meta analysis is how rapidly they respond to *STM*.

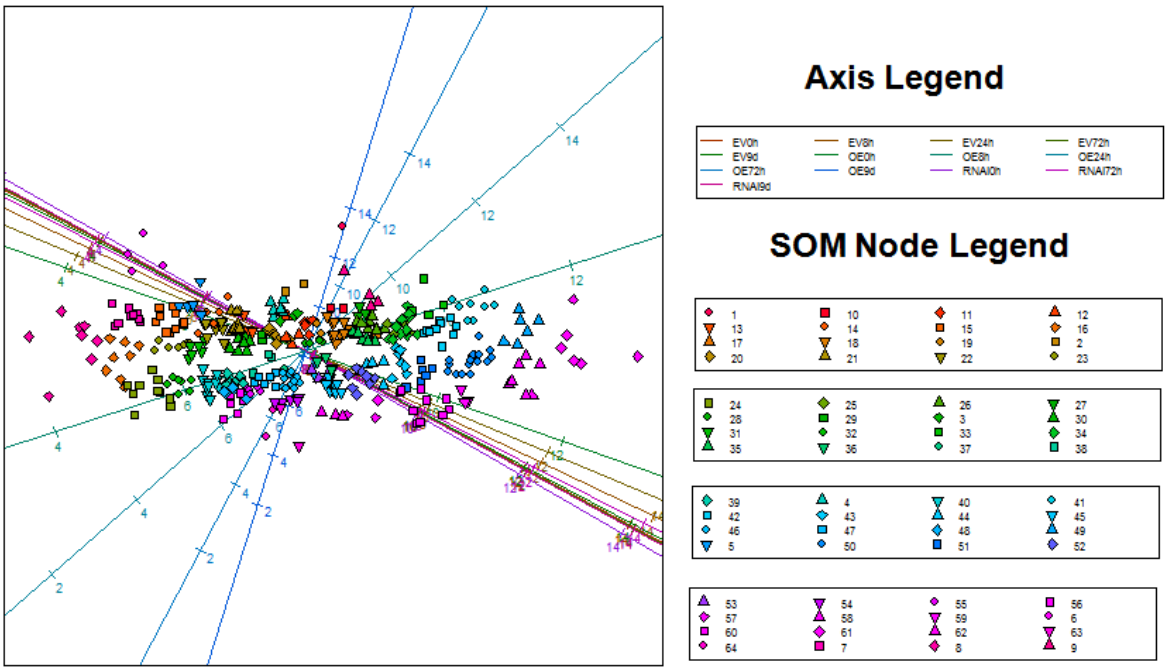


Figure 4.2 – Principal Components Analysis of Robust Response Meta Analysis genes.
Biplot of genes in robust response meta-analysis constructed from average expression along each time point and line. Lines running from left to right indicate greater expression in those lines as indicated in the axis legend. Each gene has been assigned a different colour and shape depending upon its node in the SOM as shown in the SOM Node Legend (See also Figure 4.1). The only gene from Group 1 is *STM*. Genes from groups, 10, 13, 22, 26, 31 and 33 were strongly positively correlated with *STM*'s node. Genes from Groups 24, 39, 50, 51, 52 and 55 were strongly negatively correlated with *STM*'s node.

As can also be seen in Figure 4.2, each data point has been assigned a colour and shape dependent upon the node to which it was assigned on the SOM. Those genes assigned to the same node on the SOM tend to be located quite closely via PCA as well, suggesting that the SOM has probably captured a similar underling data structure and that the approach of treating SOM nodes as co-regulated genes may be valid as the differences between assignment to nodes are clearly related to the speed and extent to which they respond to *STM* induction.

4.3 Bayesian Network Structural Inference

While clustering tools were useful in identifying broad co-regulated clusters of genes, to properly understand the relationships between genes in *STM*'s GRN, it would be necessary to derive and test predictions of direct interactions between genes. In particular, it would be interesting to infer relationships between TFs, as *STM* induction has been shown to rapidly perturb the expression of a statistically significant number of transcription factors.

Bayesian Networks are directed acyclic graphs encoding conditional dependency relationships between their nodes. They are specified by a graph structure and a joint conditional probability distribution over the connected nodes of the graph. Importantly, a number of scoring metrics have been devised which allow the likelihood of different graph structures to be compared given a dataset of experimental values for each node.

Bayesian network structural inference has been applied to gene expression data to infer a conditional dependency network between the genes within those datasets. Early examples such as (Friedman et al, 2000) recapitulated a number of known relationships from yeast datasets (Spellman et al, 1998), whereas more recent examples have reproduced more complex networks such as Imoto et al (2003) or Hartemink (2005).

As Arabidopsis is a model organism, a large number of transcriptomic microarray datasets from varying experimental conditions are freely available to download. As such there is a readily available set of independent datasets from which a static Bayesian network could be derived. Datasets were obtained from the EBI's ArrayExpress platform (Kaufmann, 2012), corresponding to Affymetrix *ATH1* microarray .cel files, selecting a subset of 2,373 datasets which had been annotated for "seedling" or "shoot" tissue, and "seedling developmental stage". Manual screening was used to remove samples

incorrectly annotated to minimize the amount of samples which would contain tissues in which *STM* was not expressed. Thus a large independent dataset of gene expression data under varying experimental conditions, enriched for seedling tissue was obtained (Appendix 6). Expression data from these was calculated and normalized using the RMAExpress software with Quantile Normalization. Finally, a discretised dataset was produced where the expression of each gene in each condition was classified as either 0 – downregulated (expression at least 1 unit lower than average for this gene on the $\log_2$ scale), 1 - unchanged, 2 – upregulated (expression at least 1 unit higher than the average for this gene on the $\log_2$ scale - i.e. whether genes are 2-fold up or-down regulated relative to their average expression.

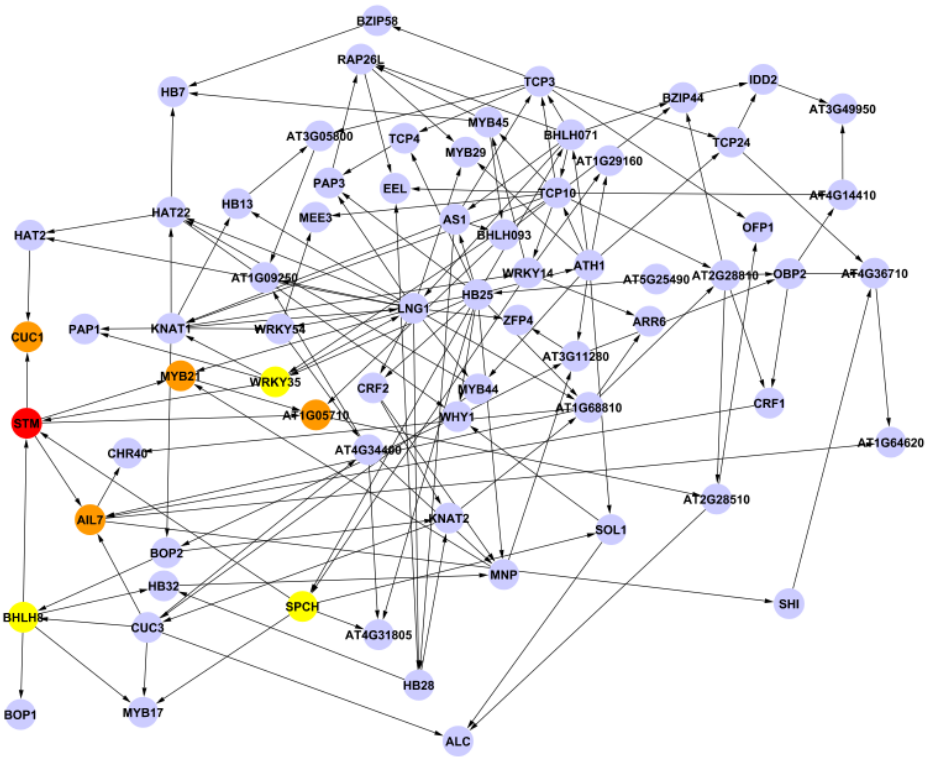


Figure 4.3 - Consensus network of transcription factors regulated by STM overlaid by predicted direct targets of *STM.*

The consensus network describes conditional dependency relationships (shown as edges with arrows indicating child nodes) between transcription factors (depicted as nodes) identified in the robust response meta analysis from a range of datasets (see 2.1.5). The network was inferred using BANJO as described in 2.1.6. No edges were permitted from *STM* to transcription factors not present in the rapid response meta analysis and no node was permitted to have greater than 5 parents. *STM* has been highlighted in red, transcription factors predicted to be direct targets of *STM* have been highlighted in orange, transcription factors predicted to target *STM* have been highlighted in yellow.

TFs had been identified as an early statistically significant component of *STM*'s core transcriptional response, suggesting that transcriptional control is an important early response to *STM* induction. In order to establish *STM's* relationship to downstream TFs Bayesian network structural inference was applied to the TFs and other genes involved in transcriptional regulation (such as *LONGIFOLIA 1 (LNG1)* (Koung Lee et al, 2006)) identified in the robust response meta-analysis. This method should identify the conditional dependencies extant between these TFs in the datasets used, thus enabling the construction of a network of relationships between TFs and thus identify putative direct targets of *STM* and interactions between downstream genes.

In order to identify direct targets of *STM,* nodes directly connected to *STM*'s node were considered potential direct targets. These are the nodes which have the strongest dependency upon STM's expression in the dataset used to infer the network. The further away from *STM* we go on the downstream network, the more likely we are to have missed important nodes in the network due to insufficient time points or experimental noise when determining the input set of genes, thus downstream connections as less likely to imply direct relationships and more likely to represent indirect dependencies between TFs. As I expected the majority of direct targets of *STM* to be significantly differentially expressed in the rapid response meta-analysis, all edges connecting *STM* directly to a TF not present in the rapid response meta analysis were banned.

Simulated annealing, a Markov chain Monte Carlo search method based around the properties of matter cooling from a hot state to the most likely cold state, was used to search efficiently through the sample space. As the final network configuration is not selected deterministically, 20 networks were generated and a consensus network was used to identify confident relationships, with the consequence that the final network is

no longer necessarily a Directed Acyclic Graph (DAG – a graph with no cycles and directional edges).

4.3.1 Initial Network Inference

As can be seen in Figure 4.3, 7 genes were linked directly to *STM's* node, suggesting that they could be direct targets of *STM*, especially as they include the only known direct target, *CUC1.* However, as can also be seen, a number of genes potentially identified as direct targets, do not show the behaviour anticipated for direct targets over the *STM* OE time course. Genes like *SPEECHLESS* or the GRAS family transcription factor encoded at locus At3g49950 are not significantly differentially expressed at 8 hours and only become so as we move further through the time course to 24 hours.

As can be seen in Figure 4.3, several known relationships from the literature have been recapitulated by the consensus network. Direct relationships have been shown between *AS1* and *TCP3* (Koyama et al, 2010), and *KNAT1* and *AS1* (Byrne et al, 2002). As mentioned, both our experiments and other work has shown *CUC1* to be a direct target of *STM* (Spinelli et al, 2011). These relationships are all present in the consensus network, giving greater confidence in the validity of the inferred relationships. However, *AS1* is also known to directly regulate *KNAT2* (Byrne et al, 2002), and this relationship is not captured. This may be due to the a weaker relationship being observed in the input datasets, but demonstrates that we cannot expect the Bayesian network to extract 100% of the true relationships in the data.

However the Bayesian network has clearly been successful in inferring some downstream relationships. Interestingly, we can derive hypotheses regarding the importance of several downstream genes by ranking genes according to their degree (the number of connections in either direction between that gene and another). As we

see in Figure 4.4, *LNG1, HB25, KNAT1 and TCP10* all have degrees greater than 10, compared to the average network degree of 2.275. All of these genes also have much higher out-degree (edges with them as apparent node) than in-degree (edges with them as a target), making them candidates as putative transcriptional hubs downstream of *STM*.

Label	In-Degree	Out-Degree	Degree
LNG1	3	12	15
HB25	1	13	14
TCP10	2	9	11
KNAT1	3	7	10
ATH1	1	8	9
TCP3	3	5	8
BHLH071	2	6	8
At1G68810	3	4	7
AIL7	4	3	7
CUC3	2	5	7
STM	3	4	7

Table 4.4 - Degree of nodes in initial Bayesian network.
The In, out and total degree of all nodes with degree 7 or greater in the consensus network shown in Figure 4.3.

KNAT1/BP in particular is known to be able to substitute for *STM* when ectopically expressed; it also has a broader expression domain than *STM*. Thus it is difficult to judge whether the large out-degree may be a duplication of *STM*'s function, in which case the data may have favoured the construction of a network where *KNAT1* was connected to target genes over *STM* due to its higher expression levels revealing clearer conditional dependencies, or whether part of *STM*'s function is duplicated by *KNAT1*. Additionally, *TCP3* has a high degree, with a higher out-degree than in-degree and is also closely functionally related to *TCP10* – this again suggests that the downregulation of *TCPs* is an important part of *STM*'s downstream GRN.

4.3.2 Network localization of Rapidly Responding TFs

Despite the fact that the network was constrained so that only TFs occurring in the rapid response meta analysis could be assigned as direct targets of *STM*, only 4 out of a possible 14 were assigned as such in the consensus network. Nevertheless almost all of the 14 rapidly responding TFs are located on the network in close proximity to each other and to *STM* (Figure 4.5.)

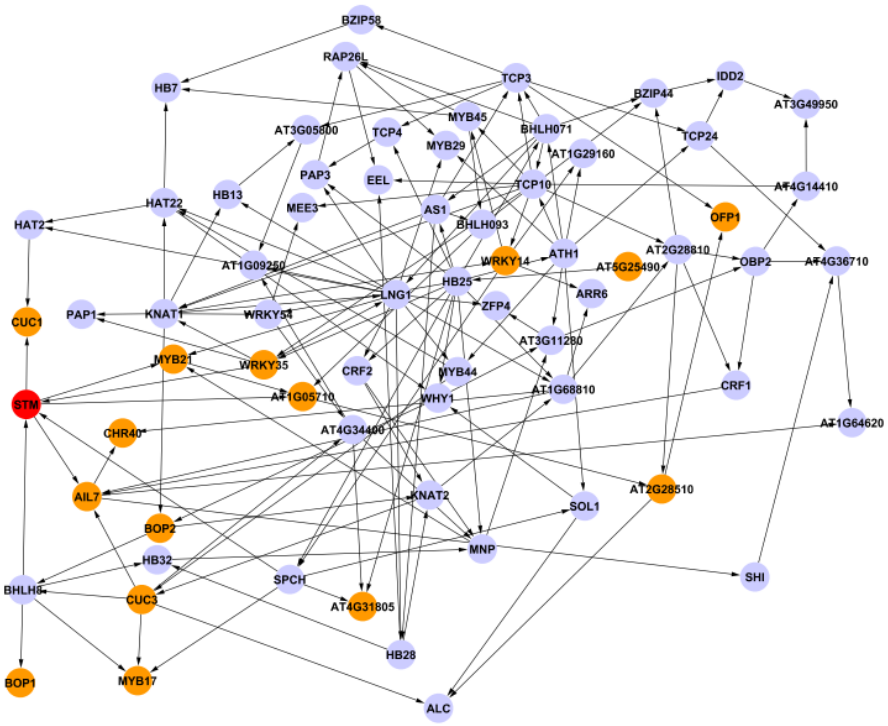


Figure 4.5 - Consensus network of transcription factors regulated by STM overlaid by significance in the rapid response meta analysis.

The consensus network describes conditional dependency relationships (shown as edges with arrows indicating child nodes) between transcription factors (depicted as nodes) identified in the robust response meta analysis from a range of datasets (see 2.1.5). The network was inferred using BANJO as described in 2.1.6. No edges were permitted from *STM* to transcription factors not present in the rapid response meta analysis and no node was permitted to have greater than 5 parents. *STM* has been highlighted in red, transcription factors statistically significant in the rapid response meta analysis have been highlighted in orange.

It is important to note that this consensus network was constructed from independent datasets to those used to determine the subset of genes under analysis. Thus the close

proximity of the rapidly responding TFs suggests that similar relationships can be observed from independently obtained datasets, which were not addressing questions regarding the effects of *STM* perturbation. This also provides support that the network has captured genuine information as it appears to corroborate the temporal dynamics observed over the *STM* OE time course.

4.3.3 Correlation of network structure and expression dynamics following *STM* induction

Of the transcription factors in this time course, only *TCP24, WRKY54* and *EEL* do not show consistent up or downregulation across the OE time course up to 72 hours. However, at the time points where these genes are inconsistent, they are not significantly differentially expressed, thus it is possible to unambiguously assign an up- or downregulated category to each TF at all time points where differentially expressed at a significant level. Additionally, all but 9 TFs increase the magnitude of their up or down regulation over the same time period. Thus, the 72 hour time point appears to provide a valid snapshot of the magnitude and direction of perturbation of the TFs within the *STM* time course.

From the data shown in Figure 4.6, we can see the observed direction of fold changes at 72 hours overlaid as either blue (negative) or orange (positive) on the consensus network. It can be seen that most genes in the close vicinity of *STM* show positive fold changes, while those genes further away usually show smaller fold changes. From the raw data from the time course experiment (appendix 13), we also note that as expected nodes closer to STM usually represent nodes which have larger absolute fold changes. *RAP2.6L* is an exception as it is has the largest fold change of the robustly responding TFs at 72 hours, but is located further from *STM* on the network. We also note that the smaller number of negatively regulated nodes are also closely interconnected. This

suggests that, unsurprisingly, the biggest effect on *STM*'s downstream network following perturbation is immediately downstream of it, which provides some reassurance that the phenotypes observed up to 72 hours are to a large extent a consequence of *STM* and not perturbations further downstream. Common grouping of genes into up/down-regulation clusters suggests they may be under common regulation.

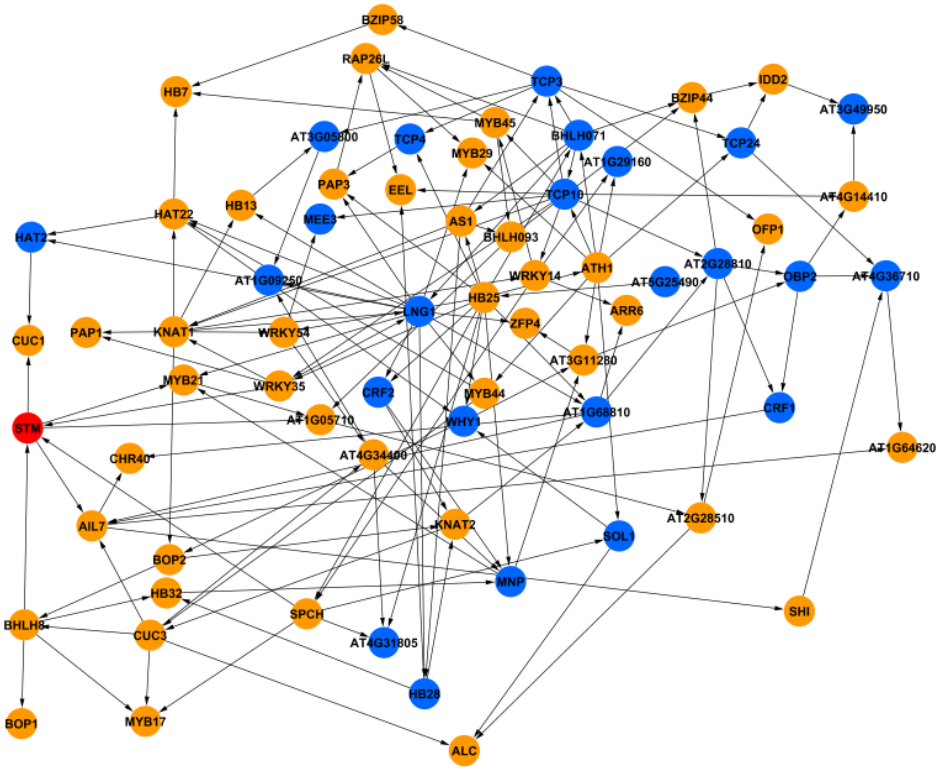


Figure 4.6 - Consensus network of transcription factors regulated by STM overlaid by upregulation/downregulation observed at 72 hours of *STM* over-expression.

The consensus network describes conditional dependency relationships (shown as edges with arrows indicating child nodes) between transcription factors (depicted as nodes) identified in the robust response meta analysis from a range of datasets (see 2.1.5). The network was inferred using BANJO as described in 2.1.6. No edges were permitted from *STM* to transcription factors not present in the rapid response meta analysis and no node was permitted to have greater than 5 parents. *STM* has been highlighted in red, transcription factors upregulated at 72 hours have been highlighted in orange, transcription factors downregulated at 72 hours have been highlighted in blue. As can be seen no downregulated genes are directly connected to *STM*'s node and most downregulated nodes cluster at the opposite end of the network to *STM*.

4.3.4 Reanalysis of the Yadav spatial dataset reveals co-ordination of similarly expressed genes following *STM* induction

Yadav et al (2009) produced an interesting dataset evaluating the expression of all genes on the ATH1 array, using FACS sorted cells from meristems of transgenic lines expressing fluorescent markers for *CLV* (a stem cell marker), *WUS* (organizing centre marker) or *FIL* (primordium marker). The dataset showed how gene expression varied between different regions of the SAM, thus their study was considered a useful starting point for evaluating the expression patterns of TFs within the *STM* GRN.

The authors used an expression value threshold to assign genes to each specific region of the meristem based upon whether expression in the appropriately marked cells exceeded this value. This approach was felt to be somewhat lacking, in that genes with low average expression, which may still have shown differences in expression pattern, may have been uncategorized, whereas genes which showed differences in expression pattern, but were expressed at an higher base level would often have been classified as present in all zones of the meristem. Thus, using the same limma-based data analysis procedure a re-analysis was produced contrasting the expression of each gene in each domain of the SAM, against the expression of that gene in each other domain of the SAM. Genes significantly differentially expressed in any comparison between *FIL* (Primordium), *CLV3* (Central Zone) and *WUS* (Organizing Centre) domains were categorized as follows;

- **Central Zone (CZ) –** Significantly differentially expressed in CZ relative to P and OC, or significantly differentially expressed relative to P or OC with no significant difference in expression between P and OC.

- **Primordium (P) -** Significantly differentially expressed in P relative to OC and CZ, or significantly differentially expressed relative to OC or CZ with no significant difference in expression between OC and CZ.
- **Organising Centre (OC) -** Significantly differentially expressed in OC relative to P and CZ, or significantly differentially expressed relative to P or CZ with no significant difference in expression between P and CZ.
- **Non-P** – Significantly differentially expressed in OC and CZ relative to P.
- **Non-OC** – Significantly differentially expressed in CZ and P relative to OC. This category may thus more closely correspond to the L1 layer of the SAM
- **Non-CZ** – Significantly differentially expressed in OC and P relative to CZ. This category may thus more closely correspond to the L3 layer of the SAM
- **Ubiquitous** – No significantly differential expression between the zones of the SAM.

While this categorization did omit some information, such as genes which showed a gradient in expression across the three zones simply being classified as native to the strongest zone of expression, this simplified scheme was easy to overlay onto the consensus network to evaluate the spatial effects of *STM* induction.

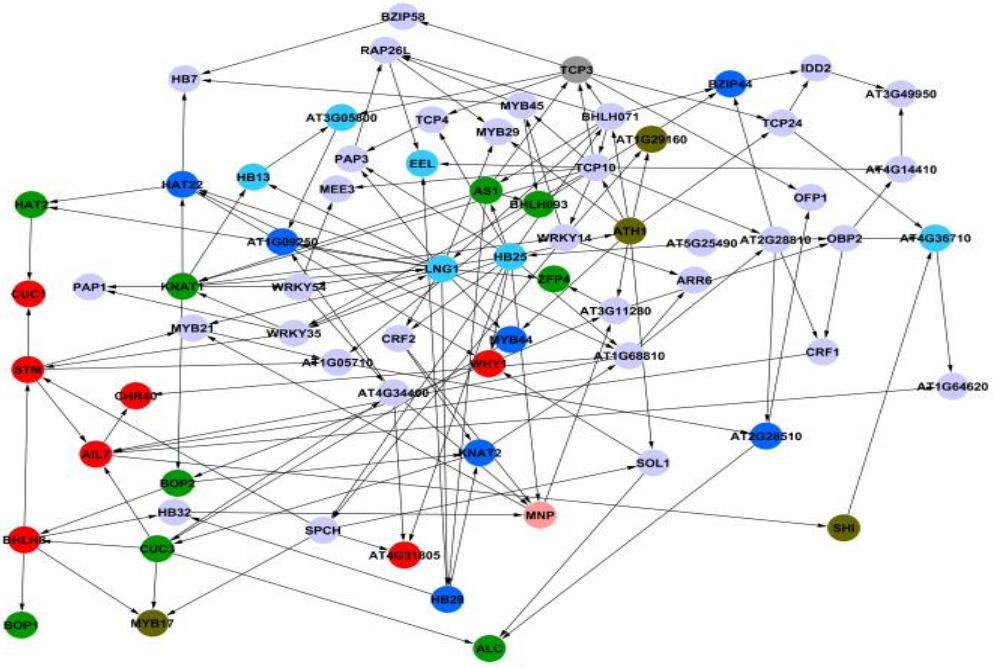


Figure 4.7 - Consensus network of transcription factors regulated by STM overlaid by predicted expression domain within the SAM given the reanalysis of the Yadav et al (2009) dataset.

The consensus network describes conditional dependency relationships (shown as edges with arrows indicating child nodes) between transcription factors (depicted as nodes) identified in the robust response meta analysis from a range of datasets (see 2.1.5). The network was inferred using BANJO as described in 2.1.6. No edges were permitted from *STM* to transcription factors not present in the rapid response meta analysis and no node was permitted to have greater than 5 parents. Nodes have been coloured according to predicted spatial domain given the reanalysis of the Yadav et al (2009) dataset described in 4.3.5. Red – CZ specific, Blue, OC specific, Green – primordium specific, Cyan – Excluded from CZ, Brown – Excluded from OC, Purple – Excluded from Primordia. Grey indicates no data is available and light blue indicates no specific expression domain.

As can be seen, in Figure 4.7 the genes nearest to *STM* which show significantly differential expression in any zone of the SAM tend to be CZ specific genes. Interestingly by this analysis, *CUC1* is categorized as a CZ specific gene, suggesting that it is expressed on the CZ side of the boundary. The genes furthest away from *STM* on the network tend to be those which are either OC or Non-CZ specific or p-specific, suggesting that the clearest split between *STM* downstream genes is between CZ and OC/OC+P genes. Interestingly the PZ specific or Non-OC genes do not show a clear pattern of localization.

4.3.5 Refining the network using correlation across microarray time course experiment

While the network appears to have captured a number of known relationships and shows clear patterning according to spatial information and time/magnitude of response to *STM* induction, there are likely to be a number of missing nodes, particularly further downstream of *STM* . As the network is also dense, which hampers the ability to find a clear visualization of the data, methods were sought to refine the network by restricting less likely edges.

In order to further refine the network, information was used from the time course experiment presented in chapter 3 –which via the meta-analysis was from the dataset used to identify genes to place in the Bayesian network, thus relationships with support in both datasets are more likely to be trustworthy. Thus by combining information across experiments a higher-confidence consensus network could be generated. Pearson's correlation co-efficients were computed (using all points and lines on the OE and RNAi time courses, *stm*-2 and wt lines, and the Mock and DEX treated *STM-GR* lines). for all pairs of TFs connected by edges in the Bayesian network, at an r value of > +/- 0.5 the known relationships between *AS1/KNAT1* and *TCP3/AS1* were present in the Bayesian

network, as such this was selected as an appropriate threshold and any edges which did not meet this threshold were deleted. The direction of the arrow representing the inferred direction of repression or induction.

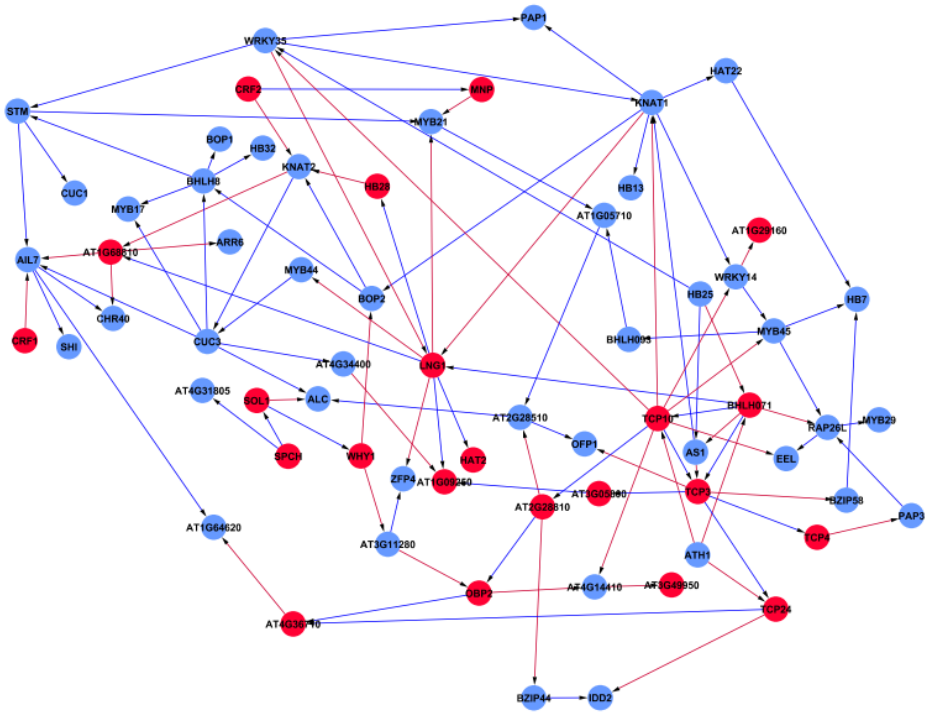


Figure 4.8 - Consensus network of transcription factors regulated by *STM* limited by correlation, overlaid by fold change at 72 hours.

The consensus network describes conditional dependency relationships (shown as edges with arrows indicating child nodes) between transcription factors (depicted as nodes) identified in the robust response meta analysis from a range of datasets (see 2.1.5). The network was inferred using BANJO as described in 2.1.6. No edges were permitted from *STM* to transcription factors not present in the rapid response meta analysis and no node was permitted to have greater than 5 parents. Pearson's Correlation coefficient between each TF pair was calculated over all time course contrasts, and for the *stm-2* mutant and *STM*-GR experiments presented in Chapter 3. All edges showing between -0.5 and 0.5 R were removed. Edges are coloured according to whether the Pearson's Correlation coefficient was negative (red) or positive (blue). Nodes have been coloured according to whether they are up (blue) or downregulated (red) at 72 hours of *STM* over-expression.

50 edges failed to meet the threshold, as can be seen in Figure 4.8, and removing these creates a far less densely connected network Some genes in particular (such as *HB25* which drops to a degree of 3 from 14) become far less critical to the overall

network structure, indicating that many of their inferred relationships may not be that strong. Other genes, such as *KNAT1* (which loses only 1 edge) remain densely connected. One of the predicted direct targets of *STM* – At1g50710 – is filtered by this process, and *MYB21*, which has a correlation coefficient of just 0.51 is significantly less confidently predicted as a direct target than *AIL7* or *CUC1* - with 0.7 and 0.71 correlation coefficients respectively. Thus, we regard *AIL7* and *CUC1* as the most likely direct targets of *STM*.

As can be seen in Figure 4.8, downstream TFs which function in a repressive manner downstream of *STM* can be identified. For example, *TCP* genes, *LNG1* and *AS1* mostly have downstream genes negatively correlated with them. Interestingly, the further away from *STM* on the network, the greater the density of repressive edges, a factor which corresponds well to the higher number of genes distant to *STM* on the network showing a negative fold change at 72 hours of *STM* induction. The particularly high out-degree of *TCP* genes and *LNG1* suggests they may be repressive hubs downstream of *STM*.

The spatial relationships downstream of *STM* are more apparent within the high-confidence network (Figure 4.9**)**. Early targets of *STM* are predominantly CZ specific, with genes further downstream of *STM* tending more to be OC or OC+Primordium specific. Primordium only genes are distributed throughout the network with the majority situated between the CZ and OC clusters.

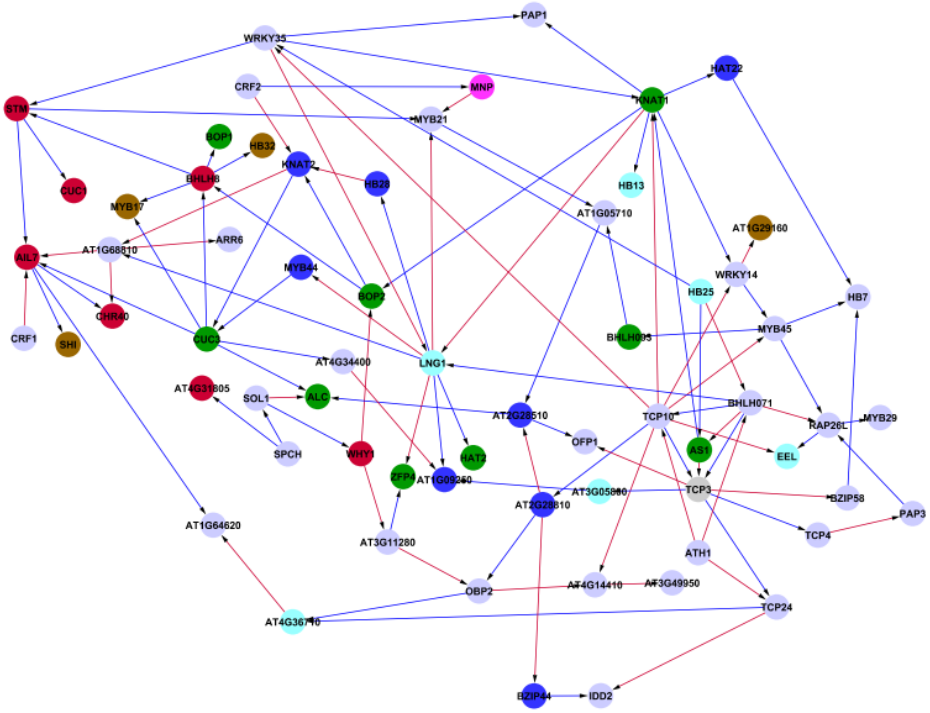


Figure 4.9 - Consensus network of transcription factors regulated by *STM* limited by correlation, overlaid by predicted expression domain within the SAM given the reanalysis of the Yadav et al (2009) dataset.

The consensus network describes conditional dependency relationships (shown as edges with arrows indicating child nodes) between transcription factors (depicted as nodes) identified in the robust response meta analysis from a range of datasets (see 2.1.5). The network was inferred using BANJO as described in 2.1.6. No edges were permitted from *STM* to transcription factors not present in the rapid response meta analysis and no node was permitted to have greater than 5 parents. Pearson's Correlation coefficient between each TF pair was calculated over all time course contrasts, and for the *stm-2* mutant and *STM*-GR experiments presented in Chapter 3. All edges showing between -0.5 and 0.5 R were removed. Edges are coloured according to whether the Pearson's Correlation coefficient was negative (red) or positive (blue). Nodes have been coloured according to predicted spatial domain given the reanalysis of the Yadav et al (2009) dataset described in 4.3.5. Red – CZ specific, Blue, OC specific, Green – primordium specific, Cyan – Excluded from CZ, Brown – Excluded from OC, Purple – Excluded from Primordia. Grey indicates no data is available and light blue indicates no specific expression domain.

4.4 Direct Target Prediction Validation

4.4.1 Preliminary Validation of Direct Target Predictions Suggests that the constrained consensus network correctly identified several valid relationships

While the network has generated a large number of potentially interesting relationships, it was desirable to test some of the predictions in order to ascertain the confidence in its detection of novel information. The easiest predictions to test were the potential direct targets of *STM*. In a 35S::*STM-GR* line (obtained from Rüdiger Simon, Dusseldorf, Germany) *STM* is expressed throughout the plant, but is unable to translocate into the nucleus without addition of DEX. In the presence of both DEX and the translational inhibitor CHX, *STM* is translocated to the nucleus, where it can affect the expression of its direct target genes, however due to the presence of CHX, targets further downstream of direct STM targets will not be upregulated as their translation is blocked. Thus by comparing the expression of a gene following CHX+DEX treatment vs CHX alone it is possible to determine whether a gene is being directly upregulated by *STM*.

First, a pilot experiment was performed using q-RT PCR to measure the expression levels of *STM*, *CUC1* and *AIL7* and *OFP1* at various time points after induction. At the time this work was done, *STM*'s connection to *CUC1* was novel and it had been shown to respond very strongly to *STM* induction by our microarray analysis, *AIL7* was a potential direct target which showed moderate induction, *OFP1* was not predicted to be a direct target by the Bayesian Network, but it did show rapid response to *STM* induction, and has been previously shown to directly target *GA20OX1* and repress it (Wang et al, 2007). As a direct antagonistic relationship between *STM* and *GA20OX1* had not been found, although the maize *GA20OX1* has been shown to be a direct target of *STM*'s maize ortholog *KN1* (Bolduc et al, 2012), *OFP1* was also investigated in the pilot experiment as a potential mediator of *STM* function which had not been

identified by the Bayesian Network.

STM was selected as a negative control as unless it is autoregulatory, we would expect its expression should be broadly unchanged following addition of DEX, since only the localization is changed. It was uncertain how long induction with DEX was required in order to observe an effect, or whether too lengthy an exposure to CHX would trigger a transcriptomic response. Thus, the expression of each gene was compared in the CHX+DEX vs CHX contrast against a DEX vs Mock (DMSO) treated pair of samples at 1, 3 and 5 hours of exposure. The intention was to obtain as close a match between the DEX vs. Mock and DEX+CHX vs. CHX samples as possible. As can be seen Figure 4.10, it appeared that *AIL7* and *CUC1* were both likely to be direct targets as they were changed by 2-fold in each comparison; however by 5 hours of induction the results were a lot noisier, and *AIL7* which had been significant at all other time points was no longer significantly upregulated – this was interpreted as possibly being an effect of toxicity following longer treatment. In 1 hour the observed fold changes were smaller. Thus 3 hours was selected as the optimum length of treatment. *OFP1* was also identified as responsive to DEX in the presence of CHX indicating that it may be a direct target.

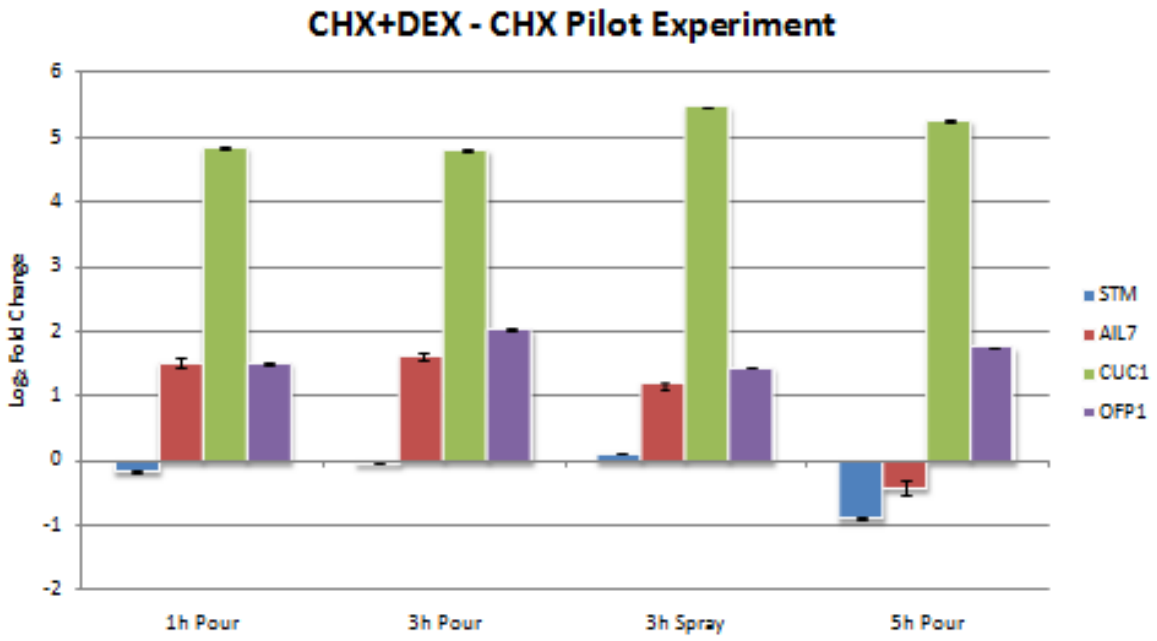


Figure 4.10 - CHX+DEX – CHX Pilot qRT-PCR experiment.

qRT-PCR validation , *AIL7, CUC1* and *OFP1* following 1, 3 and 5 hour long treatments of 60μM CHX + DEX compared to 60μM CHX alone. Plants were either treated by spraying or pouring induction mixture onto the agar plates the plants were grown on. Fold changes computed via ΔΔCt Method normalized against an *ACTIN2* control. Each contrast is composed of 3 technical replicates with error bars indicating standard error of the mean.

4.4.2 Shortlisting of significant genes from CHX-DEX Microarray Experiment

It appeared the Bayesian network approach had successfully identified at least 2 direct targets (*AIL7* and *CUC1*). In order to establish the effectiveness of the predictions in an efficient manner, a microarray experiment was performed using the same lines and conditions as the RT-PCR experiment, this served both as additional validation of predicted targets and to identify as many of *STM*'s direct targets as possible.

In the CHX-DEX microarray experiment, 564 genes were significantly differentially expressed against the CHX treated sample ($p \leq 0.01$). This is almost twice as many as identified in the DEX-Mock experiment described in Chapter 3 using the same line. This

suggests that CHX may introduce significant variability in gene expression levels that may obfuscate changes in response to STM and that a scheme for filtering CHX responsive genes is necessary when attempting to identify true significantly differentially expressed probesets. Alternatively, it is possible that the presence of CHX blocks negative feedback loops that otherwise mask direct targets. However, as has been shown for the maize homolog *KN1* (Bolduc et al, 2012), TF binding is often more widespread than would be predicted from gene expression data. In the *STM-GR* system, where a large burst of *STM* enters the nucleus at the same time, and under the effects of CHX where potential negative regulatory factors are not subsequently being translated it is conceivable than genes would be differentially expressed which would not be differentially expressed following DEX treatment alone.

To control for spurious results, and given the larger size of the CHX-DEX dataset, two methods were used to filter for putative positive targets.

1. Only genes which responded in the Mock-DEX experiment described in chapter 3 were considered in this analysis. This was to filter out targets which were merely responding to CHX treatment or to the abnormally high concentration of STM.
2. Secondly, a more stringent p-value threshold for the both datasets was used in order to remove as many false positives as possible.

In order to select the appropriate p-value, the significance of the overlap between the CHX-DEX vs. CHX and Mock vs. DEX datasets was compared for various p-values (Figure 4.12**)**. At all possible p-value combinations, the overlap was highly significant. As can be seen, there was only a small difference in the number of overlapping genes between $p<0.001$ and $p<0.0005$, so decreasing the p-value threshold below 0.001

would not be likely to eliminate many false positives, and since this represents a highly stringent significance level, p<0.001 was selected as a p-value threshold for both datasets.

P-Value	CHX+DEX - CHX	DEX - Mock	Overlap
0.05	1051	485	196
0.01	564	207	93
0.005	416	157	71
0.001	206	78	36
0.0005	161	65	33
0.0001	70	29	17

Table 4.11 - Overlap between DEX-Mock and CHX+DEX vs CHX.
At differing p-values (column 1), the numbers of significantly differentially expressed genes in CHX+DEX vs CHX and DEX vs Mock direct target microarray experiments (columns 2 and 3). The overlap column provides the size of each overlap.

This resulted in 36 genes being classified as potential direct targets. To these, two further filtering steps were performed. First in order to filter for biological relevance, a 2 fold-change threshold in the DEX-Mock dataset was applied, eliminating genes which did not reach this threshold. Finally, the variability in the CHX microarrays was considered by producing a Mock vs. CHX dataset. Genes which showed more than a 2-fold change in any direction in this dataset were considered suspect and also eliminated from the high-confidence shortlist. In the final shortlist 21 genes remained. (Figure 4.13).

4.4.3 CHX-DEX Microarray confirms direct target validations and suggests additional potential relationships

The list of potential direct targets identified by the CHX-DEX microarray is shown in Figure 4.14. Only 8 of these were identified as significantly differentially expressed in the 8 hour OE dataset, and 11 were significant in the 24 hour dataset.

CUC1 and *AIL7* were confirmed as being direct targets as both were present in the high-confidence shortlist. Three other TFs were identified within this dataset, *HB25, LOL1* and *BOP2*. Of these three, *LOL1* had not been identified in the robust response meta analysis and thus was not present when constructing the Bayesian Network.

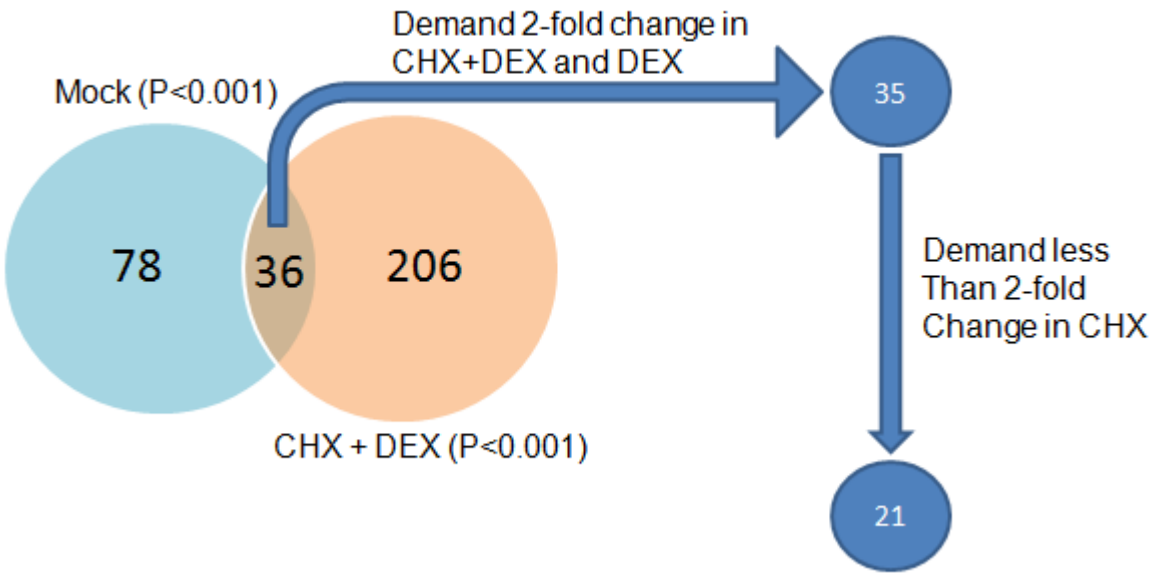


Figure 4.12 - Shortlisting Protocol for Direct Target Microarray Experiment.
Schematic of shortlisting procedure applied to CHX + DEX experiment to generate high-confidence list of likely *STM* direct targets. The number of genes at each shortlisting phase is given inside the blue circles. The venn diagram indicates the overlap between the DEX-Mock contrast and the CHX+DEX – CHX contrast.

However *BOP2* had also been suggested as a possible direct target by the SOM analysis. Overall the direct target analysis confirmed that the consensus network had captured two direct target relationships between *STM* and downstream TFs. It had also

identified three additional possible direct target TFs, as well as suggesting 16 additional possible direct targets.

AT Number	8h P-Value	24h P-Value	Symbol
AT3G22550	0.73	0.0088	
AT1G10070	0.045	1.14×10^{-6}	*BCAT-2*
AT1G22160	0.27	0.00038	
ATAG16080	0.00039	2.83×10^{-14}	*CXE17*
AT3G24450	0.57	0.037	
AT5G63140	0.14	0.43	*PAP29*
AT5G65510	0.0012	0.00034	*AIL7*
AT3G15170	9.65×10^{-7}	8.96×10^{-8}	*CUC1*
AT1G67040	0.26	0.2	
AT5G02760	0.97	0.85	
AT5G00570	0.012	0.16	*CALS1*
AT2G38400	0.61	0.12	*AGT3*
AT3G16180	1.17×10^{-5}	1.96×10^{-7}	
AT5G51290	0.84	0.82	*LOL1*
AT5G61290	0.0056	0.017	
AT1G58360	0.25	0.001	*AAP1*
AT1G44760	0.086	0.00048	
AT5G65410	0.103	3.65×10^{-5}	*HB25*
AT2G39130	0.59	0.08	
AT2G29340	0.99	0.87	
AT2G41370	7.85×10^{-5}	2.25×10^{-9}	*BOP2*

Table 4.13 – Shortlisted Genes from Microarray Direct Target Experiment.
The 21 genes shortlisted from the 3h CHX+DEX vs. CHX microarray experiment. TFs have been highlighted in orange. The corrected p-value at 8 and 24 hours on the over-expression time course described in Chapter 3 is given along with gene name (Symbol) if available.

4.4.4 Chromatin Immunoprecipitation provides additional validation of direct targets

As conflicting results for *OFP1* had been produced, and the direct target microarray experiment had proven noisy, a ChIP experiment using an anti-GR antibody was performed in the Murray Lab order to further validate direct targets (Scofield S, personal communication.) As can be seen in Figure 4.15, **t**his experiment showed that the *OFP1* promoter is directly bound by *STM*. Additionally, *CUC1, AIL7, BOP2* and *HB25* were found to be directly bound by *STM*. Thus all proposed direct TF targets of *STM* have been shown to be direct targets by at least 2 methods (microarray or qRT-PCR of CHX direct target experiments, or ChIP)

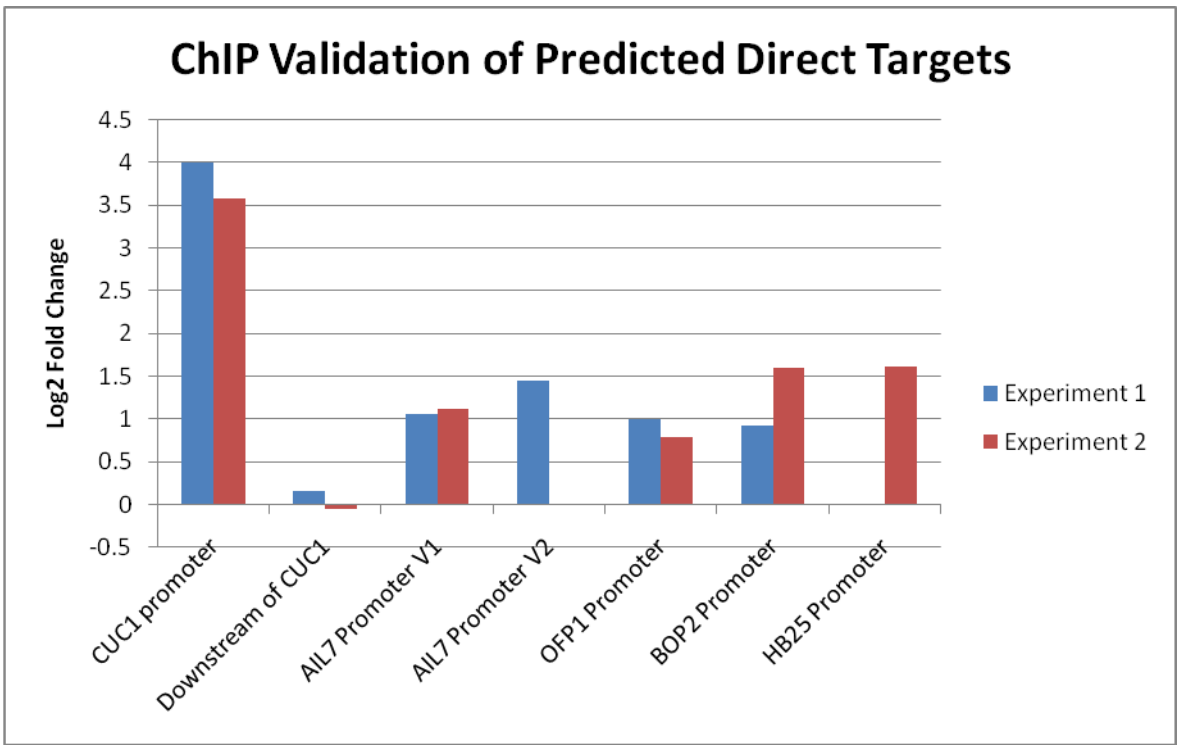


Figure 4.14 - ChIP Validation of Predicted Direct Targets.
Log_2 fold change observed by ΔΔCt analysis of qRT-PCR data between immunoprecipitated and input samples using primers designed against ACT2 to normalize. Antibody against the GR domain (Abcam, Cambridge) was used to pull down regions of the genome bound by *STM-GR* upon nuclear translocation. *CUC1, AIL7, OFP1, BOP1* and *HB25* promoters refer to primers for 500bp regions containing TGAC cores. *HB25* was only tested in one experiment, two regions of the *AIL7* promoter were tested, one was only tested in one experiment. Downstream of *CUC1* refers to primers for a location outside of the *STM* binding site within the *CUC1* promoter described by Spinelli et al (2012).

4.5 Refinement of Bayesian Network

4.5.1 Direct target data used to refine network through additional constraints

Having demonstrated that *STM*'s direct targets within the TF dataset are *BOP2, OFP1, CUC1, HB25* and *AIL7*, a new network structure was inferred using the same procedure as for the initial consensus network, but whereas previously *STM* direct targets were restricted to those in the rapid response dataset, *STM*'s direct targets were forced to be only the five known TF targets, and all other edges from *STM* were forbidden. There was a surprisingly large number amount of changes between the two networks, with 49% of the edges present in the unrefined network being still present in the refined network, resulting in 47.5% of edges in the refined network being maintained from the original consensus network. This suggests modification of at least half the GRN and that the constraints upon *STM*'s direct target has led to changes elsewhere in the network. Finally, the same thresholding procedure described in §4.3.6 was applied to filter low-confidence edges, and in total 47 edges were removed.

Both consensus networks share some nodes with very high degree, these nodes for the refined network before and after thresholding are shown in Figure 4.16. As can be seen by comparison to Figure 4.4, the hub genes are identical between the two, and the extent to which their out-degree is higher than their in-degree remains large. Two of the more connected hub genes are direct targets of *STM* (*AIL7*, and *HB25)*. However *HB25* in particular loses a substantial number of its predicted connections following thresholding, along with *LNG1*. As can be seen in Figure 4.17**,** when visualizing the graph prior to thresholding, the high connectivity of the hub genes creates a very bushy network, despite the fact that many of the hub nodes are directly connected to one another or separated by a relatively small number of intervening nodes. The known relationships between *TCP3* and *KNAT1* with *AS1* are preserved in the refined network

both before and after thresholding.

Gene	In-Degree	Out-Degree	Degree
CUC3	2/2	5/4	7/6
AT1G68810	3/3	4/3	7/6
STM	2/1	5/5	7/6
ATH1	2/2	6/1	8/3
AIL7	3/3	6/6	9/9
WRKY14	1/1	8/4	9/5
KNAT1	2/2	8/7	10/9
TCP3	2/2	8/8	10/10
HB25	3/3	9/2	12/5
TCP10	1/1	13/12	14/13
LNG1	3/3	12/6	15/9

Figure 4.15 – Degree of Nodes in Refined Network.
For each node with overall degree 7 or more before thresholding in the refined network, the In-degree, out-degree and overall degree before and after thresholding (in the form: degree before/ degree after) is shown.

As can be seen in Figure 4.18**,** the same patterning between up and down-regulation can be observed, with the *TCP* genes, *AS1* and *LNG1* acting as repressive hubs in the network. The spatial patterning is less clear in this refined network (Figure 4.19), which is partially due to the fact that *BOP2* is a direct target which is differentially expressed at a higher level in the primordia than the rest of the SAM. This has brought some of the genes which are dependent on *BOP2* closer to *STM* on the network resulting in a greater concentration of primordia specific genes in its neighbourhood. However, as *STM* is known to target *BOP2* from this thesis, this raises the question of whether *STM* actually acts positively on primordium specific genes to a greater extent than previously suspected. However, it is very clear that genes in the immediate vicinity of *STM* are CZ enriched.

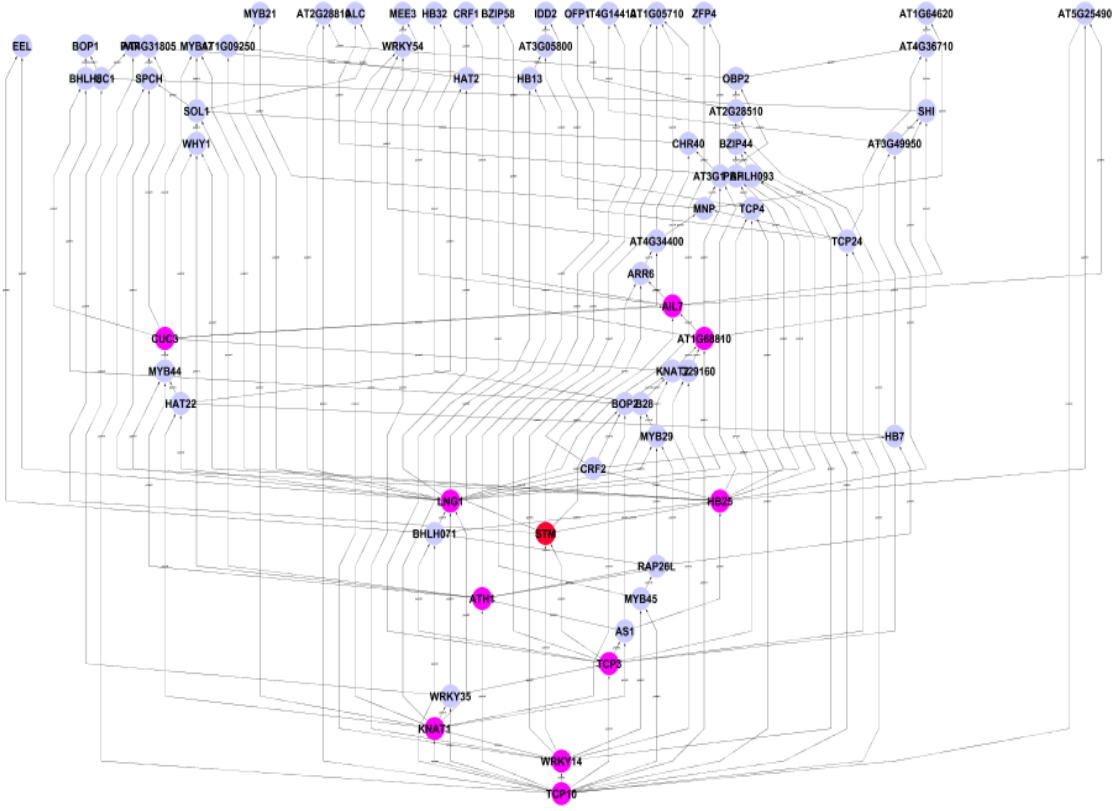


Figure 4.16 - Refined Consensus network of transcription factors regulated by STM with high degree nodes overlaid.
The consensus network describes conditional dependency relationships (shown as edges with arrows indicating child nodes) between transcription factors (depicted as nodes) identified in the robust response meta analysis from a range of datasets (see 2.1.5). The network was inferred using BANJO as described in 2.1.6. No edges were permitted from *STM* to transcription factors not present in the rapid response meta analysis, known direct targets of *STM* were forced to be children of *STM* and no other nodes were permitted to be children of *STM* and no node was permitted to have greater than 5 parents. *STM* has been highlighted in red, transcription factors with degree 7 or greater have been highlighted in purple.

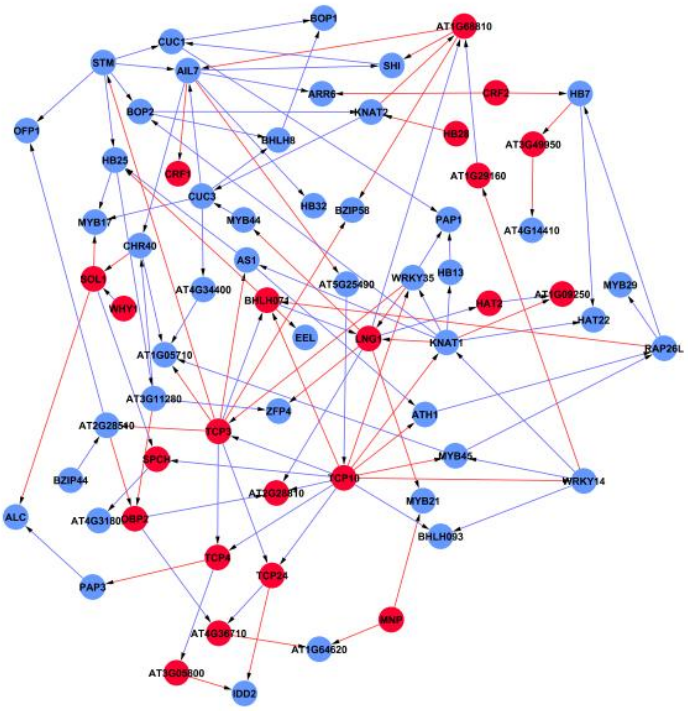


Figure 4.17 - Refined Consensus network of transcription factors regulated by *STM* limited by correlation, overlaid by fold change at 72 hours.
The consensus network describes conditional dependency relationships (shown as edges with arrows indicating child nodes) between transcription factors (depicted as nodes) identified in the robust response meta analysis from a range of datasets (see 2.1.5). The network was inferred using BANJO as described in 2.1.6. No edges were permitted from *STM* to transcription factors not present in the rapid response meta analysis, known direct targets of *STM* were forced to be children of *STM* and no other nodes were permitted to be children of *STM* and no node was permitted to have greater than 5 parents. Pearson's Correlation coefficient between each TF pair was calculated over all time course contrasts, and for the *stm-2* mutant and *STM*-GR experiments presented in Chapter 3. All edges showing between -0.5 and 0.5 R were removed. Edges are coloured according to whether the Pearson's Correlation coefficient was negative (red) or positive (blue). Nodes have been coloured according to whether they are up (blue) or downregulated (red) at 72 hours of *STM* over-expression.

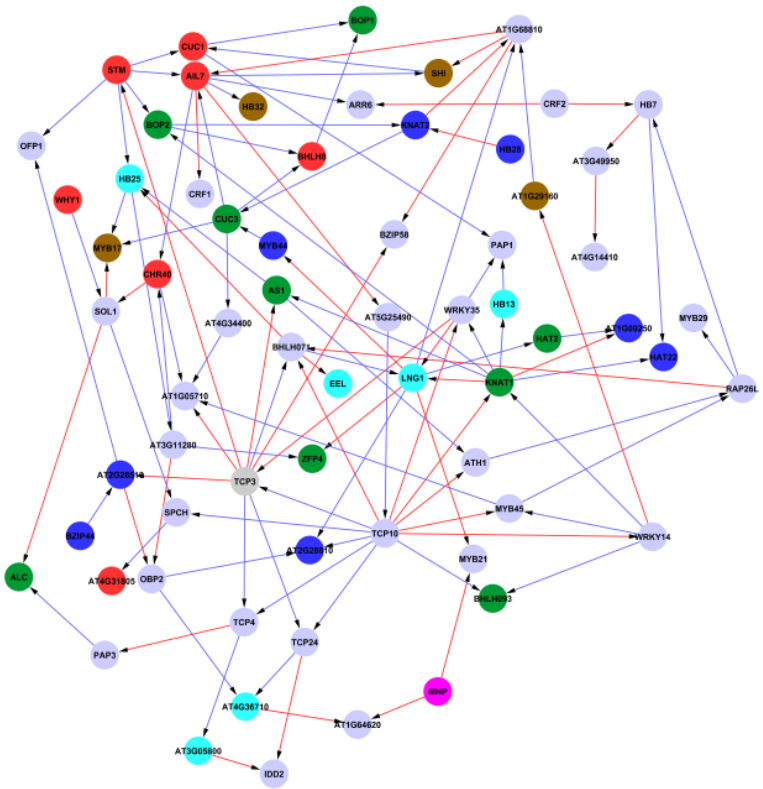


Figure 4.18 Refined Consensus network of transcription factors regulated by *STM* limited by correlation, overlaid by predicted expression domain within the SAM given the reanalysis of the Yadav et al (2009) dataset.
The consensus network describes conditional dependency relationships (shown as edges with arrows indicating child nodes) between transcription factors (depicted as nodes) identified in the robust response meta analysis from a range of datasets (see 2.1.5). The network was inferred using BANJO as described in 2.1.6. No edges were permitted from *STM* to transcription factors not present in the rapid response meta analysis, known direct targets of *STM* were forced to be children of *STM* and no other nodes were permitted to be children of *STM* and no node was permitted to have greater than 5 parents. Pearson's Correlation coefficient between each TF pair was calculated over all time course contrasts, and for the *stm-2* mutant and *STM*-GR experiments presented in Chapter 3. All edges showing between -0.5 and 0.5 R were removed. Edges are coloured according to whether the Pearson's Correlation coefficient was negative (red) or positive (blue). Nodes have been coloured according to predicted spatial domain given the reanalysis of the Yadav et al (2009) dataset described in 4.3.5. Red – CZ specific, Blue, OC specific, Green – primordium specific, Cyan – Excluded from CZ, Brown – Excluded from OC, Purple – Excluded from Primordia. Grey indicates no data is available and light blue indicates no specific expression domain.

4.6 Identification of interesting target subsets without prior microarray data

The previous experiments confirm that there is sufficient publically available data to use Bayesian networks to identify true associations between genes in Arabidopsis with a reasonable level of accuracy. This poses the question of whether it is possible to use

the existing publically available data to identify subsets of genes, appropriate for subsequent analysis using Bayesian network structural inference. This would allow for large cost savings to be made in terms of microarray or sequencing experiments that would need to be performed.

One way to identify potential subsets of genes would be to use co-expression analysis, and take as a subset for analysis all genes with a correlation co-efficient with the gene of interest above a certain threshold. When performed using the same (undiscretized) datasets which were used for the Bayesian network structural inference, the highest r^2 obtained was 0.35 (excluding *STM* with itself). The most strongly correlated TF which is a known direct target of *STM* was *HB25* with r^2 of 0.21. These low correlation co-efficients may be due to the large number of control datasets within the data. While a Bayesian network may infer a relationship between genes which have low correlation as it is evaluating the optimization of a network structure over the genes, to find the best fitting edge structure, these relationships may be missed by correlation analysis. Additionally, it is clearly difficult to justify using such low correlation co-efficients.

As such, a separate experiment was performed using the CressExpress tool as detailed in Materials & Methods. The intention was to use a smaller, more focussed subset of experiments for co-expression analysis. In this case, *AIL7* and *CUC1* had r^2 values with *STM* of 0.85 and 0.77 respectively, whereas *HB25* had r^2 of 0.57. *OFP1* and *BOP2* could not have been detected with any reasonable threshold as they had r^2 values with *STM* of 0.1 and 0.37 respectively. Additionally, while 2,540 genes had r^2 value of at least 0.5, only 398 had higher than 0.75. Thus a focussed subset of experiments with an intuitive r^2 threshold or 0.75 would be able to identify the same two direct target TFs which were predicted by the Bayesian network.

4.6.1 Using a subset of genes predicted by co-expression analysis, the Bayesian network recapitulates several known direct target relationships with *STM*.

Figure 4.20 shows a Bayesian network for all genes with r^2 of greater than 0.75 with *STM* annotated for sequence specific DNA binding activity by Gene Ontology. As can be seen, the number of predicted targets for *STM* is higher than in the constrained initial network used for this model. However, it does predict that *AIL7* and *CUC1* are direct targets of *STM*, although it also places *AGAMOUS, AGAMOUS-LIKE42, At2G35310, At1G88360, At4G00870, HOMEOBOX33, At3G06220, At5G60200, HOMEOBOX21* and *At2G35430* as connected to *STM*, and there is no evidence that these genes are direct targets. Thus as judged from the microarray direct target experiment, its false positive rate is far higher than the Bayesian network derived from the time course of *STM* induction.

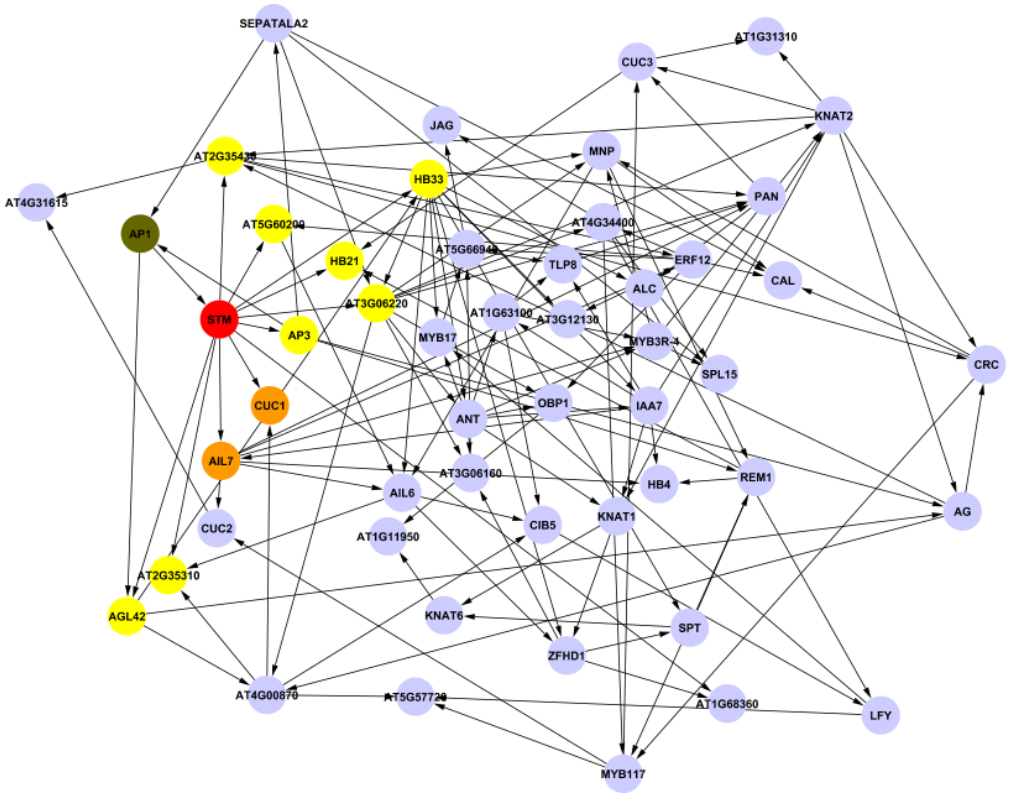


Figure 4.19 - Consensus network of transcription factors co-expressed with *STM* overlaid by predicted and known direct targets of *STM*.
The consensus network describes conditional dependency relationships (shown as edges with arrows indicating child nodes) between transcription factors (depicted as nodes) co-expressed with *STM* ($r^2 > 0.75$ in CressExpress as per 4.6.1) from a range of datasets (see 2.1.5). The network was inferred using BANJO as described in 2.1.6. No edges were permitted from *STM* to transcription factors not present in the rapid response meta analysis and no node was permitted to have greater than 5 parents. *STM* has been highlighted in red, transcription factors predicted to be direct targets of *STM* have been highlighted in red, transcription factors predicted to target *STM* have been highlighted in yellow, known direct targets of *STM* have been highlighted in yellow. The gene predicted to directly target *STM* has been highlighted in brown.

However, without having to perform any initial gene expression experiments, we could derive both correct predictions regarding the two *STM* direct targets which were identified by the initial Bayesian network approach. One flaw with this method may be that we are less likely to have detected those genes further away from *STM* direct targets on the Bayesian network as these are likely to be less correlated with *STM*. It also clearly has less specificity that the previous network predictions, with only 20% of the predictions being validated by direct target microarray experiment. An additional

complication is that some of the predicted direct targets, like *AP3* (*APETALA3)* and *AGL42* (*AGAMOUS-LIKE42)* are involved in floral transitioning (Dorca-Fornell et al, 2011, Wuest et al, 2012). While this project has focussed on the role of *STM* in vegetative tissue, it is known that *STM* also promotes carpel and formation when ectopically expressed in floral tissues, while *STM* knock-down by RNAi causes developmental defects in flowers which can be as severe as a complete loss of carpel development (Scofield et al, 2009). Thus another possibility is that this network may have captured *STM* targets which were not expressed within the OE time course or *STM-GR* direct target experiment due to the developmental stage of the plants in question.

4.7 Discussion

The SOM and the Bayesian Network were both relatively accurate at identifying likely direct targets of *STM*. Two of the five upregulated TFs closely correlated with *STM* on the SOM were identified as direct targets – *OFP1* and *BOP2*. Meanwhile two of four of the predicted direct targets of *STM* on the Bayesian network were identified as direct targets – *AIL7* and *CUC1*. *HB25* was the only TF not identified as a possible direct target by either method.

Interestingly the two methods effectively identified the genes that the other missed, suggesting these are useful complementary approaches. One possibility is that the discretisation of genes for the Bayesian network structural inference masked the detection of some relationships which it may have otherwise detected. Alternatively it could be that some relationships were not as easily identified from the larger dataset as from our time course data, and vice versa. In this instance the two methodologies appear to have complemented each other, though this does not imply that this would always be the case.

Finally, it is possible to use only generic, publically available datasets to identify appropriate subsets of genes for Bayesian network structural inference and that known relationships can be recapitulated when provided this subset of genes. As a consequence it is possible, though less accurate to use entirely publically available datasets to make reasonable inferences regarding genes of interest without having to perform wet lab experiments.

CUC1 is a very interesting direct target, particularly as we are aware that *STM* is also a target of *CUC1* (Mitsuhiro Aida, personal communication). This suggests that the two genes are in a positive feedback loop with one another; however, it has been shown in the literature that, while *CUC1* is expressed earlier in embryonic development than *STM,* following expression of *STM, CUC1* is excluded from most of *STM*'s expression domain (Aida et al, 1999). An interesting open question would be to identify the mechanism by which *CUC1* is excluded from *STM*'s domain, as it must clearly be sufficiently robust to overcome the putative positive feedback loop in place. This relationship will be investigated in greater detail in Chapter 5

OFP1 was not initially identified as a likely direct target of *STM* as it was not identified via the high-throughput direct target experiment or the Bayesian network. However, ChIP and qRT-PCR both suggest that it is a direct and positively regulated target of *STM*. As has been previously noted, *OFP1* is known to directly repress the *GA20OXIDASE* which is seen to be strongly repressed by *STM*. Thus, since *STM* does not appear to directly target the *GA20OXIDASE* genes as *KN1* does in maize– *OFP1* is a very likely candidate for mediating this effect in Arabidopsis. This result would show that the known repressive effect of *STM* on GA levels (Hay et al, 2002; Rosin et al, 2003; Jazinski et al, 2005; Ikezaki et al, 2010) is mediated at least in part through upregulation of *OFP1*, which then represses *GA20OXIDASE*. This is in contrast

GA20OXIDASE to the mechanism by which *KN1* represses GA levels in tobacco through direct repression of *GA20OXIDASE* (Sakamoto et al, 2001).

Very little is known about the specific function of *HB25.* In the Bayesian network prior to thresholding, it had a high degree of connectivity, suggesting it may be a transcriptional hub downstream of *STM*, although many of its predicted targets are lost following thresholding, which suggests that it may either be correlated with a number of genes rather than regulating them, that many of the predicted interactions were incorrect or that there is competitive regulation of its targets weakening the strength of the conditional dependency relationship observed. *STM* also appears to target *BOP2* – however given that it is located in a different region of the SAM in vivo (Ha et al, 2004) it is unclear whether this is an association only found when *STM* is expressed ectopically, or whether it may be due to protein movement, a requirement for additional co-factors for *STM* binding to the *BOP2* promoter, or other factors which may keep *STM* and *BOP2* in separate domains.

AIL7 is by contrast better studied. *AIL7* is closely related to the *AINTEGUMENTA* and *PLETHORA* genes, with extensively documented roles in controlling aerial organ and root growth (Elliot et al, 1996, Nole-Wilson, 2006). Prasad et al (2011) have demonstrated a role for *AIL7* in phyllotaxis, as triple *ail7plt3plt5* mutants show some phyllotactic aberrations which they attributed to misexpression of *PIN1* and thus auxin flow. Mudunkothge et al (2012) demonstrated that *AIL7* is most strongly expressed within the CZ, in agreement with the Yadav et al (2007) dataset. These authors also showed that mutation of *AIL7* can partially recover the phenotype of the strong *STM* mutant *stm-1* – as a proportion of *stm-1 ail7* plants went on to produce leaves and flowers, suggesting that when *AIL7* is not being regulated by *STM* its misexpression may have serious consequences, or that possibly due to its proposed effects on auxin

localisation in contrast to the role of *STM* in promoting cytokinin levels and responses, *AIL7* may antagonize *STM* at the functional, rather than transcriptional, level.

Taken together, the known direct targets corroborate well with the fact that *STM*'s most early and robustly enriched GO categories related to organogenesis and regionalization as *BOP2*, *CUC1* and *AIL7* all play important roles in proper organ and boundary formation. This suggests that understanding how boundaries are formed is critical to understanding *STM* function. *CUC* genes are critical genes in boundary specification, often being used to define the location of the boundary in localization experiments. Thus, *STM*'s direct positive induction of *CUC1,* coupled with their interactions early during embryogenesis, suggests that understanding how this loop is formed may be critical to understanding how the boundary between the SAM and emerging primordia is formed and thus, this was identified as a critical downstream module of *STM's* GRN to probe in greater detail.

Chapter 5 – Elucidation of the *CUC1-STM* regulatory module

5.1 Introduction

CUP-SHAPED COTYLEDON1 encodes a NAC family transcription factor, first described by Aida et al (1997) and identified along with its closely related homologue *CUC2* as having a striking fused cotyledon phenotype in the *cuc1 cuc2* double mutant. Furthermore this phenotype was associated with the absence of the embryonic and hence also the vegetative SAM, indicating a vital developmental role for these transcription factors. Single mutants showed at low frequency (<1%) of fusion along one side of the cotyledons, but otherwise display a wild type phenotype (Aida et al, 1997, Aida et al, 1999) which implies functional redundancy. A genetic link between *CUC* genes and *STM* has been known for some time, as the *cuc1* and *cuc2* mutations enhance both *stm-1* (strong mutant) and *stm-2* (intermediate) phenotypes (Aida et al, 1999). Both the work in Chapter 4, and results of Spinelli et al (2011) have shown that *CUC1* is a direct target of *STM*. Work from Mitsuhiro Aida's lab has suggested that *STM* is also a direct target of *CUC1* (M. Aida, personal communication). Aida's group performed qRT-PCR on 7 DAS RPS5Ap:*CUC1-GR* plants using a CHX+DEX experiment, which is analogous to the CHX+DEX experiment performed on 35S:*STM*-GR in Chapter 4). This work demonstrated that *STM* was upregulated 3-fold relative to a *TUBULIN4* control following CHX+DEX treatment (M. Aida, unpublished data). It is also known that in the *cuc1 cuc2* double mutant no *STM* expression is detected, however since no SAM is formed the absence of *STM* expression could either be because *STM* lies directly downstream of *CUC1* and its expression is dependent upon correct *CUC* gene expression, or may be due to developmental epistasis since the SAM is absent.

Interestingly, these results would suggest a positive feedback relationship between

CUC1 and *STM*. However, while there may be some overlap around the boundary, in the wild type SAM, *STM* and *CUC1* at best share only partially overlapping expression domains. However *CUC1* is expressed in the early stages of embryogenesis throughout the presumptive SAM, and following the detection of *STM* mRNA expression throughout the SAM, it becomes restricted to the boundary zone (Aida et al, 1999). Schematically this is shown in Figure 5.1.

This poses the question of what would lead to the exclusion of *CUC1* from the CZ. Data suggests that the microRNA family miR164 may be responsible, since the expression of a miR164-resistant *CUC1* mRNA under the expression of the *CUC1* promoter was detected in the CZ by Sieber et al (2007). The miR164 family consists of 3 members, miR164a, b and c, each of which have different expression patterns in the wild type. In looking at inflorescence meristems Sieber et al (2007) showed that miR164c is present throughout the CZ, whereas miR164a and miR165b are excluded from the CZ. miR164a had consistent expression on the primordia side of the boundary between emerging organ primordia and the IM.

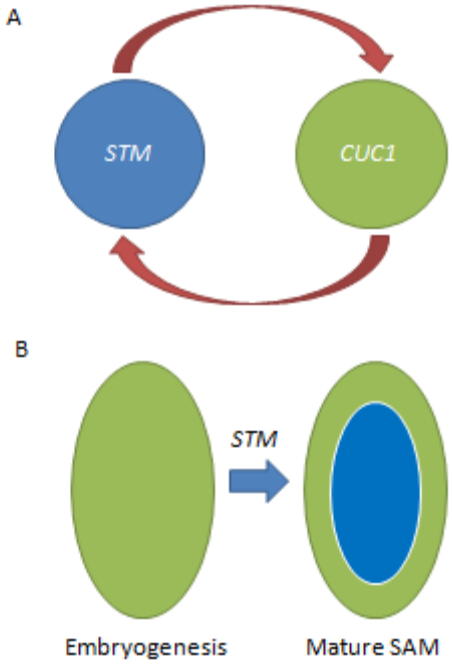


Figure 5.1 – A schematic of changes in *CUC1* expression as they relate to *STM.*

A) A schematic of direct transcriptional relationships between *STM* and *CUC1* which we have shown. Positive transcriptional responses are shown as red arrows. Genes are shown as nodes. As can be seen *STM* should drive *CUC1* to saturation in any domains they are co-expressed if no other factors are involved.

B) A schematic showing the evolution of the *CUC1* expression domain during embryogenesis (left) and following the expression of *STM* (right). As can be seen *CUC1* becomes excluded from the centre of the SAM.

Spinelli et al (2011) examined the effect of *STM* on miR164a, using GUS reporter lines for pmiR164a crossed to *STM* over-expression lines. They suggested that miR164a is overexpressed in the region of the leaf primordium most proximal to the stem following *STM* induction. However, they could not confirm this by qRT-PCR. The expression domain of miR164a however suggest that it cannot be the factor which is excluding *CUC1* from the CZ (as it is not expressed there) unless it is trafficked or has different expression in the vegetative and reproductive SAM. Thus it was hypothesized that *miR164c* the most likely miR164 to downregulate the expression of *CUC1* in the *STM* domain, as it is expressed throughout the CZ.

5.2.1 Feedback between *CUC1* and *STM* is rapid and robust

The relationship between *CUC1* and *STM* is extremely interesting since a direct positive feedback loop between two genes, should lead to extremely rapid saturation of the levels of both. However, as illustrated in Figure 5.1 this is not the case and *CUC1* is excluded from a significant proportion of *STM*'s native domain.

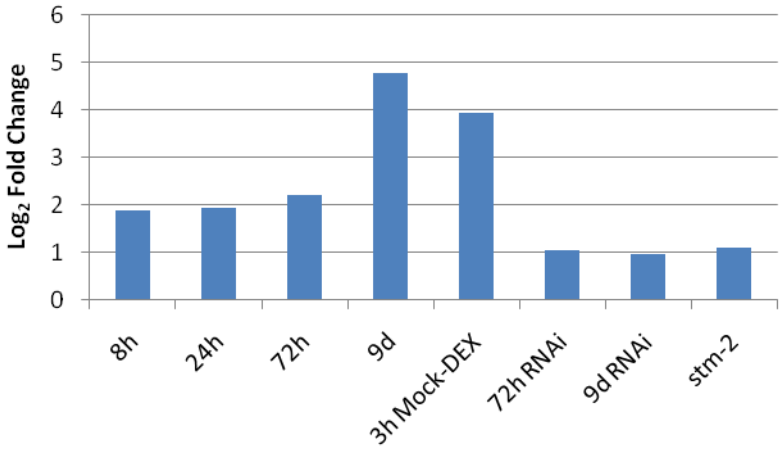


Figure 5.2 – Expression of *CUC1* over STM-related microarray experiments.
Y axis shows point estimates of Log_2 Fold-Changes of *CUC1* expression calculated using LIMMA from microarrays of *STM* over-expressing plants and *STM* RNAi lines versus Mock treated controls, and the *stm-2* mutant versus wild type. X axis labels correspond to lines as follows; 8h, 24h, 72h and 9d represent the TGV 2-component system at 8, 24, 72 hours and 9 days as described in 3.2.1, 3h Mock DEX represent the *STM*-GR line described in 3.2.7, 72h and 9d RNAi represent the *STM* RNAi lines described in 3.2.1.and *stm-2* represents the *stm-2* vs wt contrast described in 3.2.1.

I have previously shown that when expressed ectopically, an increase in the expression level of *STM* leads to an increase in the level of *CUC1* mRNA by microarray. Under multiple expression systems, we are able to observe a rapid induction of *CUC1* which further increases confidence that this is a genuine effect. The fold changes observed are shown in Figure 5.2. However, this does not capture the spatial aspects of *CUC1*

upregulation following *STM* induction. In order to investigate this further I crossed the line S^{38} (the TGV two-component system driving *STM* mRNA previously used for the time course experiment), to both *pCUC1*:*CUC1-GUS* and *pCUC1:GUS* lines. The former is a fusion of the entire *CUC1* promoter and coding region to *GUS*, and contains the mi164 target site, which is located within the coding region of *CUC1*, and the latter lacks this site since it contains only the *CUC1* promoter fused to *GUS*.

Figure 5.3 shows that in 5 DAS uninduced *pCUC1:GUS* x S^{38} plants (a,b), p*CUC1* promoter activity is only observed around the SAM, associated with emerging primordia. The expression at the tips of emerging leaves is due to leaky expression of *STM* and is visible in the parental *pCUC1-GUS* line. In contrast, in plants which have been grown on DEX (c,d), the entire region around the SAM and the primordia express GUS. This shows that long term induction of *STM* is sufficient to drive expansion of the *CUC1* expression domain.

Figure 5.3 also shows the effect of transient induction of *pCUC1:GUS* x S^{38} plants. As can be clearly seen, following 24 hours of *STM* induction, *CUC1* promoter activity can be observed in an expanded domain to the uninduced plants. The expression of *GUS* is no longer detected just in the primordia, but has expanded into emerging organs. Unlike in the long term inducted plants, following 24 hours of *STM* induction *GUS* activity is not detected in the region surrounding the SAM,

On the assumption that expression of miR164a and miR164c is consistent between inflorescence meristems and the vegetative SAM, *CUC1* promoter activity overlaps the expected expression domain of miR164a (Sieber at al, 2007). Thus it is expected that we would see a difference between expression of *pCUC1:CUC1-GUS* and *pCUC1:GUS* specifically we would expect expression of *pCUC1:CUC1-GUS* to be more restricted, as some of the mRNA will be targeted by miRNA164a before the *CUC1-GUS* can be

translated.

Figure 5.3 shows the pattern of GUS activity in *pCUC1:CUC1-GUS x S*38 plants. As can be seen in (a,b), in uninduced plants we only detect *pCUC1:CUC1-GUS* in the tips of emerging leaves, irrespective of the length of *GUS* staining we used. However, in (c,d, e, f) plants which have been grown on DEX we detect *pCUC1:CUC1-GUS* throughout the leaf margins and in emerging primordia. Since the plants grown on DEX are strongly phenotypic at this point, plants transiently induced with DEX for 48 hours were also examined. Here (g,h) an expansion of *pCUC1:CUC1-GUS* expression is seen matching that of the plants grown on DEX with GUS detected around the leaf margins and in emerging primordia consistent with the expanded region of expression in the plants grown on DEX . However, it is clear that when under control of the *CUC1* promoter the expansion in expression of the CUC1 protein-GUS fusion is far smaller than the expansion in expression of the GUS reporter alone when *STM* is induced. This is consistent with the hypothesis that miR164 or other primordium-specific factors may be responsible for down regulation of *CUC1* mRNA before translation can occur.

5.2.2 qRT-PCR analysis of putative *STM*-dependent regulation of miR164

Since it is known from the literature that miR164 is sufficient to exclude *CUC1* from the CZ of the IM (Sieber et al, 2007) and that *CUC1* is excluded from the CZ of the SAM following *STM* expression, it was natural to ask whether *STM* was upregulating one of the miR164 family. As miR164c has the closest expected expression pattern in the SAM to *STM* the changes in expression levels of miR164c following DEX induction of the S^{38} line and an empty vector control were examined (the same lines as for the microarray time course).

Mature miR164 levels were measured via TaqMan probes following 6 and 24 hours of

induction. The probe for miR164c was exclusive to that mature miRNA, and a second probe measuring combined miR164a and miR164b was used, thus using this system

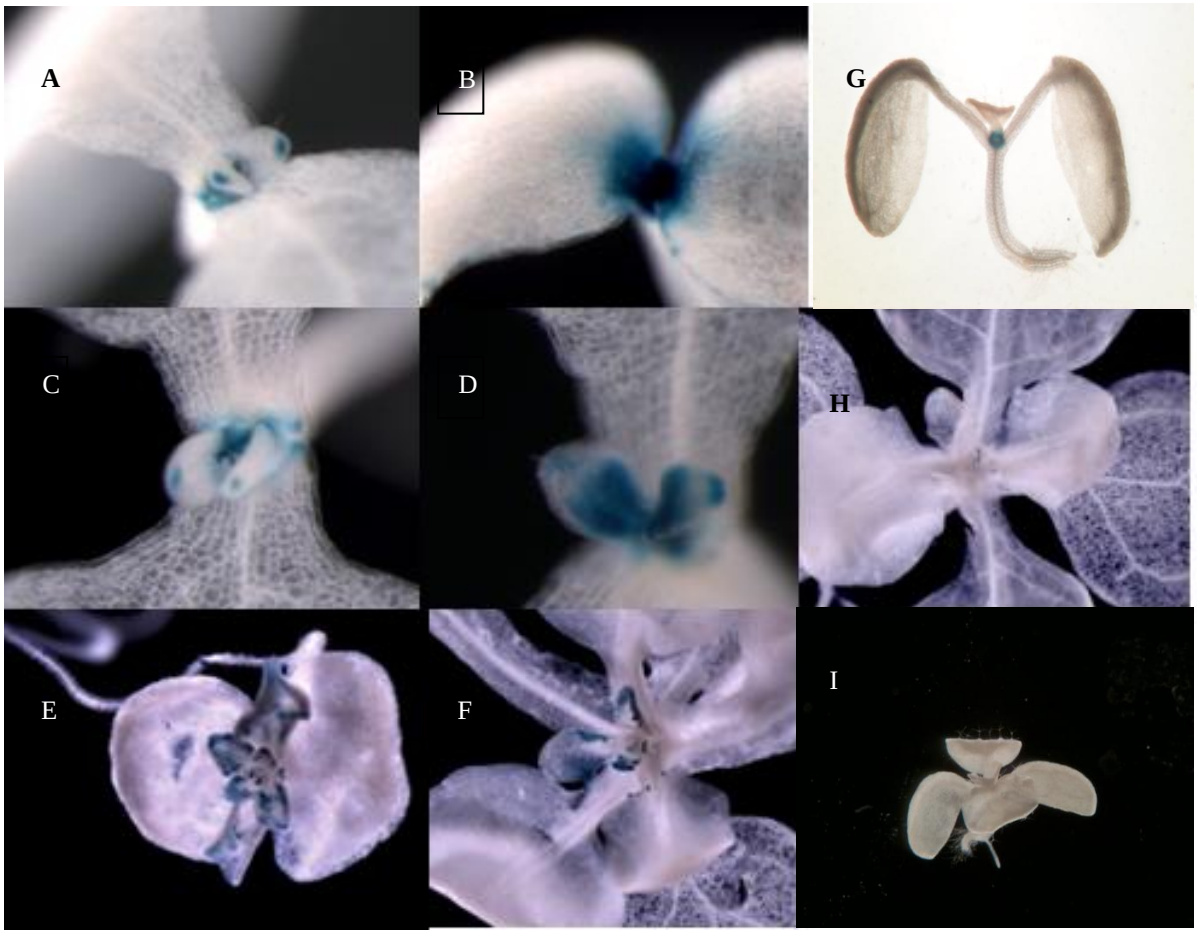


Figure 5.3 – Changes in *CUC1* expression in response to *STM* induction.

(A) The expression of *PCUC1:GUS* (blue pigment) in uninduced *pCUC1:GUS* crossed with S38 (Inducible *STM* over-expression line using the TGV system explained in Chapter 3) plants, 5 days after sowing. (B) The same line grown on GM agar with 60µM DEX. (G) GUS expression in the parental *pCUC1:GUS* line . (C) The expression of *PCUC1:GUS* (blue pigment) in uninduced *pCUC1:GUS* crossed with S38 (Inducible *STM* over-expression line using the TGV system explained in Chapter 3) plants, 5 days after sowing. (D) The same lines grown on GM agar with 60µM DEX added 24 hours before harvesting. (H) The expression of *PCUC1:CUC1-GUS* (blue pigment) in uninduced *pCUC1:CUC1-GUS* crossed with S38 (Inducible *STM* over-expression line using the TGV system explained in Chapter 3) (E) The same lines grown on GM agar with 60µM DEX. (F) The same lines grown on GM agar with 60µM DEX with DEX added 48 hours prior to harvesting. (G) GUS expression in the parental *pCUC1:CUC1-GUS* line grown on GM agar.

we could unambiguously detect changes to *mir164c* expression, but not measure the levels of *miR164a* or *miR164b* individually.

6 and 24 hours were selected as suitable time points to observe expression of the miRNA. As can be seen in Figure 5.4, miR164c expression was very strongly induced (4.9-fold on the $\log_2$ scale) in one of the 24 hour replicates, and induced (by 3.6 on the $\log_2$ scale) in the other. The change in expression of miR164c was less than 1 on the log-2 scale at 6 hours, though it appeared to have been slightly downregulated.

In the same experiment using the TaqMan reporter for combined miR164a and b, no induction of miR164a/b was observed at either time point. This could in principle be due to reciprocal effects of *STM* on miR164a and miR164b, as the probe would detect both microRNAs, though this would require the expression of both microRNAs to show very similar reciprocal dynamics. Thus, in agreement with the qRT-PCR results of Spinelli et al (2011) we do not see evidence of *STM* upregulating miR164a at the transcriptomic level, although these authors did claim to see an increase in *GUS* expression from a miR164a reporter.

In order to provide additional validation of the changes in *miR164c* expression, an RT-PCR experiment was designed to detect changes in miR164c expression without using TaqMan probes, which would permit use of ACTIN controls as in previous qRT-PCR experiments described in this work. To do this, rather than examining mature miR164c levels, primers were designed complementary to the miR164c precursor RNA. Using cDNA samples of total RNA from previous *STM* induction experiments. I examined the level of miR164c induction following 24 hours of *STM* over-expression in S^{38} versus an Empty Vector line, and in a time course contrasting a separate inducible TGV 2-component *STM* over-expression line (S34) induced for 6, 24 and 72 hours against a 0 hour control. The results are shown in Figure 5.4 and clearly demonstrate that except in

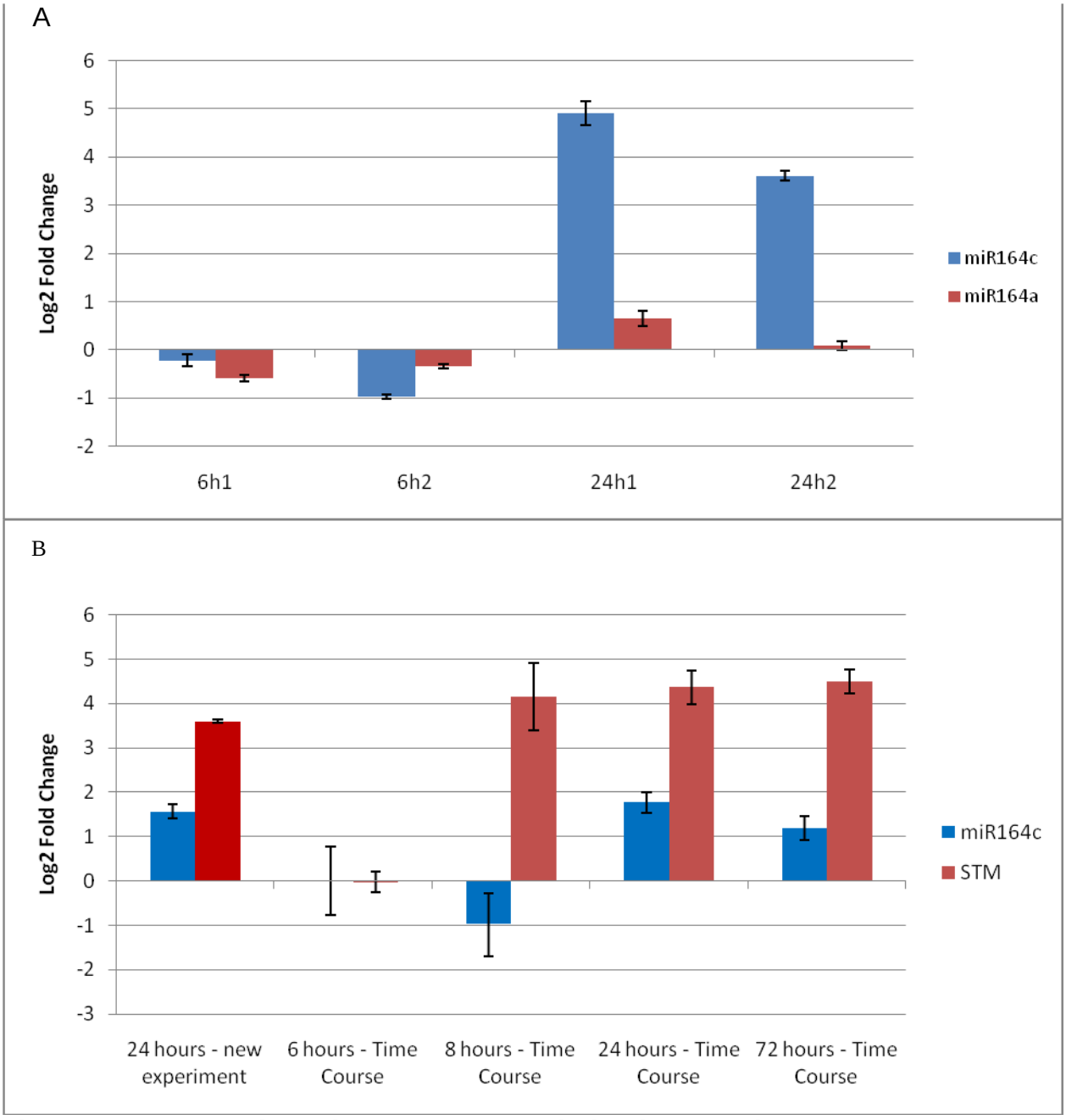


Figure 5.4 – Transcriptomic Effects of *STM* induction on *miR164*.
A) Log_2 fold change measured by qRT-PCR of miR164c Taqman probe (blue) and miR164a/b Taqman probes (red) in *STM* over-expressing 9 day old plants after 6 and 24 hours of induction. Fold changes computed via ΔΔCt Method normalized against an *SnoR85* control. Each contrast is composed of 3 technical replicates with error bars indicating standard error of the mean.
B) Log_2 fold change measured by qRT-PCR of mature miR164c (blue) precursor and *STM* (red) in *STM* over-expressing 9 day old plants at various lengths of induction Fold changes computed via ΔΔCt Method normalized against an *SnoR85* control. Each contrast is composed of 3 technical replicates with error bars indicating standard error of the mean.

the 6 hour sample it is possible to detect miR164c precursor induction by *STM*.

In order to confirm that the miR164c precursor primers were not detecting miR164a or miR164b, products from the 24 hour vs. Empty Vector experiment were sequenced using the miR164c precursor primers. The sequence matched the miR16c precursor exactly and did not match the precursors of miR164a or miR164b.

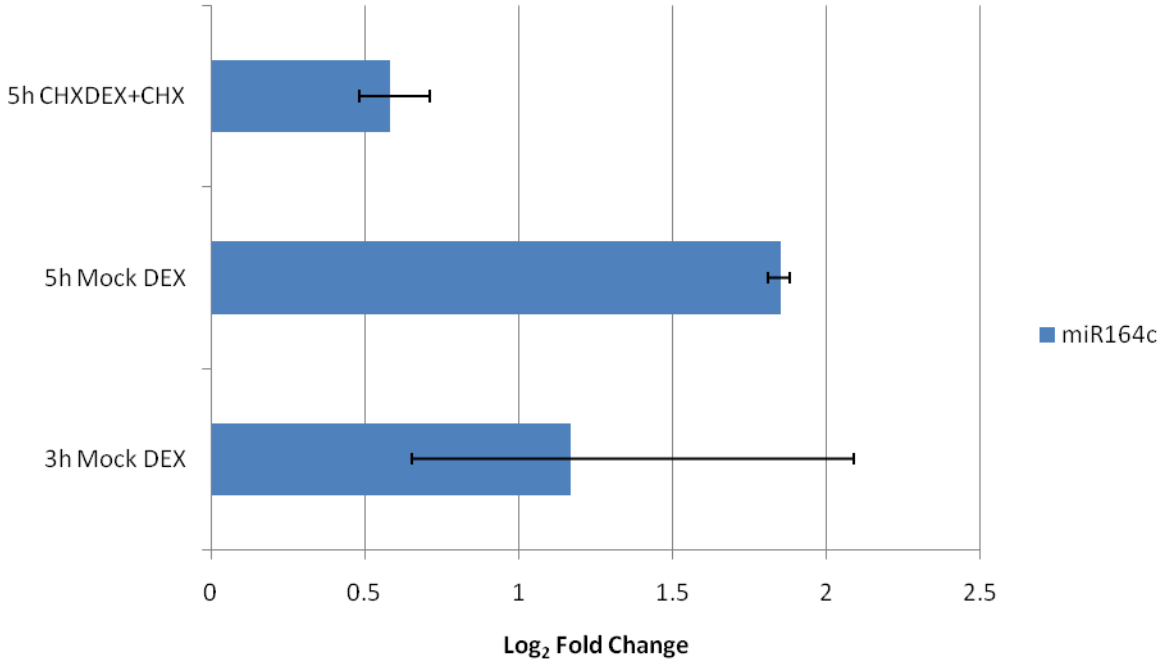


Figure 5.5 – *miR164c* Direct Target Experiment.

Log_2 fold change measured by qRT-PCR of miR164c in *STM-GR* lines treated with either DEX or DEX and CHX for 3 or 5 hours vs Mock or CHX treated controls. Fold changes computed via ΔΔCt Method normalized against an *ACTIN2* control. Each contrast is composed of 3 technical replicates with error bars indicating standard error of the mean.

Since these results suggests that *miR164c* is specifically upregulated by *STM*, I investigated whether it could be a direct target of *STM* using the samples generated for the CHX pilot experiment described in Chapter 4. As a fold change of only +1.17 could

be detected for *miR164c* in the 3h DEX-Mock sample, only the 5 hour CHX+DEX and CHX spray-treated samples were used, where the DEX-Mock contrast showed a fold change of +1.85, which suggests that there is likely to be some observable upregulation of *miR164c* following induction of *STM* for this length of time, When comparing the CHX+DEX treated 5 hour sample to the equivalent CHX treated sample a fold change of +0.58 was detected. As can be seen in Figure 5.5 there was no observable induction of the miR164c precursor in the CHX+DEX lines, suggesting that miR164c is not a direct target of *STM*. This is also consistent with the fact that we do not see an induction of miR164c prior to 24 hours.

5.2.3 *STM* over-expression triggers ectopic activation of the miR164c promoter

Having obtained evidence that miR164c is indirectly upregulated by *STM* at the transcriptional level, I proceeded to investigate how the spatial distribution of *miR164c* is affected by *STM*. p*miR164c:VENUS* lines used in Sieber et al (2007) were obtained from NASC and crossed to S^{38} lines, in order to investigate how the localization of *miR164c* changed within the SAM when *STM* levels are perturbed.

To date, in the literature there are no reports of the expression of the p*miR164c:*VENUS lines used within the SAM. As such, it was necessary to first investigate the unperturbed expression of miR164c within the SAM. Figure 5.6 shows that expression of p*miR164c:VENUS* in the SAM of uninduced plants. p*miR164c:VENUS* is only detected within the L1 layer, Since Sieber et al (2007) only showed the surface of the IM this is a novel result, suggesting that localization of *miR164c* is limited by cell layer as well as zone.

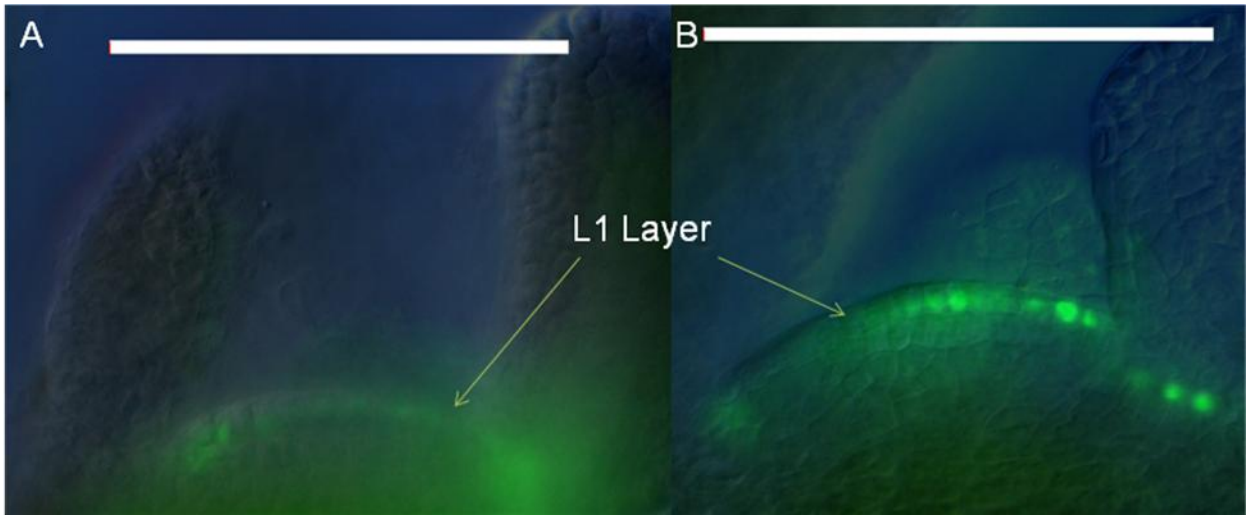


Figure 5.6 – Fluorescence microscopy images of miR164c promoter activity in the CZ.

A) pmiR164c:VENUS promoter activity (green) in the SAM of 5 day old pmiR164c:VENUS x S38 plants grown on GM agar imaged with a leica fluorescence microscope.

B) pmiR164c:VENUS promoter activity (green) in the SAM of 5 day old) pmiR164c:VENUS x S38 plants grown on GM agar with 60μM DEX imaged with a leica fluorescence microscope.

White bar indicates 1μm scale. L1 layer is marked with a white arrow. Images representative of at least n>3 plants.

In order to investigate how the over-expression of *STM* may affect the distribution of *miR164c* in the SAM, a long-term induction experiment was performed. From the images showing the expression of pmiR164c:VENUS in the SAM of S*38 x* p*miR164c:VENUS* plants grown on DEX (Figure 5.6) it did not appear that *miR164c* expression domain changes within the SAM. Its expression remains restricted to the L1 layer in both *STM*-induced and uninduced plants, but the intensity of GFP fluorescence appeared to increase, although this assay is not quantitative. However, as the SAM is such a small component of aerial plant tissue, any increase is unlikely to explain the large increase in the observed expression level of miR164c at the transcriptional level. Thus, I investigated how *STM* induction affected expression of *miR164c* outside the

SAM. VENUS expression is observed in the region surrounding the SAM in both uninduced and induced plants (Figure 5.7). As can be seen, outside the SAM, the domain in which signal is observed is far broader following induction of *STM,* with signal spreading into emerging leaves and the cotyledons.

The normal expression of miR164c is only seen only outside the boundary and within the L1 layer. These are tissues through which *STM* is believed to be able to move, and so this result is consistent with regulation of miR164c by *STM*. The over-expression seen in the region surrounding the SAM is also consistent with *STM* induction of *miR164c*; however it is not clear whether *STM* induction produces an increase in *miR164c* expression within the CZ itself as it may already be at saturation.

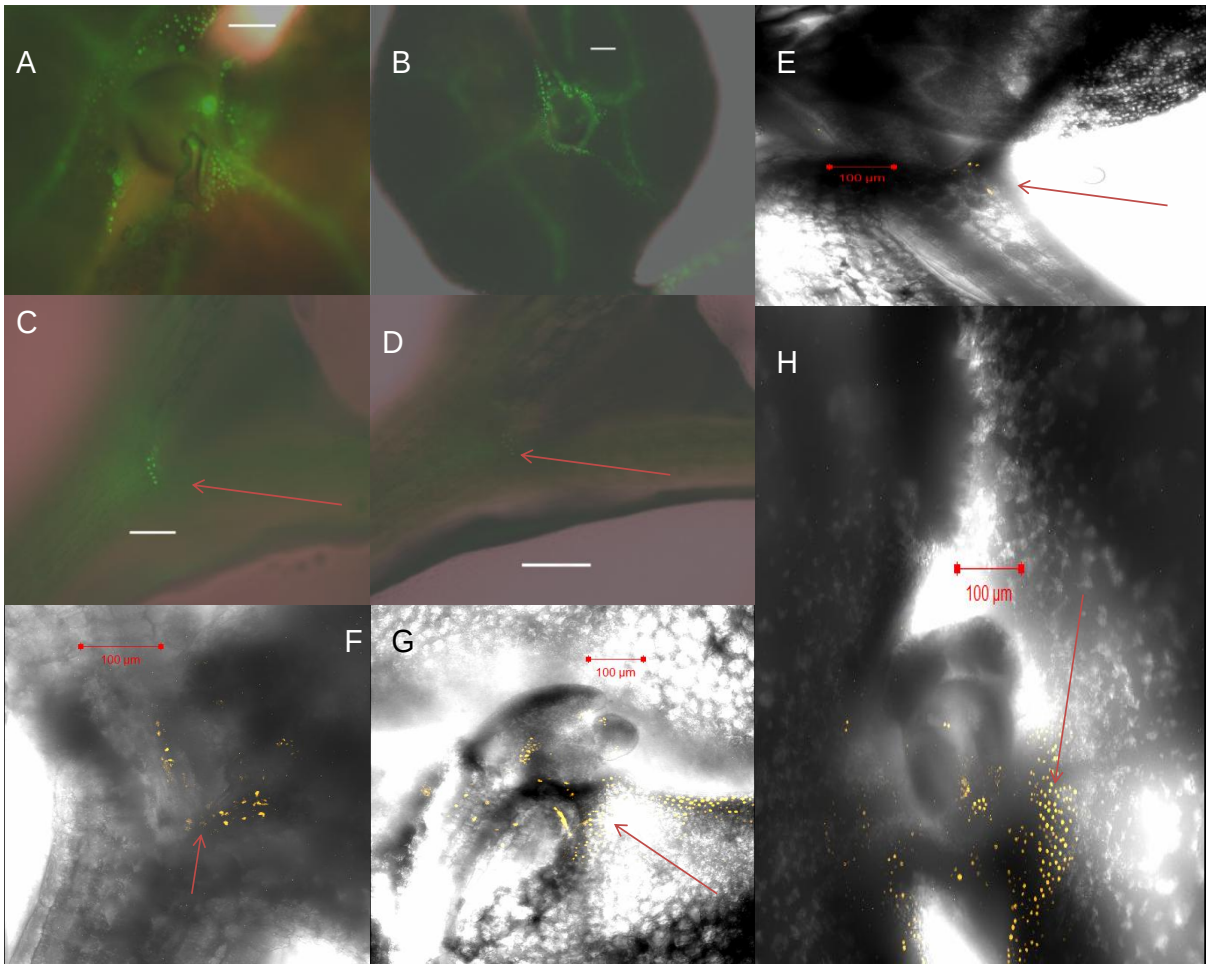


Figure 5.7 – Effects of *STM* induction on *miR164c.*
Expression of p*miR164c:VENUS* (Green) in induced (A,B) and uninduced (C,D) in S38 x pmiR164c:VENUS plants.
Expression of pmiR164c:VENUS (Yellow) in the region surrounding the SAM in uninduced (E), transiently –DEX induced (24 hours, F) and 2 projections of the region surrounding the SAM (G, H) in S38 x pmiR164c VENUS plants grown on DEX. All images representative of at least n>2 replicates.

The temporal dynamics of *STM*-mediated induction of pmiR164c:VENUS outside the SAM in plants grown on DEX was explored in greater depth using confocal microscopy. Z-stacks of the region surrounding the meristem were taken of p*miR164c:VENUS* x S[38] plants grown on DEX for differing lengths of time. As can be seen in Figure 5.7, these reveal that a similar region of cells outside of the SAM express VENUS in both wild type and short-term induced *STM* over expressing plants, with little change detected by 24

hours. However, for plants grown on DEX, a very broad range of cells in the vicinity of the SAM were expressing VENUS.

This suggests that since, after 24 hours of induction we do not observe a dramatic expansion of the p*miR164c* expression domain, much of the increase may be taking place within its native domain. As these images also represent a longer time-period than was examined by qRT-PCR, they strongly suggest that the upregulation of miR164c observed at 24 and 72 hours would be observed following longer induction periods.

5.2.4 *STM* induction of miR164a may be observable using fluorescent reporter lines, though no increase is observed at the transcriptional level

Spinelli et al (2011) observed upregulation of the miR164a promoter in leaves following *STM* induction, however this is at odds with the known data regarding the *TCP* genes directly inducing miR164a (Koyama et al 2010), and *STM*'s repression of the *TCP* genes (Chapter 3). It is also clear that *miR164a* is not present within the *STM* domain in the IM, and thus the only *STM* responsive induction of the miR164a promoter observed by the authors was in a region where *STM* is not normally expressed. Like the authors, upregulation of miR164a was not seen by qRT-PCR thus it was interesting to consider whether we could observe upregulation by *STM* of the miR164a promoter using crosses to the p*miR164a:VENUS* reporter, as these contradictory results could have been due to mischaracterization of their expression system or flawed experimental procedures/interpretation.

The effects of *STM* induction on p*miR164a:VENUS* plants crossed to the S38 *STM* inducible over expression line were examined. In uninduced plants at 5 days after germination, the strongest signal observed was in the tips of emerging leaves (Figure

5.8). However, in induced plants, signal could be detected at the base of the cotyledons in a similar pattern to the signal observed by Spinelli et al (2011) in the leaves of older plants. This suggests that upon *STM* induction, the miR164a promoter is activated at least in this zone, though since *STM* has a clear negative effect upon the expression of the *TCP* genes which are known to upregulate it, this suggests that miR164a is positively regulated by additional unknown factors which are responsible for mediating this effect upon *STM* induction. However, due to time constraints the lines examined were heterozygous for *STM* over-expression as such it was not possible to conclude absolutely in this generation of lines that *STM* was producing a consistent pattern of induction or that all plants were responsive.

Thus *STM* induction may have been observed to produce a change in the expression pattern of miR164a, using a more sensitive expression system than has been previously described. However, as no change in the expression of the mature microRNA for miR164a was detectable by qRT-PCR upon *STM* induction, all we can conclude is that the promoter appears to be more broadly expressed in induced lines than in the wild type. We cannot conclude that there is a change in the actual levels of miR164a, only that the activity of the promoter may be perturbed, although one possibility might be that some form of post transcriptional regulation is taking place.

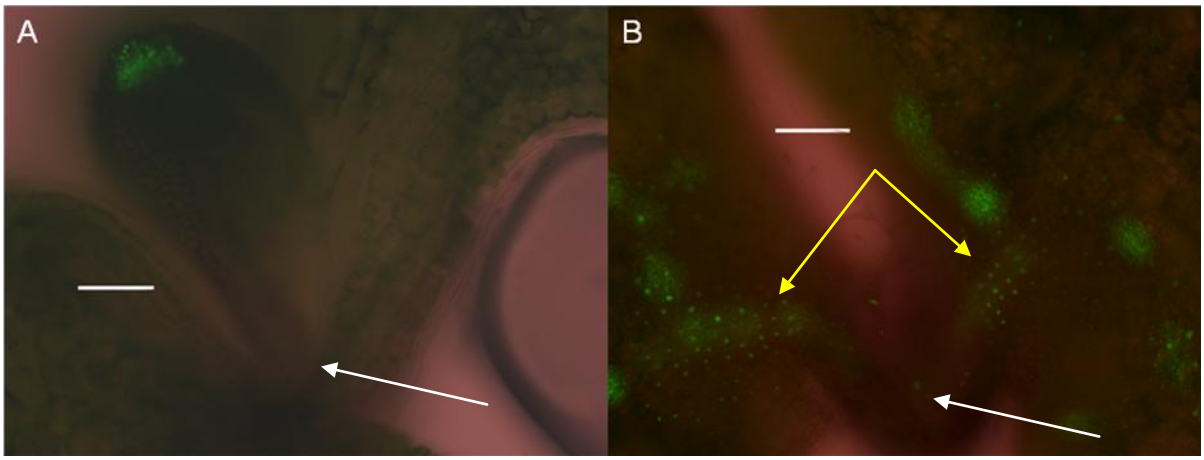


Figure 5.8 Effects of *STM* induction on *miR164a.*
Expression of pmiR164a:VENUS (Green) in uninduced (A) and responsive pmiR164a:VENUS x S38 plants grown on DEX (B) No VENUS expression is detected in the region containing the SAM (marked by an white arrow.) However, as can be seen in B VENUS expression is detectable in cotyledons whereas it is absent in the uninduced plants (yellow arrow). Scale bar indicates 100μm.

5.2.5 A Model of *STM* and *CUC1* behaviour including *miR164c*

To further explore the dynamics of *STM, miR164c* and *CUC1,* mathematical modelling approaches were explored. In Figure 9, I demonstrate the initial setup of a simplified ODE model in the situation in which the expression of *CUC1* is driving *STM* and *STM* is driving the expression of *CUC1*, via the following ordinary differential equations within a single compartment of size 1 ml:

Parameters have been selected so that degradation is occurring slower than mRNA synthesis each k=Θ=1 and each degradation rate term, γ is 0.5, thus leading to a slower rate of degradation than synthesis of both species and it is assumed to be taking place in a 1ml compartment. Unless degradation occurs rapidly enough that under initial conditions it is higher than production, such a model will rapidly drive both species to saturation.

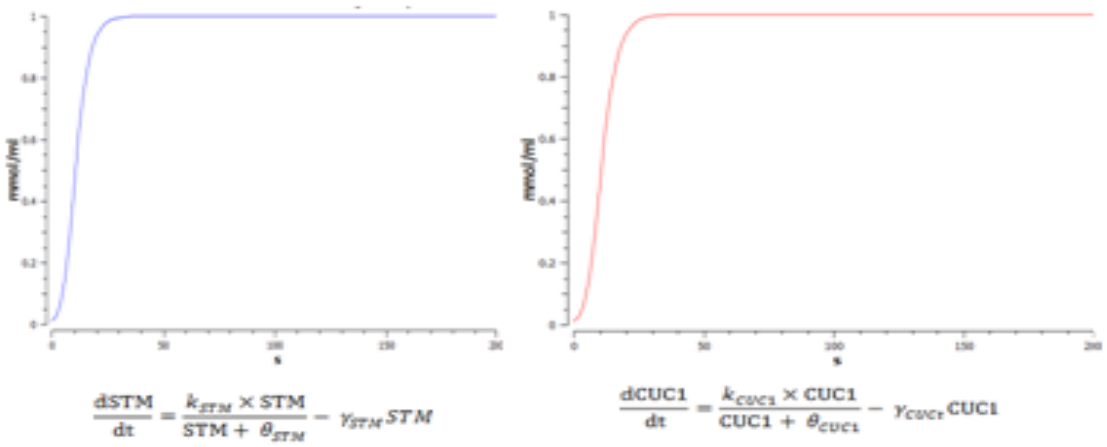


Figure 5.9 – Demonstration of ODE approach to model *CUC1* and *STM* in the absence of negative feedback.
Equations governing change in *STM and CUC1* over time and simulation of effects on *STM* (left) and *CUC1* (right) levels in a single compartment of volume 1. As can be seen with parameters selected such that degradation (RHS of each equation) is slower than induction (LHS of each equation), both species rapidly drive themselves to saturation

STM induction of *miR164c* should presumably have the effect of lowering the point at which *STM* and *CUC1* drive each other to saturation. However, we know from the literature that *CUC1* and *STM* do not share an expression domain, thus it is necessary to ask what is causing the symmetry of *CUC1* and *STM* expression dynamics between the CZ and emerging primordia to be broken. One possibility is *STM* movement, as *STM* has been shown to be capable of moving from the L1 to the L2/L3 layers of the meristem (Kim et al, 2003), this opens the possibility that *STM* may migrate from a site of production where *miR164c* is not expressed, to another zone where *miR164c* is expressed, thereby excluding *CUC1* from that region.

A

$$\frac{d([CUC1\_CZ] \cdot V_{CZ})}{dt} = -V_{CZ} \cdot (0.5 \cdot [CUC1\_CZ]) + V_{CZ} \cdot \left(\frac{1 \cdot [STM\_CZ]^1}{[STM\_CZ]^1 + 1^1} \right) - V_{CZ} \cdot (0.75 \cdot [miR164])$$

$$\frac{d([STM\_CZ] \cdot V_{CZ})}{dt} = -V_{CZ} \cdot (0.5 \cdot [STM\_CZ]) + (0.5 \cdot [STM\_Boundary]) + V_{CZ} \cdot \left(\frac{1 \cdot [CUC1\_CZ]^1}{[CUC1\_CZ]^1 + 1^1} \right)$$

$$\frac{d([miR164] \cdot V_{CZ})}{dt} = -V_{CZ} \cdot (0.5 \cdot [miR164]) + V_{CZ} \cdot \left(\frac{0.5 \cdot [STM\_CZ]}{[STM\_CZ] + 1} \right)$$

$$\frac{d([STM\_Boundary] \cdot V_{Boundary})}{dt} = -V_{Boundary} \cdot (0.5 \cdot [STM\_Boundary]) - (0.5 \cdot [STM\_Boundary]) + V_{Boundary} \cdot \left(\frac{1 \cdot [CUC1\_Boundary]^1}{[CUC1\_Boundary]^1 + 1^1} \right)$$

$$\frac{d([CUC1\_Boundary] \cdot V_{Boundary})}{dt} = +V_{Boundary} \cdot \left(\frac{1 \cdot [STM\_Boundary]^1}{[STM\_Boundary]^1 + 1^1} \right) - V_{Boundary} \cdot (0.5 \cdot [CUC1\_Boundary])$$

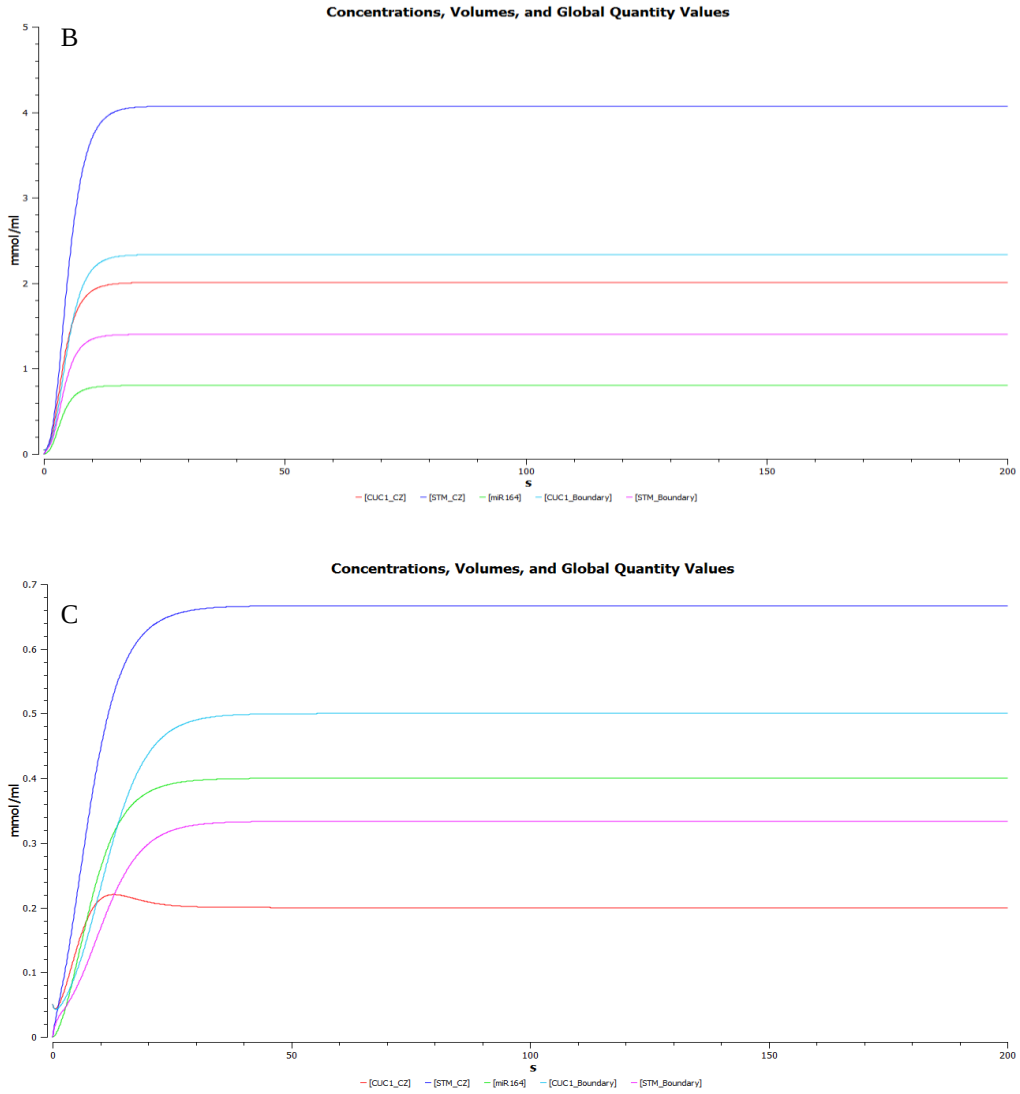


Figure 5.10 – 2 compartment model of *STM* and *CUC1* with negative feedback.

A) Equations governing changes in expression over time of all species in the model separated by CZ and boundary compartments. All parameters set to 1, except degradation terms and movement rate (all 0.5)
B) Predicted levels of CUC1 in CZ (red) STM in CZ (blue) miR164 in CZ (green) CUC1 in boundary (cyan) and STM in boundary (purple) for default parameters.
C) as above, but with lambda production terms doubled.

To test this hypothesis, I created a simple 2-compartment model, governed by the differential equations shown in Figure 5.10**.** These equations define *CUC1* in the central zone as being produced by *STM,* but degraded by miR164c along with a natural rate of decay. *STM* analogously is induced by *CUC1* but not degraded by an miRNA. *miR164c* is induced in the CZ by *STM* and degrades naturally. In the boundary, *CUC1* and *STM* both drive each other's expression and degrade naturally. A fixed proportion of *STM* moves from the boundary to the CZ each time unit.

With parameters selected to be as simple as possible (1 wherever possible, 0.5 where the rates of degradation and movement need to be lower in order to prevent species being completely eliminated), we obtain the steady state shown in Figure 5.10. In this model, *CUC1* is at much lower levels in the CZ than in the boundary, which is the primary site of *STM* production. The model is sensitive to the balance between movement/ degradation parameters, and production parameters which cause a species to be driven to extinction if degradation and movement occur too quickly, while levels of *CUC1* can become relatively high in the CZ if production terms predominate. The latter is demonstrated in Figure 5.13B.

This model predicts that *CUC1* and *STM* share an expression domain within the SAM. Many studies of *STM* expression have shown a relatively broad domain of expression within the SAM, strongly expressed in the central zone but excluded from organ primordia or emerging cotyledons (Long and Barton, 2000, Aida et al, 1999, Barton and Long 1998, Long et al, 1996). However, there is evidence from some fluorescent reporter lines that the *STM* promoter may be more strongly expressed in the boundary region (Laufs et al, 2004), than is the case for *STM* promoter fusions – though these still showed expression throughout the SAM with the exception of incipient primordia. *CUC1* transcript accumulation appears to be restricted to the boundary within the SAM

(Raman et al, 2007), though fluorescent reporters suggest a broader region of expression (Cary et al, 2002). Although there are less clear images available for *CUC1* in the literature than for *STM,* there is much more agreement about their expression domain. As can be seen in Figure 11, there is some evidence that *CUC1* and *STM* may have partially overlapping domains even in the mature SAM due to STM having stronger expression at the boundary zone.

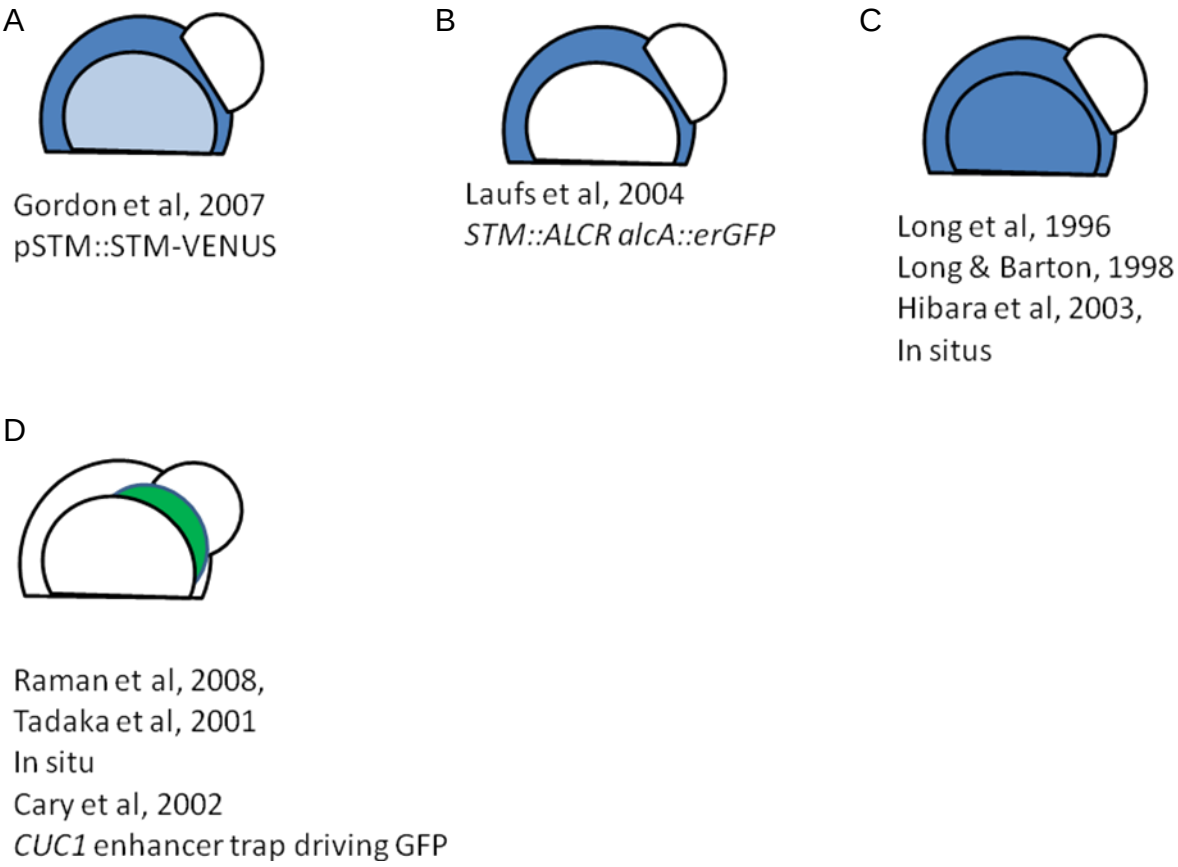


Figure 5.11 – Expression patterns of STM and CUC1 from the literature.

Observed expression patterns of *STM* (darker blue indicating stronger expression) in the A) CZ (outer circle), B) primordia (outgrowth) and C) rest of meristem (inner circle). Papers and lines in which patterns were observed are listed underneath
D) shows as above but for *CUC1* (green) - all papers describing *CUC1* suggest it is it limited to the boundary between the CZ and emerging primordia.

5.2.6 Stochastic Modelling leads us to speculate that the abort-retry aspect of *STM*'s mutant phenotype may be explainable through observed expression dynamics

The spatial localization of *STM* is not the only aspect of *STM* over-expression or mutation which may potentially be explained via *STM-CUC1* positive feedback. Using a simple stochastic model, which treats the time to a subsequent event as Poisson distributed, and relative probabilities of each event as a product of rate constants and concentrations of dependent species**,** I investigated the effects of reducing *STM*'s likelihood of binding to the *CUC1* promoter, as a means of simulating the effects of an *STM* mutation.

This simplified model was set up as shown in Figure 12 as a configuration which results in stable expression of *STM* and *CUC1* proteins given a small amount of initial *CUC1*. For this simplified model, *miR164c* expression and multiple compartments were not considered, just the double-feedback between *CUC1* and *STM*. As *CUC1* is expressed before *STM* during embryogenesis, a basal rate of *CUC1* expression is included.

Parameters were selected in the default model such that parameters governing the same process (such as the parameters governing protein translation) would all have the same value, and such that degradation terms were sufficient to produce a steady state but not too high that the no *STM* or *CUC1* protein production was predicted.

A

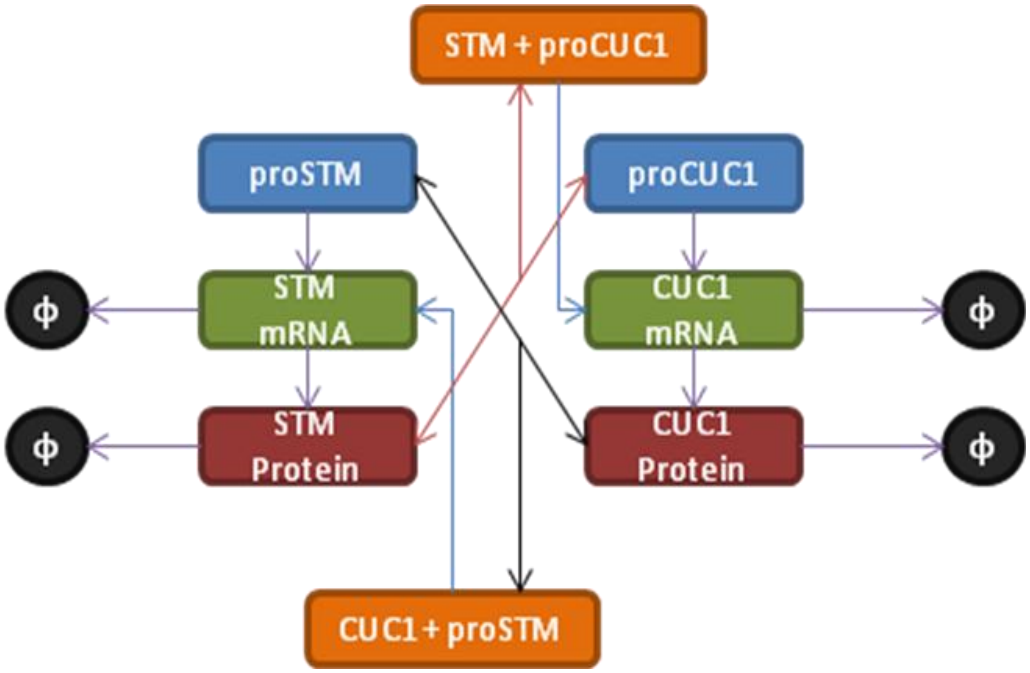


Figure 5.12 – *STM* stochastic model.

A - Schematic of relationships modelled. Φ represent degradation terms. Arrows indicate rates which are dependent upon source entities and affect the quantity of target arrows. Double headed arrows indicate reversible relationships (between protein and promoter where 1 protein and 1 promoter become a protein promoter complex or disassociate)
B - List of processes modelled, dependent quantities and rates in the default model.
C - List of initial quantities of all species present in the model.

B

Process	Rate	Depends on quantity of
Basal STM Expression	0	Nothing
Basal CUC1 Expression	0.0001	Nothing
Ectopic STM expression	0	Nothing
STM translation rate	0.5	STM mRNA
CUC1 translation rate	0.5	CUC1 mRNA
STM mRNA degradation	0.03	STM mRNA
CUC1 mRNA degradation	0.03	CUC1 mRNA
STM protein degradation	0.03	STM protein
CUC1 protein degradation	0.03	CUC1 protein
STM binding to CUC1 promoter	0.1	STM protein, CUC1 promoter
CUC1 binding to STM promoter	0.1	CUC1 promoter, STM protein
STM Disassociating from CUC1 promoter	0.05	STM-proCUC1 complexes
CUC1 Disassociating from STM promoter	0.05	CUC1-proSTM complexes
Induction of CUC1 by STM	0.3	STM protein
Induction of STM by CUC1	0.3	CUC1 protein

C

Species	Initial Quantities
CUC1 promoter	2
STM promoter	2
CUC1 mRNA	0
STM mRNA	0
CUC1 protein	300
STM protein	0
CUC1-proSTM complexes	0
STM-proCUC1 complexes	0

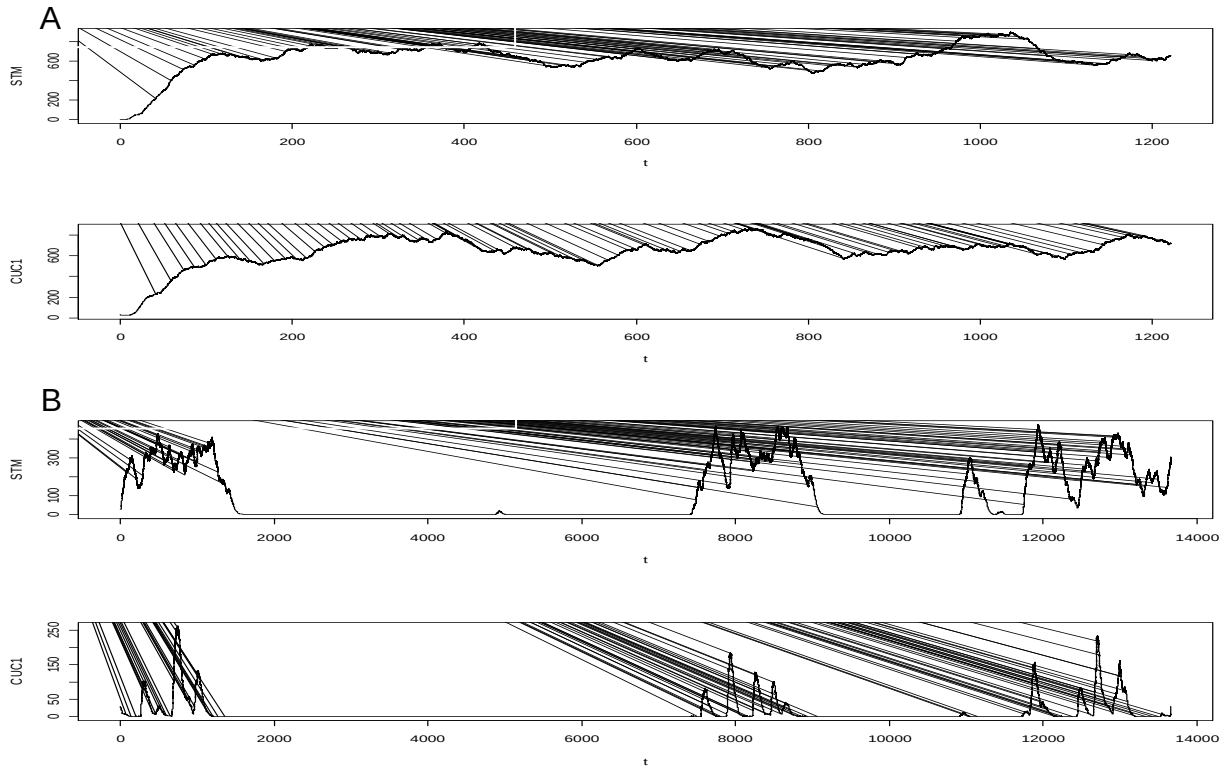


Figure 5.13 – Stochastic Model of *STM* functional impairment.

A) Predicted levels of *STM* (top) and *CUC1* (bottom) over time (horizontal axis) for default parameters of the stochastic model.

B) Predicted levels of *STM* (top) and *CUC1* (bottom) over time (horizontal axis) for stochastic model with *STM* protein binding to *CUC1* promoter reduced by 1,000-fold.

As can be seen, in Figure 13 reduction of the *STM* protein's ability to bind to the *CUC1* promoter sufficiently can cause periodic arrest of *STM* and *CUC1* expression. This is reminiscent of the phenotype observed in various *STM* knockdown lines, an example of which is shown in Figure 5.14, in which the SAM sometimes prematurely aborts before reinitiating. This model allows us to speculate that the random termination of the meristem may be the result of *STM* being less able to bind to the *CUC1* promoter, thus leading to occasional fluctuations in *CUC1* levels.

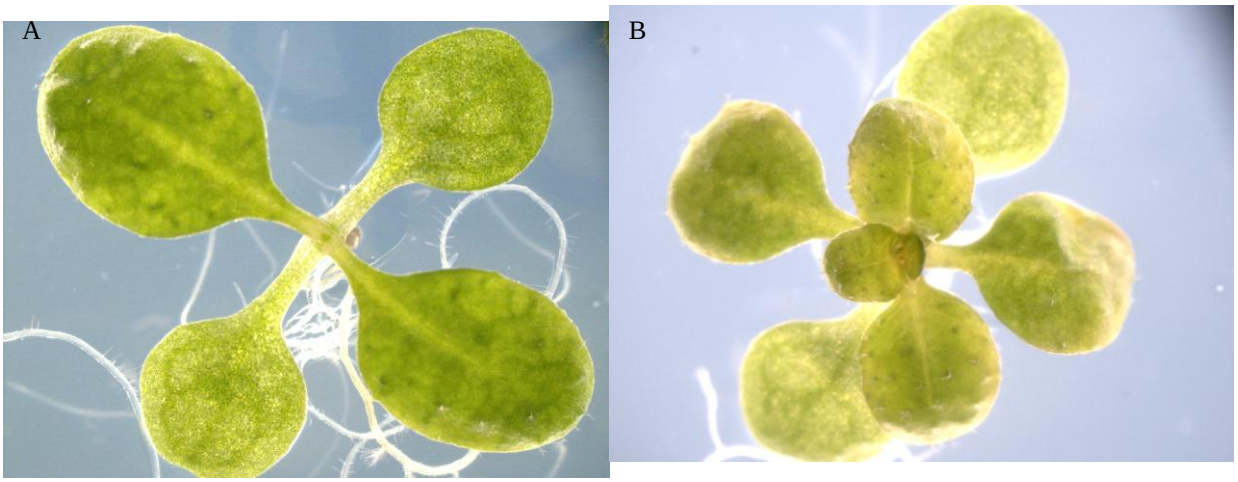


Figure 5.14 –Aberrant phyllotaxis and meristematic abort-retry in lines with STM knockdown or mutation.

A) Arabidopsis with constitutive knockdown of *STM* via RNAi (SP1-1) showing a meristem arrest phenotype (Scofield et al., 2008) grown for 15 days on GM Agar. B) Wild Type Landsberg *erecta* grown under same conditions.

5.3 Discussion

A positive feedback loop between *STM* and *CUC1* is proposed and shown that it may be disrupted through miR164c, as shown by Sieber et al (2007), since disruption of

miR164-mediated *CUC1* degradation is sufficient to allow *CUC1* enter *STM*'s domain. The work reported here shows through a combination of reporter analysis and gene expression analysis that *STM* promotes the expression of miR164c, though it does not do so directly. This is a novel finding as it has only been previously shown by Sieber et al (2007) that *CUC1* expression in the centre of the IM depends upon the activity of *miR164c* and these authors did not demonstrate the transcriptional events leading to activity of *miR164c.*

Here I have shown that *CUC1* expression is thus prevented in the CZ by the *STM*-mediated upregulation of miR164c. However, it is also prevented from entering emerging primordia. While presumably this is partially mediated by miR164a, microRNA resistant *CUC1* was not detected in primordia (Sieber et al, 2007), thus it is clear that miR164a is not sufficient for suppressing *CUC1* expression in primordia. Despite this, though induction of miR164a could not be shown by reporter studies following *STM* induction, expression of p*miR164a:VENUS* was sometimes observed in the proximal region of leaf primordia, where p*CUC1:CUC1-GUS* expression could not be detected following *STM* induction. This suggests that if miR164a is actually being induced by *STM* then it is sufficient but not necessary for *CUC1* repression.

We have shown by modelling that it is sufficient that *STM* shares a partially overlapping domain with *CUC1* and is able to move, in order to break what would otherwise be a uniform distribution of *STM* in the SAM. This results in the expected higher expression of *STM* in the CZ than in the boundary. Microscopy of *STM* and *CUC1* reporter lines from previous studies suggest a possible areas of overlap in the boundary, and data in Yadav et al (2009) suggests that *CUC1* should be classified in the same manner as *STM* or that *STM* encompasses the *CUC1* expression domain. Additionally there is evidence in the literature supporting *STM* movement (Kim et al, 2003, Jackson D, 2002)

as such these assumptions are both plausible. However it is clear that further detailed studies on reporter crosses for *STM* and *CUC1* will be needed in order to validate that their expression domains do coincide, and that *STM* does traffic in the SAM in the manner required by the model.

Finally, I have speculated that the stochastic nature of the abort-retry phenotype in moderate strength *STM* mutants and in the *STM* RNAi, may be due to the strength of the positive feedback loop between *STM* and *CUC1*. A severe reduction in the ability of *STM* to bind to the *CUC1* promoter is required to produce modest downregulation of *STM* – this is consistent with both the stochastic and deterministic models – and suggests that only infrequently will sufficient reduction of *STM* levels occur for stem cells to be consumed into organ primordia. Even under these circumstances, a small induction of *CUC1* or *STM* can trigger a rapid recovery of both proteins, thus initiating a retry phase.

Chapter 6 - Discussion

6.1 *STM* as a core regulator of the SAM

As a core regulator of stem cell fate in Arabidopsis, *STM* clearly induces a large number of phenotypic changes when its expression is perturbed (Gallois et al, 2002, Brand et al., 2002, Lenhard et al, 2002; Scofield et al. 2007; 2013). As a consequence, the exact mode of function of *STM* has proven difficult to pin down in previous studies, which have mostly endeavoured to focus on specific roles of *STM*'s genetic interactions with specific genes such as *AS1, WUS* or *CUC1*, (e.g. Endrizzi et a., 1996; Tadaka et al, 2001; Byrne et al, 2002; Gallois et al, 2002). The microarray study performed by Spinelli et al. (2011) was hampered by low power, which has been further constrained by the authors' subsequent choices of statistical analyses restricting potential downstream targets to only those detected in the lowest power experiment.

I have demonstrated that through the use of time course gene expression data it is possible to get a broad picture of *STM*'s GRN, and to place the phenotypic data in a temporal context. This has led to novel insights about the sequence of gene expression changes that lead to the *STM* overexpression and knockout/knockdown phenotypes, several of which have been validated by direct target experiments guided by data mining.

6.1.1 Boundary specification and regionalization

In particular, it is clear from both the time course data and from the identified direct targets that one of *STM*'s most immediate effects is to up-regulate a number of genes involved in boundary specification and regionalization. Of the confirmed TF direct targets, *CUC1* is a member of a family (NAC) often used to define boundary zones in

reporter analyses (Aida et al, 1997), *BOP2* is involved in specifying the polarity of emerging primordia (Ha et al, 2007), and *AIL7* regulates phyllotaxis through *PIN* proteins (Prasad et al, 2011). Thus, a large number of known direct target TFs are in some way related to boundary specification and other aspects of meristem organisation. Moreover this role of *STM* appears to have a large degree of redundancy, as all 3 of these direct targets have closely related homologues (*CUC3, BOP1* and *AIL6*) which I have also shown to be upregulated (indirectly) via *STM* (Figure 6.1), although for *AIL6* and *BOP1* the upregulation required a longer period of time to be statistically significant: 72h for *AIL6*, and 24h for *BOP1*.

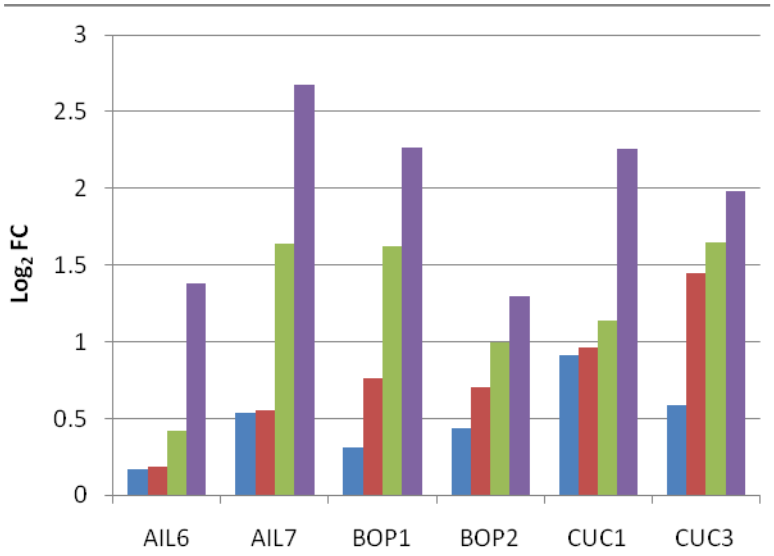


Figure 6.1 Expression of CUC1, *CUC3, BOP1, BOP2, AIL7 and AIL6.*

Observed Limma-derived point estimates of $\log_2$ fold changes for STM direct targets with redundant homologues also significantly differentially expressed over the *STM* over-expression time course at 8h (Blue), 24h (Red), 72h (Green) and 9d(Purple). *AIL6* is not statistically significantly differentially expressed at 8 or 24h. *BOP1* is not statistically significantly differentially expressed at 8h.

It seems somewhat counterintuitive that *STM* should have such a strong association with boundary specification, given that it is excluded from emerging primordia (Heisler et al, 2005), and the effect of over-expressing *STM* ectopically appears to include

inducing tissue to maintain or enter a meristematic state (Lenhard et al, 2002, Brand et al, 2002, Gallois et al, 2002). However, as discussed in Chapter 1, there is some suggestion that expression of its promoter may be stronger around the boundary. As such the ODE model described in Chapter 5 suggests that this discrepancy between *STM* promoter activity and protein expression can be explained if the boundary is the primary site of *STM* production, as the STM protein movement appears to be sufficient to produce this spatial arrangement of *STM* as discussed and shown previously in Figure 5.11 (Long et al, 1996; Long and Barton, 1998; Kim et al, 2002; HIbara et al, 2003; Laufs et al, 2004; Gordon et al, 2007).

The proposed higher levels of *STM* expression in the boundary zone would be intimately linked to the direct positive feedback loop between *CUC1* and *STM.* Without sufficient negative feedback, this should lead both proteins to saturation in any cells where they are present. As had been shown via modelling, this positive feedback, combined with *STM* movement (Kim et al, 2002), and the indirect promotion of *miR164c* expression by *STM* shown in Chapter 5 (and discussed below) is sufficient to create a pattern of expression where the boundary zone operates as a primary *STM* factory, producing high levels of *STM* throughout the CZ, while restricting *CUC1* to the boundary. *STM* thus may be argued to play as important a role in boundary specification as *CUC1*, though it does so in a non-cell autonomous manner, reminiscent of feedback from *WUS* upon the CZ (Groβhardt and Laux, 2003).

6.1.2 *STM* and *CUC1* rapid mutual induction is broken by miR164c

A microRNA is an ideal means to disrupt a strong positive feedback loop, since it inhibits the translation of the proteins which cause the loop. Repressing a loop via post-transcriptional regulation would have to over-come the high rate of production such a positive feedback loop entails. However, transcript level regulation enables the

feedback loop to be broken before translation can occur. As was shown by Sieber et al (2007), microRNA resistant *CUC1* is able to move into the centre of the L1 layer of the SAM, showing that normally miR164 expression is sufficient to prevent *CUC1* expression in this region. One suggestion is that following its initial induction early during embryogenesis, further *CUC1* expression is *STM*-dependent and as *STM* is absent from the miR164a expression domain, *CUC1* remains restricted, though may be expand or remain expressed in organ primordia (Sieber et al, 2007).

The phenotypic significance of *CUC1* mis-expression in the SAM is unclear. The *EARLY EXTRA PETALS1* (*eep1)* mutant, which is caused by mutation of *miR164c*, displays floral phyllotactic aberrations (notably, additional petals in early flowers), however the plants are able to grow normally otherwise (Baker et al, 2005), and there were additional phyllotactic defects in plants expressing microRNA resistant *CUC1* – notably accessory side shoots formed in cauline leaf axils (Raman et al, 2008) . This demonstrates that *CUC1* mis-expression does have a phenotypic effect, however it is also clear that *CUC1* mis-expression inside the SAM is far from lethal, although outside the SAM, Hibara et al, (2003) showed that *CUC1* can trigger adventitious SAM formation reminiscent of ectopic *STM* expression. Correct phyllotaxis is clearly a trait likely to be selected for, and understanding how *STM* and *CUC1* combine to create a dynamic boundary region that allows the specification of proper phyllotaxis may be important in determining how we can better engineer useful phenotypes.

Additional validation of the importance of correct *CUC1* positioning in maintaining phyllotaxis comes from outside of the vegetative SAM. During flowering, it is also known that *CUC1* regulation by miR164c via the transcription factor *RABBIT EARS* (*RBE* – which has been shown via ChIP to directly repress miR164c expression) is essential for maintaining proper floral organogenesis (Huang et al, 2012). Not only are the fusion

events in sepals of the single *rbe* mutant reminiscent of the *cuc1 cuc2* double mutant (Krizek et al, 2006), but the *eep1* mutation rescues this sepal fusion phenotype (Baker et al, 2005; Huang et al, 2012). Thus, *CUC1* regulation by *miR164c* is clearly important for maintaining correct boundary formation.

6.1.3 The Boundary Zone is a dynamic region in the SAM

One core conclusion is that the boundary zone should not be thought of as a static barrier between the SAM and organ primordia. The interplay between *CUC1, STM* and *miR164* suggests that it is in fact a dynamic region of the SAM. Genetically, the continued expression of *CUC1* within this zone makes it comparable to the proto-meristematic region at the apex of the developing embryo which begins to express *STM* before *miR164c* expression excludes *CUC1* from all other regions (Aida et al, 1999; Sieber et al, 2007). By acting as a source of *STM* production, the boundary zone participates in defining the transcriptome of the rest of the SAM and ensuring that production of *STM* is high enough that non-primordium identity is resistant to perturbations.

However, the primordium side of the boundary is clearly under separate regulation to the CZ side as *STM* is not present on this side of the boundary because *STM* is unable to move to the primordium side of the boundary (possibly due to symplastic isolation as *STM* was not observed moving out of the meristem in the Kim et al (2002) experiment), thus preventing the initiation of *STM*-mediated *CUC1* expression. However we know this would only limit *CUC1* expression if the unknown factor which drives its expression in the embryo has been disabled (Aida et al, 1997), since *CUC1* expression is clearly not completely dependent upon *STM* expression.

In this light is interesting to note the repression of *TCP* genes by *STM*. *TCP3* is a known promoter of *miR164a* and thus a strong candidate for inducing its expression on the primordium side of the boundary, especially as work in Chapter 3 has shown that it and its close homologues *TCP4* and *TCP10* are all downregulated by *STM* and excluded from its domain (Koyama et al, 2010). Interestingly, in some *STM* induced plants, *miR164a* expression expands into the proximal region of leaf primordia, even though *STM* is presumably downregulating multiple *TCP* genes. Confounding the issue, a change in the expression level of *miR164a* could not be detected by qRT-PCR either in this study or by Spinelli et al (2012). This could be due to the change only occurring in a minority of plants, and thus being undetectable in pooled seedling tissue, or possibly due to domination of *STM* and *TCP* responses in different tissues resulting in no net detectable change in *miR164a* expression

We do not know what is repressing *TCP* genes in STM's GRN. The repression is broad and is not part of *STM*'s initial transcriptional response, and an interesting avenue of further study would be to investigate the downstream targets leading to *TCP* repression. The *WRKY35* gene and Ran BP2/NZF zinc finger-like superfamily member TF At5g25490 are putative parents of *TCP* genes on the network and would be an interesting place to begin investigation.

6.1.4 Organogenesis through the lens of a dynamic boundary zone

Organogenesis is a complex process, involving the interplay between numerous genes and hormones (Carraro et al, 2006). Given the evidence we have presented that the boundary is a more dynamic zone than previously envisioned, we now need to ask how this affects our knowledge of primordium initiation.

New primordia form at auxin maxima around the periphery of the SAM with regular spiral phyllotaxis (in wild type plants.) One of the earliest events following the formation of this auxin maxima has been shown to be down-regulation of *STM* (Long et al, 1996; Long & Barton, 1998; Heisler et al, 2005) in the area coincident with an auxin maxima as deduced by *PIN1* localization. Given our model of *STM* and *CUC1* interactions, we would predict this would lead to a concomitant down-regulation of *CUC1* in those cells unless it is driven sufficiently high by *STM*-independent *CUC1* expression (which we know to occur as *CUC1* precedes *STM* during embryogenesis; Aida et al, 1997). Additionally Guo et al (2005) have shown that auxin induces the expression of *miR164a*, though these authors did not subsequently show that this affects *CUC1*. Their work would suggest that at the site of primordium initiation, we may expect up-regulation of *miR164a* through auxin and thus rapid down-regulation of *CUC1*, with the result that the boundary between the organ under formation and the SAM would be disrupted thus preventing *STM* induction by *CUC1*.

Thus interplay between *STM* and *CUC1* could be the key to creating proper boundary formation during organogenesis in the SAM. This is important as both the emerging primordia and the rest of the SAM require different cellular characteristics for proper function, such as different rates of cell elongation and mitotic cyclings (Grandjean et al, 2004). The specification of each into separate transcriptional regions is an essential part of organogenesis and may explain the phyllotactic defects observed in the *eep1* mutant and *miR164* resistant *CUC1* lines described by Baker et al (2005) and Raman et al (2008), since essential genes may be mis-expressed or inappropriate cell division rates may cause cells to be misaligned during development.

6.1.5 STM directly regulates organ polarity

We have shown *BOP2* to be a direct target of STM, which along with the indirectly upregulated *BOP1* may play a role in organogenesis as the *BOP* genes regulate expression of leaf polarity along both axes of developing primordia (Hepworth et al, 2005). This is in direct contradiction to Jun et al (2010) who argue that *STM* represses *BOP2* – however, they based their argument around genetic studies – that in the absence of BOP proteins, in the strong *stm-11* mutant a new shoot meristem was able to form. There are however numerous possibilities to explain why this may be the case without *STM* repressing *BOP* genes. One possibility is that, since other *KNOX* genes can substitute for *STM,* in the absence of *BOP* genes which Jun et al (2010) showed activates *AS2* which in turn represses other KNOX genes, these may become ectopically expressed and prove sufficient to reinitiate a new meristem in place of *STM*.

Man Ha et al (2010) have also shown that *bop1 bop2* mutants depend upon *KNAT1, 2* and *6* for their phenotype (to varying extents). They also observed that in *bop1-1* plants mis-expression of *KNAT1*, *2* and *6* genes is indeed observed in the boundary zone. It may thus be that the BOP genes are necessary to restrict KNOX gene expression in the boundary zone, by up-regulating *AS2* and therefore repressing *KNOX* gene expression in leaves.

It has previously been argued (Byrne et al, 2000; Byrne et al, 2002) that STM directly represses *AS1* and *AS2* which in turn represses *KNAT1* and *KNAT2* - however, as BOP genes are known to promote *AS1, 2* and we observe slight up-regulation of *AS1* (ranging from 0.34-fold at 8h, to 0.49-fold at 72 hours, though always upregulated between these ranges) following STM induction, this model is clearly unsatisfactory. Rather the data support the model argued for in Scofield & Murray (2006) in which *STM*

and *AS1, 2* competitively regulate the expression of *KNAT1, 2,* although it also suggests that *STM* indirectly promotes *AS1*, albeit to a low level, via the *BOP* genes.

6.1.6 *STM* contributes to the positioning of auxin maxima by regulating the expression of *AIL7*

STM directly promotes the expression of *AIL7* (which shares an overlapping expression domain with *STM* in the centre of the SAM) and which knockout studies have shown is involved in the correct spatial localization of *PIN1*, although only in combination with other *ail* mutants (Prasad et al, 2011). *PIN* proteins control polar auxin transport, and in triple mutants of *AIL7, 6* and 5, fail to develop localized *PIN*1 at putative auxin maxima as such their joint knockout perturbs the correct formation of new organ primordia, causing new organs to be frequently distributed at 180^{o} angles rather than the wild type 137.5^{o} (Pinon et al, 2013).

However, there is a large degree of redundancy among genes in this clade (since all three needed to be knocked out to obtain a phenotype), and *STM* is acting in an inductive manner upon *AIL7.* A recent study by Pinon et al (2013) suggests that the *AIL* genes may up-regulate auxin biosynthesis via promotion of *YUC1* and *YUC4* genes. Interestingly, as we showed in Chapter 3, these are the only *YUCCA* genes observed to be upregulated by *STM* (albeit only later on in the time course; other perturbed *YUCCA* genes are downregulated). However the Pinon et al (2013) study showed this up-regulation using *AIL5* rather than *AIL7*, and the up-regulation was observable in our study by *STM* was only in the 9 day over-expressor lines where pleiotropic effects could confuse the direct effects of *STM* modulating *AIL7* expression to a greater degree than earlier time points.

Thus, through promotion of *AIL7*, *STM* might play a critical role in establishing correct auxin flows in the SAM. This gives *STM* a key role in ensuring correct positioning of new organs as well as boundary maintenance and organ polarity specification, however, further research is needed to discover both how *AIL7* affects phyllotaxis and what components of *STM*'s phenotype are mediated by *AIL* genes. As *STM* is excluded from the sites of organ formation in the SAM, which occur at auxin maxima, this raises the possibility that *STM* regulates its own expression domain via modulating auxin flows (or that auxin maxima might form following *STM* depletion.

6.1.7 The non-boundary zone specific consequences of *STM* expression

Given the predominance of *STM* direct target TFs in boundary zone maintenance, it is arguable that this is the most important mode of *STM* function. However, it is clear that *STM* has a more broad effect upon correct SAM formation, as several other processes are disrupted upon perturbation of *STM* expression. In particular, we have observed strong overlap with hormone datasets, and a clear effect is observed on the cell cycle. Given the identification of subsets of important genes regulating these processes in *STM*'s GRN we can proceed to investigate how these processes may be indirectly regulated by *STM*.

6.1.8 *STM* induces a broader CK response than reported in the existing literature

Our time course data for CK responses is consistent with the data presented by other groups. Jasinski et al (2005) identified up-regulation of around 2-fold of *IPT5, 7* and *ARR5* following 24 hours of induction in a 35S::*STM-GR* inducible system, with no induction of *IPT5* after 10 days grown on DEX media, but between 3 and 8 fold induction of *IPT7* and around 3 fold induction of *ARR5* (log_2 scale). Yanai et al (2005) observed increase of *IPT7* after 2 and 24 hours in the same line. We observed

statistically significant induction of *IPT5* and *IPT7* at 9 days, as well as strong and increasing induction of *ARR5* throughout the time course which is statistically significant from 72 hours. While the response we observed is slower than Jasinski et al (2005)or Yanai et al, (2005), the dynamics observed in these two studies do not agree with each other completely either – suggesting there is some stochasticity in the speed of the response, a result backed up by Scofield et al (2013).

One possibility is that the TGV expression system used in the study reported here is a two component system which will take longer to take increase levels of *STM* as the protein has to be transcribed and translated, as opposed to a GR based system where DEX triggers an immediate increase in *STM* levels in the nucleus through immediate translocation. However, the results of Yanai et al (2005) were not exactly reproduced with the *35S::STM-GR* line, who using the same *35S::STM-GR* system were able to observe up-regulation of *IPT7* after 2 hours. We did not observe any statistically significant fold change of *IPT7* after 3 hours of DEX induction in *35S::STM-GR* lines. They also observed induction of *ARR5*, by approximately 2 fold, and we did not observe statistically significant induction of *ARR5* after 3 hours. Again, this may be due to experimental differences or differences in statistical power or analysis method. They used two week old seedlings, whereas my experiments were performed on 9 DAG seedlings. Thus the plants used were at different developmental stages.

Despite the differences, we observe a clear CK response, which is broader than that predicted in the literature. *IPT3* was the strongest responding *IPT* gene in our analysis, however, CK biosynthesis was not the only CK-related category perturbed by *STM* expression. A broad response was observed among *A-TYPE ARR* genes. *ARR5* and *ARR6* showed consistent and steady up-regulation across the OE time course. *ARR7* and *ARR15* showed early up-regulation before falling back to wild type levels, *ARR9*

showed late up-regulation. All of this suggests that while different A-Type ARRs were responsive at different points in the time course, some CK feedback from this gene family was observed from 24 hours onwards. As Yanai et al (2005) and Jasinski et al (2005) produced partially conflicting data, it may be that while there is a CK response induced by *STM*. It is not always induced through the same genes.

CK catabolism was also downregulated by *STM*. From 72 hours, the cytokinin oxidases *CKX1* and *CKX6* were both downregulated. However, the most rapid response to *STM* among core CK response or biosynthesis genes was observed from *WOL* - which is the principal CK histidine kinase receptor required for transduction across the plasma membrane of CK response signalling (Mähönen et al, 2000; Inoue et al, 2001).

Thus, we suggest that although the specific genes responding at specific times appear to vary, *STM* exerts a clear positive effect upon both CK biosynthesis and signalling, while also repressing CK degradation. In particular the variability of expression among *ARR* and *IPT* genes suggests that while the specific genes upregulated may vary according to developmental stage and induction time, we consistently see some positive induction of CK. This may explain some of the specific differences with (Yanai et al, 2005) in particular, although our data strongly concurs on the broader detail of CK induction by *STM*.

6.1.9 Hormone responses to *KNOTTED1-LIKE HOMEOBOX* genes may have evolved differently between Arabidopsis and maize

(Bolduc et al (2012) identified a number of core hormone responses directly regulated by *STM*'s ortholog *KN1* in maize. In almost all cases, we have detected differences between the hormone response in Arabidopsis to *STM* perturbation and their results. The most striking difference is that in maize, much of the hormone regulation appears

to be performed directly by *KN1*, whereas in Arabidopsis, we could not identify any evidence of direct targeting of hormonal regulation TFs by *STM*.

As direct regulation of *GA20OXIDASE* genes is not only known in maize, but also in tobacco (Sakamoto et al, 2001), it was surprising to find that there were no rapidly responding *GA20OXIDASES* to *STM* induction in Arabidopsis. *GA20OXIDASE1*, the only downregulated *GA20OXIDASE* downstream of Arabidopsis, is not affected until 9 days. This is inconsistent with direct repression of *GA20OXIDASE* genes by *STM*. Importantly, however, we have shown that *OFP1* is a directly induced target of *STM*, and this is known to directly repress *GA20OXIDASE1* (Wang et al, 2007). Thus, it appears that *STM* has diverged evolutionarily from several other plant species, and indirectly regulates *GA20OXIDASE1* through *OFP1* in Arabidopsis.

Perhaps more striking, given the number of *AUX/IAA* genes identified as direct targets of *KN1* by Bolduc et al (2012) but the total absence of any for *STM*, the response of *AUX/IAA* genes is very different to *STM* than *KN1,* as these genes are only statistically significantly differentially expressed in the *STM* OE at later time points. And while we have argued that the long-term induced lines are not necessarily comparable with the earlier time points, the only other time point where multiple AUX/IAA genes are significantly differentially expressed is 72 hours. At this point, three are upregulated and three downregulated. Only *PAP1*/*IAA18* shows a consistent up-regulation following *STM* induction, and it is only upregulated from 24 hours. Given this, and the identification of no *AUX/IAA* genes in the shortlisted direct target genes in the direct target microarray experiment, it does not appear likely that any of *STM*'s auxin response is directly mediated through *AUX/IAA* genes unlike in maize. As has been discussed earlier, the clearest candidate for mediating *STM*'s auxin response is *AIL7* - a direct target of *STM* - which regulates phyllotaxis via promotion of *PIN1* and *YUCCA*

genes (Prasad et al, 2011; Pinon et al, 2013). It is also possible that IAA expression may be below level of sensitivity in our experiments, as IAA6 is known to be rapidly repressed by *STM* (Scofield S, personal communication.)

Given the differences observed for GA and Auxin regulation between Arabidopsis and other plants, it is interesting that CK regulation appears to have a greater degree of consistency. In particular, we have shown that in Arabidopsis as in maize, *WOL* and its orthologue respectively respond rapidly to *STM* or *KN1* (Bolduc et al, 2012). As mentioned previously, *WOL* is a receptor responsible for transmitting CK responses across the plasma membrane (Inoue et al, 2001), and its early regulation across both species suggests that it is an important component of *STM* and *KN1* GRNs even if it is not a direct target of *STM*.

The overall hormone response observed following *STM* induction is consistent with what we would have expected from the literature. Thus, although we do not show the same direct targeting of hormone regulation, *STM* does clearly strongly impinge on phytohormone levels in a similar manner to reported for *KN1* in maize. We must this conclude that unlike in maize, hormones are not directly regulated by *STM*, and instead the response is mediated through intermediary proteins, unless your experiment lacks the necessary power to detect true changes (which, given its success in generating validated conclusions regarding *STM*'s GRN seems less likely). This raises the interesting possibility of disentangling *STM*'s hormone response pathways via crossing *STM* over-expressing lines to *AIL7, OFP1* and possibly *WOL* mutants. In each case we could subsequently establish which portions of *STM*'s downstream GRN were likely to be regulated by auxin patterning or GA repression respectively, although as both genes are TFs, we would expect other components of *STM*'s GRN to be perturbed as well.

6.1.10 *STM* and the *WUS-CLV* Loop

The regulatory relationship between *STM* and *WUS-CLV* has been unclear. Genetic studies have shown that the two are genetically independent (expression of neither gene is capable of rescuing the other mutant, both promote distinct subsets of genes and gain-of-function phenotypes are independent (Endrizzi et al, 1996; Lenhard et al, 2002). However, as lack of *STM* results in meristem termination, and over-expression of both genes in mature tissue produces synergistic effects (Lenhard et al, 2002; Gallois et al, 2002), they appear to be interconnected. How these two core regulatory loops are related has thus remained unclear. We have confirmed by comparing the GRN regulated by *STM* against the published downstream GRNs of *WUS* (Liebfried at al, 2005; Busch et al, 2010) that there is very little overlap between the two datasets. The overlap observed with the Busch et al (2010) dataset is likely to have occurred due to random chance, although the overlap with the dataset from Liebfried et al (2005) and the later time points on the *STM* OE time course is statistically significantly unlikely to have occurred by chance.

Intriguingly, I have shown that one mode of interaction between *STM* and *WUS-CLV* is through the CK pathway. I have shown that *STM* exerts a broad up-regulation of CK biosynthesis and signalling. This includes the up-regulation of several *A-TYPE ARR* genes, which are directly repressed by *WUS* (Liebfried et al, 2005). Importantly, these genes provide negative feedback to the CK response, and are induced by CK levels.

Thus *STM* promotes an increase in CK levels, whereas *WUS* dampens negative feedback to CK responses directly by repressing *ARRs* (Liebfried et al, 2005; Busch et al, 2010). This synergistic interaction towards CK would result in higher CK levels in the SAM, coupled with an enhancement of CK activity in the *WUS* domain. It also would assist in producing a difference in CK levels through different regions of the SAM, as

CK negative feedback would only be inhibited within the *WUS* domain. However, we have previously shown that while *STM* promotion of CK is constant, the exact subset of promoted genes varies between time points and developmental stages. This contrasts with *WUS* which appears to target specific *ARRs* (*ARR5, ARR6, ARR7* and ARR15) and hence repress negative feedback of the CK response (Liebfried et al, 2005; Busch et al, 2010).

6.1.11 *STM* and the cell cycle

Although genes which regulate progression through the cell cycle and cell growth inhibition are not directly targeted by *STM,* phenotypically the cell cycle can be clearly observed to be perturbed by STM over-expression, as evidenced by inhibited differentiation and cell growth, along with an inhibition of endoreduplication (Lenhard et al, 2002; Brand et al, 2002; Gallois et al, 2002; Scofield et al, 2013). This is consistent with the properties of stem cells which are small, slowly dividing and undifferentiating and thus suggests that in the *STM* over-expressor, cells are acquiring similar characteristics to stem cells.

The first observed effect of *STM* induction is observed on cell growth. At 8 hours of induction, multiple expansins (*EXPA5, EXPA11* – and the expansin-like *EXLB1)* were downregulated, which promote cell growth by loosening of the extra-cellular matrix (Cho and Cosgrove 2000). This is an extremely rapid down-regulation targeting multiple genes within the same family - thus suggesting that it is a relatively important component of *STM*'s early function. From a mechanical point of view, it is easy to appreciate why it may be evolutionarily important to have a rapid and robust inhibition of cell growth in the SAM - if the meristem is to maintain an uniform structure in order to produce regular phyllotaxis, then it is absolutely necessary that growth rates be controlled across the SAM in order to prevent cells being pushed out of position. Goh et

al, (2012) have shown that repression of multiple *EXPANSIN-A* genes (1,5,3 and 10 simultaneously) in Arabidopsis leaves, reduced leaf growth – and the rate of leaf expression was reduced. *EXPA5*, which is repressed by *STM* is also leaf specific, (Goh et al, 2012) which raises the possibility that exclusion from the SAM may be an *STM*-mediated function.

Cell cycle inhibition was observed at 72 hours, at which point *STM* broadly down-regulates almost all genes in both *CYCD* and *CYCB* gene families. As these genes control passage through the G_1-S and G_2-M checkpoints respectively, this corresponds to a broad inhibition of all stages of the cell cycle (Menges et al, 2005). However, while this is not consistent with the observed phenotype of *STM* over-expression in which differentiation is inhibited and increased numbers of meristem cells observed, at 9 days the same effect upon the cell cycle could not be observed as almost all *CYCD* and *CYCB* genes were in contrast upregulated. Additionally the fold changes on all cell cycle genes at 72 hours of *STM* induction apart from *CYCD1;1* was less than 2-fold thus, although statistically significant, it may not be sufficient to have a major biological impact

Furthermore, at 72 hours an inhibitory effect upon endoreduplication could be observed at the transcriptomic level, as a number of genes involved in endoreduplication were differentially expressed at this time point. This includes *ICK* family proteins, whose expression is consistently upregulated by *STM.* The effect of *ICK* proteins upon endoreduplication is dose-dependent (Verkest et al, 2005), thus a caveat in assuming that this is responsible for inhibiting endoreduplication is made that their up-regulation by *STM* is sufficient. I found that the gene *SIAMESE* (*SIM*), which encodes a CDK inhibitor known to promote endoreduplication, is repressed by *STM* (Churchman et al, 2006).

The key difficulty in drawing strong conclusions regarding *STM*'s effects upon the cell cycle is that the 9 day time point, where *STM* is showing phenotypic differences attributable to perturbation of the cell cycle, the transcriptomic effects of *STM* show a broad but not deep up-regulation of cell cycle genes. It is thus unclear if the low fold changes are sufficient to produce a biological effect; however if it is true, it may be that inhibition of the cell cycle is predominantly being carried out by *ICK* genes by 9 days, as far more of these are upregulated by 9 days than at 72 hours, whereas initially *STM*'s inhibitory effects may be attributable primarily to down-regulation of *CYCD* and *CYCB* genes, but the degree of down-regulation is low enough that it is only the broad down-regulation of a range of cyclins which makes it plausible that there is a true biological effect. An additional caveat is that Scofield et al (2013) showed via comparison of *STM* OE and wild type leaves that *CYCD* genes were strongly upregulated (particularly, *CYCD3;1* and *CYCB1;1)*, thus differences in the extent of *CYCD* and *CYCB* gene responses may be dependent upon the tissue being considered (growing tissue vs. mature).

CYCD1;1 does not show a similar pattern of expression to the remainder of the *CYCD* genes across the time course. It is significantly differentially expressed from 24 hours onwards and remains statistically significantly upregulated at all subsequent time points. It was identified by the rapid response meta analysis as statistically significant, and by the meta analysis as being in a node whose codebook vector correlated strongly with *STM*. Thus it is the strongest candidate for a *CYCD* gene which may be mediating effects upon the cell cycle through *STM*.

6.1.12 Difficulties interpreting long term induction of *STM*

We have shown that the degree of overlap between the 9 day datasets and most earlier time points is extremely unlikely to have occurred by random chance, and we have also

shown that a large proportion of the response to *STM* induction remains consistent throughout the time course. However, with several processes, such as the cell cycle, there are marked differences in patterns of expression between genes at 72 hours and 9 days. It remains unclear how comparable the 72 hour and 9 day datasets actually are for three reasons.

First, the gap between the two time points is far larger than the gap between earlier points on the time course, as such it is difficult to ascertain what (negative) feedback may have occurred between the two time points. Second, there are far more widespread phenotypic effects in these plants. Finally, the 9 day time points represented plants which were germinated and grown on DEX, meaning that in these plants undifferentiated tissues will have been exposed to elevated or reduced levels of *STM* - which may respond differently to the differentiated tissue which is being affected by ectopic shorter term induction of *STM*. Additionally, the plants will have been exposed to elevated levels of *STM* over a wider range of developmental stages, and it is unclear whether the effects of STM induction would vary alongside developmental stage from this experiment. It is for example clear from the relationship between *STM* and *CUC1* that in part the wild type expression pattern of *STM* is dependent upon the developmental stage. It is possible that the difference in effect of *STM* perturbation could be different on plants which are just germinating to plants which are already germinated, and this may have long term effects upon their transcriptome. Since we know that *STM* does play a critical developmental role, in future experiments it may be better to induce plants for long term induction post germination, rather than prior, to account for as much of the potential differences between developmental stages as possible.

The RNAi lines show far less commonality, and have less concordance with one another than the OE lines. Meanwhile, the direction of expression for many significantly differentially expressed genes in the *stm-2* microarray is not what would have been predicted from the over-expression analysis. Moreover, as we were only affecting genes in *STM*'s native domain it is unclear whether we would have had sufficient power to detect all significant changes. In the *stm-2*, (an intermediate mutant which maintains some *STM* function) due to the abort-retry phenotype observed, it is plausible that some of the significant results may be due to an effect of increased SAMs initiated as part of this process. Additionally, in the *STM* RNAi lines, the sensitivity to detect differentially expressed genes may be confounded by its effect being limited to STM's native expression domain so genes expressed normally in leaf primordia (for example) would not show detectable increase in expression when also activated in the comparatively small SAM.

These problems justify the subsequent focus on using the Robust Response Meta analysis for data mining. As this was only calculated with the 8, 24 and 72 hour time course datasets, alongside the 3h STM-GR Mock-DEX experiment, the issue of comparability is avoided, and only those datasets where we were comfortable we had sufficient power to detect meaningful effects were used.

6.1.13 Difficulties involved with deriving predictions in function between orthologous genes

We have seen noticeable differences in the hormone response effects of *KN1* and *STM* (Bolduc et al, 2012). However, there are additional differences. While many of the same types of genes are directly regulated by both *STM* and *KN1*, we see additional inconsistencies, such as the inability to detect significant changes in expression from bound NAC-family transcription factors in *KN1,* though it must be borne in mind that

maize and Arabidopsis have evolved distinct forms of phyllotaxy. In maize, phyllotaxy is disticious, with new organs forming directly opposite the previous organ, and it is unclear what effect this distinct developmental mode would have on genetic control of organogenesis, although *KN1* does play a key role in phyllotaxy in maize (Jackson et al, 1994). While the overall downstream GRNs between the two appear to share similar features in terms of the types of genes differentially expressed, the developmental timing, and direct targets are not highly similar. Overall, *STM* direct targets seem to contain a preponderance of boundary specification and organogenesis genes, whereas *KN1* direct targets contain a preponderance of hormone control genes, although both genes regulate large numbers of transcription factors.

This difference has implications beyond this study. Whereas we might hope that results from Arabidopsis as a model organism should be applicable to other plant species which we wished to study, this suggests that - in this case - it is not possible to assume the GRN for orthologous genes will be mostly unchanged. Evolutionarily, as the same processes are regulated by both *KN1* and *STM* - this also calls into question how important the developmental timings between these processes are if these processes are regulated in a different order by orthologous genes. Alternatively, the differences between developmental timings may produce interesting phenotypic differences, but it is difficult to conclude either way without comparisons between orthologous genes in more species. However, the implications are that conclusions from Arabidopsis must be validated in other species before assuming they hold true. This would not be an unusual instance, for example, in Antirrhinum *PHANTASTICA* is the principal regulator of adaxial cell identity, whereas in Arabidopsis genes such as the HDZIPIII TF *PHABULOSA* are more important in this role (Review: Eckhart et al, 2004).

6.2 The Effectiveness of Data Mining Techniques in predicting testable relationships between genes in Arabidopsis

A second goal of this study has been to evaluate how effective several data mining and machine learning techniques are in deriving testable hypotheses between genes. In particular, it is interesting to ask how much the large amount of publically available data for a model organism such as Arabidopsis has contributed to our ability to generate such predictions. As such I shall now proceed to examine how well the various algorithms applied were in producing correct predictions.

6.2.1 Bayesian Network Structural Inference

The intention of the Bayesian network structural inference was to identify putative direct targets of *STM* - and we were able to demonstrate that 2/3 of the predicted direct targets from the initial Bayesian network inference were either responsive to *STM* in the presence of a translational inhibitor, had their promoter bound by STM, or both. However, despite this excellent level of specificity, these predictions only represented 40% of the direct targets validated in this study.

This level of sensitivity was only possible with additional information used to place constraints on the possible network structures. Unsurprisingly, the more information we were able to provide, the more accurate and sensitive the predictions became. The identification of *HB25* demonstrated that the hypothesis that we should expect all statistically significant genes to be present in the rapid response meta analysis was incorrect. However, for the overwhelming majority of direct targets it held, and the application of that restriction was sufficient to improve the quality of the overall network in the vicinity of *STM*, thus I would argue that incorporating as much prior data as possible is justified in performing these types of exploratory analysis.

Furthermore, additional evidence would be needed to demonstrate that placing constraints upon the network would improve the validity of predictions further downstream. One problem is that further downstream of *STM,* we expect that there are likely to be more missing links in the network, as it was constructed based around *STM* responsive genes - thus we would expect that missing genes will be those less well associated with *STM* itself and hence further away on the network. However, given the patterns observed in the network following unbiased separation into submodules, it appears that genes with similar spatial expression and behaviour upon induction of *STM* have been identified in the neighbourhood of one another on the consensus network. This provides some confidence that a reasonable number of predictions would be found to be accurate, though we would expect less accuracy than the direct target predictions.

With regards to the direct target predictions, for an exploratory analysis, the specificity achieved is very encouraging, as it suggests we can be relatively confident in predictions derived from this methodology prior to validating them. Importantly, the data used to generate the predictions were freely available microarrays. This suggests that the use of such freely available data to inform Bayesian networks could be used to predict direct targets in Arabidopsis in an efficient, cost-effective manner, providing an appropriate method for deriving appropriate subsets of genes. The 50% specificity demonstrates that the available microarrays for Arabidopsis cover a broad enough range of transcriptomic states to perform these analyses effectively, even on genes mostly expressed in a small number of tissues. Finally, the use of independent datasets for validating and corroborating in-house transcriptomic data enables additional confidence in predictions from noisy experimental procedures.

It remains unclear how the network could be further improved upon. The difficulty is that without more validation of downstream targets, it is difficult to evaluate the effectiveness of one network structure against another. While we can now compare the effectiveness of different procedures for their ability to predict *STM* direct targets, we do not know for certain if this comes at a cost of predictive power between genes further downstream. Furthermore, when generating these models a priori, there is no clear impartial criterion by which to judge the quality of the predictions. Overfitting of the model may have been an issue, and there are a number of different pre-processing steps, notably the discretization procedure which may have produced better results if varied.

Another possible source of error is that in this case, we may have observed a confounding effect from the existence of close homologs for many of the genes within our consensus network. If a strong association is present between two genes at otherwise different regions of the network (possibly due to differences in expression domain for example), then a better score may be obtained by missing one of the true connections. This Bayesian network may have been particularly prone to this problem, as a number of genes have closely related functions (*TCP3, 4, 10, 24; STM, KNAT1, KNAT2; BOP1, BOP2; CUC1, CUC3) (*Chuck et al, 1996; Hepworth et al, 2005; Aida & Tasaka et al, 2006; Martin-Trillo et al, 2010) and all of which either have a high degree of connectivity or are directly related to *STM*.

6.2.2 Self Organizing Maps complement the Bayesian Network for candidate direct target identification

Of the TFs identified on the SOM as being most closely positively correlated with *STM*, two were identified as direct targets (*BOP2* and *OFP1*). Interestingly, these were TFs which were not identified as direct targets by the Bayesian network. There is no specific reason why the two methods should have been complementary, and we should not

expect this would always be the case, however the two approaches were both successful in identifying different groups of direct targets. One explanation would be that the SOM and the Bayesian Network used different input datasets. In this case, it may be that different features of *STM*'s GRN were emphasized in each.

The SOM also enabled me to look at non-TFs downstream of *STM* and suggested an interesting subset of robustly regulated targets. One example is *CYCD1;1* which ongoing research suggests forms an important part of *STM*'s regulatory network.

6.2.3 The Rapid Response Meta Analysis identified *STM* direct targets successfully

While not a machine learning method, the rapid response meta analysis contained 4 of the 5 identified direct target TFs of *STM*. *HB25* was the only TF not identified; however, with an adjusted p-value of 0.07 it was still a relatively plausible candidate for a direct target. However, the rapid response meta analysis was not very specific. 13 genes were identified as having TF activity by GO enrichment, and while *BOP2* was not identified by its GO term, its homolog *BOP1* is a known transcriptional regulator. Including both *BOP* genes, the rapid response meta analysis had a specificity of 33% for identifying subsequently validated direct targets. This is good, but clearly can be improved upon by using data mining techniques such as the Bayesian network or the SOM. This provides further justification for initially constraining potential direct targets of *STM* on the Bayesian network to those genes identified by the rapid response meta analysis, as it also shows that the sensitivity of the rapid response meta analysis in detecting direct targets was 80% while also revealing the early responding components of the *STM* GRN.

The use of the meta analyses to refine microarray predictions was demonstrated to be an extremely useful technique. In this study, meta analysis techniques were used to compare datasets from different time points, however these are not the only differences identified, as the differences observed between multiple expression systems were particularly striking. We observed a core response to *STM* between our TGV system and *STM*-GR system, and between the ethanol inducible systems used by Spinelli et al (2011). However, there were also clearly responses which were different between all three. Meta analysis offers one means of statistically validly combining data from similar, but only partially consistent data sources, and in an unbiased manner determining consistent responses, which is important given that the differences between such expression systems appear to be non-trivial.

6.2.4 The use of co-expression to enable completely *in silico* predictions of direct targets

I have shown that it is possible, albeit with a decreased level of accuracy to make direct target predictions in a completely *in silico* manner. The main difficulty is in identifying the subset of input genes to use in the analysis without an initial wet-lab experiment. While co-expression analysis is suitable for indentifying some direct target genes, it only detected those genes predicted as direct targets by the Bayesian network inference in a suitably focussed subset of microarrays where developmental stage and tissue type were controlled.

Thus, the core difficulty is in selecting an appropriate subset of microarray experiments to use for the co-expression analysis. However, comparison with prior data in the literature can be used to validate that you have a plausible subset of candidate genes. For example, the presence of multiple *KNAT* and *IAA* genes in the focussed subset of *STM*-correlated genes was suggestive that an appropriate subset of genes had been

identified. Finally, as co-expression analysis identified a fairly large number of floral genes for *STM*, we cannot currently confirm whether these are or are not direct targets as experiments were performed in vegetative growth. Given the large number of strongly correlated genes, a core TF such as STM has when looking at co-expression data, data mining techniques are extremely important in refining the subset to a manageable number of predictions.

It is interesting to consider whether the process of knowledge discovery for target genes could be automated. Given a set of microarrays for specific tissue, and a subset of TFs known to be expressed at a reasonable level within that tissue, it may be possible to perform an automated prediction process to derive direct target hypotheses between co-expressed genes and the list of genes under investigation. The more interesting among these could then be tested experimentally, essentially fishing for interesting relationships between a large number of genes of interest.

Interestingly, the consensus network generated for the co-expression study predicted the same direct targets of *STM* as the initial constrained network. As such, it was more sensitive than the unconstrained network, predicting both *CUC1* and *AIL7*. This could not always be guaranteed to be the case, however, the dataset was smaller, and all genes present were known to correlate well with *STM* - though in many cases this would possibly have been due to similarity of expression domain rather than those genes being genuinely downstream of *STM*. This raises the question of whether for this subset of genes it may have been easier for the Bayesian network inference algorithm to identify associations between than for the subset of genes from the microarray experiment. For those genes which simply had similar expression domains, we would not expect networks inferring direct relationships between them to score as highly as for those where a regulatory relationship (indirect or indirect) was present. With the

microarray-generated subset it may be more likely that strong indirect regulation be mis-classified as a direct association.

6.2.5 Multiple direct target identification methods are required in order to make confident predictions

ChIP studies such as Bolduc et al (2012) already demonstrated that a positive result from ChIP analysis alone is not sufficient to conclude that a gene is directly regulated by a specific TF. Large numbers of genes identified as bound by KN1 in their studies, were not in fact responsive to KN1. Similarly, we have shown that CHX based direct target experiments can be extremely noisy. Often additive effects magnify expected fold changes, and variability due to CHX treatment can be too high to draw conclusions as to gene regulation. qRT-PCR provides an alternative readout of fold changes in CHX experiments; however, some genes which appear to be differentially expressed by qRT-PCR, such as *OFP1*, may not be identified by microarray analysis.

Given this, in order to conclude a gene is a direct target, it is necessary to combine data from multiple sources. If evidence is obtained for binding, then confirmation is required that the gene in question is actually responsive, and that thus the binding leads to modulation of gene expression. In the case that a change in mRNA expression is observed under inhibition of translation, it is necessary to conclude that the TF of interest is binding the target in order to be certain that the observed effect is not due solely to CHX variability.

In both cases, when using an ectopic over expression system, a positive result may still be due to saturation of the tissue in question with an unusually high amount of protein. Thus the biological context of direct target genes must be examined in detail in order to

conclude whether or not it is likely to be a (developmentally relevant) particularly interesting result.

6.3 The placement of *STM* in its proper context within SAM function

As a result of the data mining predictions discussed, I have been able to guide the analysis of *STM* direct targets and transcriptomic effects from statistical analysis of discrete time points to a dynamic perspective of STM function, particularly at the boundary zone. This has enabled a clear understanding of how *STM* acts as a core regulator of SAM function. Importantly, I have demonstrated that it is possible, though less efficient, to infer GRNs using Bayesian network structural inference in a completely *in silico* manner. Though, the use of purpose-generated microarray data to produce the most accurate subsets of immediately downstream genes and select appropriate network constraints improved network accuracy.

As we have seen, there is no one clear method for deducing that a gene is a definitive direct target, as such we must currently rely upon a mix of different methodologies, and accept that the more positive results, the more confidence we can place in the conclusion. Similar problems exist due to the noisiness of microarray data, in particular with differences between long term or short term inductions and the surprising number of differences between expression systems. Meta analysis has been shown to be very effective in robustly pooling data from noisy sources,

While gene expression changes can reveal a large amount of information about how perturbations grossly affect the transcriptome, the most high-throughput methods available to us are limited by temporal resolution. Thus, any transcriptomic data represents a snapshot rather than a truly continuous picture of gene expression. While it is not usually possible to directly prove hypotheses regarding gene expression

dynamics using mathematical models, these fulfil an important role in systems biology by testing the plausibility of hypotheses and allowing us to explore what dynamics are possible given the observed expression data. Using ODE and stochastic models as a supplement to gene expression data, I have been able to move from simply looking at snapshots to building a dynamic picture of the SAM.

As a transcription factor, STM does not produce a phenotype independently, but relies upon the modulation of downstream genes and regulatory modules to exert a phenotypic effect. Through careful dissection of how STM affects other core modules on a case-by-case basis, I have been able to show how STM connects to hormone regulation and boundary formation, as well as extending our understanding of how STM impinges on the cell cycle and interfaces with stem cell identity regulation via cytokinin (WUS-CLV).

Overall, although I have been able to show the clear importance of *STM* in maintaining and initiating proper SAM function, it is clear that such inference requires careful analysis of inherently noisy gene expression data. Given the large volumes of data produced by genome-wide experiments it is essential to use data mining and statistical approaches to evaluate the data in an unbiased way. I have demonstrated that by effectively applying these approaches it is possible to derive testable and novel biological predictions, at both the direct, transcriptomic scale, and to provide a system-wide view of gene function which has expanded our understanding of the dynamic behaviour of STM and to place it in its proper context as a master regulator of SAM function.

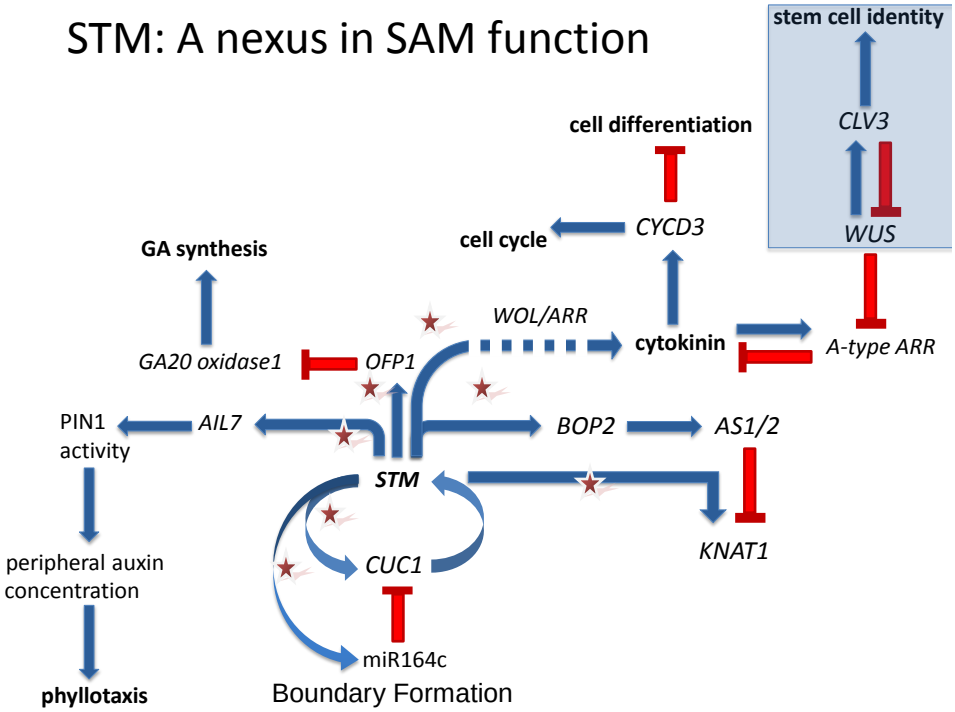


Figure 6.2 - *STM* is a nexus in SAM function.

A schematic of relationships perturbed via *STM* misexpression in the SAM. Solid blue arrows indicate positive induction of the target process or gene. Dotted lines indicate indirect mediation of a positive effect upon the target process or gene. Red bars indicate negative feedback upon the target process or gene. The box around WUS/CLV indicates that this is a separate process which communicates with the *STM* nexus via CK regulation. Red stars indicate work shown or validated as part of this project. *CUC1* direct targeting of *STM* identified via collaboration with Mitsuhiro Aida (NAIST, Japan). Sieber et al (2007) demonstrated that *miR164c* is sufficient for *CUC1* repression in the SAM. *AIL7* effect on phyllotaxis and PIN proteins was shown by Prasad et al (2011). Jun et al (2010) showed *BOP2* activates *AS1/2*, while Byrne et al (2000,2002) proposed models for *AS* gene interactions with *KNAT1*. *OFP1* has been shown to directly repress *GA20OX1* by Wang et al (2007). *WUS* was shown to directly repress the *A*-type *ARRs* by (Liebfried et al (2005), *STM*'s effect on *ARRs* has been repeated by this study and mostly agreed with the work of Jasinski et al (2005) and Yanai et al (2005). Scofield et al (2013) further elaborated on the effects of *STM* on the cell cycle, Laufs et al (1998) identified the relationship between *WUS* and *CLV.*

As I have shown, *STM* either directly or indirectly regulates a wide variety of processes - ranging from boundary formation and phyllotaxis to hormone synthesis, through to organogenesis and differentiation as shown in Figure 6.3. What is clear is that almost every major process which occurs in the SAM is in some way affected by *STM*

expression, as we might expect given the severity of its over-expression and mutant phenotypes.

Thus I argue that this study clearly shows *STM* is a nexus for SAM functions. Each core process in the SAM; the maintenance of a pool of slowly dividing stem cells and organogenesis with correct phyllotaxis are regulated by *STM*. And through ensuring correct boundary zone establishment and hormone distribution, *STM* ensures that SAM homeostasis is maintained.

References

Aida et al, 1997, Genes involved in Organ Separation in *Arabidopsis*: An Analysis of the *cup-shaped cotyledon* Mutant, *The Plant Cell*, 9, 841-857

Aida et al, 1999, Shoot apical meristem and cotyledon formation during *Arabidopsis* embryogenesis: interaction among the *CUP-SHAPED COTYLEDON* and *SHOOT MERISTEMLESS* genes, *Development* 126, 1563-1570

Aida and Tasaka, 2006, Genetic control of shoot organ boundaries, *Current Opinion in Plant Biology*, 9:72–77

Ashburner et al, 2000, Gene Ontology: tool for the unification of biology, *Nat Genet*, 25(1): 25–29

Affymetrix, 2012, *GeneChip Arabidopsis ATH1 Genome Array*, http://www.oceanridgebio.com/affy_docs/arab_datasheet.pdf

Baker et al, 2005, The *early extra petals1* mutant uncovers a role for microRNA *miR164c* in regulating petal number in *Arabidopsis*, *Current Biology*, 15(4):303-15

Barton & Poethig, 1993, Formation of the shoot apical meristem in A*rabidopsis thaliana*: an analysis of development in the wild type and in the *shoot meristemless* mutant, *Development* 119, 823-831

Barton M, 2010, Twenty years on: the inner workings of the shoot apical meristem, a developmental dynamo, *Dev Biol*, 1;341(1):95-113

Bellaoui et al, 2001, The Arabidopsis BELL1 and KNOX TALE Homeodomain Proteins Interact through a Domain Conserved between Plants and Animals, *The Plant Cell* 13, 2455–2470

Belles-Boix et al, 2006, *KNAT6*: An *Arabidopsis* Homeobox Gene Involved in Meristem Activity and Organ Separation, *The Plant Cell*, 18, 1900–1907

Benjamini and Hochberg, 1995, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B (Methodological),* 57(1), 289-300

Beuchel et al, 2009, Role of A-type *ARABIDOPSIS RESPONSE REGULATORS* in meristem maintenance and regeneration, *European Journal of Cell Biology*, 89(2-3):279-84

Bharathan et al, 1999, Phylogenetic Relationships and Evolution of the KNOTTED Class of Plant Homeodomain Proteins, *Mol. Biol. Evol.* 16(4):553–563

Bolduc et al, 2012, 2012, Unravelling the *KNOTTED1* regulatory network in maize meristems, *Genes & Development,* 26:1685–1690

Bolstad et al, 2003, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics,* 19(2), 182-193

Bonferroni, C. E, 1935, "Il calcolo delle assicurazioni su gruppi di teste." In *Studi in Onore del Professore Salvatore Ortu Carboni.* Rome: Italy, pp. 13-60

Boruc et al, 2011, Dynamics of the Plant Nuclear Envelope and Nuclear Pore, Plant *Physiology*, 158(1), 78-86

Bowman and Eshed, 2000, Formation and maintenance, of the shoot apical meristem,

Trends in Plant Science, 5(3), 1360-1385

Brand et al, 2002, Regulation of *CLV3* Expression by Two Homeobox Genes in *Arabidopsis*, *Plant Physiology*,129, 565–575

Brady et al, 2011, A stele-enriched gene regulatory network in the *Arabidopsis* root, *Molecular Systems Biology* 7;459

Busch et al, 2010, Transcriptional control of a plant stem cell niche, *Developmental Cell*, 18(5):849-61

Byrne et al, 2000, *ASYMMETRIC LEAVES1* mediates leaf patterning and stem cell function in Arabidopsis, *Nature*, 408, 967-971

Byrne et al, 2002, *ASYMMETRIC LEAVES1* reveals knox gene redundancy in Arabidopsis, *Development* 129, 1957-1965

Carraro et al, 2006, Cell differentiation and organ initiation at the shoot apical meristem, *Plant Molecular Biology*, 60(6):811-26

Cary et al, 2002, Developmental events and shoot apical meristem gene expression patterns during shoot development in *Arabidopsis thaliana*, *The Plant Journal*, 32, 867–877

Chapelle et al, 2012, Impact of the absence of stem-specific β-glucosidases on lignin and monolignols, *Plant Physiology*, 160(3):1204-17

Cheng et al, 2011, Auxin biosynthesis by the *YUCCA* flavin monooxygenases controls the formation of floral organs and vascular tissues in *Arabidopsis,* Genes and *Development* 20:1790–1799

Cho and Cosgrove, 2000, Altered expression of expansin modulates leaf growth and pedicel abscission in Arabidopsis thaliana, *PNAS*, 97(17), 9783–9788

Chuck et al, 1996, *KNAT1* induces Lobed Leaves with Ectopic Meristems When Overexpressed in Arabidopsis, *The Plant Cell*, 8, 1277-1289

Churchman et al, 2006, *SIAMESE*, a plant-specific cell cycle regulator, controls endoreplication onset in Arabidopsis thaliana, *Plant Cell*, 18(11):3145-57

Clark et al, 1993, *CLAVATA1*, a regulator of meristem and flower development in *Arabidopsis*, *Development*, 119(2):397-418

Clark et al, 1996, The *CLAVATA* and *SHOOT MERISTEMLESS* loci competitively regulate meristem activity in *Arabidopsis*, *Development* 122, 1567-1575

Clark et al, 1997, The *CLAVATA1* gene encodes a putative receptor kinase that controls shoot and floral meristem size in Arabidopsis, *Cell*, 89(4):575-85

Cole et al, 2006, Nuclear import of the transcription factor *SHOOT MERISTEMLESS* depends on heterodimerization with *BLH* proteins expressed in discrete sub-domains of the shoot apical meristem of *Arabidopsis thaliana*, *Nucleic Acids Research*, 34, 4 1281–1292

Cosgrove D, 2000, Loosening of plant cell walls by expansins, *Nature*, 407, 321-326

Cosgrove et al, 2002, The Growing World of Expansins, *Plant Cell Physiol*. 43(12): 1436–1444

Cubas et al, 1999, The *TCP* domain : A motif found in proteins regulating plant growth and development, *The Plant Journal*, 18(2), 215-222

De Jager and Murray, 1999, Retinoblastoma Proteins in Plants, *Plant Molecular Biology* 41, 295-299

Dennis Jr et al, 2003, DAVID: Database for Annotation, Visualization, and Integrated Discovery, Bioinformatics, *Genome Biology 2003,* 4:R60

DeRouille et al, 2005, Computer simulations reveal properties of the cell–cell signalling network at the shoot apex in *Arabidopsis*, *PNAS*, vol. 103, no. 5, 1627–1632

DeWitte et al, 2003, Altered cell cycle distribution, hyperplasia, and inhibited differentiation in Arabidopsis caused by the D-type cyclin *CYCD3*, *Plant Cell*, 15(1):79-92.

DeWitte et al, 2007, Arabidopsis *CYCD3* D-type cyclins link cell proliferation and endocycles and are rate-limiting for cytokinin responses, *PNAS*, 104;36, 14537–14542

Dorca-Fornell et al, 2011, The *Arabidopsis SOC1*-like genes *AGL42, AGL71* and *AGL72* promote flowering in the shoot apical and axillary meristems, *Plant Journal*, 67(6):1006-17

Eckardt et al, 2004, The Role of *PHANTASTICA* in Leaf Development, *The Plant Cell*, 16, 1073–1075

Edgar and Lash, 2003, The Gene Expression Omnibus (GEO): A Gene

Expression and Hybridization Repository, *The NCBI Handbook*, Chapter 6, http://www.ncbi.nlm.nih.gov/books/NBK21101/

Elliott et al, 1996, *AINTEGUMENTA*, an *APETALA2*-like Gene of *Arabidopsis* with Pleiotropic Roles in Ovule Development and Floral Organ Growth, *The Plant Cell*, 8, 155-168

Endrizzi et al, 1996, The *SHOOT MERISTEMLESS* gene is required for maintenance of undifferentiated cells in *Arabidopsis* shoot and floral meristems and acts at a different regulatory level than the meristem genes *WUSCHEL* and *ZWILLE*, *The Plant Journal*, 10(6), 967-979

Fletcher and Meyerowitz, 2000, Cell signalling within the shoot meristem, *Current Opinion in Plant Biology*, 3:23–30

Fozard et al, 2013, Modelling auxin efflux carrier phosphorylation and localization, *Journal of Theoretical Biology*, 319, 34-49

Friedman et al, 2000, Using Bayesian Networks to Analyze Expression Data, *Journal of Computational Biology*, 7(3/4), 601–620

Fukishima et al, 2012, Exploring Tomato Gene Functions Based on Coexpression Modules Using Graph Clustering and Differential Coexpression Approaches, *Plant Physiology*, 158, 1487-1502

Gallois et al, 2002, Combined *SHOOT MERISTEMLESS* and *WUSCHEL* trigger ectopic organogenesis in Arabidopsis, *Development* 129, 3207-3217

Gazzani et al, 2004, A Link Between mRNA Turnover and RNA Interference in *Arabidopsis*, *Science*, 306, 1046-1048

Geldner et al, 2001, Auxin transport inhibitors block *PIN1* cycling and vesicle trafficking, *Nature*, 413, 425-428

Godsey B, 2013, Improved Inference of Gene Regulatory Networks through Integrated

Bayesian Clustering and Dynamic Modelling of Time-Course Expression Data, *Plos One*, 8;7, e68358

Goh et al, 2012, Inducible Repression of Multiple Expansin Genes Leads to Growth Suppression during Leaf Development, *Plant Physiology*, 159, 1759–1770

Gordon et al, 2007, Pattern formation during de novo assembly of the *Arabidopsis* shoot meristem, *Development* 134, 3539-3548

Grandjean et al, 2004, In vivo analysis of cell division, cell growth, and differentiation at the shoot apical meristem in *Arabidopsis*, *Plant Cell*, 16(1):74-87

Groß-Hardt, and Laux, 2003, Stem cell regulation in the shoot meristem, *Journal of Cell Science* 116, 1659-1666

Guo et al, 2005, MicroRNA directs mRNA cleavage of the transcription factor NAC1 to downregulate auxin signals for *arabidopsis* lateral root development, *Plant Cell*, 17(5):1376-86

Ha et al, 2004, *BLADE-ON-PETIOLE1* encodes a BTB/POZ domain protein required for leaf morphogenesis in Arabidopsis thaliana, *Plant Cell Physiology*, 45(10):1361-70

Hackbusch et al, 2005, A central role of *Arabidopsis thaliana* ovate family proteins in networking and subcellular localization of 3-aa loop extension homeodomain proteins, *PNAS*, 12, No 13, 4908–4912

Hake et al, 2004, The role of *KNOX* genes in plant development*, Annu. Rev. Cell Dev*. Biol, .20:125-151

Hartemink et al, 2001, Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.*, 422–433

Hartemink A, 2005, Reverse Engineering Gene Regulatory Networks, *Nature*, 23, 5, 554-555

Hay et al, 2002, The Gibberellin Pathway Mediates *KNOTTED1*-Type Homeobox Function in Plants with Different Body Plans, *Current Biology*, 12, 1557-1565

Heckerman et al, 1995, Learning Bayesian Networks: The Combination of Knowledge and Statistical Data, *Machine Learning*, 20, 197-243

Hedden and Phillips, 2000, Gibberellin metabolism: new insights revealed by the genes, *Trends in Plant Science Reviews*, 5,12, 523-530

Heisler et al, 2005, Patterns of auxin transport and gene expression during primordium development revealed by live imaging of the *Arabidopsis* inflorescence meristem, *Current Biology*, 8;15(21):1899-911.

Henriksson et al, 2005, Homeodomain Leucine Zipper Class I Genes in *Arabidopsis*. Expression Patterns and Phylogenetic Relationships, *Plant Physiology*, 139, September,. 509–518

Hepworth et al, 2005, *BLADE-ON-PETIOLE*-dependent signalling controls leaf and floral patterning in *Arabidopsis*, *Plant Cell*, 17(5):1434-48

Hibara et al, 2003, *CUC1* gene activates the expression of SAM-related genes to induce adventitious shoot formation, *The Plant Journal*, 36, 687-696

Hibara et al, 2006, Arabidopsis CUP-SHAPED COTYLEDON3 regulates postembryonic shoot meristem and organ boundary formation, Plant Cell, 18(11):2946-57

Holm S, 1979, A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* 6(2): 65–70

Hoops et al, 2006, COPASI—a COmplex PAthway SImulator, *Bioinformatics*, 22(24), 3067–3074

Huang et al, 2012, *RBE* controls *microRNA164* expression to effect floral organogenesis, *Development* 139, 2161-2169

Hubbell et al, 2002, Robust estimators for expression analysis, *Bioinformatics*, 18(12), 1585–1592

Ikezaki et al, 2010, Genetic networks regulated by *ASYMMETRIC LEAVES1 (AS1)* and *AS2* in leaf development in Arabidopsis thaliana: KNOX genes control five morphological events, *Plant Journal*, 61(1):70-82

Imoto et al, 2003, Combining microarrays and biological knowledge for estimating gene expression networks via Bayesian networks, *Journal Bioinform Comput Biol*, 2(1), 77-98

Inoue et al, 2001, Identification of *CRE1* as a cytokinin receptor from *Arabidopsis*, *Nature*, 409, 1060-1063

Irizarry et al, 2003, Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *biostatistics*, 4(2), 249–264

Jackson et al, 1994, Expression of maize *KNOTTED1* related homeobox genes in the shoot apical meristem predicts patterns of morphogenesis in the vegetative shoot, *Development* 120, 405-413

Jasinski et al, 2005, KNOX Action in *Arabidopsis* Is mediated by Coordinate Regulation of Cytokinin and Gibberellin Activities, *Current Biology*,15, 1560–1565

Jeong et al, 1999, The *Arabidopsis CLAVATA2* gene encodes a receptor-like protein required for the stability of the *CLAVATA1* receptor-like kinase, *Plant Cell*, 11(10):1925-34

Jiang et al, 2004, Cluster Analysis for Gene, Expression Data: A Survey, I*EEE Transactions on Knowledge and Engineering*,16(11), 1370-1386

Jönsson et al, 2005, An auxin-driven polarized transport model for phyllotaxis, *PNAS*, 103(5), 1633–1638

Jun et al, 2010, *BLADE-ON-PETIOLE1* Coordinates Organ Determinacy and Axial Polarity in *Arabidopsis* by Directly Activating *ASYMMETRIC LEAVES2*, *Plant Cell*, 22(1), 62-76

Kayes and Clark, 1998, *CLAVATA2*, a regulator of meristem and organ development in *Arabidopsis*, *Development*, 125(19):3843-51

Kim et al, 2002, Intercellular trafficking of a *KNOTTED1* green fluorescent protein fusion in the leaf and shoot meristem of *Arabidopsis*, *PNAS*, 99(6):4103-8

Kim et al, 2003, Developmental regulation and significance of KNOX protein trafficking in *Arabidopsis*, *Development* 130, 4351-4362

Krizek et al, 2006, *RABBIT EARS* is a second-whorl repressor of *AGAMOUS* that maintains spatial boundaries in *Arabidopsis* flowers, *Plant Journal*, 45(3):369-83

Kohonen T, 1990, The Self Organizing Map*, Proceedings of the IEEE*, 79(9), 1464 – 1480

Kohonen et al, 2001, *Self-Organizing Maps*, 3rd Edition, Springer-Verlag New York, ISBN:3540679219

Koung-Lee et al, 2006, *LONGIFOLIA1* and *LONGIFOLIA2*, two homologous genes, regulate longitudinal cell elongation in *Arabidopsis*, *Development*, 133(21):4305-14

Koyama et al, 2007, *TCP* transcription factors control the morphology of shoot lateral organs via negative regulation of the expression of boundary-specific genes in *Arabidopsis*, *Plant Cell*, 19(2):473-84

Koyama et al, 2010, *TCP* Transcription Factors Regulate the Activities of *ASYMMETRIC LEAVES1* and *miR164*, as Well as the Auxin Response, during Differentiation of Leaves in *Arabidopsis*, *The Plant Cell*, 22, 3574–3588

Kusaba et al, 1998, Decreased *GA1* Content Caused by the Overexpression of *OSH1* Is Accompanied by Suppression of GA 20-Oxidase Gene Expression, *Plant Physiology*, 117: 1179–1184

Laufs et al, 1998, Cellular Parameters of the Shoot Apical Meristem in *Arabidopsis*, *The Plant Cell*, 10, 1375–1389

Laufs et al, 2004, *MicroRNA* regulation of the CUC genes is required for boundary size control in Arabidopsis meristems, *Development* 131, 4311-4322

Lee et al, 2006, *LONGIFOLIA1* and *LONGIFOLIA2*, two homologous genes, regulate longitudinal cell elongation in *Arabidopsis*, *Development* 133, 4305-4314

Lenhard et al, 2002, The *WUSCHEL* and *SHOOTMERISTEMLESS* genes fulfil complementary roles in *Arabidopsis* shoot meristem regulation, *Development* 129, 3195-3206

Li et al, 2012, TCP transcription factors interact with *AS2* in the repression of class-I KNOX genes in Arabidopsis thaliana, *The Plant Journal*, 71, 99–107

Liebfried et al, 2005, *WUSCHEL* controls meristem function by direct regulation of cytokinin-inducible response regulators, *Nature*, 438(7071):1172-5

Livak and Schmittingen, 2001, Analysis of relative gene expression data using real-time quantitative PCR and the 2(-ΔΔC(T)) Method, *Methods*, 25(4):402-8

Long et al, 1996, A member of the *KNOTTED* class of homeobox proteins encoded by the *STM* gene of *Arabidopsis*, *Nature*, 379, 66-69

Long and Barton, 1998, The development of apical embryonic pattern in *Arabidopsis*, *Development* 125, 3027-3035

Maere et al, 2005, *BiNGO*: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks, *Bioinformatics*, 21(16), 3448–3449

Mähönen et al, 2000, A novel two-component hybrid molecule regulates vascular morphogenesis of the *Arabidopsis* root, *Genes & Development,* 14:2938–2943

Man Ha et al, 2003, The *BLADE-ON-PETIOLE 1* gene controls leaf pattern formation through the modulation of meristematic activity in *Arabidopsis*, *Development* 130, 161-

172

Man Ha et al, 2007, *BLADE-ON-PETIOLE1* and 2 Control *Arabidopsis* Lateral Organ Fate through Regulation of LOB Domain and Adaxial-Abaxial Polarity Genes, *The Plant Cell*, 19: 1809–1825

Martin-Trillo et al, 2010, *TCP* genes: a family snapshot ten years later, Trends in Plant *Science*, 15(1):31-9

Mayer et al, 1998, Role of *WUSCHEL* in Regulating Stem Cell Fate in the *Arabidopsis* Shoot Meristem, *Cell*, 95, 805–815

McConnell and Barton, 1998, Leaf polarity and meristem formation in *Arabidopsis*, *Development* 125, 2935-2942

Meister et al, 2005, Multiple Protein Regions Contribute to Differential Activities of *YABBY* Proteins in Reproductive Development, *Plant Physiology*, 137, 651–662

Menashe et al 2013 Co-expression Profiling of Autism Genes in the Mouse Brain, *PLoS Comput Biol* 9(7), e1003128.

Menges et al, 2005, Global analysis of the core cell cycle regulators of *Arabidopsis* identifies novel genes, reveals multiple and highly specific profiles of expression and provides a coherent model for plant cell cycle control, *Plant Journal*, 41(4):546-66

Miyawaki et al, 2004, Expression of cytokinin biosynthetic isopentyltransferase genes in Arabidopsis: tissue specificity and regulation by auxin, cytokinin and nitrate, *The Plant Journal*, 37, 128-138

Mudunkothge and Krizek, 2012, Three *Arabidopsis AIL*/*PLT* genes act in combination to regulate shoot apical meristem function, *The Plant Journal*, 71, 108–121

Müller and Sheen, 2007, Advances in Cytokinin Signaling, *Science*, 318, 68-69

Nole-Wilson & Krizek, 2006, *AINTEGUMENTA* Contributes to Organ Polarity and Regulates Growth of Lateral Organs in Combination with *YABBY* Genes, *Plant Physiology*, 141, 977–987

Obrig et al, 1971, The mechanism by which cycloheximide and related glutarimide antibiotics inhibit peptide synthesis on reticulocyte ribosomes*, J Biol Chem*, 246(1):174-81

Okada et al, 1991, Requirement of the Auxin Polar Transport System in Early Stages of *Arabídopsis* Floral Bud Formation, *The Plant Cell*, 3, 677-684,

Olszewski et al, 2002, Gibberellin Signaling: Biosynthesis, Catabolism, and Response Pathways, *The Plant Cell*, S61–S80, Supplement 2002

Palmieri et al, 2006, Molecular Identification of an *Arabidopsis* S-Adenosylmethionine Transporter. Analysis of Organ Distribution, Bacterial Expression, Reconstitution into Liposomes, and Functional Characterization, *Plant Physiology*, 142. 855–865

Parman and Halling, 2013, *affyQCReport: A Package to Generate QC Reports for Affymetrix Array Data*, R Package version 1.24.0

Pernisová et al, 2009, Cytokinins modulate auxin-induced organogenesis in plants via regulation of the auxin efflux, PNAS, 106(9), 3609–3614

Pinon et al, 2013, Local auxin biosynthesis regulation by PLETHORA transcription factors controls phyllotaxis in *Arabidopsis*, *PNAS*, 10(3), 1107–1112

Prasad et al, 2011, Arabidopsis *PLETHORA* Transcription Factors Control Phyllotaxis, *Current Biology* 21, 1123–1128,

Raman et al, 2008, Interplay of *miR164*, *CUP-SHAPED* COTYLEDON genes and *LATERAL SUPPRESSOR* controls axillary meristem formation in *Arabidopsis* thaliana, *The Plant Journal*, 55, 65–76

Reddy et al, 2004, Real-time lineage analysis reveals oriented cell divisions associated with morphogenesis at the shoot apex of *Arabidopsis thaliana*, *Development*, 131(17): 4225-37

Reinhardt et al, 2003, Regulation of phyllotaxis by polar auxin transport, *Nature*, 426, 255-260

Rieu and Laux, 2009, Signaling pathways maintaining stem cells at the plant shoot apex, *Seminars in cell developmental biology*, 20(9):1083-8

Rieu-Khamlichi et al, 1999, Cytokinin activation of *Arabidopsis* cell division through a D-type cyclin, *Science*, 5(283), 1541-4

Rojo et al, 2002, *CLV3* is localized to the extracellular space, where it activates the Arabidopsis *CLAVATA* stem cell signalling pathway, *Plant Cell*, 14(5):969-77

Rosin et al, 2003, Overexpression of a knotted-like homeobox gene of potato alters vegetative development by decreasing gibberellin accumulation, *Plant Physiology* 132(1):106-17

Sakamoto et al, 1999, The Conserved KNOX Domain Mediates Specificity of Tobacco *KNOTTED1*-Type Homeodomain Proteins, *The Plant Cell*, 11, 1419–1431

Sakamoto et al, 2001, KNOX homeodomain protein directly suppresses the expression of a gibberellins biosynthetic gene in the tobacco shoot apical meristem, *Genes & Development* 15:581–590

Sawa et al, 1999, *FILAMENTOUS FLOWER*, a meristem and organ identity gene of *Arabidopsis*, encodes a protein with a zinc finger and HMG-related domains, *Genes & Development*, 13:1079–1088

Schoof et al, 2000, The stem cell population of Arabidopsis shoot meristems in maintained by a regulatory loop between the *CLAVATA* and *WUSCHEL* genes, *Cell*, 100(6):635-44

Schwechenheimer C, 2012, Gibberellin signalling in plants - the extended version, *Frontiers in Plant Science*, 2(107)

Scofield et al, 2007, The *KNOX* gene *SHOOT MERISTEMLESS* is required for the development of reproductive meristematic tissues in Arabidopsis, *The Plant Journal* , 50, 767–781

Scofield et al, 2008, A model for Arabidopsis class‐1 *KNOX* gene function, *Plant Signaling & Behaviour* 3(4), 257-259

Scofield et al, 2013, The *Arabidopsis* homeobox gene *SHOOT MERISTEMLESS* has cellular and meristem-organisational roles with differential requirements for cytokinin and *CYCD3* activity, *The Plant Journal*, 75, 53–66

Scofield and Murray, 2006, KNOX gene function in plant stem cell niches, *Plant Molecular Biology*, 60:929–946

Sieber et al, 2007, Redundancy and specialization among plant microRNAs: role of the *MIR164* family in developmental robustness, *Development* 134, 1051-1060

Silverstone et al, 2007, Functional analysis of SPINDLY in gibberellin signalling in Arabidopsis, *Plant Physiology*, 143(2):987-1000

Skoog and Miller, 1957, Chemical regulation of growth and organ formation in plant tissue cultures*, In vitro. Symp. Soc. Exp. Biol*, 11, 118-131, 1957

Slonim D, 2002, From patterns to pathways: gene expression comes of age, *Nature genetics supplement*, 32, 502-508

Smith et al, 2006, A plausible model of phyllotaxis, PNAS, 103(5), 1301-1306

Smoot et al, 2011, Cytoscape 2.8: new features for data integration and network visualization, *Bioinformatics,* 27(3), 431–432

Smyth, G, 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3(1)

Spellman et al, 1998, Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell*, 9(12):3273-97

Spinelli et al, 2012, A Mechanistic Link between *STM* and *CUC1* during *Arabidopsis* Development, *Plant Physiology*, 156, 894–1904

Srinivasasainagendra et al, 2008, CressExpress: A Tool For Large-Scale Mining of Expression Data from Arabidopsis, *Plant Physiology*, 147, 1004–1016

Steeves and Sussex, 1989, *Patterns in Plant Development*, Cambridge University Press, Cambridge

Storey J, 2002, A direct approach to false discovery rates, *J. R. Statist. Soc.* B 64, *Part* 3, 479–498

Storey and Tibshirani, 2003, Statistical significance for genomewide studies, *PNAS,* 100(16), 9440–9445

Su et al, 2011, Auxin–Cytokinin Interaction Regulates Meristem Development, *Molecular Plant,* 4(4), 616–625

Tadaka et al, 2001, The *CUP-SHAPED COTYLEDON1* gene of *Arabidopsis* regulates shoot apical meristem formation, *Development* 128, 1127-1135

Tamayo et al, 1999, Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, *PNAS*, 96(6):2907-12.

Tedegar and Ward, 2012, Molecular evolution of plant *AAP* and *LHT* amino acid transporters, *Front Plant Sci,* 3(21), 1-11

Tiwari et al, 2003, The Roles of Auxin Response Factor Domains in Auxin-Responsive Transcription, *The Plant Cell*, 15, 533–543

Tran et al, 2007, Functional analysis of *AHK1*/*ATHK1* and cytokinin receptor histidine kinases in response to abscisic acid, drought, and salt stress in *Arabidopsis*, *PNAS*, 104(51), 20623–20628

Uchida et al, 2007, Regulation of *SHOOT MERISTEMLESS* genes via an upstream-conserved noncoding sequence coordinates leaf development, *PNAS*, 104(40),15953–15958

Valerio O, 2000, Mapping chromosomal proteins *in vivo* by formaldehyde crosslinked-chromatin immunoprecipitation, *Trends in Biochemical Sciences*, 25, 99-104

Vandepoele et al, 2002, Genome-Wide Analysis of Core Cell Cycle Genes in *Arabidopsis, The Plant Cell*, 14, 903–916

Verkest et al, 2005, The Cyclin-Dependent Kinase Inhibitor *KRP2* Controls the Onset of the Endoreduplication Cycle during *Arabidopsis* Leaf Development through Inhibition of Mitotic *CDKA*;*1* Kinase Complexes, *The Plant Cell*, 17(6):1723-1736

Vernoux et al, Auxin at the Shoot Apical Meristem, *Cold Spring Harb Perspect Biol*, 2(4)

Wang et al, 2007, *Arabidopsis* Ovate Family Protein 1 is a transcriptional repressor that suppresses cell elongation, *The Plant Journal*, 50, 858–872

Wehrens and Buydens, 2007, Self- and Super-organising Maps in R: the kohonen package*, J. Stat. Softw*, 5(4)

Werner et al, 2003, Cytokinin-Deficient Transgenic Arabidopsis Plants Show Multiple Developmental Alterations Indicating Opposite Functions of Cytokinins in the Regulation of Shoot and Root Meristem Activity, *The Plant Cell*, 15, 2532–2550

Williams and Fletcher, 2005, Stem cell regulation in the Arabidopsis shoot apical meristem, *Current Opinion in Plant Biology*, 8:582–586

Wuest et al, 2012, Molecular basis for the specification of floral organs by *APETALA3* and *PISTILLATA*, *PNAS*, 109(33):13452-7

Yadav et al, 2007, Gene expression map of the *Arabidopsis* shoot apical meristem stem cell niche, *PNAS*, 106(12), 4941–4946

Yanai et al, 2005, *Arabidopsis* KNOXI Proteins Activate Cytokinin Biosynthesis*, Current Biology*, 15, 1566–1571

Appendices

Appendix 1 Primer Sequences

CUC1 Forward – TCCTCCGCTAAGGATGAA

CUC1 Reverse – GAGCGGGAAGGAATGTA

OFP1 Forward – GGAGTCGATGGAGGAGA

OFP1 Reverse – TGGGGTGGTGGAAGATTAAG

AIL7 Forward – TCTTCTCCCTCGGATCAAAA

AIL7 Reverse – GGCCACAAGAAAAACTCAGC

STM Forward – GGCTGGACCAGAAACAGATAA

STM Reverse – AAAGCATGGTGGAGGAGATG

ACT2 Forward – ACATTGTGCTCAGTGGTGGA

ACT2 Reverse – CTGAGGGAAGCAAGAATGGA

miR164c Precursor Forward – TAACACTTGATGGAGAAGCA

miR164c Precursor Reverse – AGACACGTGTTGGAGTAGTA

Appendix 2 Time Course

Electronic Appendix – Full scripts (Time Course.R) and Raw data (.cel) files can be found in Appendix 2. The following .csv files contain Time Course Microarray Results

8hOE.csv 8h STM OE – 8h EV

24hOE.csv 24h STM OE – 24h EV

72hOE.csv 72h STM OE – 72h EV

9dOE.csv 9d STM OE – 9d EV

72hRNAi.csv 72h STM RNAi – 72h EV

9dRNAi.csv 9d STM RNAi – 9d EV

mutant.csv Stm-2 – wt

Results are organized as follows:

Array_element_name – Probeset ID

Locus – AtG Number

LogFC – Limma point estimate of log2 fold change

Pvalue – Unadjusted P-value

adjPvalue – Corrected p-value

symbol – Gene Short Name if available

Description– Annotation if available

Appendix 3 35S::STM-GR Experiment

Electronic Appendix – Full scripts (GR.R) and Raw data (.cel) files can be found in Appendix 3. The following .csv files contain Time Course Microarray Results

Deck-Mock.csv DEX vs Mock Treated Plants

CHXDEX-DEX.csv CHXDEX vs DEX Treated Plants

Mock-CHX.csv Mock vs CHX Treated Plants

Column descriptions are as follows:

ProbeId – Probeset ID

Log2 Fold Change – Limma point estimate of log2 fold change

p-value – Unadjusted P-value

adj P-Value – Corrected p-value

Short Name – Gene Short Name if available

Annotation – Tair Annotation if available

Appendix 4 Hormone Analysis and CYCD3 OE Experiments

Electronic Appendix – Full scripts (Hormones.R) and Raw data (.cel) files can be found in Appendix 4. The following .csv files contain Hormone analysis and CYCD3OE results

3hZeatin.csv	CK treatment experiment
3hIAA.csv	Auxin treatment experiment
3hGA.csv	GA treatment experiment
CYCDyoung.csv	Young CYCD3 OE tissue vs young wt tissue
CYCDold.csv	Old CYCD3 OE tissue vs old wt tissue

Results are organized as follows:

ProbeId – Probeset ID

Log2 Fold Change – Limma point estimate of log2 fold change

p-value – Unadjusted P-value

adj P-Value – Corrected p-value

Short Name – Gene Short Name if available

Annotation – Tair Annotation if available

Appendix 5 Meta Analysis R Script

Electronic Appendix – Full scripts (meta.R) can be found in Appendix 5. The following .csv files contain Meta Analysis Results

rapid.csv	Rapid Response Meta Analysis
robust.csv	Robust Response Meta Analysis
phenotypic.csv	Phenotypic Meta Analysis

Column descriptions are as follows:

ProbeId – Probeset ID

qStouffers – Corrected Stouffer's P-Value

qFisher – Corrected Fisher's Inverse Chi squared test P-Value

Short Name – Gene Short Name if available

Annotation – Tair Annotation if available

Appendix 6 Bayesian Network .cel files

Electronic Appendix – cel_list.txt - List of .cel files used for Bayesian network structural inference

Appendix 7 SOM and PCA

Electronic Appendix – Full scripts (data_mining.R)

Som_nodes.csv contains assignments of genes to nodes in the SOM. Column headings are as follows:

Probeset	ATH1 Probe ID
Node	Number of node gene assigned to
AtG	Gene AtG Number
Symbol	Short gene name if available
Description	TAIR annotations for gene if available.

Correlation of codebook vectors 8x8.csv contains a matrix of all nodes showing their pearson's correlation coefficient (unsquared) against each other node.

Appendix 8 Stochastic Model Script

Electronic Appendix – Full scripts (Stochastic.R)

Appendix 9 COPASI Model File

Electronic Appendix – 2-compartment.copasi contains 2-compartment model described in Chapter 5 for use with COPASI

Appendix 10 Time Course Go Enrichment Tables

Electronic Appendix – All results stored as .bgo files which can be opened in Cytoscape using BINGO plugin or Microsoft Excel.

Appendix 11 Yadav Script

Electronic Appendix – R Script used to calculate significance of log_2 expression values provided in Yadav et al (2009) found in yadav.txt, input data in mas5.txt

Yadav.csv contains full results with columns as follows:-

Array_element_name	probeset id on ATH1 array
Locus	AtG number
Classification	Which of the three contrasts FIL, WUS and CLV are higher than the other expressed in the form X>Y
CLVp-FILp	p-value for CLV vs FIL contrast
CLVp-WUSp	p-value for CLV vs WUS contrast
FILp-WUSp	p-value for FIL vs WUS contrast
CLVp	Average log_2 expression in CLV domain
WUSp	Average log_2 expression in WUS domain
FILp	Average log_2 expression in FIL domain
Symbol	Common Gene Name
Description	TAIR annotations for probeset

Appendix 12 Co-Expression Analysis Results

Electronic Appendix – 260632_at.xls contains in Microsoft Excel format a spreadsheet of co-expressed genes with *STM*

Headings are as follows:

Probeset	ATH1 probeset ID
Gene	Gene Symbol
P	P-Value for correlation
R2	Pearsons r^2 correlation coefficient
Desc	Description of gene